NLP Project Part 3

Q3. 1. From among all of the coherence metrics you computed, which do you think are the most informative, and why? What scores do they provide for your transcripts?

The most informative coherence metrics from the analysis are:

- Lemma Type-Token Ratio (lemma_ttr): This metric measures lexical diversity by comparing the number of unique lemmas to the total number of lemmas in the text. A higher value indicates more varied vocabulary usage, which can enhance the richness and engagement of the conversation.
- Content Word Overlap between Adjacent Sentences (adjacent_overlap_cw_sent): This metric assesses the overlap of content words (nouns, verbs, adjectives, adverbs) between adjacent sentences. Higher overlap suggests stronger thematic continuity and cohesion, indicating that the conversation stays on topic.
- Latent Semantic Analysis Score between Adjacent Sentences (lsa_1_all_sent): This metric measures the semantic similarity between adjacent sentences using Latent Semantic Analysis. Higher scores indicate that the sentences are more semantically connected, reflecting better coherence in the conversation.

The scores for these metrics in the transcripts are as follows:

- Lemma Type-Token Ratio (lemma_ttr):
  - all_test.txt (both chatbot and user): 0.5671
  - chatbot_test.txt (chatbot only): 0.6000
  - user_test.txt (user only): 0.7755
- Content Word Overlap between Adjacent Sentences (adjacent_overlap_cw_sent):
  - all_test.txt: 0.1266
  - chatbot_test.txt: 0.1296
  - user_test.txt: 0.0000
- Latent Semantic Analysis Score (lsa_1_all_sent):
  - all_test.txt: 0.8413
  - chatbot_test.txt: 0.8395
  - user_test.txt: 0.9926

2. For the metrics you selected, which of these have the highest coherence scores:

- **(1) The transcripts with both chatbot and user utterances**
- **(2) The transcripts with just chatbot utterances**
- **(3) The transcripts with just user utterances**

Which have the lowest coherence scores? Why do you think this is?

For the selected metrics:

- **Highest Coherence Scores:**
  - **Lemma Type-Token Ratio (lemma_ttr):** The **user-only transcript (user_test.txt)** has the highest score of **0.7755**, indicating greater lexical diversity in the user's language compared to the chatbot.
  - **Latent Semantic Analysis Score (lsa_1_all_sent):** The **user-only transcript** again has the highest score of **0.9926**, suggesting that the user's sentences are highly semantically connected.
- **Lowest Coherence Scores:**
  - **Content Word Overlap between Adjacent Sentences (adjacent_overlap_cw_sent):** The **user-only transcript** has the lowest score of **0.0000**, indicating no direct overlap of content words between adjacent sentences.

The **user-only transcript** exhibits the highest coherence in terms of lexical diversity and semantic similarity. This could be because the user maintains a consistent topic and uses varied vocabulary, enhancing coherence.

The **chatbot-only transcript** has a moderate content word overlap score (**0.1296**), suggesting that the chatbot uses similar phrases or terminology across its responses, which may aid in maintaining thematic continuity but could also indicate repetitive language.

The **combined transcript (all_test.txt)** generally falls between the user-only and chatbot-only transcripts in terms of scores. The interaction between two different speakers may introduce shifts in vocabulary and topics, slightly reducing overall coherence metrics.

3. What was the most challenging part of implementing your chatbot, from among all deliverables? How did that part of the chatbot hold up in your own sample interactions?

The most challenging part of implementing the chatbot was integrating the **dependency parser** for the stylistic analysis. Setting up the Stanford CoreNLP server and ensuring seamless communication between the Python code and the server required careful configuration and troubleshooting. Managing dependencies and handling potential connection errors added complexity to the implementation.

In the sample interactions, the dependency parsing feature functioned correctly, providing insightful stylistic analysis based on the user's input. However, there were occasional delays due to the processing time of the CoreNLP server, which could affect the user experience. Overall, the integration held up well but highlighted areas where performance optimization could be beneficial.

4. If you had additional time to improve upon your final chatbot implementation, what are some potential upgrades you might make? How might you implement these?

Potential upgrades to the chatbot include:

- Enhancing Context Awareness:
    - Implementation: Develop a context management system to retain information from previous exchanges. This could involve using a dialogue state tracker or storing conversation history in a structured format.
    - Benefit: Enables the chatbot to provide more relevant and coherent responses by understanding the context of the conversation.
- Improving Natural Language Understanding (NLU):
    - Implementation: Integrate advanced NLP models like BERT or GPT-3 for intent recognition and response generation. Fine-tuning these models on relevant datasets can improve understanding of user inputs.
    - Benefit: Allows the chatbot to comprehend nuanced language and generate more accurate and human-like responses.
- Expanding Sentiment Analysis Capabilities:
    - Implementation: Train the sentiment analysis model on a larger, more diverse dataset to recognize a broader range of emotions. Incorporate multi-class sentiment classification to detect emotions beyond positive and negative.

