CMPUT 461 - Justifications for Cleaning and Transformation

Name:Ishaan Meena
StudentID: 1780950

Data Cleaning Decisions

1. Removal of Metadata

- Metadata headers (lines starting with `@`, such as `@UTF8`, `@PID`, `@Begin`, and `@Languages`) are removed because they are irrelevant to the downstream task of creating a language model.
- Regular Expression Used: ^@.*
- This ensures that only the relevant text remains for processing, focusing the dataset on actual speech content.

2. Removal of Extraneous Information

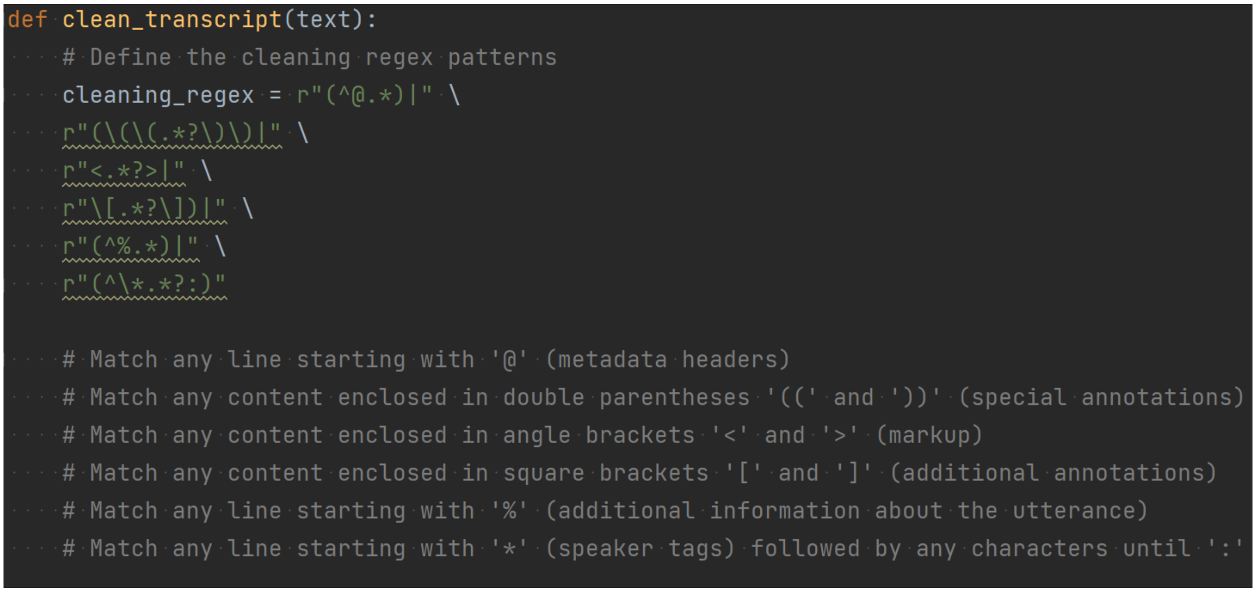
- Any information that is not part of a person's utterance (e.g., comments, special characters, non-verbal cues) are removed. This helps ensure that the model trained later is not affected by irrelevant data.
- Regular Expressions Used:
  - (\(\(.*?\)\)|<.*?>|\[.*?\]) for removing special annotations.
  - ^%.* for removing additional information about utterances.
- By removing extraneous information, we focus solely on the spoken language, improving the kind of data that the language model trains on.

3. Handling of Non-Word Utterances

- Non-word utterances (e.g., "umm", "uh") are kept because they represent real speech patterns and could potentially influence language modeling. These could be significant in understanding children's speech development.
- I choose not to filter out these utterances because of their potential contextual significance.

4. Punctuation and Contractions

- In order to prevent misunderstanding in the phonetic representation, which could compromise pronunciation accuracy, punctuation marks are removed.
- Basic punctuation is removed using regex patterns, while contractions (such as "don't") are kept intact as they are commonly used in everyday speech.



```
def clean_transcript(text):
    # Define the cleaning regex patterns
    cleaning_regex = r"(^@.*)|" \
    r"(\(\(.*?\)\))|" \
    r"<.*?>|" \
    r"\[.*?\])|" \
    r"(^%.*)|" \
    r"(^\*.*?:)"

    # Match any line starting with '@' (metadata headers)
    # Match any content enclosed in double parentheses '((' and '))' (special annotations)
    # Match any content enclosed in angle brackets '<' and '>' (markup)
    # Match any content enclosed in square brackets '[' and ']' (additional annotations)
    # Match any line starting with '%' (additional information about the utterance)
    # Match any line starting with '*' (speaker tags) followed by any characters until ':'
```

Specific Regex Patterns

Pros:

- Efficiency: The use of regex allows for fast pattern matching and cleaning of large amounts of text-based data.
- Flexibility: Regular expressions can be adapted for various patterns in the data without needing multiple lines of code.
- Simplicity: Combining multiple cleaning operations into a single regex pattern simplifies the code and improves readability.

