Context-Aware Movie Chatbot:

Multi-Turn Conversations with Sentiment-Driven Responses

Anand Fernandes, Saad Saeed, and Outhai Xayavongsa

Master of Science in Applied Artificial Intelligence, University of San Diego

AAI-520: Natural Language Processing

Professor Kahila Mokhtari, Ph.D.

October 21, 2024

Author Note

This project was developed for AAI-520: Natural Language Processing at the University of San Diego. Team members include Outhai Xayavongsa (Team Leader), Saad Saeed (Lead Assistant), and Anand Fernandes (Team Member). The project code can be accessed at https://github.com/oxayavongsa/NLP-Chatbot.

Abstract

This work focused on developing a generative chatbot using the Cornell Movie Dialogs Corpus, which contains over 220,000 film dialogues. The aim was to build a context-aware, multi-turn conversational agent capable of generating meaningful, coherent responses. After evaluating various models, including LSTM, GPT-2, GPT-3, and GPT-4, the T5 model was selected for its robust sequence-to-sequence capabilities, effectively overcoming issues of coherence and context retention (Raffel et al., 2020). The training strategy included splitting the dataset into 80% training, 10% validation, and 10% testing. The model achieved an accuracy of 87%, with precision, recall, and F1-scores in the mid-80s. Training showed consistent improvements over 20 epochs, mitigating challenges such as GPU memory limitations and overfitting through batch size adjustments, regularization techniques, and expanded beam search.

Exploratory analysis indicated character dominance impacting response diversity, addressed by preprocessing techniques like tokenization and rare-word replacement. Visual tools like scatterplots and word clouds evaluated conversational consistency. Future improvements include using a more diverse training set, sentiment-driven responses, and reinforcement learning to enhance interaction quality (Vozna, 2024).

Keywords: Generative chatbot, T5 model, Cornell Movie Dialogs Corpus, multi-turn conversations, model evaluation, dialogue diversity, context-aware responses.

Context-Aware Movie Chatbot: Multi-Turn Conversations with Sentiment-Driven

The development aimed to create a chatbot capable of managing context-aware, multi-turn conversations using the Cornell Movie Dialogs Corpus, which includes over 220,000 lines of dialogue. Various models—LSTM, GPT-2, GPT-3, and GPT-4—were assessed but struggled with coherence and context management in longer conversations. Ultimately, the T5 model, known for its robust sequence-to-sequence capabilities, was selected to generate coherent responses across multiple turns. Key challenges included data preprocessing, model training, and conversation generation, which were addressed to achieve an effective chatbot.

Challenges Faced and Solutions Implemented

Developing the chatbot required overcoming several challenges, particularly around data preprocessing. The raw dataset's mixed metadata complicated the extraction of meaningful conversation pairs, necessitating the development of custom parsing functions and preprocessing techniques like tokenization, stopword removal, and lemmatization. Rare words were replaced with an `<UNK>` token to improve the model's generalization capabilities. Early experiments with LSTM and GPT-2 models faced coherence issues, which led to adopting the T5 model for its superior sequence-to-sequence capabilities (Raffel et al., 2020).

During the training phase of T5, GPU memory limitations and risks of overfitting were significant challenges. These issues were addressed by adjusting batch sizes, utilizing PyTorch's "torch.cuda.set_per_process_memory_fraction()" to effectively manage GPU memory, and applying regularization methods to promote generalization. Hyperparameters such as beam search width were adjusted to enhance response fluency, and while some challenges, like generating contextually varied responses, remained, these steps significantly improved overall performance.

Model Architecture and Rationale

The T5 model, based on a transformer encoder-decoder architecture, was chosen for its advanced sequence-to-sequence processing capabilities, making it effective for conversational tasks (Raffel et al., 2020). Previous models like LSTM struggled with context retention, resulting in incoherent responses, while GPT-2 and GPT-3 had difficulty maintaining coherence in multi-turn interactions. T5 excelled due to its pre-trained language capabilities and adaptability to text-to-text tasks. Fine-tuning T5 on the Cornell dataset enabled high-quality dialogue generation, making it an ideal choice for creating dynamic and natural conversations in a chatbot.

Evaluation Results and User Feedback

The T5-based chatbot's performance was evaluated using metrics such as accuracy, precision, recall, and F1-score, achieving an overall accuracy of 87%, with precision and recall in the mid-80s and an F1-score of approximately 85% (Figure 1). These metrics indicate effective generation of relevant responses for straightforward queries. Scatterplots and word clouds were used to evaluate response diversity and consistency, highlighting areas needing improvement. Users appreciated the chatbot's coherence in short conversations but noted limitations in handling abstract or complex queries, often resulting in repetitive answers.

Future Improvements and Scalability Options

Enhancing the chatbot's performance will involve expanding the training dataset to include dialogues from broader sources to improve topic diversity. Incorporating sentiment analysis can help tailor the chatbot's responses to user emotions, adding empathy to interactions (Vozna, 2024). Real-time learning capabilities and cloud-based deployment, such as on Google Cloud AI, could enhance scalability and response time. Integrating reinforcement learning with human feedback can further refine the model, ensuring it remains adaptive and capable of producing nuanced dialogue.

References

Chidananda, R. (2016). *Cornell Movie-Dialog Corpus* [Data set]. Kaggle. https://www.kaggle.com/datasets/rajathmc/cornell-moviedialog-corpus

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). *Exploring the limits of transfer learning with a unified text-to-text transformer*. Journal of Machine Learning Research, 21(140), 1-67. https://huggingface.co/docs/transformers/en/model_doc/t5

Vozna, A. (2024, June 13). *AI Chatbot development: A complete guide.* https://gloriumtech.com/ai-chatbot-development-a-complete-guide/