CONTEXTUAL SPELL PATCHER USING A SEQ2SEQ MODEL

•••

CS6030 NLP PROJECT

GOKUL S - 2018103026

STEVEN FREDRICK GILBERT - 2018103071

Introduction

- This project aims at performing contextual spelling correction of misspelled words.
- Contextual spelling corrections are more challenging and complex as the relationship between the word to be corrected and the surrounding context must be understood to make the correction.
- The type of corrections we are dealing with here are typographical i.e misspelt words which include missing, extra or wrong characters in a word.

Objectives

- To use a Seq2Seq model with an added Bahdanau Attention layer to better improve the performance over a conventional encoder-decoder architecture to understand the contextual information of sentences.
- Using the learnt contextual information, we employ the trained model to predict the correct spelling of the word based on the context of the given sentence.

DATASET

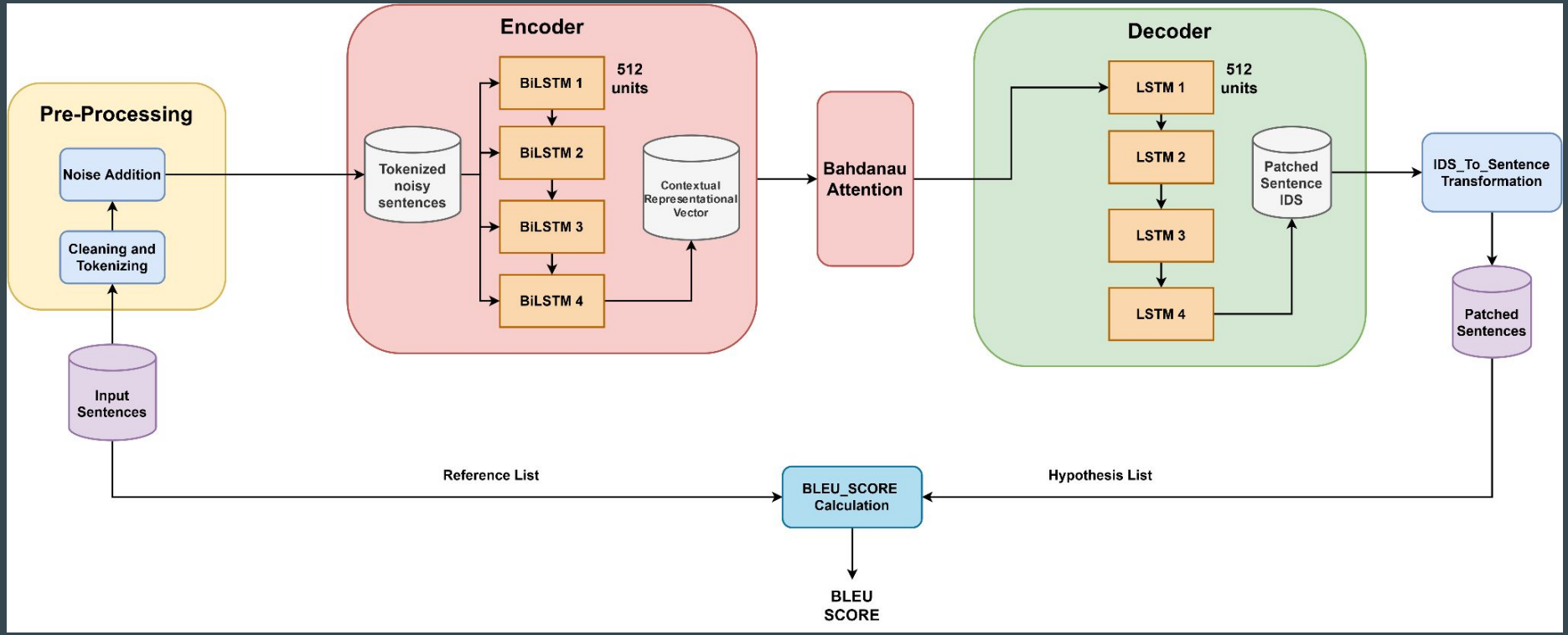
Project Gutenberg

- Project Gutenberg is a digital collection of full texts of books both classic and modern.
- These books will provide a diverse set of sentences for the model to train on to learn contextual information of words.
- 20 popular books have been used to train the model.

  Eg: Sherlock Holmes, David Copperfield, Frankenstein.

- There are a total of 132,287 sentences with 2,993,652 words available.

Methodology Diagram



OUR CONTRIBUTIONS

- We've used a **Bahdanau self-attended decoder** that weighs each word with its importance instead of a conventional decoder that looks at every single word in the sentence equally.
- We've used a **bidirectional LSTM encoder** which has significant improvements over previous models that uses a unidirectional LSTM encoder because it considers the context of a particular word in a sentence from both directions.
- Instead of a single BiLSTM/ LSTM layer, we've used a **stacked BiLSTM/LSTM architecture consisting of 4 layers**.

