



Automated Grammar and Language Error Correction System



Introduction

Objective: Create an automated system for correcting grammatical errors in English sentences while providing user-friendly feedback.

Motivation:

Non-native English speakers often struggle with complex grammar rules.

Existing tools like Grammarly are effective but proprietary, making them inaccessible for customization or research.

This project bridges the gap by offering an open-source alternative with educational feedback.

Key Challenges:

Identifying and correcting grammatical nuances such as:

- Subject-verb agreement (*he go* → *he goes*).
- Tense consistency (*was cooking and eat* → *was cooking and eating*).
- Prepositional usage (*on the car* → *in the car*).

Balancing correction accuracy with meaningful user explanations.

Impact: This system has potential applications in education, aiding language learners to understand and correct their errors effectively.

Related Work

Existing Solutions:

Grammarly and *Microsoft Word Grammar Checker*:

- Proprietary tools with advanced algorithms for grammar correction.
- Lack transparency and are inaccessible for academic or open-source replication.

Open-source approaches like *GECToR* (Grammatical Error Correction Transformer):

- Utilize large-scale datasets to implicitly learn grammar rules.
- Effective in correcting complex errors but limited to specific text types (e.g., formal or academic).

Related Work

Research Foundations:

Transformer-based models like *BART* and *T5*:

- Proven success in understanding and generating human-like text.
- Adaptable for grammar correction tasks with minimal rule-based programming.

Studies show the limitations of existing systems in handling informal or conversational English.

Innovations in This Project:

- Combines transformer-based grammar correction with detailed error feedback, using OpenAI API.
- Focus on educational value, enabling users to learn from corrections.
- Addresses gaps by testing on a variety of sentence types, including casual and conversational language, expanding usability beyond formal contexts.

Objectives

Primary Goal: Develop an automated grammar correction system that not only identifies and fixes grammatical errors but also provides educational feedback to help users understand their mistakes.

Specific Objectives:

- **System Development:**
Create a robust NLP system capable of correcting grammatical errors in English sentences.
- **Methodology Exploration:**
Compare traditional rule-based methods with advanced machine learning models, particularly transformer-based models like T5, BART.
- **Performance Evaluation:**
Measure system accuracy using the GLEU score.
Evaluate on diverse sentences to ensure generalizability and effectiveness.
- **Educational Value:**
Incorporate a feedback mechanism to explain grammatical corrections, aiding language learners in improving their skills.
- **Benchmarking:**
Compare the system's performance against existing tools like Grammarly, demonstrating the potential of open-source alternatives.

End Goal: Deliver a comprehensive grammar correction tool that performs well on unseen sentences, provides meaningful insights to users, and contributes to advancements in automated grammar correction.

Selected Datasets

Dataset Used: *JFLEG Combined.csv*

Designed specifically for grammatical error correction tasks.

Contains two main columns:

- Sentence: Original sentences with grammatical errors.
- Corrections: Corrected versions of the sentences.

Features of the Dataset:

Includes a variety of sentence types (simple, structured, casual).

Covers diverse grammatical errors, including:

