Project Report

Shawn C. Pan

CS 6320.001

1. URLs:

Main demo:

https://youtu.be/JfUjIlrOqXI

Extended contents:

https://youtu.be/PI0NBitx5jU

Github Repo:

https://github.com/cpshawn/CS6320_project

2. Summary

In this project, I created an AI-assisted tool to detect and fix typos and grammar issues in user-input sentences. The tool is based on fine-tuned OpenAI model 'gpt-4o-2024-08-06', and the training dataset comes from C4_200M Synthetic Dataset for Grammatical Error Correction [1].

3. Design

Environment: Python 3.11

Although the original OpenAI models already can do a good job fixing grammar and typo issues. To improve its accuracy, I applied C4_200M dataset to fine-tune the model. C4_200M dataset consists of a series of examples, each containing an input (original sentence, with grammar issues or typos), and a corresponding output (corrected sentence).

I took a total of 500 examples and split them into 80% of training data (400) and 20% of testing data (200), based on the suggestion of multiple resources [2][3][4]. The training dataset will be applied for fine-tuning, and we can either test it on testing dataset, or user-input sentences.

To evaluate the model performance, I compare the output result with both the original sentence and target sentence (if testing on dataset, the target sentences are provided). I will evaluate four aspects: grammar similarity, semantic similarity, sentence contradiction and Keyword Overlap.

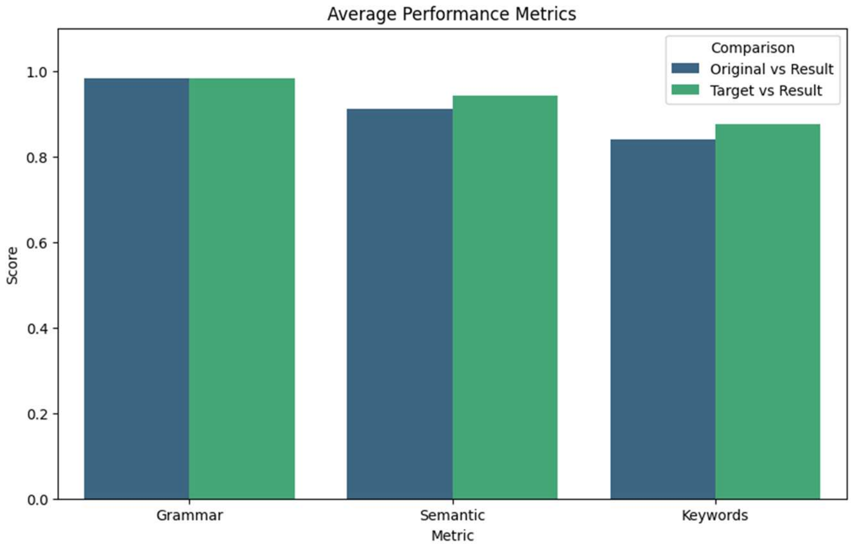
(1) **Grammar similarity:** The first and foremost is grammar similarity. We want to know how close the two sentences are, in sentence structure and grammar.
(2) **Semantic similarity:** To know if the two sentences show the same meaning, I use SentenceTransformer's all-MiniLM-L6-v2 model to measure their cosine similarity.
(3) **Sentence contradiction:** It is not enough to know two sentences have close words and structures. For example: 'Cat eats mouse' and 'Mouse eats cat' are grammatically and semantically correct, but their meanings are in contradiction, which can be detected by this step with roberta-large-mnli model for contradiction likelihood score.
(4) **Keyword Overlap**: In case the model changes the original sentences too much, I calculate keyword overlap between the two sentences. It can guarantee that corrections don't drop or replace critical words.

