Proposal for a Conversation Chatbot using Cornell Movie Dialogs

Natural Language Processing

Fall 2019 Section 56

John Barker

Northwestern University

Introduction

The Presidential Historical Society is requesting our team to implement a prototype of a conversation chatbot to construct a foundation and test feasibility for a question answering chatbot with a knowledge base of the presidential corpus. They want us to use the Cornell movie dialog corpus to build the conversation chatbot. This will provide a well-structured corpus to allow for experimentation of the encoder and decoder components and a sequence to sequence model (seq-to-seq) From this study the society wants to determine the type of computer systems and the storage facilities to handle the data they hope to encompass in the virtual teacher.

Data Preparation

The Cornell Movie Dialogs consist of over 220,000 conversational exchanges which will be used to build a conversational chatbot. The dialog will be parsed into input output pairs representing a response reply interaction between two characters from a movie to train the chatbot in the art of conversation. The length of the dialog lines will be studied in order to determine key statistics about their size to limit padding in the input and output pairs. Bucketing may be employed to limit the padding if there are disparities in the line length. The dialog will be selected based on a reduced character set dialog 100 or fewer tokens. From the input text and the output text a unique char vocabulary set will be used to improve training times.

Review Research and Methods

A thorough search of the internet for other research using the Cornell movie dialogue will be done to find the best methods of creating a conversation chatbot. “Natural Language Processing In Action” will be a vital source in implementing the conversation chatbot. An encoder decoder architecture will be used to tokenize the input into thought vectors and then use

the thought vector as input to the decoder to generate response to the original input. The LSTM neural network will be used to configure the encoder and the decoder. The model will be trained using the movie corpus data. The data will split 20 percent of the data for validating purposes during training.

Review Results Evaluate Models

The model will be tested using various hyper parameter combinations. Different optimizers will be used, different epoch sizes will be used, different network layer and neuron counts will be used to train the model. Once the model is trained, It will be used to predict outcomes in order to see how effective the training was at generating good responses from the chatbot. A conversation will then be created between the chatbot and input dialog to see how much of a conversationalist the chatbot has become.

Implementation and Programming

The conversation chatbot will be implemented using Python, Keras with TensorFlow as the backend and the Cornell Movie Dialogue as the input. The encoders and decoders will be built using the LSTM neural network. The pipeline will follow a flow of data preparation, parameter setup, encoder, decoder configuration, model training and finally model predictions.

Conclusion

Management wants to know if it is possible to build a conversational chatbot which can be used as the start of a knowledge chatbot for the presidential corpus We will build a chatbot infrastructure using the encoder decoder architecture and seq to seq embedding. We need to find out if we have enough processing power to effectively train and retrain our chatbot as we grow the chatbot’s knowledge base. This should give management the proper impetus to move forward with the question answer chatbot.

References

Northwestern School of Professional Studies (2019) Movie dialogues for chatbots. Retrieved from https://canvas.northwestern.edu/courses/101742/pages/movie-dialogs-for-chatbots?module_item_id=1263816

Geron, A. (2017). Hands-on machine learning with scikit-learn and tensorflow. Schastopol, CA: O'Reilly Media, Inc.

Lane H, Cole H., Hapke H. (2019). Natural language processing in action. Shelter Island, NY: Manning Publications, Co.