A Comparative analysis of Abstractive Dialog Summarization using fine tuned LLMs

ANM Zahid Hossain Milkan
ID 200041202
Computer Science and Engineering
Islamic University of Technology
Dhaka, Bangladesh
zahidmilkan@iut-dhaka.edu

Amit Bin Tariqul
ID 200041214
Computer Science and Engineering
Islamic University of Technology
Dhaka, Bangladesh
amit20@iut-dhaka.edu

Sahab al Chowdhury
ID 200041255
Computer Science and Engineering
Islamic University of Technology
Dhaka, Bangladesh
sahab@iut-dhaka.edu

***Abstract*—Summarization refers to the compression of voluminous data into a shorter version for easy interpretation and grasping main content of data while maintaining context and relevance. The tremendous bulk of information available today from various sources tends to be overwhelming to understand, thus emphasizing the necessity of summaries or briefings. They contribute to effective understanding of subject matter and influence quick decision making which can lead to higher dynamism altogether. Over the years the mainstream has been divided into two main categories based on representation - Extractive Summarization and Abstractive Summarizatrion. Furthermore based on the approach taken to produce summaries there are two divisions - Supervised Learning and Deep Reinforcement Learning. In the supervised abstractive approach, varying datasets, both existing and new are used to analyze various existing LLMs to gain deeper understanding of dataset, model characteristics and real life scenario. And in the other approach, human judged specific evaluation criteria utilized for training and better analysis perspective. This project takes an attempt to understand the nature of dialog based data, their structure and effect on various models, their interpretation by two contemporary popular LLMs namely BART and PEGASUS and their performance on such data. Based on the result, a rough idea can be obtained as to model performance on specific criteria data and enhanced performance of transfer learning based approach as to how it improves over raw processed results. Moreover, various metrics to evaluate performance, how they vary and why also can be grasped. Furthermore it helps gain further insight how recent newer models outperform previous models and how impactful the model architecture can be to produce better results overall.**

***Index Terms*—summarization, SAMSUM, BART, PEGASUS, abstractive summarization, reinforcement learning, extractive summarization**

I. Introduction

The rapid growth of information communication technology acts as the infrastructure of the modern century. The degree of personal communication has massively increased owing to easy availability and access to technology giving rise to huge amount of information. Newer information is being utilized, produced and processed everyday. While the huge amount, easy access and availability serves a great purpose for the futuristic prediction or quality assurance purposes, interpretability wise, comprehending such huge levels of information is a cumbersome job for humans on a personal level. A summarized version of the data maintaining the context is therefore necessary.

Techniques of Summarization

Among recent techniques and approaches taken to summarize data to bring out meaningful information, two broader categories are noticed. Extractive and Abstractive Summarization [1].

A. Extractive Summarization

In this form of summarization, the golden rule often followed is *Stick to the Script*. So the summary is prepared by taking bits and parts of the sentences and words from within the voluminous given data itself. It does not aim to producing new data or any novel representation, but rather a strict reduction approach to produce meaningful summary from data. The advantage of this method of summarization is the simplicity, the preservation of original text and easier computation and accuracy higher. However as this approach does not paraphrase, it is susceptible to redundancy, has lower quality and lower level of abstraction as it is only selective, not generative.

