ANLY 580 Project Proposal

1. Group Members:

Jieqian Liu, Jieyi Sun, Zidong Xu, Minglei Cai

2. Topic & Background:

Popular approaches for dialogue systems include Sequence-to-Sequence (Seq2Seq) response generation, knowledge graph-driven question answering (KG-QA) , context-sensitive response retrieval and RL methods. There are three kinds of popular chatbots now divided by the usage and principles behind: customer service chatbots that only answer for required keywords, task oriented chatbots that can answer and complete specific tasks, and open domain chatbots that can chat with users without restriction to tasks.

In this project, we will focus on fine-tuning pre-trained dialoGPT models and deploy our open domain chatbot to the public with a virtual phone number thus everyone can have a conversation with our chatbot. Depending on the time, we will decide whether or not to have trials on the other methods.

3. Datasets:

According to our aim, we know we need to collect data that includes multi-turn conversations, so we think maybe the conversational dataset will be a good choice. According to our search, we found the dataset Movie Dialogue Corpus, which contains a metadata-rich collection of fictional conversations extracted from raw movie scripts.
https://www.kaggle.com/datasets/Cornell-University/movie-dialog-corpus

This dataset only includes movies that had more than 5 IMDB (The Internet Movie Database; data interfaces available at http://www.imdb.com/interfaces) votes, which includes 220,579 conversational exchanges between 10,292 pairs of movie characters, 9,035 characters from 617 movies, and a totally of 304,713 utterances.

4. Methods & Main Procedures:

1) Data Cleaning:
   Because the dataset includes multiple tables, so we plan to join these tables at the beginning to get the true utterance in the conversation of the

movie, and then we may use nltk, Gensim, or other methods to do data cleaning, remove stopwords, tokenization and so on.

2) Loading Pre-trained Models and Fine-tuning
   Based on our need to train an open-domain chatbot, our limited amount of data and training resources, we need a pre-trained model, which we will select from open source platforms such as huggingface. At present, the model of the DialoGPT series is a good choice (such as DialoGPT-small / DialoGPT-medium / DialoGPT-largel), and we will use the resources of google colab pro+ to fine-turn the pre-trained model. In addition, although there are many ways of fine-tuning, we currently plan to use vanilla fine-tuning, and will test different values of hyperparameters of the model used for the training process to find the appropriate values.

3) Model Evaluation
   Human evaluation on some matrix is already very popular for evaluating the performance of dialog systems, but since we do not have access to experts or specific human resources to perform the evaluation, we will perform human evaluation ourselves. We intend to evaluate the models by the training loss as part of our standards for the hyper-parameter selection. Also we will test through the replies generated by our chatbot many times until we get a good result, that is, the response given by our chatbot can basically be meaningful and various.

4) Deployment
   We plan to use Google voice API to receive messages through a virtual number. Considering the price of google colab, it is not 24-hour deployed, only running when we want to make interactions with the chatbot.

5. Presentation:

We will record several live demons and upload them to youtube for demonstration purposes, and create a web page to show our training process and results in the form of a report. We will also explain in the form of ppt in class and show our demo and interact with classmates, so that students can use our chatbot through Google Voice on the spot.

6. Unknown:

- About dataset: Although we have found a movie dataset that is very suitable for training dialogue chatbots, since we still need to join multiple

tables and filter dialogues that match the topics, and the language habits contained in movie dialogues are uncertain, we cannot guarantee A sufficient dataset can be obtained after data preprocessing.

- About Model fine-tuning: We need to perform experiments on DisloGPT of different sizes to find a proper size. If the model is too small, the conversations may be low quality. If the model is too large, it will take too much time to wait.
- Compute resources constraints: And because Our data set is large, so we can't be sure whether it will go well in the parameter tuning phase of the model. Training a well-performing model may encounter multiple attempts, which requires a lot of training resources. We are not sure about our time and computing resources. Is it enough.

7. TimeLine:

- Data preparation: Nov 10 - Nov 17
- Model Fine-turning: Nov 17- Nov 26
- Presentation Preparation: Nov 26 - Nov 30
- Final Deliberable polish: Nov 30 - Dec 8

8. References:

I. Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., & Dolan, B. (2020, May 2). *Dialogpt: Large-scale generative pre-training for conversational response generation*. arXiv.org. Retrieved November 8, 2022, from https://arxiv.org/abs/1911.00536

II. Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Shuster, K., Smith, E. M., Boureau, Y.-L., & Weston, J. (2020, April 30). *Recipes for building an open-domain chatbot*. arXiv.org. Retrieved November 8, 2022, from https://arxiv.org/abs/2004.13637