Module 4: Portfolio Milestone

NLP Chatbot

Ryan Thompson

Colorado State University - Global

CSC 525

Dr. Banerjee

11 May 2025

**Dataset Selection and Training Strategy for an NLP Chatbot**

To build an effective NLP chatbot, I selected two complementary datasets, the Cornell Movie Dialogue Corpus and the Customer Support Tweets on Twitter. These datasets were chosen for their diversity in conversational structure and their relevance to real-world human interactions. The Cornell Movie Dialogue Corpus offers a broad range of informal, multi-turn conversations derived from film scripts, which provides excellent training material for natural dialogue flow. In contrast, the Customer Support Tweets dataset contains focused, real-world customer service interactions, making it well-suited for task-oriented dialogue modeling.

The Cornell dataset contains over 220,000 conversational exchanges between more than 10,000 pairs of movie characters. Each line is tagged with metadata such as character ID, movie ID, and conversation ID, allowing for structured sequence modeling. On the other hand, the Customer Support Tweets dataset includes labeled tweets between users and companies such as Apple, Google, and Amazon, with contextually rich support requests and responses.

To train the chatbot, I will utilize the Hugging Face Transformers library and fine-tune a pretrained model like DialoGPT or GPT-2, depending on performance trade-offs and compute constraints. The training process will begin with cleaning and formatting the datasets into prompt-response pairs. The movie corpus will be used to fine-tune the base model for generating natural, coherent responses, while the support tweets will be used to further train the model for more goal-oriented interactions.

The training will follow a two-phase approach. First, general fine-tuning will be done on the Cornell Movie Dialogue Corpus using a sequence-to-sequence training format. This phase will teach the model how to maintain context and flow in multi-turn dialogue. The second phase will focus on domain adaptation using the Customer Support Tweets dataset. Here, supervised fine-tuning will reinforce patterns in technical support language, intent recognition, and empathetic tone.

Once trained, the chatbot will utilize the learned dialogue generation capabilities to engage users in both casual conversation and task-specific scenarios. The system will also incorporate intent detection and response reranking components to ensure it selects contextually appropriate replies. Additionally, feedback loops will be implemented to allow for continuous learning from user interactions. By combining open-domain and task-specific data, this dual-dataset approach ensures the chatbot is not only conversationally fluent but also practically useful for customer engagement tasks.

**References**

Danescu-Niculescu-Mizil, C., & Lee, L. (2011). Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 76–87.

Twitter Support Data. (n.d.). *Customer Support on Twitter*. Retrieved from<https://www.kaggle.com/thoughtvector/customer-support-on-twitter>