Leveraging Multilingual Models for Robust Grammatical Error Correction Across Low-Resource Languages

Divesh Kubal
Trinka AI
divesh.kubal@trinka.ai

Apurva Nagvenkar
Trinka AI
apurva.nagvenkar@trinka.ai

Abstract

Grammatical Error Correction (GEC) is a crucial task in Natural Language Processing (NLP) aimed at improving the quality of user-generated content, particularly for non-native speakers. This paper introduces a novel end-to-end architecture utilizing the M2M100 multilingual transformer model to build a unified GEC system, with a focus on low-resource languages. A synthetic data generation pipeline is proposed, tailored to address language-specific error categories. The system has been implemented for the Spanish language, showing promising results based on evaluations conducted by linguists with expertise in Spanish. Additionally, we present a user analysis that tracks user interactions, revealing an acceptance rate of 88.2%, as reflected by the actions performed by users.

1 Introduction

GEC is a critical task within the field of NLP that focuses on identifying and rectifying grammatical inaccuracies in text. This task has gained significant attention in recent years due to its potential to enhance the grammaticality and overall readability of user-generated content. This is particularly beneficial for non-native speakers who often produce text containing various grammatical errors.

GEC systems traditionally depend on large annotated datasets to learn linguistic structures and errors, with model accuracy highly dependent on data quality and volume. While research has focused mainly on English language, GEC applies to multiple languages, with the LANG-8 Learner Corpus (Koyama et al., 2020) being a key resource featuring contributions from 80 languages. However, this corpus is highly imbalanced, skewed towards Japanese and English, which limits robustness of model development for low-resource languages. Additionally, uncontrolled data collection leads to issues like excessive paraphrasing and incomplete corrections, complicating training. Most approaches create language-specific models, limiting their multilingual applicability.

In this paper, we propose a novel architecture capable of addressing the GEC problem across multiple languages using a single model. Our approach aims to establish a more efficient and scalable solution for grammatical error correction. This approach will particularly help for low-resource languages. This paper leverages the M2M100 model (Fan et al., 2021), a multilingual encoder-decoder (seq-to-seq) framework trained for many-to-many multilingual translation. This model supports translation in 100 languages across 9,900 language pairs using a single architecture. By fine-tuning the M2M100 model for the GEC task, we harness its multilingual capabilities to address grammatical errors in various languages.

Our approach incorporates a synthetic data preparation pipeline, which we found to be crucial for generating high-quality GEC data. Insights from language-specific experts on grammar error categories significantly enhance the quality of this synthetic data generation, allowing the pipeline to be applied repeatedly for any selected language. We implement this entire architecture for the Spanish language and demonstrate its applicability across multiple languages, showcasing the potential of our proposed solution to advance GEC research.

2 Literature Survey

GEC systems are primarily categorized into two divisions: Text-to-Text (T2T), which rewrites entire input sentences, and Edit-based, which focuses on detecting and correcting specific errors.

Edit-based Approaches: Seq2Edit models, such as LaserTagger (Malmi et al., 2019) and PIE (Awasthi et al., 2019), predict token-level operations, including insertion, deletion, and swapping.

Proceedings of the 31st International Conference on Computational Linguistics: Industry Track, pages 505–510
January 19–24, 2025.

Seq2Edits (Stahlberg and Kumar, 2020) extends this by targeting sequences of edit operations, while GECToR (Omelianchuk et al., 2020) introduces custom transformations alongside standard edits.

Seq2Seq Approaches: The Seq2Seq paradigm encodes erroneous sentences and generates error-free outputs. This approach, explored in various works ((Liu et al., 2020); (Wang et al., 2021); (Li et al., 2022); (Fang et al., 2023a)), is noted for producing more fluent sentences, albeit at a slower decoding speed. (Zhao et al., 2019) enhance this framework with a copy mechanism, while (Kaneko et al., 2020) incorporate pre-trained knowledge. Pseudo dataset construction has emerged as a critical technique in GEC, allowing for the effective generation of error-free sentences with injected noise (Zhao et al., 2019; Liao et al., 2023; Kiyono et al., 2020; Yasunaga et al., 2021; Fang et al., 2023b).

Multilingual Approaches: Recent advancements in massively multilingual machine translation have led to the development of notable models such as M2M-100 (Fan et al., 2021), NLLB (Costa-jussà et al., 2022), and MADLAD-400 (Kudugunta et al., 2024). Additionally, large language models have demonstrated promising capabilities in error correction through prompting techniques ((Loem et al., 2023; Fang et al., 2023c; Coyne et al., 2023).