    - Benefit: Provides more personalized and empathetic interactions by accurately identifying the user's emotional state.

- Optimizing Performance:
    - Implementation: Streamline the dependency on external servers by exploring lightweight parsing alternatives or local processing options. Implement asynchronous processing to reduce response times.
    - Benefit: Enhances user experience by minimizing delays and potential connectivity issues.

Q4 (50): Recruit two participants to interact with your chatbot.

Participant 1

1. Participant's Success at Completing the Target Tasks

Participant 1 successfully completed both the sentiment and stylistic analyses with some guidance. They knew how to proceed at each step because you explained the process and assisted them during the interaction. However, the sentiment analysis did not work as expected for this participant. When they said, "I just had McDonald's," the chatbot incorrectly inferred that they were feeling down. This unexpected response led to some confusion, possibly due to the limited responses or training data for the sentiment analysis in the chatbot.

2. Length of the Participant's Exchange with the Chatbot

Participant 1 chose to repeat both the stylistic and sentiment analyses. They repeated the sentiment analysis because the initial result was unexpected, and they wanted to see if a different input would yield a more accurate response. They also repeated the stylistic analysis out of curiosity to see how the chatbot would respond to different inputs. This led to a longer exchange, indicating engagement with the chatbot and an interest in testing its capabilities.

Participant 2

1. Participant's Success at Completing the Target Tasks

Participant 2 also knew how to complete both analyses, thanks to your explanations and guidance. The sentiment analysis worked as expected for this participant, accurately reflecting their emotional state based on their input. The stylistic analysis functioned correctly as well, providing appropriate feedback. This smooth interaction suggests that the chatbot performed well when the input aligned with its training and response patterns.

2. Length of the Participant's Exchange with the Chatbot

Participant 2 did not choose to repeat any of the tasks. Their exchange with the chatbot was concise, completing each analysis once without seeking additional interactions. This might indicate that they were satisfied with the initial responses or less inclined to explore further due to time constraints or a lack of interest in testing different inputs.

Survey Results

Participant 1 Survey Responses:

- Understanding of Dialogue Options: 4
- System Performance as Expected: 2
- Naturalness of Dialogue: 3

Participant 2 Survey Responses:

- Understanding of Dialogue Options: 3
- System Performance as Expected: 4
- Naturalness of Dialogue: 2

Average Scores:

1. In this conversation, I knew what I could say or do at each point of the dialogue.
    - Average Score: (4 + 3) / 2 = 3.5
2. The system worked the way I expected it to in this conversation.
    - Average Score: (2 + 4) / 2 = 3
3. The dialogue produced by this system seemed natural.
    - Average Score: (3 + 2) / 2 = 2.5

Analysis of Results

The average score for the first question was 3.5, indicating that participants generally knew what they could say or do during the conversation. This is slightly higher than expected, suggesting that the guidance provided was effective in helping them navigate the chatbot's functionalities.

For the second question, the average score was 3, reflecting a moderate level of satisfaction with how the system performed compared to expectations. Participant 1 rated this lower due to the unexpected result in the sentiment analysis, while Participant 2 had a higher score, indicating a disparity based on individual experiences.

The average score for the naturalness of the dialogue was 2.5, which is lower than ideal. This suggests that participants felt the chatbot's responses were somewhat unnatural. This could be due to limited response variations or the chatbot's inability to handle certain inputs effectively.

Additional Insights:

- Participant 1 was intrigued enough to retry tasks, indicating engagement but also highlighting potential areas for improvement in the chatbot's response accuracy.
- Participant 2 had a smoother experience but still found the dialogue less natural, possibly due to rigid or scripted responses from the chatbot.
- Both participants benefitted from guidance, implying that the chatbot's interface might not be entirely intuitive for new users.
- The disparity in the sentiment analysis outcomes points to a need for enhancing the model's ability to interpret a wider range of inputs accurately.

Participants' Additional Feedback:

- Participant 1 expressed curiosity about how different inputs would affect the chatbot's responses, suggesting that increased variability and adaptability in responses could enhance user engagement.
- Participant 2 mentioned that while the system worked as expected, they felt the conversation lacked a natural flow, indicating room for improvement in making interactions feel more organic.