Personalized Writing Coach

Ankit Mahto

About the Dataset

JFLEG: English Grammatical Error Benchmark

The dataset was collected to address limitations of existing grammatical error correction (GEC) datasets by emphasizing fluency corrections over minimal edits. It provides multiple human-annotated corrections for each sentence, focusing on fluency and naturalness

For each incorrected sentence, there are 4 corrected versions of it

Models used

T5 Small

Sequence-to-Sequence Transformer

Trained as a text-to-text model, making it ideal for grammar correction tasks

Converts input into an encoded form, then decodes it into output

GPT2

Single-layer decoder model

Can generate fluent and coherent text, making it useful for grammar correction through text completion.

Models Used

DistilBERT

Bidirectional Transformer (Encoder-only)

Can be used in error detection rather than correction (e.g., classifying whether a sentence is grammatically correct or incorrect)

Preprocessing data

Prefix like grammar or grammar correction is added with the data

One of the four corrected versions is selected for the labels

Tokenizer is used to tokenize the words and padding is also applied to it

with tokenizer.as_target_tokenizer() is used to tokenize the labels

Hyperparameters

Epoch set low (3) due to large training time

Different learning rates were tested and the best result was on 2e-5

The beam length during the text generation is set to 5 which yielded the best results

Results and Output

To generate the output, the sentence we give is tokenized with the presaved tokenizer

Model.generate is then used with the appropriate hyperparameters for each model

Tokenizer is again used to decode the generated text which is our output

The model is successfully able to correct the grammar and also pin out where the error in the sentence is

Upcoming changes

Backend and Frontend will be setup for the app

The models will be trained for more epochs to see if we can further enhance their accuracy