3 Proposed Approach

In this section, a detailed architecture for developing a robust GEC system tailored for multiple languages is proposed. The approach encompasses several core components designed to systematically extract, manipulate, and process linguistic data to enhance error correction capabilities. The architecture depicted in figure 1 comprises the following core components: Text Corpus Extraction, Identification of Language-Specific Grammar Error Categories, Introduction of Grammar Errors, Construction of a Parallel Corpus, Selection of a Transformers-Based Encoder-Decoder Model, Fine-tuning the Model, and Tweaking the Inference Mechanism.

3.1 Text Corpus Extraction

This first step involves selecting the languages for which the GEC system is to be built, followed by defining the domain of the corpus. The chosen domain can vary, encompassing general, academic, technical, or specialized areas such as medical literature. For this research, the focus will be primarily on general language-specific data. To facilitate the extraction of publicly available text corpus in multiple languages, we utilize resources from the Leipzig Corpora Collection, which offers access to a wide array of text sources, including news articles, web pages, and Wikipedia entries. Specifically, we will employ the most recent year's Wikipedia data for comprehensive coverage.

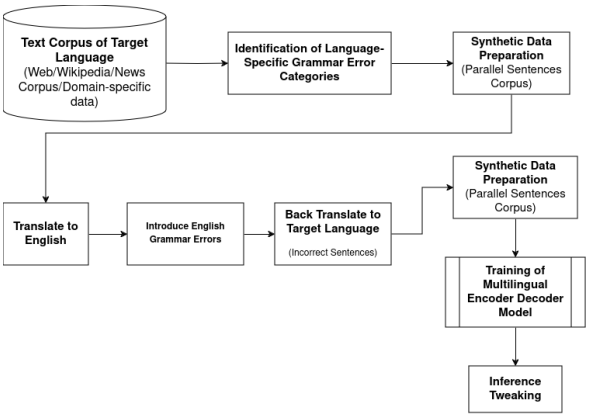

Figure 1: Unified Architecture for Multilingual Grammar Error Correction

Additionally, in cases where domain-specific or in-house data is available, this information can be appended to enrich the corpus further. This augmentation will allow the GEC system to adapt to specialized vocabulary and nuances related to styles of a laguage, thereby enhancing its applicability across different contexts. Hence, this will facilitate the trained model to be versatile and effective in correcting grammatical errors across a range of language families and subject areas. Once the text corpus is extracted, it will be processed by segmenting paragraphs into individual sentences utilizing a language-specific sentence segmenter, thereby preparing the data for next steps.

3.2 Identification of Language-Specific Grammar Error Categories

Following the corpus extraction, the next step is to identify the specific categories of grammatical errors corresponding to each language selected in the previous step. This step is of utmost importance, as it lays the groundwork for introducing synthetic errors into the text corpus. To enhance the Spanish GEC capabilities, we have collaborated with a linguist specializing in Spanish to curate a comprehensive list of fine-grained error categories. The fine-grained identification of these categories

is helpful, as it directly influences the nature of the grammatical errors introduced in the next step, ultimately affecting the quality and effectiveness of the synthetic data generated.

3.3 Introducing Grammar Errors in the Correct Text Corpus

The aim of this step is to systematically introduce grammatical errors into the correct text corpus extracted from Step 3.1. A primary challenge in training a unified model capable of correcting grammatical errors across multiple languages is the scarcity of annotated data. Specifically, most GEC systems require paired examples of incorrect and correct sentences. As established in Step 3.1, we have a downloaded corpus of correct sentences (in target languages). To generate erroneous counterpart for each correct sentence, we use a back-translation approach.

The procedure consists of the following steps:

1. **Translation to English**: The correct sentences in the target languages are translated into English using the open-source Opus-MT models available on the Hugging Face Model Hub. The quality of translation is not a primary concern at this stage, as the objective is to introduce errors into the text.

2. **Introduction of Grammatical Errors**: In this step, rule-based grammatical errors are introduced into the English sentences obtained from the previous step. This is achieved through the use of the errorify function from the PIE toolkit (Awasthi et al., 2019).

3. **Back-Translation**: The error-laden English sentences are subsequently back-translated into the original target languages utilizing the same Opus-MT models (Tiedemann and De Gibert, 2023) from the Hugging Face Model Hub (Jain, 2022).

