Answer Generation Task Report

Jay Saraf

1. Introduction

Generating answers based on the question answer pair is a task in Natural Language Processing, which requires certain aspects of question answering and certain aspects of text generation. In this case study, two models were used one being BART and other being FlanT5 (one of the varients of T5 model). Both of these models are useful in text generation task incomparison to other models. BERT in particular is useful for extractive summarization. But in the given case study, the model should be able to answer any new question, which might be a bit related to the existing dataset on which the model is trained on.

2. Literature Review

Large pretrained models like BART and T5 have been found use for sequence to sequence text generation tasks. BART is a combination of BERT and GPT[2]. BERT is a bidirectional encoder model and GPT is a unidirectional transformer. BERT alone cannot be used for text generation tasks.

3. Methodology

3.1 Data Exploration and Cleaning:

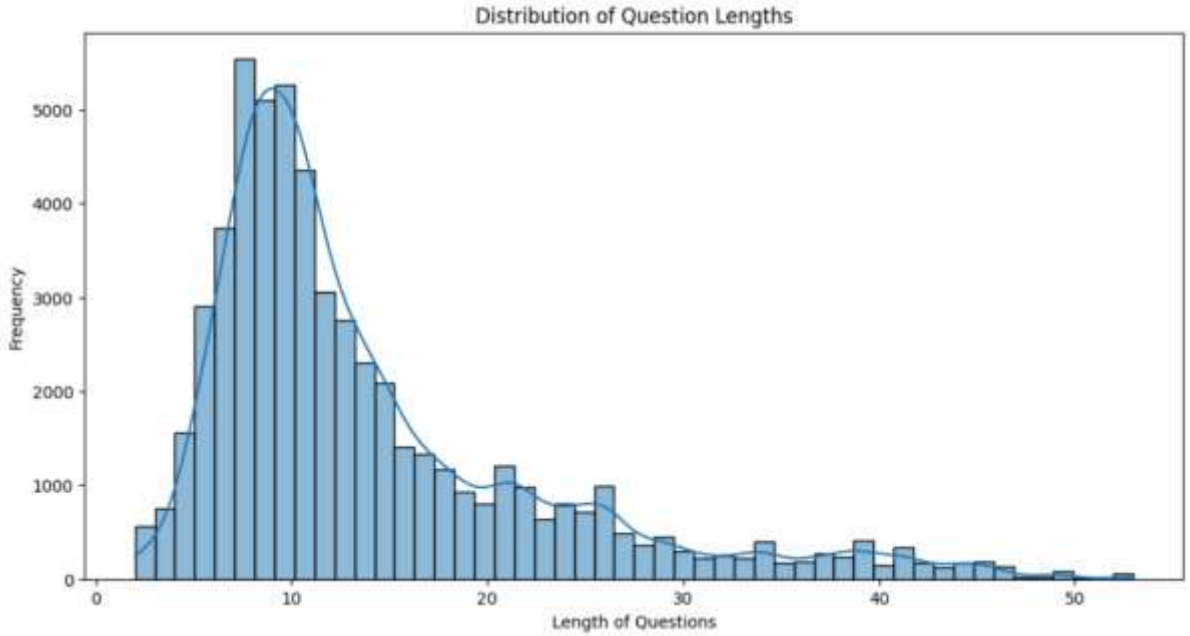
The dataset was loaded using the dataset library. The dataset can also be used by first downloading from the hugging face website.

For exploring the dataset I printed out the format of the input data, the data type being used. There are two columns :- question and answer. For the training dataset there are a total of 56402 question-answer pairs.