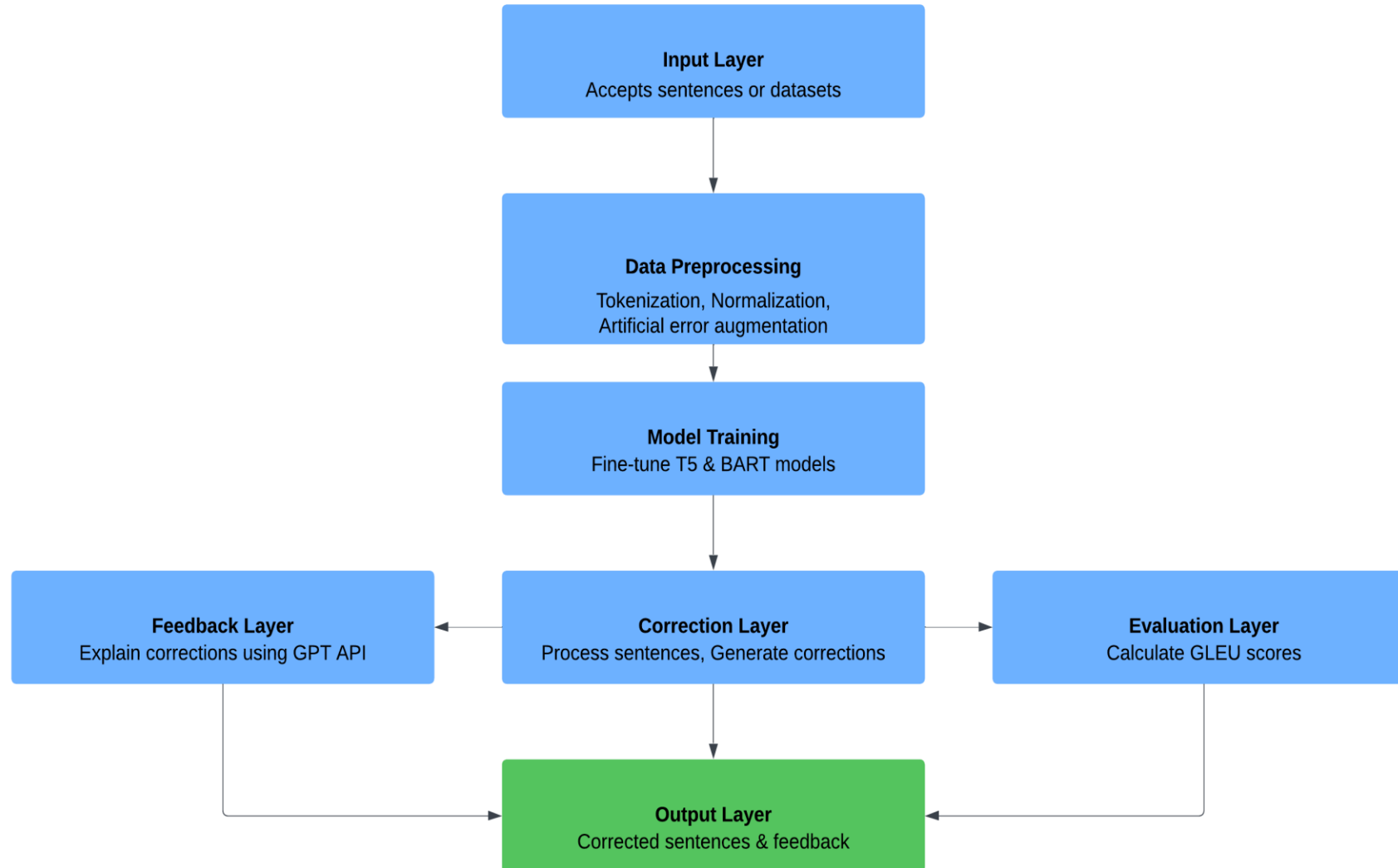
- Articles: *a* vs. *the*.
- Subject-Verb Agreement: *he go* → *he goes*.
- Tense Consistency: *was cooking and eat* → *was cooking and eating*.
- Preposition Usage: *on the car* → *in the car*.

Relevance:

Ensures the system can handle a broad range of grammatical structures.

Provides examples for both training and testing the model's accuracy.

System Architecture



Baseline Solution

Initial Approach

Developed a custom grammar correction model trained from scratch using the JFLEG dataset.

Challenges

Low GLEU Score: Achieved only 0.08, indicating poor fluency and accuracy.

Complex Sentences: Struggled with corrections in structurally complex sentences.

Idiomatic Expressions: Failed to handle idiomatic and context-dependent phrases effectively.

Conclusion

Highlighted the need for more advanced models to address complex grammatical structures and idiomatic usage.

Key Takeaway

The baseline solution provided insights into limitations, guiding the transition to pre-trained transformer models.

Proposed Solutions

Dataset Selection

Dataset: JFLEG for grammar error correction.

