**Portfolio Project**

Viet Ngo

Colorado State University Global

CSC525: Principles of Machine Learning

Dong Nguyen

08/11/2024

**Portfolio Project Milestone 4**

The first dataset that will be used to train the chatbot is DailyDialog. The DailyDialog dataset is a large-scale corpus designed for training dialogue systems. It features conversations that cover a variety of daily communication topics, making it well-suited for general conversational models. The dataset contains dialogues with a focus on general conversational skills rather than domain-specific or task-oriented interactions (Devastator, 2024). In addition to daily conversation, it is important that the chatbot be trained to understand basic questions and answers. The second dataset that will be used to train the chatbot is Stanford Question Answer Dataset (SQuAD). The SQuAD is a prominent benchmark dataset designed for evaluating machine reading comprehension and question answering systems. It provides a collection of questions and corresponding answers based on a set of passages, enabling the development and assessment of models that can understand and extract relevant information from text. The dataset consists of paragraphs or passages sourced from Wikipedia articles. Each passage provides a context within which the questions are framed. Each passage is paired with multiple questions. Questions are designed to be answerable based on the information contained within the passage (Stanford University, 2019). The next important consideration is that the chatbot understands the emotional tone of the conversation. The third dataset that will be used to train the chatbot is EmpatheticDialog. The EmpatheticDialog dataset is a resource designed to train and evaluate conversational models on their ability to generate empathetic responses. It was created to help improve how dialogue systems handle emotionally sensitive and supportive conversations. The dataset centers on dialogues where one participant expresses empathy, understanding, and support towards the other. These conversations are typically based on everyday situations where empathy is crucial (Rashkin et al., n.d.).

It is important to train the chatbot on multiple datasets. Training a chatbot on multiple datasets can be highly beneficial and often necessary to develop a robust and versatile conversational agent. Multiple datasets provides diverse knowledge, improved performance, and higher adaptability. The chatbot will be sequentially trained with each dataset in the same manner. The first step is preprocessing, which includes tokenization and normalization. The second step is to split the data. The typical split is 70% for training, 15% for validation, and 15% for testing. The next step would be the actual training of the chatbot. This includes selecting an appropriate loss function, optimization to minimize the loss function and adjust model parameters, and tuning hyperparameters, such as learning rate, batch size, and number of training epochs. The last step of the training process is evaluation. Evaluate model performance using metrics such as perplexity, BLEU score, ROUGE score, and human evaluation for dialogue quality (Santhosh, 2023).

The chatbot will utilize the model's ability to understand and generate responses based on dialogue history. This allows the chatbot to maintain context and coherence throughout the conversation. The trained model will be used to generate responses to user inputs. The model will create responses based on the input text and the context provided by previous utterances. In order to ensure continuous learning, the chatbot will learn from new interactions and feedback. This can involve retraining or fine-tuning the model periodically with new data to adapt to evolving user needs and preferences.

**References**

*Devastator. (2024, January). DailyDialog: Multi-turn dialog+intention+emotion. Kaggle.* [*https://www.kaggle.com/datasets/thedevastator/dailydialog-multi-turn-dialog-with-intention-and*](https://www.kaggle.com/datasets/thedevastator/dailydialog-multi-turn-dialog-with-intention-and)

*Rashkin et al. (n.d.). EmpatheticDialogues. Dataset | Papers With Code.* [*https://paperswithcode.com/dataset/empatheticdialogues*](https://paperswithcode.com/dataset/empatheticdialogues)

*Santhosh, S. (2023, April 16). Understanding Bleu and Rouge score for NLP evaluation. Medium.* [*https://medium.com/@sthanikamsanthosh1994/understanding-bleu-and-rouge-score-for-nlp-evaluation-1ab334ecadcb*](https://medium.com/@sthanikamsanthosh1994/understanding-bleu-and-rouge-score-for-nlp-evaluation-1ab334ecadcb)

*Stanford University. (2019, November 17). Stanford Question Answering Dataset. Kaggle.* [*https://www.kaggle.com/datasets/stanfordu/stanford-question-answering-dataset*](https://www.kaggle.com/datasets/stanfordu/stanford-question-answering-dataset)