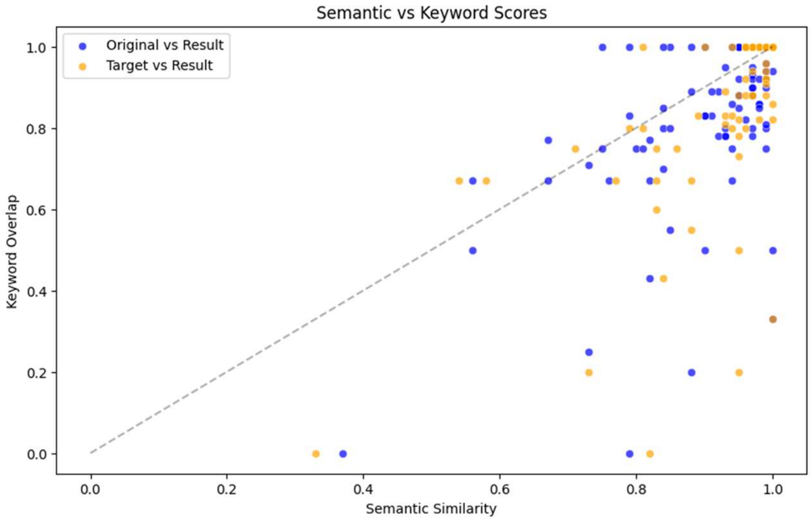
All four scores range from 0 to 1. Grammar, Semantic and keyword Overlap are better with higher scores, while Contradiction Risk is lower with a lower score. The results will be saved into a csv table and can be plotted by `plot_performance.py` for a better view.

4. Result

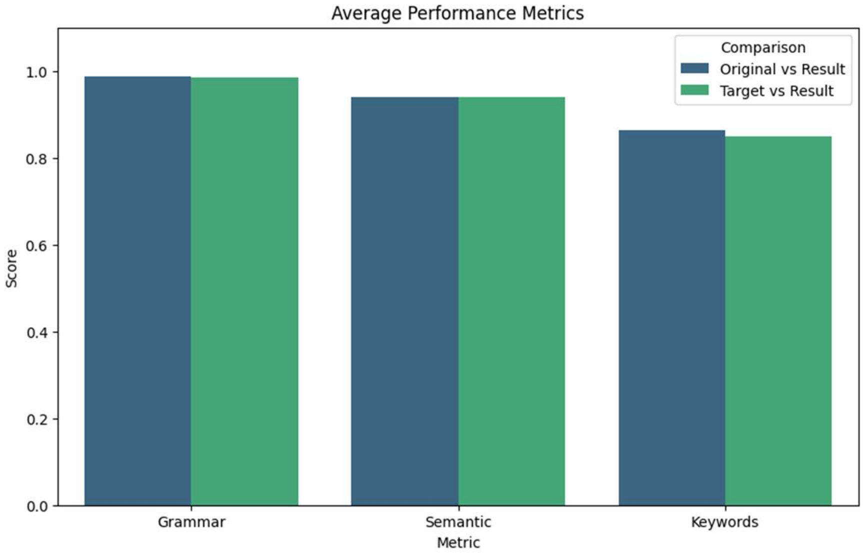
For testing datasets from C4_300M, I did a total of 5 tests. The model is only trained once with the first training dataset. But each time of the test, I randomly generated 100 testcases, and the outputs are attached with this report. Below are some graphs from the testcases:

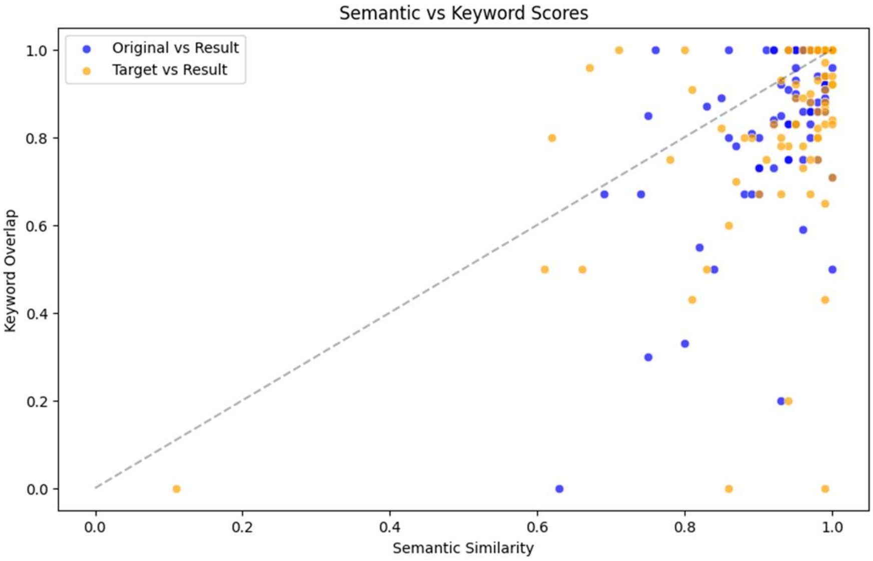
(1) Test 1:


Average Performance Metrics
Comparison
Original vs Result
Target vs Result
Score
1.0
0.8
0.6
0.4
0.2
0.0
Grammar
Semantic
Keywords
Metric

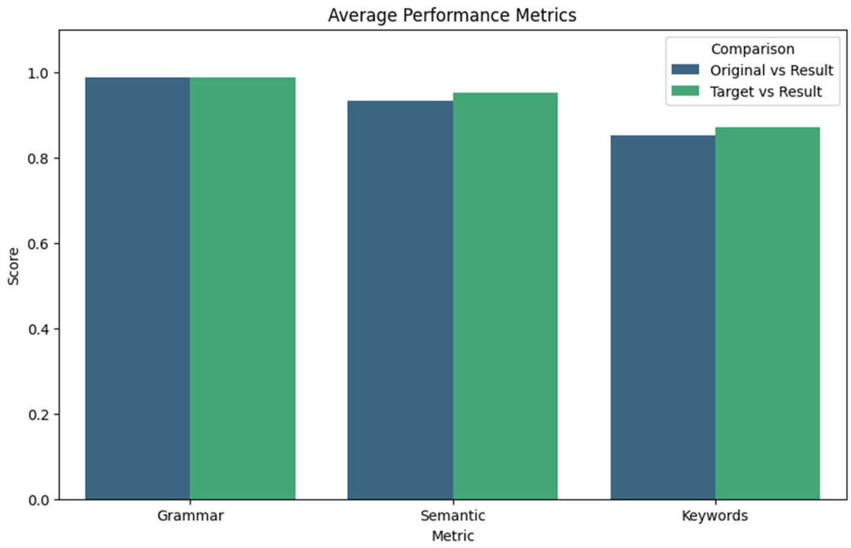

Semantic vs Keyword Scores
Original vs Result
Target vs Result
Keyword Overlap
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Semantic Similarity

(2) Test 2:

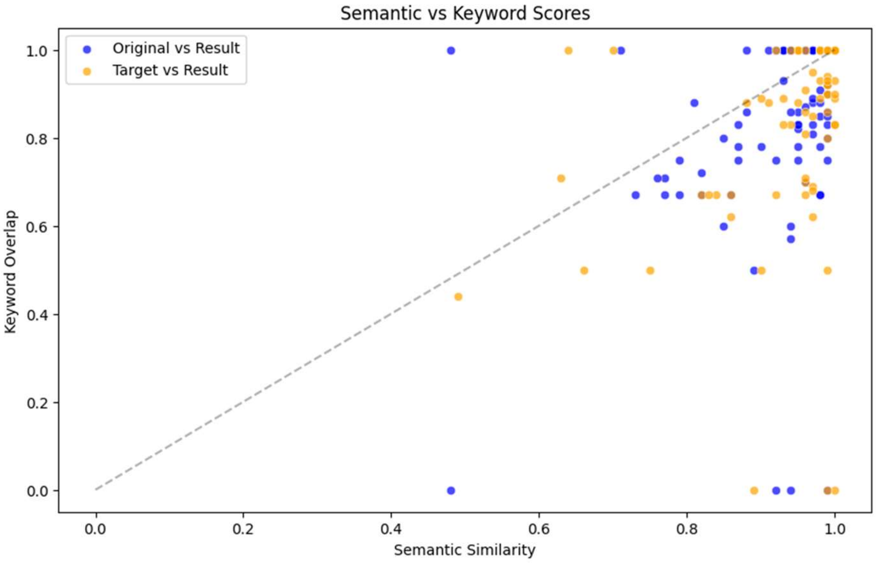

Average Performance Metrics
Comparison
Original vs Result
Target vs Result
Score
1.0
0.8
0.6
0.4
0.2
0.0
Grammar
Semantic
Keywords
Metric


Semantic vs Keyword Scores
Original vs Result
Target vs Result
Keyword Overlap
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Semantic Similarity

(3) Test 3:

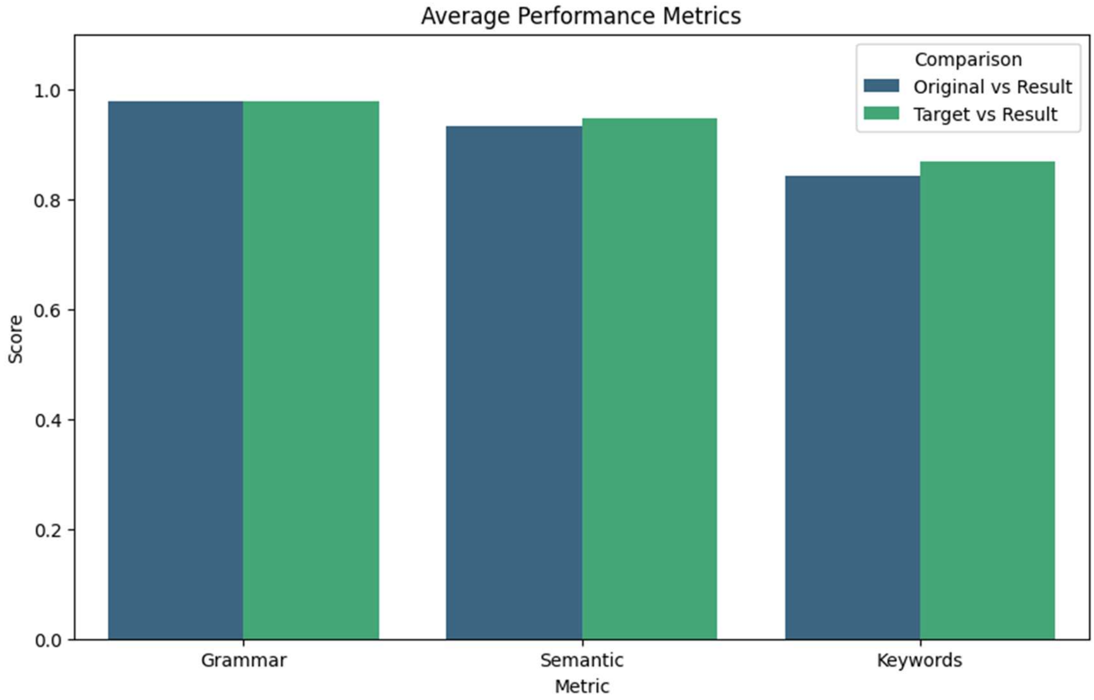


Average Performance Metrics
Comparison
Original vs Result
Target vs Result
Score
1.0
0.8
0.6
0.4
0.2
0.0
Grammar
Semantic
Keywords
Metric