Augmentation: Artificial errors introduced (e.g., article misuse, tense inconsistencies).

Data Preprocessing

Normalization: Ensures consistency in input formats.

Tokenization: Converts sentences to model-compatible inputs with a "fix:" prefix.

Model Training

Models: T5 and BART fine-tuned on augmented datasets.

Hyperparameters: LR: 5e-5, Batch Size: 8, Epochs: 5.

Grammar Correction

Real-time sentence correction using **beam search** decoding for accuracy.

Proposed Solutions

Feedback Generation

OpenAI GPT API: Generates detailed explanations for corrections (e.g., articles, tense, prepositions).

Evaluation

Metric: GLEU score to measure fluency and correctness.

User Interaction

Interactive Interface: Real-time corrections and feedback for user inputs.

Batch Processing: Supports dataset-level corrections.

Key Features

- Accurate grammar correction.
- Educational feedback for learning.
- Interactive and batch processing support.

Analysis and interpretation

Dataset Analysis

Dataset: JFLEG, with 604 training, 151 validation, and 748 test sentences.

Common Errors: Article misuse, subject-verb disagreement, tense inconsistency, and preposition misuse.

Augmentation: Artificially introduced errors improved model robustness.

Training and Validation Loss

T5 Model:

Validation loss stabilized at 0.11 after the third epoch.

Efficient learning and strong generalization.

BART Model:

Higher validation loss indicates slight overfitting or difficulty in generalizing.

Model Performance

GLEU Scores:

T5 Outperformed BART, achieving higher fluency and grammatical accuracy.

Augmented dataset significantly improved results compared to the baseline.

Analysis and interpretation

Feedback Quality

Feedback addressed grammatical issues effectively:

- Article misuse
- Subject-verb agreement
- Tense consistency
- Preposition misuse

Detailed, user-friendly explanations supported grammar learning.

System Evaluation

Interactive Mode: Real-time corrections and feedback for user input.

Batch Processing: Handles datasets, enhancing versatility for large-scale use cases.

Key Insight

T5's superior performance and high-quality feedback establish it as the preferred model for this grammar correction system.

Conclusions

Successfully developed an **automated grammar and language error correction system**.

Integrated fine-tuned **T5** and **BART** models with **OpenAI's GPT API** for feedback generation.

Addressed grammatical issues such as:

- Article misuse
- Subject-verb agreement
- Tense consistency
- Preposition misuse

Performance Highlights:

T5 Model: Achieved superior GLEU score of **0.8049**, excelling in fluency and accuracy.

Provided **user-friendly feedback**, enhancing educational value for language learners.

Future Work

Handling Idiomatic Expressions: Improve robustness for idioms and complex phrases.

Enhanced Readability: Explore sentence restructuring for better clarity.

Larger Datasets: Incorporate diverse datasets for improved generalization.

Advanced Models: Experiment with models offering stronger semantic understanding.

Evaluation Metrics: Introduce comprehensive metrics beyond GLEU for nuanced scenarios.

Performance Optimization: Focus on faster response times and consistent outputs.

This future roadmap aims to make the system more versatile, robust, and user-friendly for broader applications.

Acknowledgements

We extend our heartfelt thanks to:

Dataset Providers

For granting access to the JFLEG dataset, which was instrumental in training and evaluating our grammar correction system.

Transformer Model Developers

To the creators of T5 and BART, whose pre-trained models provided a robust foundation for our solution.

Open Source Contributors

For tools and libraries like Hugging Face Transformers, PyTorch, NLTK, and others, which were pivotal for model implementation, evaluation, and data preprocessing.

Dr. Lindi Liao

For their invaluable mentorship and expertise in natural language processing and machine learning frameworks, guiding our project to successful completion.