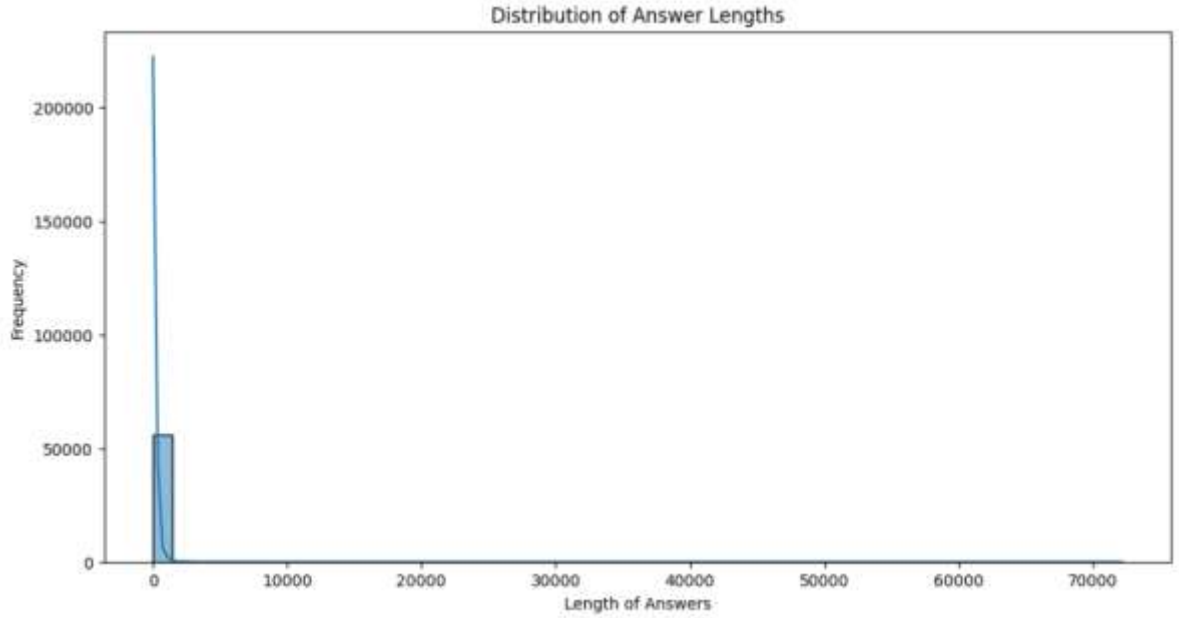
I have checked whether there were any missing values.

For data preprocessing, I first done expansion on contractions, after which I have lower cased all the text, removed html tags, removed non alphabetic character by blank spaces and the removed extra spaces from the text. Stop words are quite essential when text generation tasks are done this is because stopwords provide context to the text and this is the reason I haven't done stopword removal and for similar reason I haven't done stemming or lemmatization.

3.2 Data visualization



The above is the distribution of the length of the questions in the training dataset.



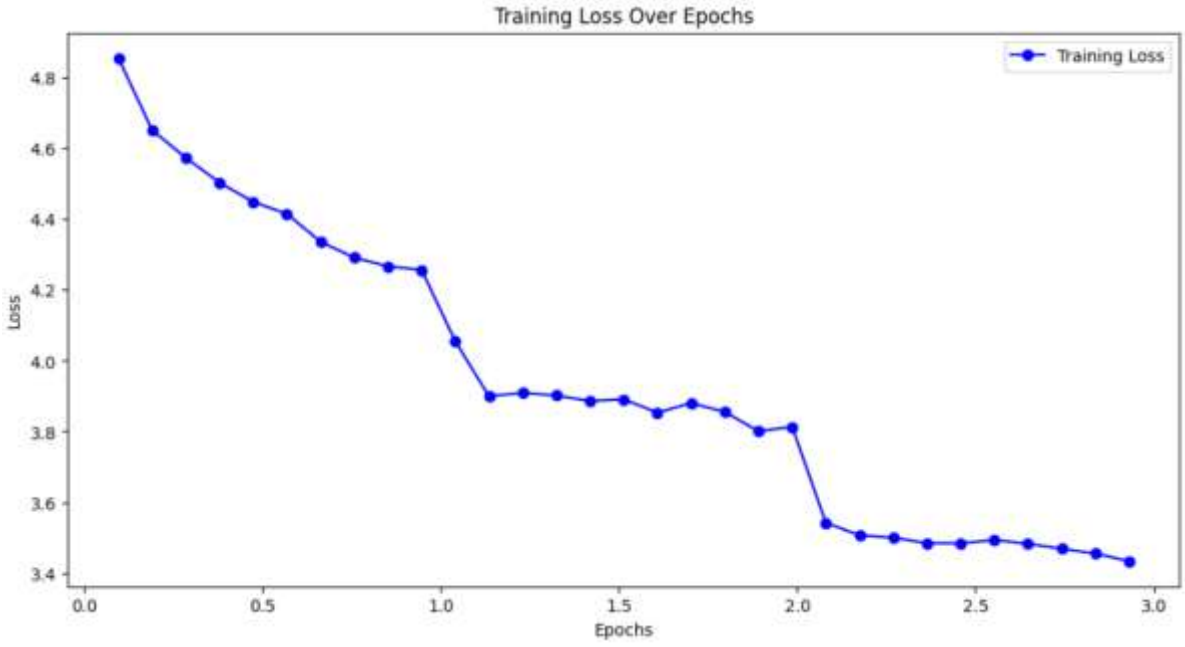
The above is the distribution of the length of the answers in the training dataset.

The above image is of the word cloud of the words used in the questions of the training dataset. This show the most frequently used words in the dataset.

The above image is of the word cloud of the words used in the answers of the training dataset. This show the most frequently used words in the dataset.

4. Results

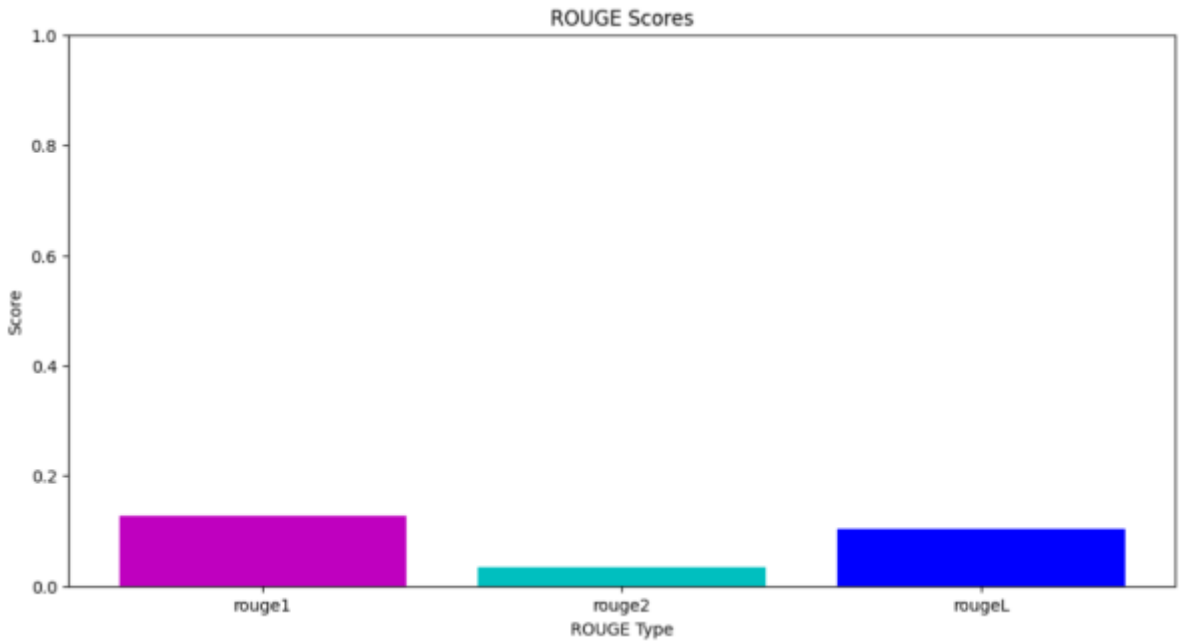
1. For BART Model



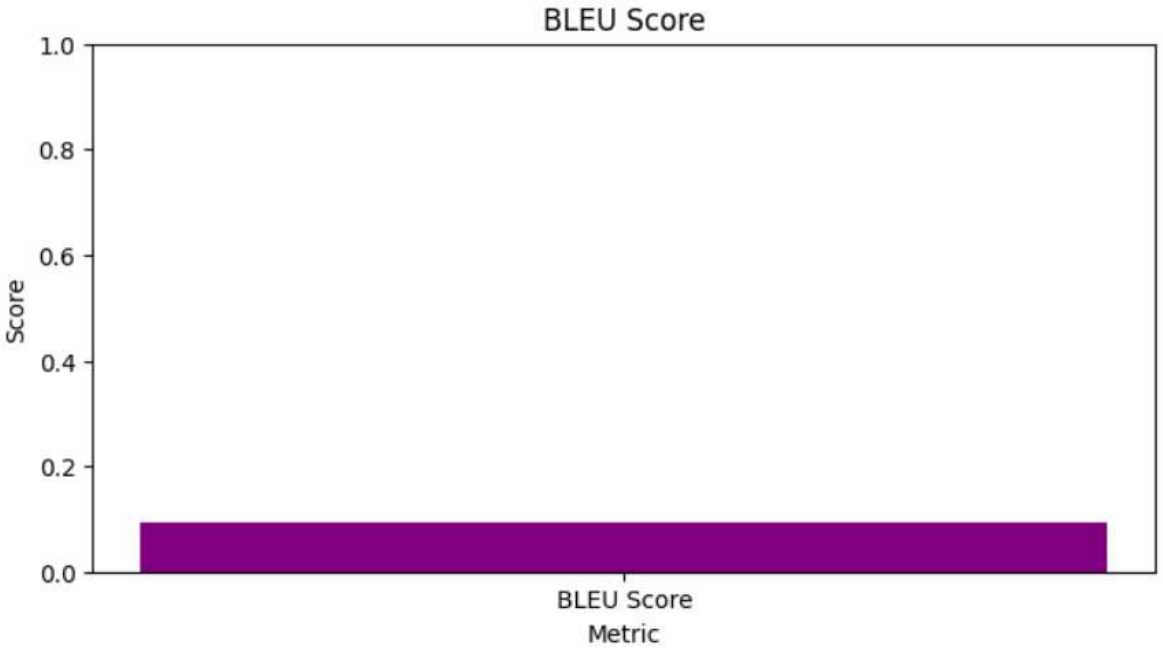
The above graph is the training loss of BART base model over 3 epochs.

Epoch	Training Loss	Validation Loss	Rouge1	Rouge2	Rougel	Bleu	F1
1	4.256600	4.030489	0.122870	0.028983	0.097937	0.093535	0.002318
2	3.813400	3.832710	0.127091	0.031535	0.100770	0.098249	0.002653
3	3.433600	3.762687	0.128153	0.033619	0.102480	0.094127	0.002721

The above table contains all the metrics like Rouge score, Bleu score and f1 score, which are required to evaluate the performance of the model.