Cons:

- Some regex patterns may inadvertently remove relevant data. For instance in my code, removing all lines starting with % also eliminates useful context about the utterances.
- Limited Context Awareness: Regex does not account for the semantic meaning of the text, which may lead to unintended omissions. (Something more sophisticated like a context-based expression matching system, may be beneficial)
- If regex patterns are too complex, it can be challenging to troubleshoot and ensure they are working as intended.

Data Transformation Decisions

1. Use of CMU Pronunciation Dictionary

- The CMU Pronunciation Dictionary serves as an source for mapping words to their ARPAbet phonemes, which is critical for building a robust language model.
- **Handling Unknown Words**: For words not found in the dictionary, I decided to retain the original word in the output.
- This decision maintains the context in which the words are used, which may be relevant for various linguistic analyses.

2. Lexical Stress Markers

- Lexical stress markers (e.g., AA1, AA0) are removed to simplify the phoneme representation. This reduces complexity while still allowing for accurate sound representation.
- Regular Expression Used: \(\d+\) to remove numeric suffixes from words with multiple pronunciations (e.g., HELLO(1)).

3. Multiple Pronunciations

- When words have multiple pronunciations, maintain a frequency table to track how often each pronunciation occurs. This allows us to select the pronunciation that is most frequently used.
- By choosing the most common pronunciation based on this frequency data, we ensure that the transformation process is more aligned with **real-world usage**, enhancing the accuracy and relevance of the representations in the transformed data. This approach improves consistency and supports more effective language modeling.

4. Choice of Output Format

- The cleaned data has been structured to maintain full sentences, as this retains context and coherence.
- By choosing to transform the data into full sentences, we aim to improve the model's ability to learn from context, which is critical for effective language modeling and downstream tasks.

5. Manual Stopword Removal

- The manual stopword removal helps refine the dataset by eliminating common words that do not carry significant meaning in the context of the language model. (like, “a”, “the”, “or”)
- This technique enhances the model's ability to focus on meaningful content and improve its predictive performance by reducing noise.

Bias and Limitations

Non-Word Utterances (Bidirectional Influence):

- Bias: I retained non-word utterances like "umm" and "uh" to capture natural speech patterns, acknowledging their potential significance in children's language development.
- Limitation: This choice might introduce noise in more formal contexts, where such utterances could be less relevant.

Extraneous Information Removal (Unidirectional Influence):

- Bias: The decision to remove metadata and annotations was based on the need to focus solely on spoken content for language modeling.
- Limitation: This may overlook contextual clues that could help in understanding specific analyses.

Retention of Original Words for Unknown Entries:

- Bias: The decision to keep out-of-dictionary words maintains contextual integrity.
- Impact: In technical or specialized domains, failing to substitute unknown words with a placeholder could lead to confusion and misinterpretation in analyses focused on specific terminology. (Research/academic domains)

Generalization

The cleaning and transformation techniques applied in this assignment are specifically designed for transcripts of child-adult interactions. However, these methods may not seamlessly transfer to different contexts, such as news programs or academic domains. For example:

- Language Formality: In news transcripts, the language tends to be more formal, and non-word utterances like "umm" may not appear frequently. Retaining such utterances could introduce noise into the dataset and reduce model performance. Different cleaning rules might be necessary.
- Punctuation Handling: The removal of punctuation is aimed at simplifying phonetic representation. However, in transcripts with more complex syntactic structures, such as academic discussions, preserving punctuation could be vital for understanding meaning and intent.

Summary of NLP Techniques Used

- Regular Expressions (Regex): Used for cleaning and preprocessing text data.
- CMU Pronunciation Dictionary: Maps words to **ARPAbet** phonemes for pronunciation.
- Stopword Removal: Eliminates **common** words to **reduce noise** in the dataset.
- Non-Word Utterance: Retains speech fillers (e.g., "umm") for **linguistic significance**.
- Frequency-Based Pronunciation: Selects the **most common** pronunciation from multiple options.
- Punctuation and Contraction: **Removes punctuation** for clarity; preserves contractions.
- Sentence Structuring: Maintains full sentences for **contextual** coherence.

Future Improvements

Improvement	Description
Context-Aware Regex	Implement context-based regex to differentiate relevant and irrelevant data.
Semantic Analysis	Use machine learning to analyze and retain contextually significant data.
User Feedback Loop	Establish a feedback mechanism where users can highlight errors, allowing continuous improvement.
Incorporate NLP Techniques	Utilize NLP techniques to understand context better and refine data extraction methods.

Conclusion

This assignment's cleaning and transformation techniques seek to improve the caliber of linguistic data utilized for model training. It is made sure that the dataset appropriately reflects spoken language by employing phonetic mapping using the CMU Pronunciation Dictionary and regular expressions for data cleaning. This document entails the many decisions and their implications to the accuracy of the linguistic-mapping model.