This approach enables the generation of synthetic data across multiple languages. The quantity of data generated is dependent upon the specific use case and the computational resources available.

3.4 Construction of Parallel Corpus

The objective of this stage is to construct a parallel corpus containing pairs of incorrect and correct sentences. Each pair will serve as a single data point within the training dataset, with the correct sentences extracted from Step 3.1 and their erroneous counterparts generated in Step 3.3 (previous step). Once these incorrect-correct sentence pairs are aligned, instructions will be prepended to the incorrect sentences to guide the model during training. For instance, an instruction such as "Correct all the Grammatical Errors: " will be appended to English data points. Experiments indicates that instructions tailored to the target language yield superior outputs compared to generic instructions in English, enhancing the model's contextual understanding. For each language, a language-specific instruction is used.

The highlight of our proposed approach is the training of a single model on a diverse dataset of multiple languages created by appending and randomly shuffling parallel sentences across multiple languages. This shuffling strategy mitigates potential biases (gradient-biases) during gradient-based training and promotes a more generalized learning capability across the languages involved.

3.5 Selection of a Transformers-Based Encoder-Decoder Model

In this step, the objective is to select an appropriate transformers-based Encoder-Decoder model that has been pretrained on multiple languages. The choice of model is critical to ensuring that the GEC system can effectively leverage the vast linguistic knowledge encapsulated within these pretrained frameworks. Models such as mBART (Liu, 2020), T5 (Raffel et al., 2020), MarianMT (Junczys-Dowmunt et al., 2018), M2M100 (Fan et al., 2021), etc. have demonstrated efficacy in multilingual settings and will be considered based on their architecture, performance benchmarks, and compatibility with our dataset requirements. The objective of this selection process is to maximize the model's ability to generalize across various languages while maintaining high performance on the specific GEC tasks.

3.6 Fine-tuning the Multilingual Encoder-Decoder Model

Once the multilingual Encoder-Decoder model is selected, the next phase involves fine-tuning the model using synthetically generated GEC data from the constructed parallel corpus. This fine-tuning was performed on high-performance infrastructure equipped with dual Nvidia A30 GPUs, each with 24GB of VRAM. The training process is designed to balance efficiency and thoroughness,

with careful optimization of batch sizes, learning rates, and epoch durations to achieve optimal performance. Detailed training metrics were logged to evaluate the model's convergence and generalization capabilities, ensuring that the final output is both robust and reliable.

4 Manual Evaluation

The manual evaluation was conducted on three test sets, comprising general Spanish data and academic texts. Test sets 1 and 2 were derived from the COWS-L2H corpus (Yamada et al., 2020), which contains Spanish learner writing, evaluated over two rounds by a Spanish language expert. Test set 3 consisted of academic data sourced from research papers. Table 1 depicts the overall results of manual evaluation. Testset 1 (containing 91 sentences) demonstrates the highest performance, yielding an impressive F1 score of 95.71%. Testset 2 which consists of 25 sentences shows slightly lower overall performance, with F1 score of 93.33%. Testset 3 comprising of 100 sentences, focused on academic writing, had 66 TP, with a notable F1 score of 87.50%. Overall, the evaluation highlights that while the system performs well across varied datasets, the model requires further refinement for optimal enhancement of scholarly text.

5 User Analysis

5.1 Interface

Figure 2 shows an interface where the Spanish Multilingual GEC model is deployed. Since the task is GEC, the corrections generated by the model are presented in spans, requiring the user to perform actions on each span rather than the entire sentence. The user has two simple operations to choose from: Accept and Reject.

- Accept: The user has high confidence in the correction, likely indicating true positives (TPs).
- Reject: The user has low confidence in the correction, likely indicating false positives (FPs).

5.2 Analysis

After deploying the Spanish GEC model within our product, we initiated a data collection phase where data was systematically gathered from our database, ensuring that only specific information was accessed while safeguarding the critical components of users' data. We exclusively collected information on the actions performed by users and the categories associated with the corrections. This approach ensures that no sensitive or personal information was used for analysis, maintaining strict data confidentiality.

The purpose of this data collection was to gain an initial understanding of the model's performance for users, without examining the domain or content of the documents uploaded by them. We were particularly diligent in ensuring that the data used for analysis did not include any information from sensitive data plans.