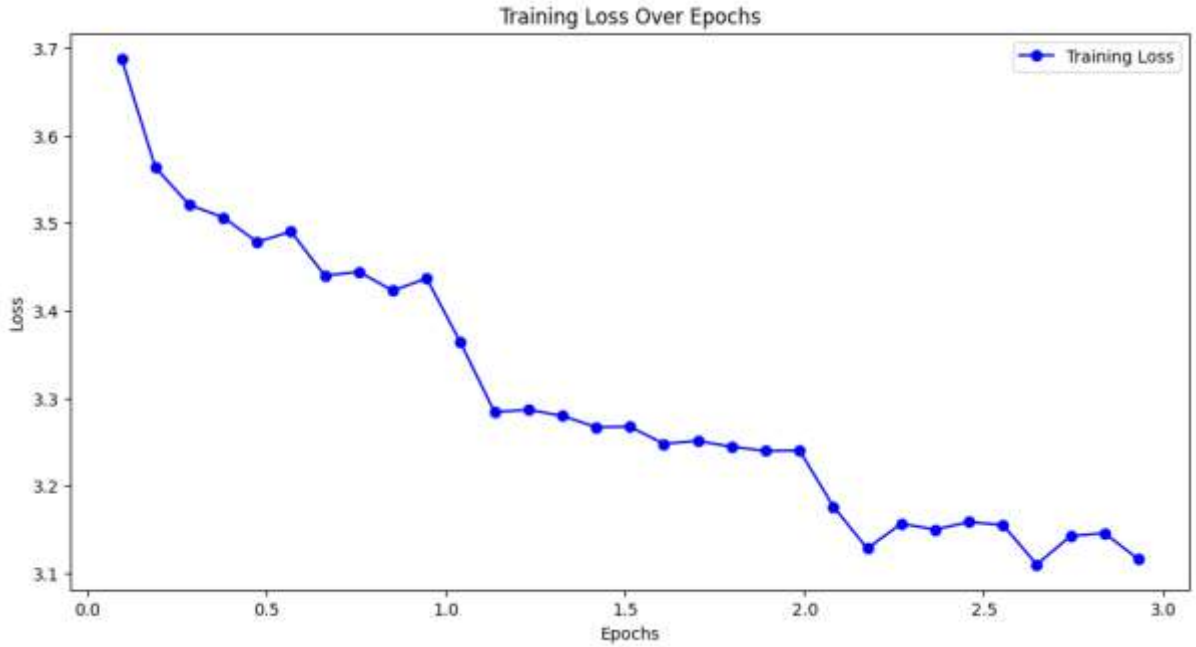


The above graph depicts the rouge score values of BART model.

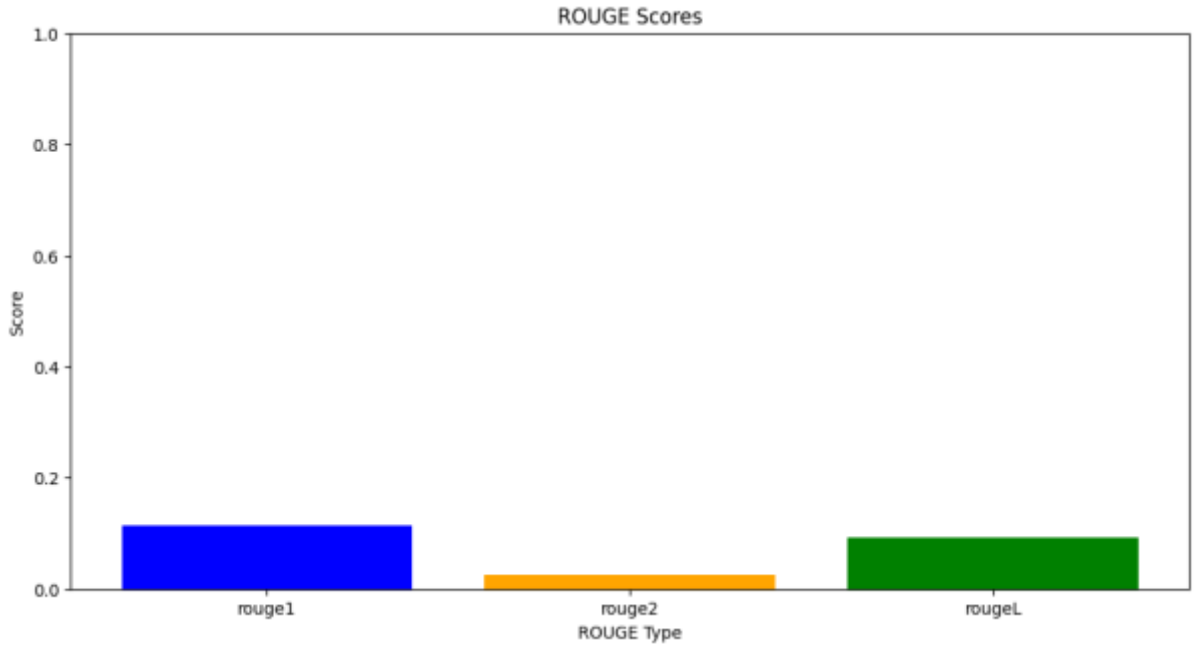


The above graph depicts the bleu score for the BART model.

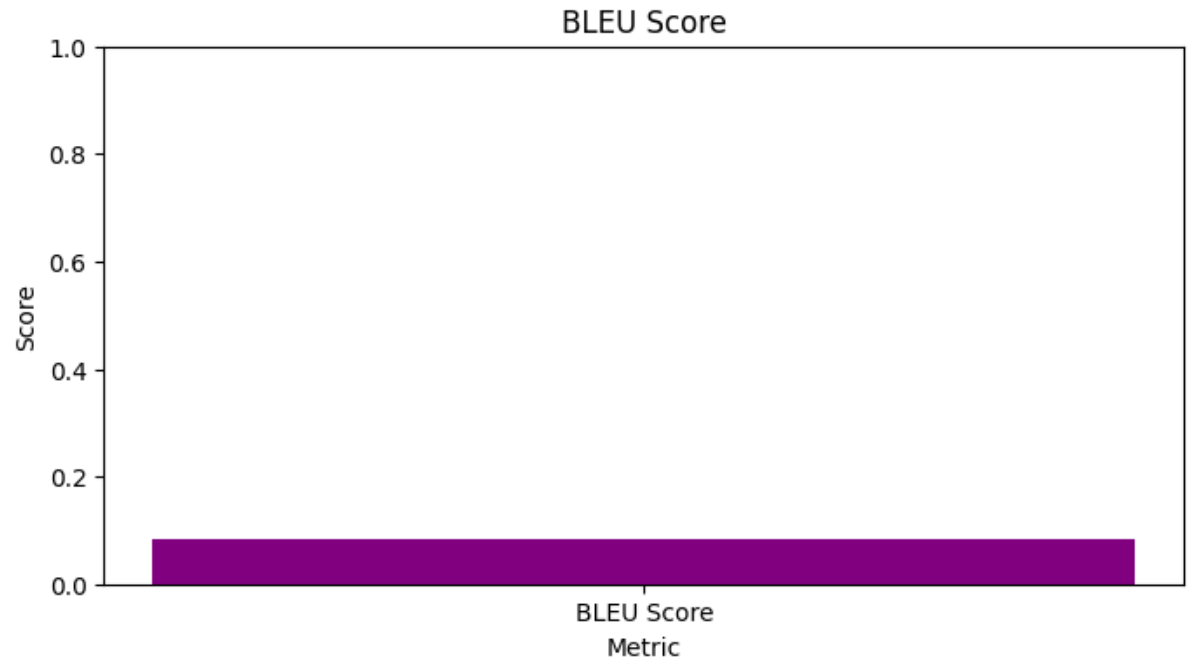
2. For Flan T5 Model



The above graph is the training loss of Flan T5 model over 3 epochs.



The above image is of rouge scores of Flan T5 model.



The above image shows the Bleu score for Flan T5 model.

5. Conclusion

Evaluation results are as follows:

	eval loss	rouge1	rouge2	rougeL	bleu	f1

BART	3.76	0.12	0.03	0.10	0.090	0.0027
FLAN T5	3.12	0.11	0.02	0.09	0.084	0.0023

BART model gave better results in terms of the output. The answers generated by BART were more coherent in comparison to T5.

6. **Insights and Recommendations:**

Latest versions of BERT and GPT can be combined together to get better results. It is not necessary to remove stopwords. I tried not removing any non-alphabetic character and results were still good. If models are trained for more epochs than the results will come out to be even better. Human in the loop can be consider by upvoting or downvoting the answer generated by the model. Extractive method could be applied before training generative models.

7. References:

[1] https://arxiv.org/pdf/2311.02961

[2] https://arxiv.org/pdf/1910.13461

[3] https://arxiv.org/pdf/2002.10832