Module 6: Portfolio Milestone

Brady Chin

Colorado State University Global

CSC525-1: Principles of Machine Learning

Dr. Dong Nguyen

April 27th, 2025

Portfolio Milestone

My chatbot project is currently a functional, text-based conversational bot implemented in Python and using the Hugging Face Transformers library. It's a pre-trained GPT-2 model that was fine-tuned on the Cornell Movie-Dialogs Corpus (Danescu-Niculescu-Mizil, C., 2011). The chatbot is able to support simple dialogues and give casual responses in line with the tone of the training data. A command-line interface for testing is provided along with simple fallback responses are also provided in the event that the model gives non-sequiturs or irrelevancies.

Areas for Improvement

The chatbot can still be improved in some areas. First, I want to improve the overall response quality of the chatbot. While the chatbot is able to understand some of the more basic prompts, it is unable to understand more unique or longer prompts. Furthermore, Some of the response that are produces are incoherent.

Planned Improvements

After several tries modifying the training techniques, I looked into the Cornell Movie-Dialog Corpus dataset that I am using to train my model and the conversations here are incoherent and do not make sense. Table 1 shows the dialog for one of the conversations that is in the dataset:

Table 1: Dialog in Cornell Movie-Dialog Corpus dataset

Line Code	Character	Dialog
L194	BIANCA	Can we make this quick? Roxanne Korrine and Andrew Barrett are having an incredibly horrendous public break- up on the quad. Again.
L195	CAMERON	Well, I thought we'd start with pronunciation, if that's okay with you.
L196	BIANCA	Not the hacking and gagging and spitting part. Please.
L197	CAMERON	Okay... then how 'bout we try out some French cuisine. Saturday? Night?

I realized that the dataset that I am using is causing the issues since I am training the model on incoherent responses. Since this dataset has tens of thousands of lines, the cleaning process will not be worth the time as most of the data will need to be rewritten.

To solve this, I am exploring new datasets. I have done some research into the Ubuntu Dialogue Corpus but it has been difficult to find good dataset with dialog. I am leaning towards using the PersonaChat dataset (Jairath, A., 2022). This consists of conversation data and I will only need small modifications to remove the Cornell dataset and implement the PersonaChat dataset. While there is still some conversations that are incoherent, it is definitely an improvement on the current one.

Conclusion

Though my current chatbot project is currently at a basic level of conversation, its shortcomings mostly stem from the source quality of training data. The nonsensical dialogue of the Cornell Movie-Dialogs Corpus has had a direct impact on response relevance and quality. In the future, shifting over to a properly structured dataset such as PersonaChat will rectify these errors and significantly improve the coherence, tone, and flow of conversation of the chatbot for the final submission.

Reference

Danescu-Niculescu-Mizil, C., (2011). *Cornell movie-dialogs corpus.* Cornell University. https://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html

Jairath, A., (2022). *Facebook ai - personachat (8784 examples).* Kaggle. https://www.kaggle.com/datasets/atharvjairath/personachat