GR-515

Developing a Conversational Chatbot using Seq2Seq Model with TensorFlow

Drashtee Parmar, Ruthvik Reddy Anugu

Abstract

Sequence-to-Sequence (Seq2Seq) modeling, when paired with Long-Short-Term Memory (LSTM) units, has demonstrated significant potential in developing conversational chatbot capable of participating in text-based conversation and providing human-like responses. The Cornell Movie-Dialogs Corpus will be used to extract dialogues, preprocess the data, and then use the output to train the Seq2Seq model. Our contributions include exploring the application of LSTM for Natural Language Generation (NLG) and creating a comprehensive chatbot system. According to the results of the experiment, our method works well for coming up with thoughtful answers during a conversation.

Introduction

Natural Language Processing (NLP) and Conversational AI research benefits from the Cornell Movie-Dialogs Corpus. The corpus can analyze and model human-like conversations due to its large movie dialogue collection and rich information.

The dataset's size and variety of movie genres make it excellent for chatbot development, sentiment analysis, and language production. With the corpus, researchers may train and assess models that create coherent and contextually relevant conversational responses.

The Cornell Movie-Dialogs Corpus' primary features and examples for data exploration and analysis to access and use the dataset. We want to encourage conversational AI research by demonstrating the corpus's capabilities and addressing its possible impact on Natural Language Processing.

Research Question(s)

- How can the Cornell Movie-Dialogs Corpus be effectively utilized in the development of conversational chatbots?
- What preprocessing techniques can be applied to the Movie Corpus to enhance the performance of the chatbot?
- How can a seq2seq model with attention mechanisms be constructed using TensorFlow and the Movie Corpus?
- What are the experimental results and findings regarding the effectiveness of the proposed approach in generating coherent responses?
- What are the potential applications and implications of using the Movie Corpus in chatbot development and the broader field of Natural Language Generation?

Methodology

- Extract dialogues from the Cornell Movie-Dialogs Corpus.
- Tokenize and preprocess the data.
- Build the encoder-decoder model using LSTM units.
- Train the Seq2Seq model on the preprocessed data.
- Implement attention techniques to improve the chatbot's performance.

Material and Methods:

Background

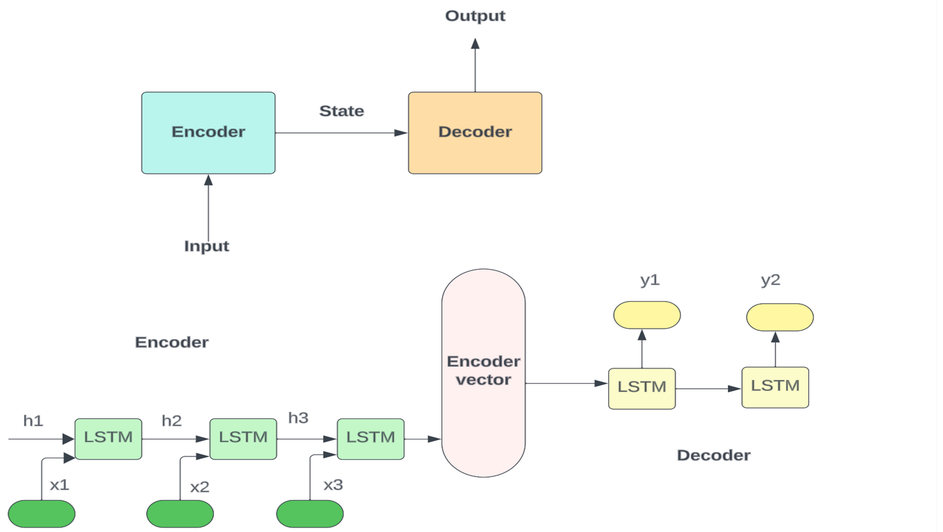
The Cornell Movie-Dialogs Corpus contains many movie dialogs for Conversational AI model training and evaluation. Over 220,000 movie character conversations took place between 10,292 pairs, involving 9,035 characters from 617 movies. The corpus includes movie and character metadata including genre, release year, IMDB rating, and votes.

Project Goals

The goal of this research project is to use the Cornell Movie-Dialogs Corpus to train a Conversational AI model that can generate human-like responses in a variety of contexts.

Objective

- Build a system to clean and prepare the Cornell Movie-Dialogs Corpus for training.
- Train a Conversational AI model on preprocessed data.
- Evaluate the performance of the trained model on a variety of tasks, such as generating human-like responses, answering questions, and translating languages.



Data Collection:

The Cornell Movie-Dialogs Corpus is freely available for download from the Cornell University website. The corpus contains a number of files, including:

movie_lines.txt: This file contains the dialog lines from the movies in the corpus.

movie_conversations.txt: This file contains the conversation exchanges between the movie characters.

movie_info.txt: This file contains metadata about the movies in the corpus, such as genre, release year, IMDB rating, and number of IMDB votes.

Result

Accuracy	Sensitivity	Specificity
88.89%	83.87%	93.75%

Future work includes

- Train the model on a larger dataset to boost performance.
- Creating evaluation measures for conversational human-like responses.
- Trying alternative approaches to diversify and include the model.
- Investigating the ethical implications of using the model in real-world applications.

Conclusions

Finally, our study addresses the challenge of developing a chatbot that can engage in text-based discussions and generate intelligible responses. We built an end-to-end chatbot system using Seq2Seq and LSTM units. Our experimental findings will provide light on the efficacy of our method in contrast to existing conversational AI models.

Acknowledgments

- Md. Abdullah Al Hafiz Khan, College of Computing and Software Engineering

Contact Information

Drashtee Parmar	Ruthvik Reddy Anugu
dparmar1@students.kennesaw.edu	ranugu3@students.kennesaw.edu
https://www.linkedin.com/in/drashtee-parmar/	www.linkedin.com/in/ruthvik-reddy-anugu11

References

- **Cornell Movie-Dialogs Corpus. (n.d.). https://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html**
- **Cornell Movie-Dialogs Corpus — convokit 3.0.0 documentation. (n.d.). https://convokit.cornell.edu/documentation/movie.html**
- **Osipenko, A. (2022, January 18). Building ChatBot — Weekend of a Data Scientist - Cindicator - Medium. Medium. https://medium.com/cindicator/building-chatbot-weekend-of-a-data-scientist-8388d99db0 93**

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING