Name: Avirat Belekar

CWID: 10454332

Course Name: CS584-A Natural Language Processing

Assignment 5 : Machine Translation

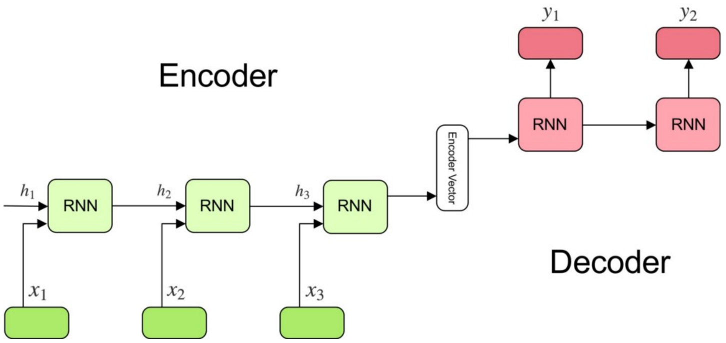
1. Encoder-Decoder model without attention

Problem Statement:

To translate Czech sentences into English sentences using Sequence to Sequence

Model:

.



Encoder:

A stack of several recurrent units (LSTM or GRU cells for better performance) where each accepts a single element of the input sequence, collects information for that element and propagates it forward.

Decoder:

A stack of several recurrent units where each predicts an output y_t at a time step t. Each recurrent unit accepts a hidden state from the previous unit and produces and output as well as its own hidden state.

Model Summary:

Encoder Model Summary:

First layer = GRU

Second Layer = GRU

Hidden Units = 12

Epochs = 25

Batch_size = 256

Opitmizer = Adam’s optimizer

Loss = Categorical cross entropy

Decoder Model Summary:

First layer = GRU

Second Layer = GRU

Hidden Units = 256

Epochs = 17

Batch_size = 256

Opitmizer = Adam’s optimizer

Loss = Categorical cross entropy

Activation = softmax

Decoder with Attention Model Summary:

First layer = GRU

Second Layer = GRU

Hidden Units = 256

Epochs = 17

Batch_size = 256

Opitmizer = Adam’s optimizer

Loss = Categorical cross entropy

Activation = softmax

Dataset and Experimental setup:

Data was collected from the European Parliament Proceedings Parallel Corpus 1996-2011.

The downloaded pair of language is Czech-English.

Since the dataset is too large I have cut the dataset into 100 lines so that the model doesn’t give a memory error.

Dataset summary :

645 English words.

268 unique English words.

10 Most common words in the English dataset:

"of" "(vote)" "the" "Minutes" "see" "and" "for" "on" "European" "sitting"

632 Czech word.

315 unique Czech words.

10 Most common words in the czech dataset:

"(hlasování)" "viz" "zápis" "a" "na" "o" "Dohoda" "pro" "(kodifikované" "znění)"

Results:

Encoder Decoder model without Attention	Bleu Score 8.96
Encoder decoder model with Attention	Bleu Score 9.53

Czech : na a a hlasování

Ground-truth English: on and on voting

Translation from seq2seq model: on on voting

Translation from seq2seq plus attention: on and on voting