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

Chatbots are the start of AI being able to process thoughts and information as human-like as possible. In this report it will be discussed how a dataset on character movie lines was used to develop a chatbot. This chatbot only ends up being good for a user to utilize for practicing their lines and getting into character. Due to the lack of data outside of the movie lines, deviation from the input lines in the models current trained state wouldn't be much and if so, would result due to errors. For a first iteration, being able to input and output the lines as trained to is best. Future iterations with additional data and teaching would be where this chatbot could potentially evolve into a more robust and adaptive chatbot, however, everything must start somewhere and for something like this, it's a fantastic start. Now as someone goes through this report, it will be revealed the thought process behind the preprocessing, model selection and training. Now for a report it is vague in comparison and would need to go hand and hand with the code this is written too to fully understand. However, from a base comprehension, the reader will learn the mind-set behind the development and it will be revealed what challenges were faced and several recommendations for any future iterations pursued.

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]. 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.

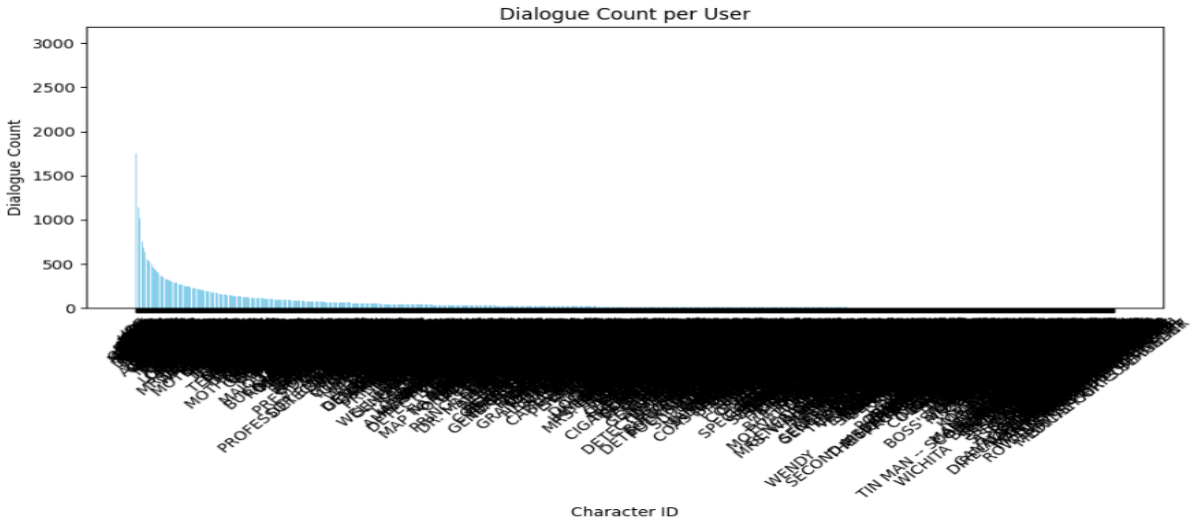
Pre-Processing

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

[1] https://kaggle.com/

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



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.

	character	dialogue
0	BIANCA	did you change your hair?
1	CHASTITY	no.
2	BIANCA	you might wanna think about it
3	PATRICK	i missed you.
4	MISS PERKY	it says here you exposed yourself to a group o...
5	PATRICK	it was a bratwurst. i was eating lunch.
6	MISS PERKY	with the teeth of your zipper?
7	MICHAEL	you the new guy?
8	CAMERON	so they tell me...
9	MICHAEL	c'mon. i'm supposed to give you the tour. so ...
10	CAMERON	north, actually. how'd you ?
11	MICHAEL	i was kidding. people actually live there?