We conducted two types of analyses:

- **Overall analysis:** This evaluated the total number of actions performed by the model.
- **Category-level analysis:** This involved evaluating the model's performance based on specific categories of corrections.

5.2.1 Quantitative Insights from the User Data

The quantitative analysis of user interactions with the Spanish GEC model provided valuable insights into both user behavior and the system's effectiveness. These metrics indicate a high level of user engagement with the system, which is notable given that the Spanish GEC system was launched only recently. This highlights its relevance and utility in real-world applications.

As shown in Table 2, we extracted 161083 unique sentences of which Spanish GEC model triggered on 83868 (52.06%). Total number of spans obtained are 161083 and user performed action on 30897 (24%).

The analysis of these actions provides the following key insights as shown in Table 3:

- Acceptance Rate: A significant 88.2% of the model's suggestions were accepted by users, indicating a high level of confidence in the model's corrections.
- Rejection Rate: On the other hand, 11.8% of suggestions were rejected, which points to areas where the model's performance can be improved, especially in handling certain grammar rules.

Testset	TP	FP	FN	Recall	Precision	F1-score
Testset 1	134	4	8	94.37%	97.10%	95.71%
Testset 2	28	1	3	90.32%	96.55%	93.33%
Testset 3 - Academic	77	10	12	86.52%	88.51%	87.50%
Overall	239	15	23	94.09%	91.22%	92.63%

Table 1: Summary of manual evaluation metrics for different test sets.

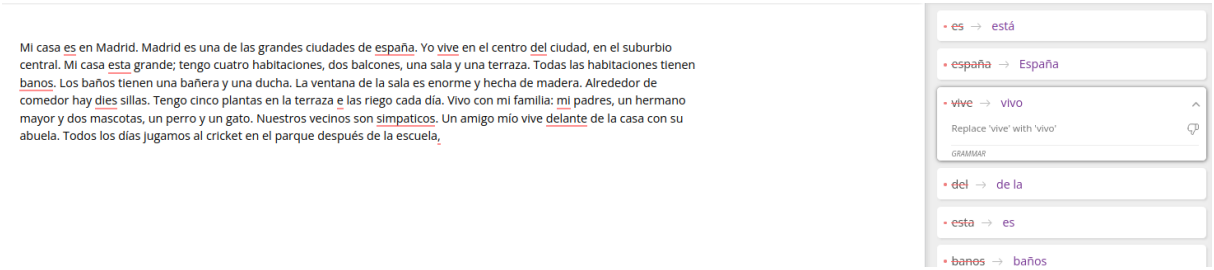

Figure 2: Interface of Spanish GEC Engine

Details	Number
# Sents	161083
# Sents: Model triggered	83868
# Spans	124441
# Spans: Actions performed	30897
# Spans: No Actions performed	93544

Table 2: Statistics of information extracted from User database

Action	Percentage
Accept	88.2%
Reject	11.8%

Table 3: Distribution of User Actions with their percentages.

6 Conclusion and Future Work

This paper presents a scalable GEC architecture for low-resource languages using the M2M100 multilingual transformer model. Our evaluation shows strong performance, with an 88.2% acceptance rate from real-time users, affirming the system's reliability. However, as shown in Table 2, a significant portion of the model's suggestions i.e. 93,544 firings/edits were ignored where no actions were performed by users. This discrepancy highlights the need for further investigation into the reasons behind these ignored suggestions. In future work, we will prioritize understanding user behavior and preferences more deeply to ensure our system becomes increasingly aligned with user needs. We aim to conduct a detailed analysis to identify the root causes of ignored suggestions and implement concrete improvements to address them. Furthermore, we plan to extend the proposed architecture to support a fully multilingual setup, enabling efficient GEC across various languages. This expansion will enhance the system's accessibility and effectiveness in multilingual environments, fostering broader adoption and utility.

References

Abhijeet Awasthi, Sunita Sarawagi, Rasna Goyal, Sabyasachi Ghosh, and Vihari Piratla. 2019. Parallel iterative edit models for local sequence transduction. *arXiv preprint arXiv:1910.02893*.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Steven Coyne, Keisuke Sakaguchi, Diana Galvan-Sosa, Michael Zock, and Kentaro Inui. 2023. Analyzing the performance of gpt-3.5 and gpt-4 in grammatical error correction. *arXiv preprint arXiv:2303.14342*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Tao Fang, Jinpeng Hu, Derek F Wong, Xiang Wan, Lidia S Chao, and Tsung-Hui Chang. 2023a. Improving grammatical error correction with multimodal feature integration. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9328–9344.

Tao Fang, Xuebo Liu, Derek F Wong, Runzhe Zhan, Liang Ding, Lidia S Chao, Dacheng Tao, and Min Zhang. 2023b. Transgec: Improving grammatical error correction with translationese. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3614–3633.

Tao Fang, Shu Yang, Kaixin Lan, Derek F Wong, Jinpeng Hu, Lidia S Chao, and Yue Zhang. 2023c. Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation. *arXiv preprint arXiv:2304.01746*.

Shashank Mohan Jain. 2022. Hugging face. In *Introduction to transformers for NLP: With the hugging face library and models to solve problems*, pages 51–67. Springer.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, et al. 2018. Marian: Fast neural machine translation in c++. *arXiv preprint arXiv:1804.00344*.

Masahiro Kaneko, Masato Mita, Shun Kiyono, Jun Suzuki, and Kentaro Inui. 2020. Encoder-decoder models can benefit from pre-trained masked language models in grammatical error correction. *arXiv preprint arXiv:2005.00987*.

Shun Kiyono, Jun Suzuki, Tomoya Mizumoto, and Kentaro Inui. 2020. Massive exploration of pseudo data for grammatical error correction. *IEEE/ACM transactions on audio, speech, and language processing*, 28:2134–2145.

Aomi Koyama, Tomoshige Kiyuna, Kenji Kobayashi, Mio Arai, and Mamoru Komachi. 2020. Construction of an evaluation corpus for grammatical error correction for learners of japanese as a second language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 204–211.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2024. Madlad-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems*, 36.

Bei Li, Quan Du, Tao Zhou, Yi Jing, Shuhan Zhou, Xin Zeng, Tong Xiao, JingBo Zhu, Xuebo Liu, and Min Zhang. 2022. Ode transformer: An ordinary differential equation-inspired model for sequence generation. *arXiv preprint arXiv:2203.09176*.

Junwei Liao, Sefik Eskimez, Liyang Lu, Yu Shi, Ming Gong, Linjun Shou, Hong Qu, and Michael Zeng. 2023. Improving readability for automatic speech recognition transcription. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(5):1–23.

Xuebo Liu, Longyue Wang, Derek F Wong, Liang Ding, Lidia S Chao, and Zhaopeng Tu. 2020. Understanding and improving encoder layer fusion in sequence-to-sequence learning. *arXiv preprint arXiv:2012.14768*.

Y Liu. 2020. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*.

Mengsay Loem, Masahiro Kaneko, Sho Takase, and Naoaki Okazaki. 2023. Exploring effectiveness of gpt-3 in grammatical error correction: A study on performance and controllability in prompt-based methods. *arXiv preprint arXiv:2305.18156*.

Eric Malmi, Sebastian Krause, Sascha Rothe, Daniil Mirylenka, and Aliaksei Severyn. 2019. Encode, tag, realize: High-precision text editing. *arXiv preprint arXiv:1909.01187*.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. Gector–grammatical error correction: tag, not rewrite. *arXiv preprint arXiv:2005.12592*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Felix Stahlberg and Shankar Kumar. 2020. Seq2edits: Sequence transduction using span-level edit operations. *arXiv preprint arXiv:2009.11136*.

Jörg Tiedemann and Ona De Gibert. 2023. The opus-mt dashboard–a toolkit for a systematic evaluation of open machine translation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 315–327.

Yu Wang, Yuelin Wang, Kai Dang, Jie Liu, and Zhuo Liu. 2021. A comprehensive survey of grammatical error correction. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(5):1–51.

Aaron Yamada, Sam Davidson, Paloma Fernández-Mira, Agustina Carando, Kenji Sagae, and Claudia Sánchez-Gutiérrez. 2020. Cows-l2h: A corpus of spanish learner writing. *Research in Corpus Linguistics*, 8(1):17–32.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2021. Lm-critic: Language models for unsupervised grammatical error correction. *arXiv preprint arXiv:2109.06822*.

Wei Zhao, Liang Wang, Kewei Shen, Ruoyu Jia, and Jingming Liu. 2019. Improving grammatical error correction via pre-training a copy-augmented architecture with unlabeled data. *arXiv preprint arXiv:1903.00138*.