Screenshots of outputs and code snippets

Original Sentence: i have an idea to visit the europe next summer.
Corrected Sentence (T5): I have an idea to visit Europe next summer.
Feedback: The sentence 'I have an idea to visit Europe next summer.' is grammatically correct. Here are the elements in focus:

- Incorrect use of articles: The sentence is accurately using the article 'an' before 'idea,' which is a singular, non-specific countable noun. Moreover, 'Europe' usually does not take an article, which is correct here.
 - Subject-verb agreement: The subject 'I' correctly matches with the verb 'have.' It is correct for singular first-person use.
 - Tense consistency: The sentence remains in the present tense throughout ('I have an idea'), and there is no change of tense that could cause inconsistency.
 - Preposition misuse: There are no prepositions misused in this sentence. 'To' correctly indicates direction or intended result in the context it is used in the sentence.
- =====

Original Sentence: The players in the team is practicing.
Corrected Sentence (T5): The players in the team are practicing.
Feedback: The sentence, 'The players in the team are practicing,' is grammatically correct.

- Incorrect use of articles: There's no misuse of articles here. The definite article 'the' correctly refers to a specific group of players and a specific team.
- Subject-verb agreement: The subject 'The players' correctly corresponds to the plural verb 'are practicing'. The subject and the verb are in agreement here.
- Tense consistency: The sentence is consistent in presenting its idea in the present continuous tense.
- Preposition misuse: The preposition 'in' is used correctly to denote that the players belong to the team.

No improvements are necessary as the sentence is grammatically sound as it is.

=====

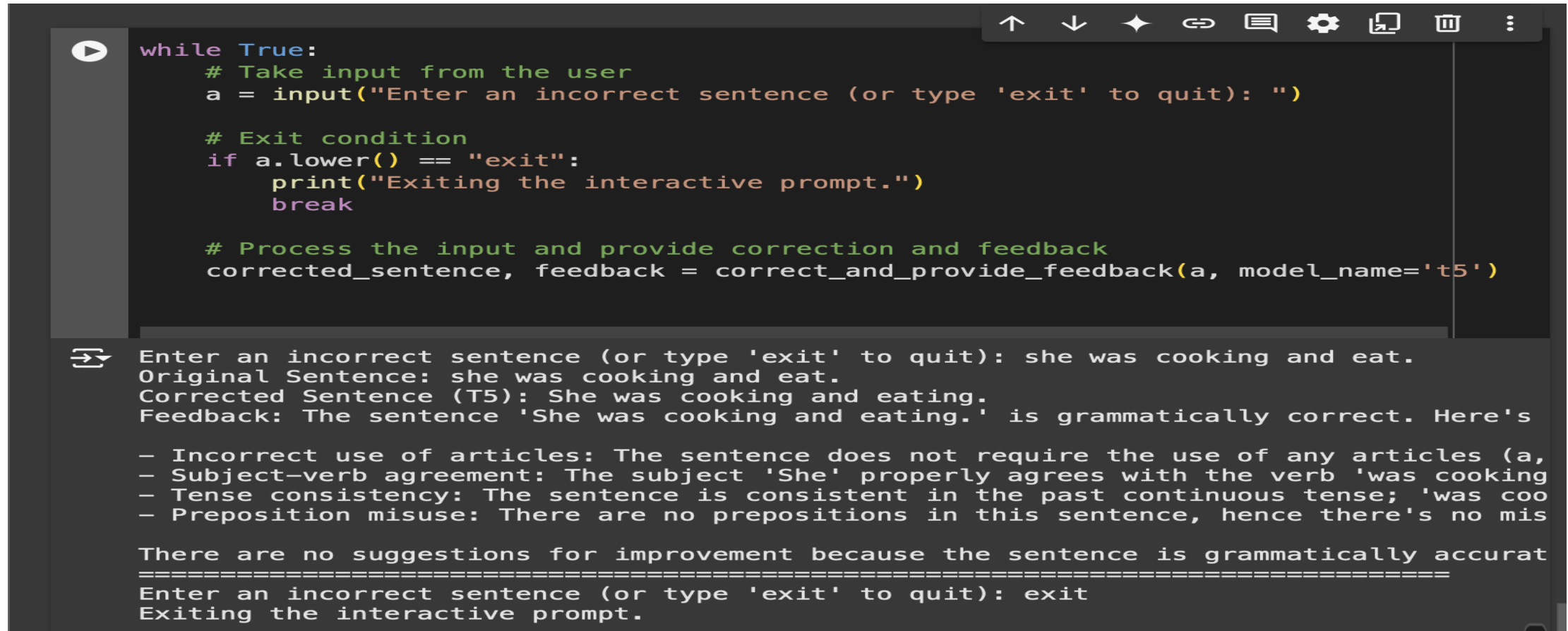
Original Sentence: He is good on mathematics.
Corrected Sentence (T5): He is good at mathematics.
Feedback: The sentence 'He is good at mathematics.' is grammatically correct. Here's why:

- Incorrect use of articles: The sentence does not require an article, so there is no issue here.
- Subject-verb agreement: The singular subject 'He' correctly corresponds with the singular verb 'is'.
- Tense consistency: The entire sentence is in the present tense, so there are no tense consistency errors.
- Preposition misuse: The preposition 'at' is correctly used to indicate proficiency in some area, in this case, mathematics.

Original Sentence: she was cooking and eat.
Corrected Sentence (T5): She was cooking and eating.
Feedback: The sentence: "She was cooking and eating," does not contain any grammatical errors based on those areas mentioned.

- There's no incorrect use of articles. The sentence doesn't require any articles "a," "an," or "the."
- There's correct subject-verb agreement. The past continuous tense verbs "was cooking" and "was eating" align properly with the singular subject "she."
- Tense consistency is maintained. Both actions ("cooking" and "eating") are in the past continuous tense.
- There's no misuse of prepositions. The sentence doesn't include any prepositions.

Screenshots of outputs and code snippets



The image shows a code editor window with a dark theme. The code is written in Python and is a while loop that takes user input, checks for an exit condition, and processes the input to provide feedback. The output terminal below the code shows the execution of the program with a sample input sentence and the resulting feedback.

```
while True:
    # Take input from the user
    a = input("Enter an incorrect sentence (or type 'exit' to quit): ")

    # Exit condition
    if a.lower() == "exit":
        print("Exiting the interactive prompt.")
        break

    # Process the input and provide correction and feedback
    corrected_sentence, feedback = correct_and_provide_feedback(a, model_name='t5')
```

Enter an incorrect sentence (or type 'exit' to quit): she was cooking and eat.
Original Sentence: she was cooking and eat.
Corrected Sentence (T5): She was cooking and eating.
Feedback: The sentence 'She was cooking and eating.' is grammatically correct. Here's

- Incorrect use of articles: The sentence does not require the use of any articles (a,
- Subject-verb agreement: The subject 'She' properly agrees with the verb 'was cooking
- Tense consistency: The sentence is consistent in the past continuous tense; 'was coo
- Preposition misuse: There are no prepositions in this sentence, hence there's no mis

There are no suggestions for improvement because the sentence is grammatically accurat
=====

Enter an incorrect sentence (or type 'exit' to quit): exit
Exiting the interactive prompt.

References

1. Center for Language and Speech Processing @ JHU. (n.d.). JHU-CLSP/jfleg · datasets at hugging face. jhu-clsp/jfleg · Datasets at Hugging Face. <https://huggingface.co/datasets/jhu-clsp/jfleg>
2. T5 Model Paper Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." Journal of Machine Learning Research (JMLR).
<https://arxiv.org/abs/1910.10683>
3. Hugging Face Transformers Library:
Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Brew, J. (2020).
"Transformers: State-of-the-Art Natural Language Processing." EMNLP 2020: System Demonstrations.
<https://github.com/huggingface/transformers>
4. OpenAI GPT Documentation:
OpenAI. (2023).
"API Reference: Chat Completions Endpoint." <https://platform.openai.com/docs/>
5. Evaluation Metric - GLEU Score:
Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016).
"Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation." <https://arxiv.org/abs/1609.08144>