Grammar Error Correction using BERT

Sunil Chomal | sunilchomal@gmail.com

Use of BERT Masked Language Model (MLM) for Grammar Error Correction (GEC), without the use of annotated data

Background

Determiners and prepositions are among the most frequent errors made by learners of English, and we aim to correct those.

We use Bidirectional Encoder Representations from Transformers (BERT) as a Language Model
The pre-trained (BERT) model can be finetuned with just one additional output layer to create state-of-the-art models for a wide range of tasks

We would train BERT to detect errors
We would use BERT to suggest corrections & check the correctness of the results

Enablers

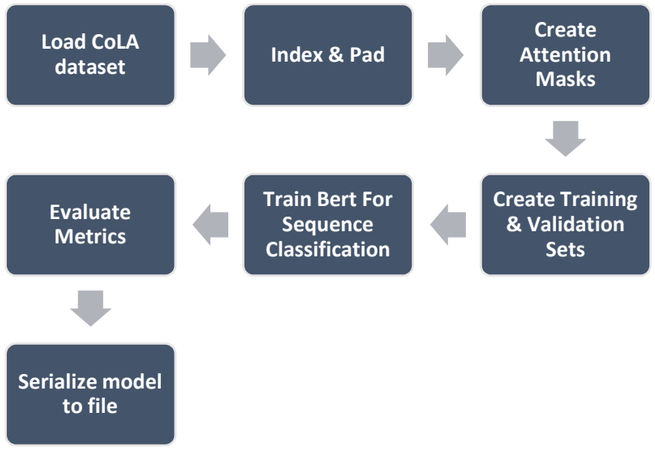
Transformers	Pytorch Transformers
BERT	Spacy
BERT Classification Task	Hunspell
Bert Masked Language Model	CoLA
	CoNLL-2013 Dataset

Source Code

https://drive.google.com/drive/folders/1BoxpDeWZaNC8O3M3222pEAH-aW69sZ14

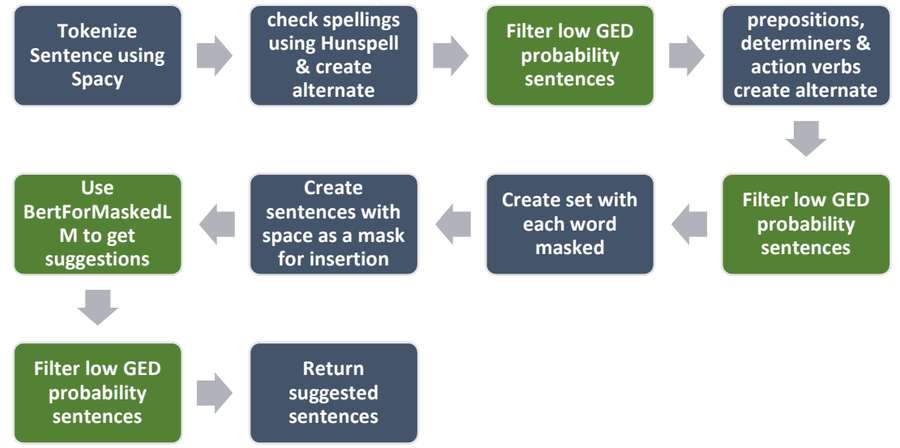
The code has been tested on Google Colab with runtime as GPU. All required files are downloaded by the code, and no additional uploads are required.

Grammar Error Detection



- We use the *bert-base-uncased* as the pre trained model. It consists of 12-layer, 768-hidden, 12-heads, 110M parameters and is trained on lower-cased English text.
- For fine-tuning we have used CoLA dataset for single sentence classification.
- BertForSequenceClassification is a BERT model transformer with a sequence classification/regression head on top (a linear layer on top of the pooled output).
- We trained the network for 4 epochs, and on Google Colab with a Tesla K80 GPU, it takes about 25 minutes.
- After training we get a training loss of 0.1 and a validation accuracy of 0.81.
- Using the out of domain validation data to calculate the Matthews correlation coefficient, we achieve a value of 0.44

Grammar Error Correction



- Tokenize the sentence using Spacy
- Check for spelling errors using Hunspell
- For all preposition, determiners & helper verbs, create a set of probable sentences
- Create a set of sentences with each word "masked", deleted or an additional determiner, preposition or helper verb added
- Used BERT Masked Language Model to determine possible suggestions for masks
- Use the GED model to select appropriate solutions

<u>Tweaks</u>

- BERT MLM would suggest alternate words for existing nouns. The resultant sentence would have a valid grammar, but that is usually not the purpose of this exercise.
- In case of nouns, we SequenceMatcher from python difflib to only allow suggestions which are similar to the word being replaced.
- We restrict addition & deletion for only prepositions, determiners & helper verbs
- From the logits at the softmax layer, we calculate the probability of the sentence being grammatically correct, and use that to filter out the possible suggestions

Results

Source	*CSCI S-89A Lecture Notes*
Original Sentence	They drank the pub.
Reference Response	-
Our suggestions	they drank at the pub.

Source	*Language Model Based Grammatical Error Correction without Annotated Training Data [6]*
Original Sentence	**I am looking forway to see you soon.**
Reference Response	I am looking forward to see you soon. I am looking forward to seeing you soon.
Our suggestions	I am looking for a way to see you soon. I am looking at Norway to see you soon. **I am looking forward to seeing you soon.** **I am looking forward to see you soon.** Am I looking forward to see you soon.

Source	*The CoNLL-2013 Shared Task on Grammatical Error Correction – m2scorer – examples - score_gold [1]*
Original Sentence	The cat sat at mat. Giant otters is an apex predator.
Reference Response	Giant otters are an apex predator. The cat sat on the mat. / The cat sat at the mat.
Our suggestions	**giant otters are an apex predator** **the cat sat at the mat**

Source	*The CoNLL-2013 Shared Task on Grammatical Error Correction – official-preprocessed [1]*
Original Sentence	There is no a doubt, tracking system has brought many benefits in this information age.
Reference Response	There is no doubt, tracking systems have brought many benefits in this information age.
Our suggestions	there is no doubt, the tracking system has brought many benefits in this information age. there is no doubt, tracking the system has brought many benefits in this information age.