CS4372 Project 4 Report

Salma Khalfallah
SMK210009

November 21, 2025

1 Introduction

Text summaries are one of the most important applications of a Transformer, which are able to take an input of text and output a shorter version of the same text while retaining the most valuable information from the text. Previously seeming like an impossible task, transformer models and the attention method have allowed summaries to be generated by anyone with a computer and a little bit of programming knowledge.

For the project, I have chosen the book "In the Land of Mosques and Minarets" by Francis Miltoun from Project Gutenberg. The book explores the author's journey across the North African region, specifically Algeria and Tunisia, and invites readers to explore the rich tapestry of a land often misunderstood. Due to the length of the book, I elected to summarize a single chapter as opposed to the entire book to maintain a computationally-accessible project.

2 Structure of Transformers in the Context of Summarization

In order to begin to perform modeling, it is imperative to understand what is going on underneath the hood transformer-car. (transformobile?) The engine of the transformer relies entirely on the attention method, which is a method that determines the importance of a word relative to the ones around it. This allows our summarization model to obtain the most important information from the original text. During the encoding process, the input text is examined, and key words and phrases are extracted. Finally, the decoder generates a new summary based off of the encoder's feature extraction which (hopefully) encapsulates the essence of the original text. This architecture forms the basis of the summarization model that will be utilized for the task at hand.

One important aspect to remember: as a student, I simply do not have the resources or time to be able to train a transformer model from scratch. This is where transfer learning comes into play. Transfer learning is simply utilizing

a pre-trained model from a large word corpus in order to train new text. For the task at hand, two models will be tested on: Facebook's 'bart-large-cnn' and Google's 'google/pegasus-xsum' model. These models are:

1. Trained on a large number of data
2. Contain millions of parameters
3. Is easily accessible and readily available for use

For these reasons, these models will be our choice of model for the summarization task.

3 Introducing the ROUGE Score

In order to evaluate the quality of the summaries, the ROUGE metric should be introduced. The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of metrics that rate the quality of a text summary by comparing the number of similar words in an n-long extraction of a generated summary when compared to some referential text material. There are a couple of metrics available, however they all roughly follow the same formula of a ROUGE-N formula, where:

$$Rouge - N = \frac{\sum_{i \in N} CountMatch(i)}{\sum_{i \in N} CountReference(i)}$$

There is also a ROUGE-L metric, which takes into account sentence level structure similarity and identifies the longest co-occurring n-gram sequence.

$$Rouge - L = \frac{LCS(Generated, Reference)}{Count(\text{Words in reference summary})}$$

These metrics will be utilized in order to identify the quality of generated summary from each model.

4 Modeling and Results

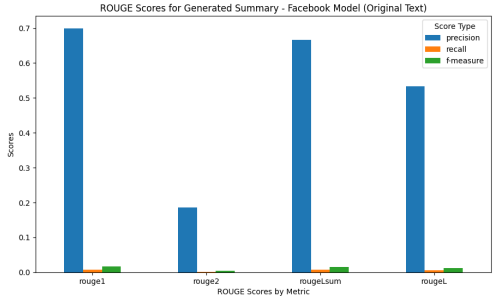
Before evaluating the model summaries, I had to develop my own human-made summary of the data to provide as a frame of reference for comparison with the models. I generated the following summary:

> "The text explores a brief summary of the state of Tunis and Algeria, including cultural norms and colonial history in the region."

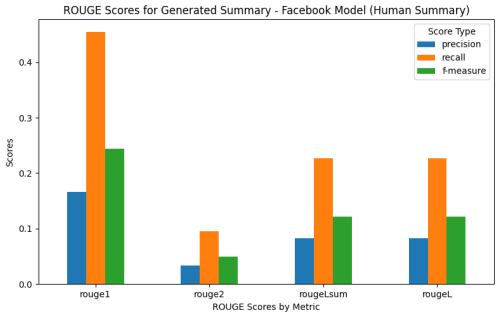
Both of the models struggled to compare to this summary. The first model, Facebook's BART model, given parameters of `min_length = 30` and `max_length = 240`, generated the follow summary:

"France, Germany, Italy, Spain, Morocco, Algeria, Nigeria, Egypt, South Africa, Nigeria and Algeria are the most prosperous nations in the world. But the continent is not without its problems, including poverty, disease, unemployment, racism, sexism, homophobia, racism and homophobia. The continent has a long way to go before it can be truly considered a "melting pot" of rich and poor."

When evaluating for the model, I decided to include both ROUGE metrics when comparing to the original text data as well as the human-made summay (see: above). The results of the model can be seen in the visual below



(a) ROUGE Scores when using original text data as reference material



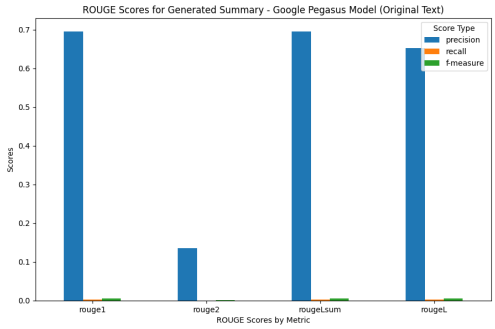
(b) ROUGE Scores when using human-made summary as reference material

Figure 1: ROUGE Scores of Modeling on Original Text and Human-Generated Summary

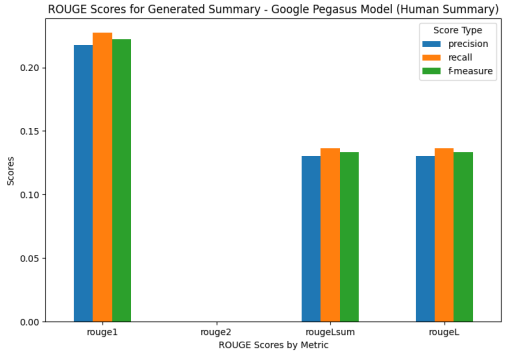
Both the human eye as well as the ROUGE scores can determine that the Facebook model did not do the best job in obtaining a good summary. Notice: the highest ROUGE score is consistently the ROUGE-1 metric in both visuals. This indicates that the model performed best with identifying some key phrases, such as "Algeria" (which is featured prominently in the original text

data. However, the model struggles (and fails) to determine how to *connect* those key phrases in a cohesive manner that captures the essence of the original text data. The following phrases do not appear in the original text data despite appearing in the generated summary: "Germany, Nigeria, South Africa, prosperous, nations, world, poverty, unemployment, racism, sexism, homophobia, "melting pot"" and more. Similar results can be visualized for the second Google-Pegasus model

> "In our series of letters from African journalists, film-maker and columnist Youcef Youcef looks at the contrasts between Europe and North Africa."



(a) ROUGE Scores when using original text data as reference material



(b) ROUGE Scores when using human-made summary as reference material

Figure 2: ROUGE Scores of Modeling on Original Text and Human-Generated Summary

I would like to immediately note: Upon a quick search, the journalist "Youcef Youcef" does *not* exist. The ROUGE scores seem to agree with me.

Notice that all of the evaluations tended to have a relatively high precision score. This means that the ROUGE metrics believe the models were relatively successful in retaining information relative to the original text. On the other hand, the ROUGE scores indicate a low recall metric. This indicates that a good chunk of the original data is missing from the final summary and is not reflective of the entire text data. Finally, we obtained low F-1 measures with a high of $\sim 20\%$ and a low of 0%. This indicates a balance between the recall and precision measure is not present, which is what is desired.

There are different reasons as to why the model did poorly, but one fundamental reason for the struggling models is due to the data in which the original Facebook and Google models were trained on respectively. For example, the Facebook dataset is trained primarily on CNN and Daily Mail article data. This is modern, news-based text data that trains on specific vocabulary and sentence structures. On the other hand, the original inputted text data from Gutenberg introduces new elements of language that the program has not yet learned. As a result, the model is able to identify key words and elements, i.e. why the 1-gram ROUGE metrics did relatively better. However, it fails at higher n-grams for this very same reason.

5 Conclusion

In conclusion, although text summarization is a useful tool, care must be taken in every step of the way in order to extract the best possible output. This includes care and consideration in model selection based on input data, computational feasibility, and more. Although the base models selected for the summarization task are generally strong in abstract summaries, they failed when introduced to new, foreign text elements such as sentence structure and vocabulary. In the future, as an extension, I would like to continue by performing more model finetuning and base model selection such that the final output is stronger and better reflects the essence of the original inputted text data.