Graphs

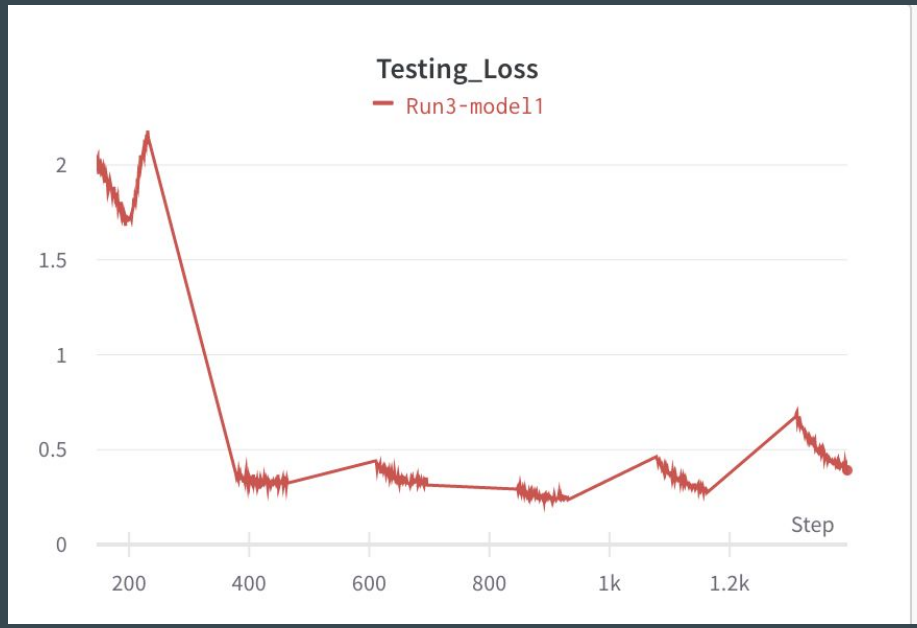
Model 1 - 2 unidirectional layers with no Attention mechanism

Graphs

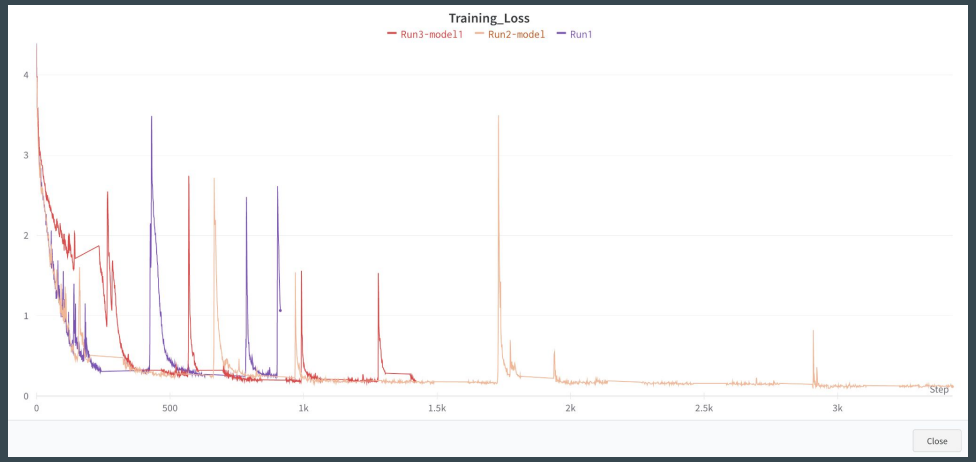
Model 2 - 2 unidirectional layers with Bahdanau Attention

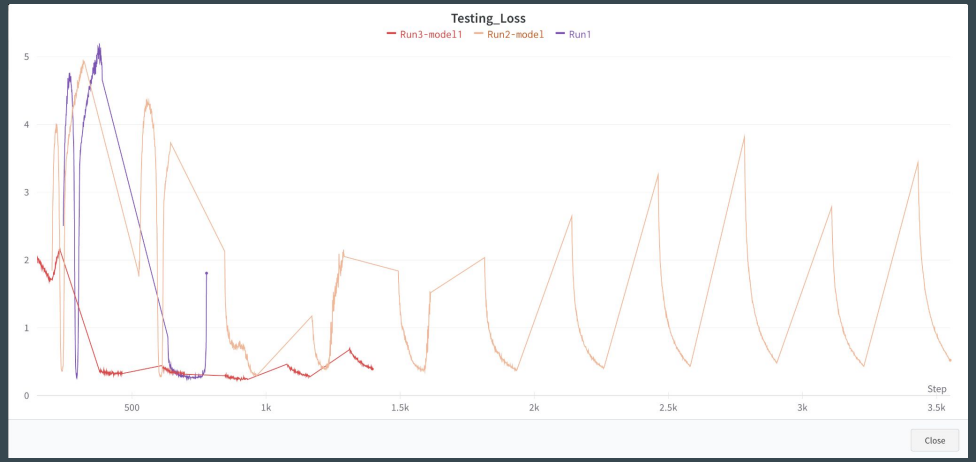
Graphs

Model 3 - 4 Bidirectional layers with Bahdanau Attention



Model Comparison





Performance

Corpus BLEU Score: Implemented by comparing n-grams of the candidate with the n-grams of the reference translation and count the number of matches. These matches are position-independent. The Weight_Vector is [0.25, 0.25, 0.25, 0.25].

```
In [422]: reference = ids_to_sentences(testing_sorted)

In [426]: hypothesis = evaluate(testing_sorted)

In [448]: from nltk.translate.bleu_score import corpus_bleu
          reference_list = list(map(lambda x: x.split(" "),reference))
          hypothesis_list = list(map(lambda x: x.split(" "),hypothesis))

          print(corpus_bleu(reference_list,hypothesis_list))

          0.825686517089071
```

Model	Corpus BLEU Score
1	0.7536534
2	0.7894657