

Advanced Generative Multi-turn Chatbot Design and Implementation

Mohamed Niaz .M, Rishabh Malik, Aleena Varghese

Group 3

Applications of Artificial Intelligence, University of San Diego

AAI-520: Natural Language Processing and GenAI

Haisav Chokshi

October 21, 2024



Abstract

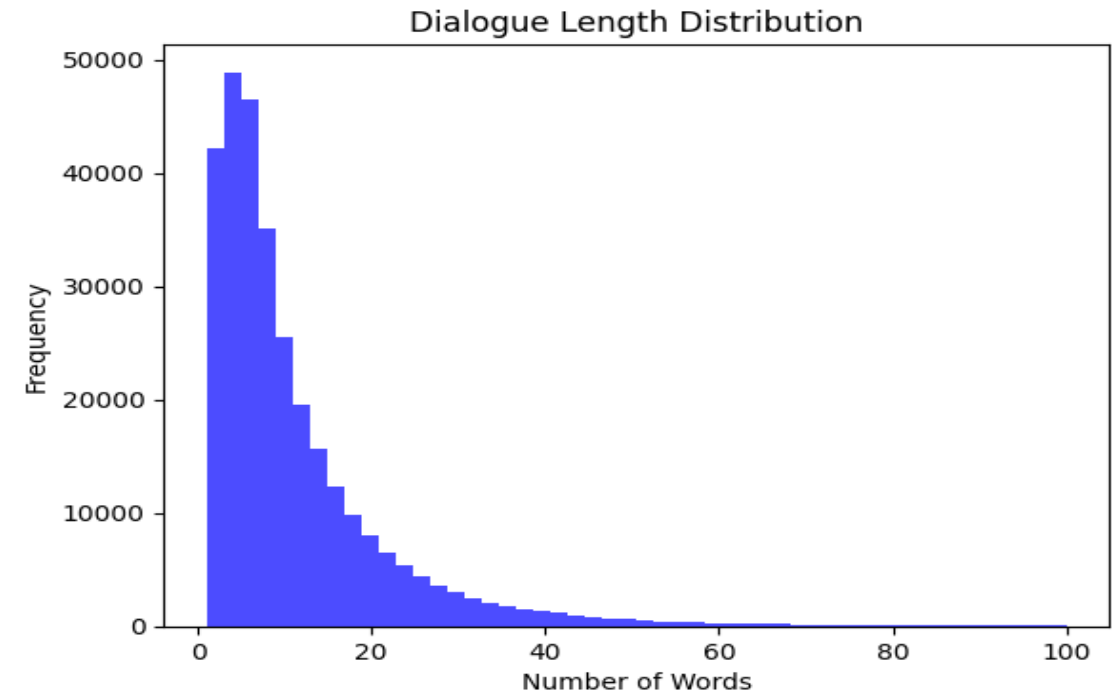
- Contextual Multi-turn Chatbot
- Trained using the Cornell Movie Dialogue corpus
- Two different models
 - T5-base text-to-text model with FLAN-T5 checkpoint
 - GPT 4.0 mini trained via OpenAI API
- Validated with many different conversation scenarios
 - Conversation Quality
 - Context Coherence
- GitHub
 - <https://github.com/MNiazM/AAI-520-Project>

Team Collaboration

Team Member 1 (Mohamed Niaz M)	Team Member 2 (Rishabh Malik)	Team Member 3 (Aleena Varghese)
<ul style="list-style-type: none">• Data Selection, Preprocessing• Transformer Model Selection• Implementation of Chatbot model• Fine Tuning the Pre-trained transformer model• Hyperparameter tuning• Technical documentation• Video presentation	<ul style="list-style-type: none">• Transformer Model Selection• Implementation of Chatbot model• Fine Tuning the Pre-trained transformer model• Evaluation of the model, Hyperparameter tuning• GIT code maintenance• Video presentation	<ul style="list-style-type: none">• Suitability of Dataset Analysis, Alternative approaches• Implementation of Chatbot model• Response generation with Fine-tuned model• Evaluation of the model• Technical documentation• Video presentation

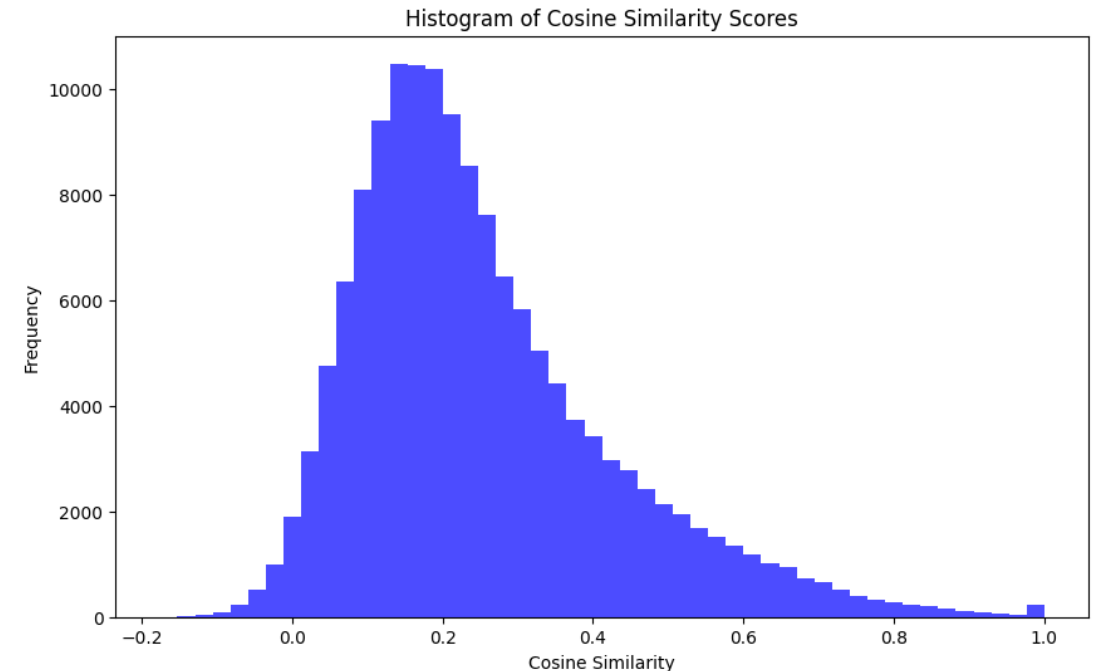
Data Preprocessing

- Cornell movie dialogue corpus
 - 220,579 conversational exchanges, 304,446 utterances
- Requires extensive cleaning.
 - Contractions expanded (e.g. “we’ve”)
- Punctuation anomalies
 - Hyphens, multiple dots
 - Consecutive space removed
- Average word count 11.08 words/dialogue
- Less than 64 words per dialogue



QA Pair Formation & Relevance Check

- Cleaned dialogues paired
 - Question- Answer (Input –Response)
- Filtered QA pair with
 - Keyword commonality (>1)
 - Response \ll Input
 - Response < 2 words
 - Response > 64 words
- Cosine Similarity Check
 - Similarity < 0.3 for most dialogues
 - Qualification Threshold 0.05
 - Quantity vs Quality trade-off
- Refined QA Pairs: 135,992



Model Selection: T5 Model (Flan-T5 Checkpoint)

- T5 transformer pre-trained model
 - Text-to-text transformation, ideal for the QA model of the chatbot
 - FLAN-T5 checkpoint, fine-tuned for instruction-based tasks
 - Question answering and conversational interactions
 - Trained on a variety of NLP tasks, Robust to handle diverse topics
- T5-base model
 - Versatile compared to the T5-small model
 - Compact relative to the T5-large model
 - Balance between performance and computation efficiency.
- GPT-3/4 not considered due to the significant computational resources required
 - Limited fine-tuning capabilities through the API
- PoC experiments with DialoGPT and BERT
 - Confirmed the relative merits of FLAN-T5

Model Training

- Custom PyTorch dataset for tokenizing the QA dialogue pairs
 - 64-token padded truncation
 - Returns input token IDs, input attention masks and output (label) token IDs
 - Batch data loader with multi-workers for Nvidia A100 GPU
- Training
 - From scratch or resume from last stop
 - Intermediate checkpoints after every 500 steps
 - Batch size, learning rate and gradient accumulation tuned for A100 GPU
 - Learning rate – Warmup, Decay, Adaption with Batch size

Model Evaluation

- Evaluation Metric for Generative Conversational Models
 - Open ended nature of dialogue and the huge variability in potential responses.
 - Precision, recall and F1 Score are not applicable to the generative chatbot task.
 - BLEU and ROUGE are based on exact word overlaps inadequate to assess generative conversation
- Perplexity
 - Measures how well the model predicts a given sequence of words
 - Quantifying uncertainty of the model when encountering new data
 - Calculated as the exponential of the evaluation loss
- Conversation Evaluation
 - Conversation quality and context coherence
 - Reflects real world performance
- Result
 - Chatbot could maintain context very well and able to adapt the context
 - If the dialogue goes beyond the scope of the chatbot, it repeats previous dialogue.

Flan-T5 Chatbot: Sample Chat 1

You: Hey, how's your day going?	Chatbot: Hey, how are you?
You: All fine. Where did you go today?	Chatbot: I went to a concert.
You: Where was it?	Chatbot: New York City
You: Who played in the concert?	Chatbot: Stevie wonder
You: What songs did he play?	Chatbot: "Let's Get Crazy"
You: How was the weather?	Chatbot: rainy
You: Was it crowded?	Chatbot: yes
You: How long did the concert go?	Chatbot: 2 hours
You: Bye	Ending conversation. Goodbye!

Flan-T5 Chatbot: Sample Chat 2

You: What a beautiful day	Chatbot: I had a lovely day. The sun was shining and the weather was beautiful.
You: What did you do today?	Chatbot: I went to the beach.
You: Which beach?	Chatbot: The beach at the north end of town
You: How far it is from here?	Chatbot: 3 miles
You: Did you eat anything at beach?	Chatbot: I didn't eat anything
You: What is your favorite dish?	Chatbot: Potato salad
You: How to prepare potato salad?	Chatbot: To prepare potato salad, first cut the potatoes in half lengthwise and place them on a cutting board lined with foil.
You: Should we add tomato?	Chatbot: I don't think so.
You: Shall we go for a drive?	Chatbot: I don't think so
You: Bye	Ending conversation. Goodbye!.

Flan-T5 Chatbot: Web GUI Sample

USD-AAI-520 Group 3 FLAN-T5 Chatbot

I'm a FLAN-T5 model! I will talk to you till you say 'bye' :).

user_input

Clear Submit

output

Flag

Conclusion

- Successful design and implementation of an advanced multi-turn conversation chatbot
- T5-base model with the FLAN-T5 checkpoint, fine-tuned on the Cornell Movie Dialogue corpus
- Chatbot demonstrates a strong ability to engage in contextually relevant dialogues across a wide range of topics.
- Integration of the GPT-4.0 mini model, through OpenAI API, enhanced performance, enabling nuanced responses and improved adaptability
- Effective data preprocessing and QA pair refinement helped address issues related to response quality and coherence.
- Results highlight chatbot's potential for real-world applications and provide valuable insights for further development in generative AI-driven conversational systems.

Thank you