Pavan Arani, PSA220002
NLP Using Transformers, Text Summarization Assignment
Book: The Project Gutenberg eBook of Frankenstein

This project makes use of natural language processing methods, which are transformer-based to produce quality summaries of the novel Frankenstein written by Mary Shelley. The data set was the entire text downloaded at Project Gutenberg in its plain text form. This project aimed to use an abstractive transformer model, to compare its performance with an easy extractive baseline, and to measure the results using ROUGE metrics to determine the quality of summarization of this model. The text file was initially stripped off the standard Gutenberg license header and footer, and then it was broken into twenty-five chapters by a rule-based parser. The results of preprocessing was about 419,000 meaningful characters and all the chapters were successfully extracted to be summarized separately.

The preprocessing step also involved sentence tokenization, normalization of whitespace and preparation of text blocks by chapters. These blocks were significant because it would be out of context to summarise the entire novel concurrently, but it would be more coherent to summarise by chapter and yield higher ROUGE scores. Following preprocessing, there were 2 summarization techniques, a basic extractive baseline, and a transformer-based abstractive model. The extractive baseline is a simple heuristic, which consists of sentence position, length and keyword relevance to select twelve important sentences in the novel. This method, fast and computationally inexpensive, generated summaries that had retained original text and in most cases lacked narrative cohesion.

The main summarization model was the BART-Large CNN transformer model which is a sequence-to-sequence model that is normally applied in the context of summarization. It is designed based around using an encoder-decoder architecture where it will encode the input chapter to high-level contextual embeddings and the decoder outputs rewritten text, which is conditioned on the encoder output and self-conditioned at past positions. The model was deployed in the Apple MPS GPU backend which is far quicker to perform inferences in the Mac system. Various hyperparameters were optimized to get good quality summaries, including maximum length (200 to 250 tokens) and minimum length (50 to 70 tokens). Four beam search with repetition penalty of 1.5 was performed to enhance coherence and prevent looping.

The abstractive summaries that were produced by BART were far more readable and understandable than the extractive baseline. They succeeded in capturing the most important moments of each of the chapters such as the very first interest of the scientific world in Victor Frankenstein, creating the monster, the letters written by Robert Walton to Margaret Saville, the murder of William, and the emotional scandal that propels the rest of the story. The model also reflected the same themes on the final combined summary, where the model made the right references of the leading characters and plot development.

To gauge the output of the summarization, the ROUGE-1, ROUGE-2 and ROUGE-L measures were computed and stored. The scores were the comparison of overlap of the abstractive

summaries and extractive baseline. The ROUGE-1 had a precision of 0.5649, recall of 0.0571 and F1 score of 0.1037. The ROUGE-2 had a precision of 0.0923, recall of 0.0093 and an F1 score of 0.0168. The ROUGE-L had a precision of 0.3588, recall of 0.0363 and F1 score of 0.0659. Those results indicate that the recall is low because the abstractive summaries are highly compressed, but the precision and structural coherence are high. This is typical with the abstractive models that produce much shorter records compared to the reference summaries.

This project indicates the usefulness of transformer-based abstractive summarization of long-form literary text. The BART model effectively created summaries of the chapters that were concise and meaningful and it was more successful than the extractive baseline in terms of coherence and storyline. In spite of low ROUGE recall because of the compressed nature of the summaries, high precision and ROUGE-L performance demonstrate that the produced summaries were structurally sound and thematically accurate. This project verifies that transformer models with the appropriate preprocessing and parameter optimization are a robust and efficient model that can be used to summarize full-length novels.