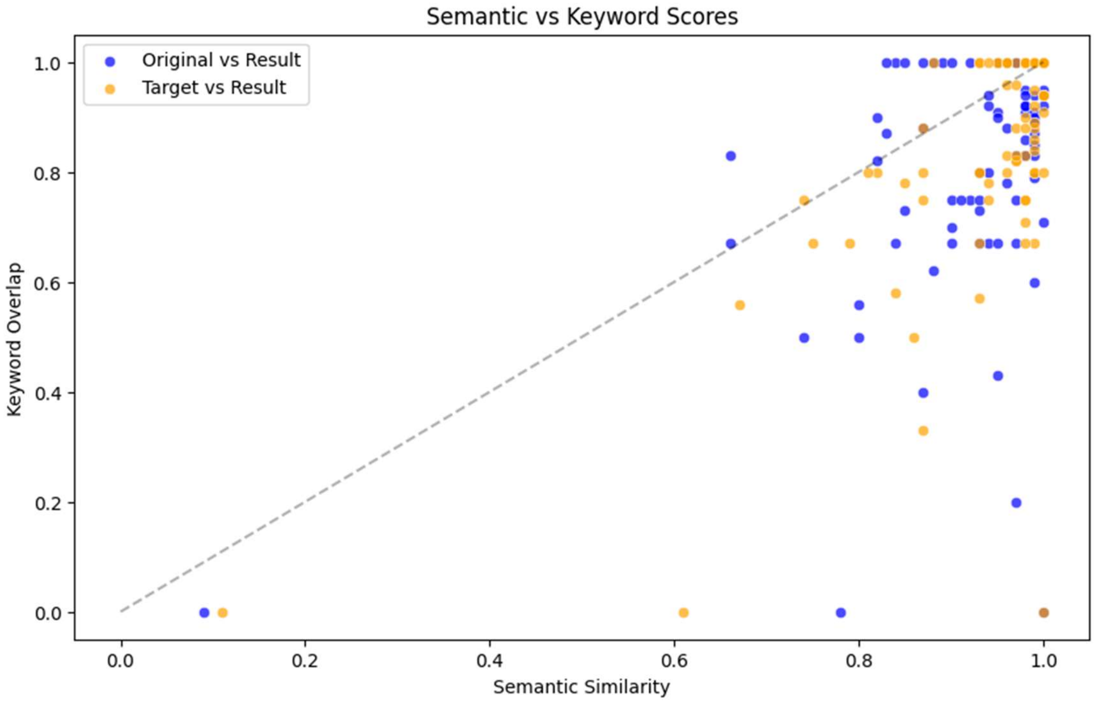


Semantic vs Keyword Scores
Original vs Result
Target vs Result
Keyword Overlap
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Semantic Similarity

(4) Test 4:

