

Chatbot Design

AAI 520 - Natural Language Processing
Final Project

Group 4:
Hani Jandali, Kay Cheung, Suvo Ganguli

Overview

- Dataset
- Method 1: GPT2
- Method 2: Seq2seq
- Deployment
- Contributions

Method 1: GPT2

Data Preprocessing & Loading

- Dataset: Cornell Movie Dialogues Corpus
 - 304,713 utterances from 617 movies
- Files:
 - **movie_lines.txt**: Individual lines of dialogue
 - **movie_conversations.txt**: Structured conversations

Model Setup

- **GPT-2 Tokenization:**
 - Used GPT-2 tokenizer, with EOS token for padding
 - Created a custom dataset class for tokenized inputs
- **Training:**
 - Fine-tuned GPT-2 using Hugging Face Trainer API
 - Early stopping and nucleus sampling employed to prevent overfitting and improve response quality
 - Training details:
 - Epochs: 3
 - Batch Size: 16
 - Learning Rate: 5e-6

Method 1: GPT2

Response Generation

- **Nucleus sampling** (`top_p=0.9`) and **temperature** control (`temperature=0.5`)
- Function to generate responses: `generate_responses()`

Results & Challenges

- **Strengths:**
 - Generated fluent, grammatically correct responses
 - Handled simple queries with coherent outputs
- **Challenges:**
 - Responses were often **generic** and **neutral**
 - Lack of depth in conversations
 - Tendency to **repeat phrases**

User: I love movies, do you? Chatbot: I love them.

User: Do you like pizza? Chatbot: I'm not a pizza guy.

User: Do you like popcorn? Chatbot: I'm not sure.

User: What is your favorite movie? Chatbot: I'm not sure.

Conclusion:

- GPT-2 struggled with conversational depth
- Highlighted the need for more task-specific fine-tuning or alternative architectures

Method 2: Sequence to Sequence

Preprocessing

- Similar preprocessing to GPT-2 Model
 - Vocabulary Construction
 - Word Filtering
 - Sentence Pair Extraction

Model

- Sequence to Sequence Architecture with Luong Attention Mechanisms
- Dot attention score calculation
- Masked Negative Loss-Likelihood
- Greedy Search Decoding

Method 2: Sequence to Sequence

Results

- Qualitative Results
 - Increased Context Capture and Response
- Quantitative Results
 - BLEU Score: 0.387
 - Perplexity: 1.343
 - Rouge-1: 0.686
 - Rouge-2: 0.565
 - Rouge-L: 0.687

Challenges

- Contiguous Sequences
- Tokenization via NLTK/SpaCy
- Computational and Time Complexity

```
> Hello!
Bot: i must speak with you miss flaemm .
> What is wrong?
Bot: i don t know . nothing . everything
> What happened?
Bot: i didn t have the guts for it .
> Did you do it?
Bot: i think so . the press .
> Did you press the button? It is urgent
Bot: the hell ? ? ?
> I told you to press it!
Bot: what was the other side ? him ?
> Yes, James was on the other side
Bot: no . it out . valentine .
> Valentine is dead
Bot: no it s not . is penelope .
> Penelope died with Valentine
Bot: i know . . murder
> Which is why I said to press the button
Bot: the day after tomorrow . the money .
> Forget the money. Did you press it?
Bot: i m sorry . . . .
> What does that mean?
Bot: it means . . . s discount .
> Is that a no?
Bot: yes . i thought so . a is .
> I cannot believe you
Bot: why not ? ?
> Because I think you are lying for Valentine
Bot: right . i ll be right back . everything
> quit
```

Deployment (Web Interface)

Description

A Flask-based chatbot was developed to interact with users. It calls the fine-tuned GPT-2 model to generate responses to user inputs. The app uses the Hugging Face pipeline for text generation and serves the model using Flask.

How to Run the Chatbot

1. Install all required libraries from requirements.txt
2. Run the Flask app: app.py from [AAI520_FinalProject/FlaskApp](#)
3. Access the chatbot by https://kay-q-mich.github.io/AAI520_FinalProject/
4. Visit [ngrok's dashboard](#) to obtain your authentication token.
5. Copy the token and paste it into the GitHub page when prompted.
6. Enter text in the textbook to chat

Deployment Tools

- **Flask:** Used for hosting the chatbot web application.
- **ngrok:** Used to expose the Flask app to the internet for testing purposes.

Contributions

- Hani Jandali
 - Data Preprocessing
 - Method 2: Seq2seq
 - Evaluation Metrics
- Suvo Ganguli
 - Data Preprocessing
 - Method 1: GPT2
 - Evaluation Metrics
- Kay Cheung
 - Web Interface
 - Integration with Model