**CSCE 689**

**Final Project: Chatbot for Animatronic Baymax**

**Ethan Barnes**

**12/5/2024**

**Methodologies**

This project consists of four main parts: speech to text, a chatbot, sentiment analysis, and text to speech.

**Speech to Text**

The speech to text (STT) portion of the project functions by recording the user’s speech and then transcribing that input to be used as the user response in the chatbot. This was implemented using AssemblyAi’s STT model, which is popular for its accuracy. In my testing, I have found that this model rarely makes errors with clear speech, so I decided it would be sufficient. The user’s speech is recorded using the pvrecorder Python library and once the user activates a keyboard interrupt, the recording stops and the audio is saved to a folder. That audio file is then used as input to the STT model to get a transcript of what the user spoke.

**Chatbot**

Next in the pipeline, the speech transcript is input into the chatbot which uses OpenAI’s GPT-4o model. A conversation system is implemented to simulate the model remembering what the user and itself have already said. With this, the user can have a conversation with the model while the memory can fit into the model’s max memory, which is set at 4096 tokens. When inputs start to exceed the token limit, messages will begin to be deleted from the memory starting from the beginning so that the model can have the context of what was most recently spoken about. This does mean that if the conversation were to carry on past the memory of the first messages exchanged, the model could repeat information already given, or possibly even forget what the user’s problem was to begin with. Given that the purpose of this project is for the chatbot to provide a diagnosis and help quickly and concisely, not actually attempting to impersonate a medical professional, the token limit should not be an issue in normal application. When the user is done, as per the movie, the user must say, “I am satisfied with my care”. When the word ‘satisfied’ is detected in the response, SentenceTransformer is used to find the cosine similarity between the embedding of that phrase and whatever the user happens to say. If the similarity is above 90%, the program terminates.

**Sentiment Analysis**

Before termination, however, the chatbot’s response each time is run through a sentiment analysis model, roberta-base-go\_emotions, a model trained from roberta-base on the go\_emotions dataset, and the most likely emotion from a list of 28 is returned. This is used in evaluation and will later be used for the animatronic to perform an animation based on this emotion (outside the scope of this project).

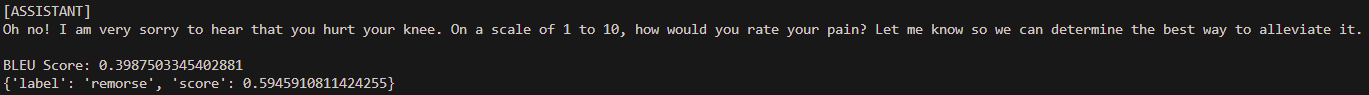
**Text to Speech**

Finally, the response of the chatbot is passed into a text to speech (TTS) model that is trained on Baymax’s voice to return an audio file of what sounds like Baymax speaking the chatbot’s response. This uses the PyHT Python library to interact with play.ht where I created a simple voice model for Baymax with only 30 seconds of audio of the character. The audio file will be played separately from the current configuration to not interfere with recording the next response of the user. The direct use of the audio recording is outside the scope of this project.

**Results/Analysis**

After using this chatbot, results show two important takeaways: the chatbot mimics the tone and mannerism of Baymax well, shown by the consistent sentiment scores, but does not fully capture the character, shown by the BLEU scores.

Starting with BLEU score, as mentioned in my project proposal, BLEU (BiLingual Evaluation Understudy) score can be used with existing lines from the character to evaluate if the LLM’s responses are similar to references. Though this is used to evaluate the accuracy of translation using LLMs, I used this to determine if the chatbot would speak like the character (without fine-tuning) simply based on what it knows of the character do to its popularity. I implemented this evaluation by taking every line that Baymex speaks in the movie Big Hero 6 and computed the BLEU score of each response of the chatbot with the entirety of Baymax’s quotes. This is not standard practice, but it would give me an idea of when the chatbot is acting like Baymax, or simply acting as a medical helper.

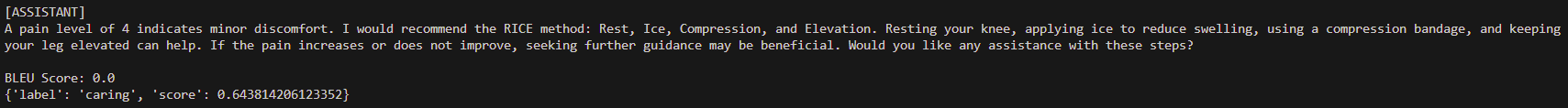


[ASSISTANT]

Oh no! I am very sorry to hear that you hurt your knee. On a scale of 1 to 10, how would you rate your pain? Let me know so we can determine the best way to alleviate it.

**BLEU Score: 0.3987503345402881**

{'label': 'remorse', 'score': 0.5945910811424255}



[ASSISTANT]

A pain level of 4 indicates minor discomfort. I would recommend the RICE method: Rest, Ice, Compression, and Elevation. Resting your knee, applying ice to reduce swelling, using a compression bandage, and keeping your leg elevated can help. If the pain increases or does not improve, seeking further guidance may be beneficial. Would you like any assistance with these steps?

**BLEU Score: 0.0**

{'label': 'caring', 'score': 0.643814206123352}

These two contrasting responses show that the BLEU score is higher, or even nonzero when it says something similar to what Baymax would normally say. In this case, the chatbot asks me to rate my pain on a scale of 1 to 10, which is exactly what the character does in the movie. However, when the chatbot is simply providing medical assistance, the BLEU score is zero. This makes sense as this is not representative of how Baymax speaks – he is a character in a movie, designed to have a personality.

A graph with blue squares and white text

Description automatically generated

*Emotion Distribution of All Quotes*

A graph with blue squares

Description automatically generated

*Emotion Distribution of All Quotes Without Neutral*

|  |  |  |
| --- | --- | --- |
| Most Common Emotions | | |
| Caring | Curiosity | Remorse |

The above histograms show the emotion of every line spoken by Baymax in Big Hero 6. Most of his lines are neutral, and this makes sense for a robotic character, but also for normal speech. In order to understand the tone and personality of Baymax, removing the neutral emotions shows that he is curious and caring. When evaluating the chatbot, caring, curiosity, and remorse were the top three emotions in the responses. Remorse is high due to the chatbot saying things such as, “I’m sorry to hear that”, which is normal for GPT-4o, but not necessarily normal for Baymax. This can be changed with fine-tuning the model as previously suggested, so that the model can better understand how Baymax would respond.

**Conclusion**

What I have learned from this project is that system prompting alone can greatly increase the performance of an LLM. I did not fine-tune this model and only set a strict system message and the chatbot adhered to the requirements. The downside of this is that the model doesn’t completely understand how to talk like Baymax. It is still trying to be a medical assistant, which Baymax is, but that is not representative of Baymax’s primary personality characteristics, which is silly and curious like a child. Regardless of this drawback, the project functions as intended and is ready to be integrated with the Baymax animatronic that is built so that users can converse with Baymax.