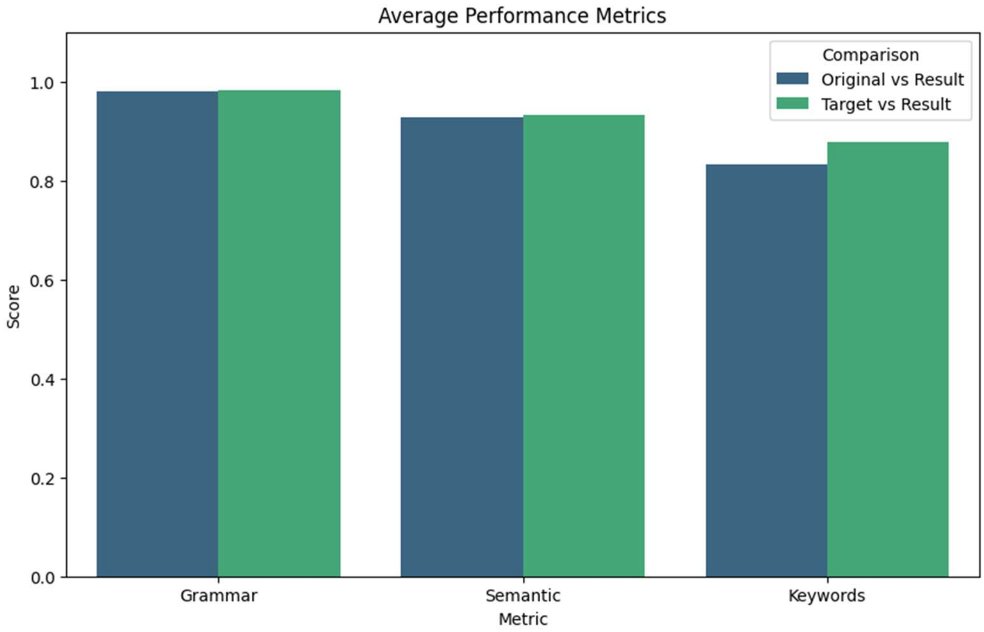


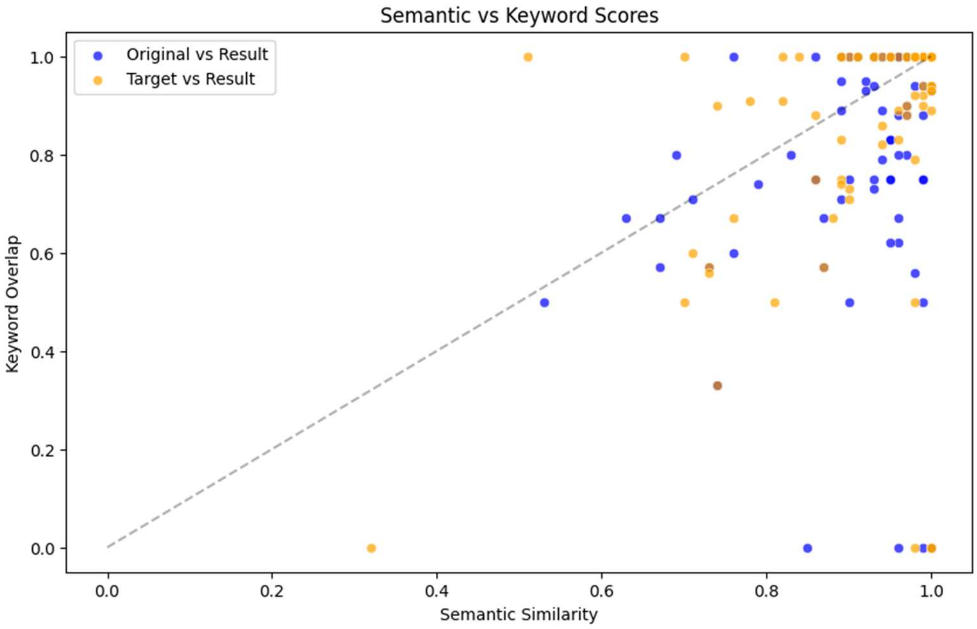
Average Performance Metrics
Comparison
Original vs Result
Target vs Result
Score
1.0
0.8
0.6
0.4
0.2
0.0
Grammar
Semantic
Keywords
Metric



Semantic vs Keyword Scores
Original vs Result
Target vs Result
Keyword Overlap
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Semantic Similarity

(5) Test 5:





From the graphs, we can observe:

Despite the examples selected, the model has an overall good performance on grammar and semantics. It improves semantic similarity while keeping a high keyword overlap, which means that the model does not change original sentence meanings. The model also has no contradiction risk, making the grammar correction process completely reliable.

Resource

[1] C4_200M Synthetic Dataset for Grammatical Error Correction: https://www.kaggle.com/datasets/felixstahlberg/the-c4-200m-dataset-for-gec

[2] https://pmc.ncbi.nlm.nih.gov/articles/PMC11419616/

[3] https://developers.google.com/machine-learning/crash-course/overfitting/dividing-datasets

[4] https://docs.aws.amazon.com/machine-learning/latest/dg/splitting-the-data-into-training-and-evaluation-data.html