Team 3: Final Project

Cesar Lucero Ortiz, Akram Mahmoud & Jeremy Cryer

Department of Applied Artificial Intelligence, University of San Diego

AAI520: Natural Language Processing

Dr. Roozbeh Sadeghian

Oct 23, 2023

Abstract

Team 3: Final Project

Introduction

What is a chatbot? As stated in this article, a chatbot is at states, “At a technical level, a chatbot is a computer program that simulates human conversation to solve customer queries. When a customer or a lead reaches out via any channel, the chatbot is there to welcome them and solve their problems. They can also help the customers lodge a service request, send an email, or connect to human agents if need be” (Shweta, 2023). Now, with this understanding, the goal for the chatbot being designed and discussed in this technical report is to be capable of comprehending the Cornell Movie-Dialog Corpus data that was obtained using Kaggle[[1]](#footnote-1). Now during this report, the methodology for designing the chatbot will be discussed, with evaluation results, along with challenges faced and future improvements recommended to be added in an additional iteration.

Methodology

In this part of this technical report, the methodology of this chatbot will be discussed. The architecture of the model and how it came to proceed in that direction will all be discussed. How the model was trained, and the evaluation results of the model will also be explored. These are critical to go into detail with, as understanding this will help anyone proceeding with a second iteration fully understand how to go about creating the model and might give insight on where to adapt and adjust if needed.

Pre-Processing

First things first, like all models you need to start with uploading your dataset you plan to use for training and testing. As mentioned earlier, multiple options of datasets were provided, and this group chose to use the movie-dialog one which can be obtained through Kaggle like so many other datasets to be used for models like this. Now this dataset came with four text files full of useful data but to get the model rolling, it was decided to use just the text file full of various movie lines. Now when uploading the raw dataset, we had to ensure to add specified separators to split the data columns correctly. Then the first preprocessing method used was to adjust all words in the dialog line column to lowercase. At this point there was a choice to either remove or keep punctuation and stop-words. From here it was unanimous to keep the stop words and punctuation to see if they could be trained in, thus allowing better conversation with the chat bot as that is a part of standard language syntax. Before more preprocessing was performed, it was thought to see how much data exists per character to establish an idea for potential bias. However, it was decided for this iteration that no dialog will be removed, to allow as much variation as possible into the model. As seen Figure 1, there is a lot of variation between data count for all characters form the low and high end of the data counts.

Figure 1

A screen shot of a screen

Description automatically generated

*Note: Graph displaying the dialogue count per character.*

Now at this point, the data frame being used only has lowercases implemented and a bunch of useless data columns that don’t mean anything for this problem unless we need to use them in the future to link the other three datasets. So now the data will be reduced to just the character and their dialogue. To do this, some structuring will take place using a taught function, this will join consecutive lines said by each character, assuming everything is in order as stated by the line ID value and reducing the total amount of dialog to be processed as you can see in Image 1.

Image 1.

A screenshot of a phone

Description automatically generated

*Note: Structured dialog data display.*

Then once this was done, the last item to complete prior to model building was to group the data by character. Now this was decided to be done as characters tend to take on a persona that if grouped together should add specific semantics and sentiment to the chat box as it should vary based on the character and their lines.

Model Architecture

It is ready to build the model. Originally it was attempted to use a BERT pre-trained model just due to familiarity with it during this course. However, after some challenges which will be discussed later in this report, a switch to a GPT model was done. This resulted in the GPT2LMHeadModel to by utilized for this modelling solution, along with its corresponding tokenizer which will be included in the model architecture instead of done individually during the preprocessing stage. Along with the tokenizer, a sequence padder was added to fill in the empty tokens. Using the loaded pre-trained model and the tokenizer, the dialog dataset that was preprocessed can be used on a training loop to fine-tune the pre-trained GPT2 model for this set of data. Then post fine-tuning, the model can be saved for use.

Training & Evaluation

After saving the model, it’s time to use it.

Learnings and Recommendations

Challenges

Improvements on Future Iterations

Conclusion

References

Chidananda, R. (2018, March 28). Cornell movie-dialog corpus. Kaggle. https://www.kaggle.com/datasets/rajathmc/cornell-moviedialog-corpus

Shweta, K. (2023, July 28). What is a chatbot? everything you need to know. Forbes. https://www.forbes.com/advisor/business/software/what-is-a-chatbot/

1. https://kaggle.com/ [↑](#footnote-ref-1)