Research Question

Ruhi Mahadeshwar
s4014456

December 2023

Anomic aphasia can occur due to damage to the brain, for example, through a traumatic injury, a brain tumour or a stroke. Anomic aphasia may also occur due to neurodegenerative diseases such as Alzheimer's. Patients with anomic aphasia (PWA) have word retrieval failures and cannot express the words they want to say. This leads to frustration in daily life. The use of assistive devices could help patients with anomic aphasia to overcome daily frustration. One way to develop such assistive devices can be through large language models (LLMs), with a focus on decoder only transformer architecture. Therefore the main research question of this project is "Can decoder-only transformers be used to assist in the completion of daily anomic aphasic speech?". One important thing to note is that the project will not look at transcribing recorded speech to sentences[1].

To answer this question, there are two steps. Firstly, a rule based model will be developed to generate speech commonly seen in PWA. This can be done using linguistic features from studies on aphasic patients, and, results from verbal tests, such as the cookie theft picture. By doing so, we will be able to generate a large scale dataset of synthetic aphasic speech from various grammatically correct datasets. This is done to overcome the lack of large scale studies of linguistic deficits in aphasia. Further comparisons of the synthetic dataset to the real-life datasets will also be done, for example, using a measure of sentence complexity or LSA. This will show if the synthetic dataset is an accurate representation of aphasic speech in terms of linguistic features commonly seen in PWA.

Then, a pretrained decoder only transformer model, such as ChatGPT, GPT-4 or LLaMA will be fine tuned using the synthetic dataset to suggest 5 corrected sentences given an aphasic sentence as input. The efficacy of the model can then be evaluated using bleu, rouge, LSA or perplexity. The model can then be evaluated on the validation set from the synthetic dataset and a possible real life dataset. This will show if the model can be used in assistive devices even if it was trained on synthetic dataset.

Further comparison of the model can be done to an encoder-decoder model (such as T5) trained on the same synthetic dataset to see if both models are able to complete aphasic speech.

---

[1] If time permits, I would also be interested in looking at this.