

Long Homework 4

Seq2Seq Chatbot

November 16, 2018; due December 2, 2018 (11:59pm)

In this homework you will be creating a Chatbot using a sequence-to-sequence model. You are allowed to work in groups of up to 2 students. This homework is of an open format; all we will be providing you with is the data. It is up to you to pre-process the data, build the seq-2-seq model with keras, and train the model. You will be submitting your code and write-up containing the three sections defined below.

1 Data

The data you will be training on is movie subtitles from the Cornell Movie-Dialogs Corpus. This corpus contains a rich collection of fictional conversations extracted from raw movie scripts:

- 220,579 conversational exchanges between 10,292 pairs of movie characters
- involves 9,035 characters from 617 movies
- in total 304,713 utterances

`movie_lines.tsv` contains the data. The file has five tab separated columns containing the following fields:

1. lineID
2. characterID (who uttered this phrase)
3. movieID
4. character name
5. text of the utterance

You may not need to use all the columns. There are various ways you can go about segmenting the turns. The simplest is to treat each turn as being uttered by different characters. You can also clean the dialogue and remove utterances that do not seem to be conversations.

2 Write-up

2.1 Data Pre-processing

We are only providing you with the utterances. It is up to you to process the utterances into a format that can be fed into the model. For information on how to process the data take a look at this paper: [A Neural Conversation Model](#). Some ideas for pre-processing may include removing infrequent words and replacing them with the <UNK> token. In your write-up include what you did to pre-process the dialogue data.

2.2 Model

You will be implementing the sequence-to-sequence model described in class where the model makes predictions using the left context and the dialogue context. More information model can be found in these [lecture slides](#), this [paper](#), or the [reading](#).

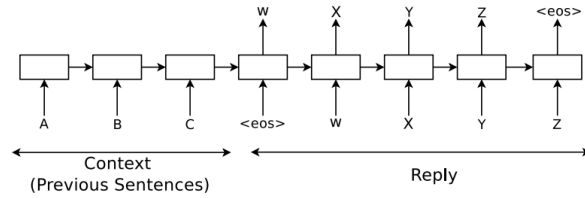


Figure 1: Sutskever et al. (2015); Vinyals and Le (2015)

We've slowly introduced you to Keras over the course of the semester. Now it is time for you to implement a model without any starter code. Keras provides an easy way to build a complex model in few lines of code. You can try building the model yourself reading documentation, but here is a good sequence-to-sequence tutorial that you can follow and adapt to the task: [Seq-2-seq Tutorial](#). You do not need to write your code in an ipython notebook. In your write-up describe the architecture of your model.

2.3 Sample Dialogue

To see how your chatbot functions you are required to have a conversation with your chatbot using the prompts below. Input the human utterances from the following two conversations into your model, replace the machine responses with your model's responses, and include the conversations with your model's responses in your write-up. Below are some sample dialogues of such human-chatbot conversations.

General Q&A

Human: my name is david. what is my name?
Machine: david.
Human: my name is john. what is my name?
Machine: john.
Human: are you a leader or a follower?
Machine: i'm a leader.
Human: are you a follower or a leader?
Machine: i'm a leader.

Morality

Human: what is moral?
Machine: what empowered humanity , what intellectual the essence is.
Human: what is immoral?
Machine: the fact that you have a child. **Human:** what is morality?
Machine: what is altruism?
Human: what is the definition of altruism?
Machine: if you don 't believe in god , then you don 't know.
Human: ok ... so what is the definition of morality?
Machine: well , the truth is , you 're not a believer in god almighty.
Human: tell me the definition of morality , i am quite upset now!
Machine: i 'm not ashamed of being a philosopher!

3 Deliverable

- Write-up describing data processing/featurization, model, and sample dialogue
- Code containing model, with a README giving instructions on how to run the code