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. To do this you just have to set the model to the saved fine-tuned one we already created and then create a function to generate responses. This function is a multi-loop function that has a response generation and a conversation loop, that way the user can type in a conversation box, and then once a response is generated, the user can respond again. This can go on and on until the user wants to cancel the cell. Now after testing the chatbot out, you can see it is as expected, designed specifically for the movie lines. What was noticed is that if you type out a sentence, even if it's not fully complete. The chatbot will find the

line that is closest to what was input, complete the sentence to what it believes is correct, then respond with the next line that falls next in sequence. Now, after doing this a few times, there are times where the response is given with an incomplete sentence or did not finish punctuation. It was also observed that it does randomize the response, so if someone inputted the word how, it would not go to the same line that starts with how each time, it would go to a different one if one was to exist. Now from an evaluation perspective, there was not enough time to get actual numerical metric values such as a Bleu value. This would have been too tedious to go into the raw data and manually input conversation lines as test values and then compare with the output. Due to time restrictions, the group stuck to manual observations as the chatbot was experimented with.

Learnings and Recommendations

With this chatbot, some challenges and possible improvements for future iterations were identified as the model was built. Here these will be discussed in depth so that the next person to design a similar model can start ahead of any possible issues, to improve their productivity and outcome.

Challenges

While creating this model, the biggest challenge that was faced was inexperience in creating a chatbot in general. Originally, all members were only familiar with BERT models from the previous assignments in the course. What ran into an issue was the way the preprocessed data syntax was set; it was causing issues using the BERT pretrained model based on the structure. After some exploration, it was found the GPT2 pre-trained model was another route that worked better for the way the data was structured. Now the main thing to take away from that is that some research should be performed prior to a modelling, just so the user fully understands the models being planned for use and understands the data format needed to be successful with those approached. Now, another observation which could be a challenge as well in future iterations that was noticed is the line sequencing. Now as seen in Image 2, it can be noticed that not all line IDs are there. It skips IDs from time to time. The challenge that could be identified with this observation is that if there are lines missing, it will make the chatbot not as accurate as the next sequential line may be wrong due to missing data. Now, a cross check was not done on this but would be necessary on any future iterations performed.

Image 2

	Line_ID	Char_ID	Movie	Char	Line
86	49	u0	m0	BIANCA	did you change your hair?
85	50	u3	m0	CHASTITY	no.
84	51	u0	m0	BIANCA	you might wanna think about it
648	59	u9	m0	PATRICK	i missed you.
647	60	u8	m0	MISS PERKY	it says here you exposed yourself to a group o...
...	...	...	...	...	...
304704	666522	u9034	m616	VEREKER	so far only their scouts. but we have had repo...
304679	666546	u9027	m616	CHELMSFORD	splendid site, crealock, splendil i want to es...
304678	666547	u9029	m616	CREALOCK	certainly, sin
304696	666575	u9028	m616	COGHILL	choose your targets men. that's right watch th...
304695	666576	u9031	m616	MELVILL	keep steady. you're the best shots of the twen...

304713 rows × 5 columns

Note: A image of the raw data frame prior to grouping.

Improvements on Future Iterations

With this first iteration, the model is off to a great start, just being able to output complete sentences and function. Most models don't always function on the first attempt, so this is above and beyond for something like a chatbot. Now, on a future iteration it would be recommended to cross-check the line IDs and sequence to make sure that nothing is missing as that could throw off the accuracy. It would also be more beneficial to not only try to explore more data input to add more variation and avoid bias, but also find a way to combine the additional three text files that are associated with this dataset. That would add more complexity and give better semantics and sentiment with the training. It might also be beneficial as experimenting with the chatbot, to go into the raw data and start identifying lines, and seeing if the raw data is incomplete or not and compare with the chatbot output. That way if the chatbot is leaving out words or punctuation, it can be identified and fixed depending on what caused that error. An evaluation metric could also be investigated such as a BLEU or ROUGE score. This will help give the user a better understanding of where and how improvements may be needed.

Conclusion

During this project a chatbot was developed using a GPT2 pretrained model and fine-tuning it to a Kaggle dataset about movie dialogue. Now after some pre-processing, modelling, training, and testing, a chatbot was developed for the movie line data. Now this chatbot isn't anything like ChatGPT mainly due to its lack of variety in available data to respond to. However, for a first iteration, it is a great start to a chatbot that can help someone practice their lines and get into character for a production. There were obviously challenges and learning to overcome in future iterations but is a great start for the beginning of development.

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/