CHATBOT

Sandesh Maurya , Sanjeev Ranjan , Gaurav Dahiya
Team 04

November 26, 2018

1 Introduction and Overview

In this project we have built a simple textual chatbot , which is a sequence to sequence model with attention decoder. An encoder is an utterance by the user and the decoder is the response to that utterance. The bot requires dataset for training itself. We have used the Cornell Movie-Dialogs Corpus for training our bot. The training dataset has been pre-processed as it contains some data which is useless for the model.

Related work

This is based on Neural Machine Traslation.

2 Methods

We have used the sequence to sequence model for building the main algorithm. This model can divided into two small sub-models. The first sub-model is called as Encoder(E), and the second sub-model is called as Decoder(D). E takes a raw input text data just like any other Recurrent Neural Network(RNN) architectures do. RNN takes multiple sequence as input and produces a multiple sequence output. At the end, E outputs a neural representation. The output of E is the input data for D. Output of E is in some encrypted form and D has the ability to decrypt this output, producing a totally different output data. Following are the steps used to build the model -

(1) Define input parameters to the encoder model.
(2) Build encoder model.
(3) Define input parameters to the decoder model.
(4) Build decoder model for training.
(5) Build decoder model for inference.
(6) Put (4) and (5) together to build the decoder model.
(7) Connect encoder and decoder models.
(8) Define loss function, optimizer, and apply gradient clipping.

3 Experimental Analyses

Datasets

Dataset used by us - Cornell Movie-Dialogs Corpus

Other available datasets -

(1) Twitter Chat Log
(2) More Movie Subtitles
(3) Every publicly available Reddit comments

Results

The training took a lot of time, but the output was not satisfactory. We trained our model for 3 days still it was able to train only 20 percent of the data. The responses of the bot are based only on the trained part and are quite commom for different queries. The responses are often gibberish. Some resposes are shown below -

Q : happy birthday have a nice day
A : thank you a lot

Q : what is your name?
A : i dont know that

Q : my name is sandesh
A : that is not cute a lot

4 Discussion and Future Directions

Our aim was to train the model completely and further enhance it. But as the training of the basic model could not complete, there was no scope for further improvements. To improve the performance we would have used the conversation between human and bot as another training dataset for training the bot.

References

[1] Siraj Raval Sequence to Sequence Model

[2] Tensorflow Tutorials

[3] Neural Machine Translation