Character Tagging Model

Apurv Kumar (14CS10006) - group leader
Shubham Sharma (14CS30034)
Aniket Choudhary (14CS30004)
Rameshwar Bhaskaran (14CS30027)
Asket Agarwal (14CS30006)

Overview

Objective

Objective is to predict the corresponding phonemes in Sanskrit from the data dcs_words_sentence_wise.txt represented as english words with capital characters denoting different phonemes in Sanskrit.

Preprocessing

1. Converted each sentence in dcs_words_sentence_wise.txt to a sentence containing lowercase letters and storing them in inp.txt.
   Ex: paYcan ratna muKya ca uparatna catuzwaya → paycan ratna mukya ca uparatna
   Catuzwaya

2. Created a training test file with each line containing words in lowercase and corresponding uppercase space separated letters.
   Ex: suramfttika s u r a m f t t i k A

3. Clean training and test data to remove erroneous words like words containing non ascii characters etc.

Methods Used

Used g2p seq2seq toolkit to train and evaluate the test data.
The tool does Grapheme-to-Phoneme (G2P) conversion using recurrent neural network (RNN) with long short-term memory units (LSTM). LSTM sequence-to-sequence models were successfully applied in various tasks, including machine translation and

Grapheme-to-phoneme. This implementation is based on python TensorFlow.

Train the model using following command :

```
g2p-seq2seq --train train.dict --model model_folder_path
```

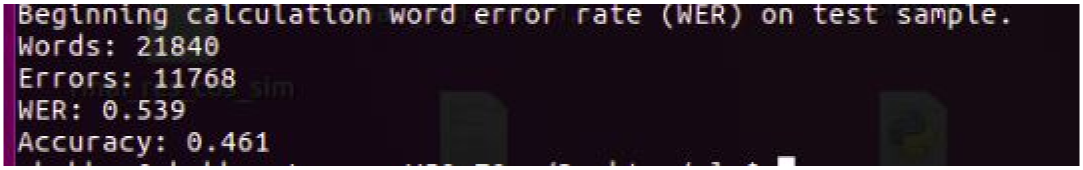
Evaluate on the test data using following command:

```
g2p-seq2seq --evaluate test1.dict --model model_folder_path
```

Results

For each line in the test data, we found the unique words and the model will evaluate by testing the phonemes sequence of how many words it can correctly predict. Word accuracies is as follows:

WER(Word Error Rate) → 0.539
Accuracy → 0.461



```
Beginning calculation word error rate (WER) on test sample.
Words: 21840
Errors: 11768
WER: 0.539
Accuracy: 0.461
```