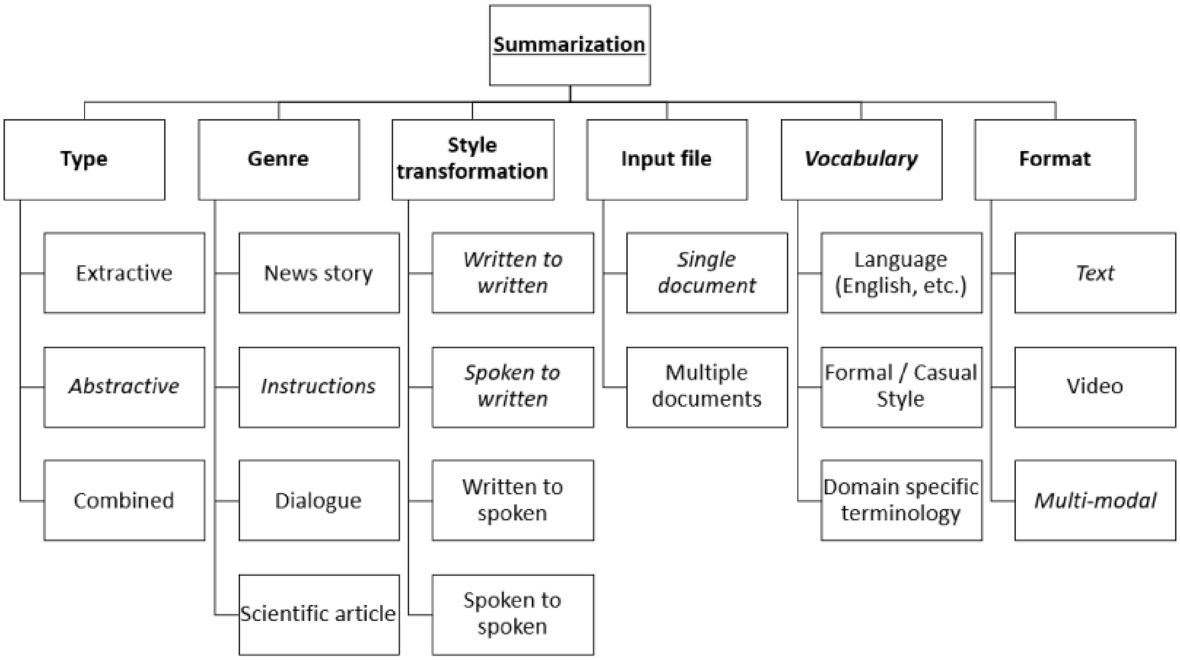
B. Abstractive Summarization

In this form of summarization, the generation of new data maintaining context is encouraged. So the summary is prepared by understanding subject matter and paraphrasing to form new interpretation of data. The advantage of this method of summarization is the higher quality, flexibility and creativity. However it has many disadvantages as well. As this approach paraphrases, it is much harder to produce and maintain context, higher computation problem and error.

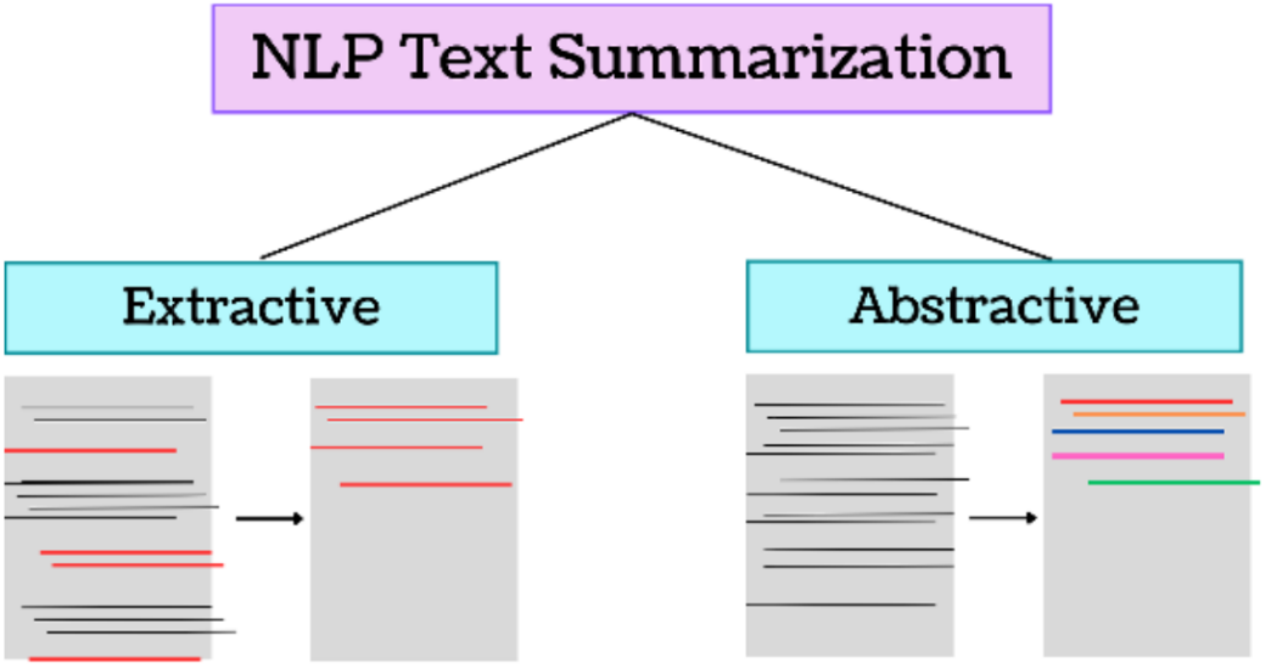
Method of learning

Supervised Approach

In this approach we take labeled datasets to utilize them and train model parameters, optimize and fine tune them to specific purpose and then use them to produce test and evaluate result of test data. Lack of good dataset or too little information is a persistent problem in this method. However if sufficient data available this is the best performing and most common



(a) Different categories of Summarization



(b) Abstractive vs Extractive summarization figurative

Fig. 1: Illustrations of Summarization Types

Though extensive categories of summarization exists, they are primarily of the two types in objectivity.Abstractive and extractive summarization can be better understood from here.

learning method available. It needs labeled data from which models learn and adjust their weights to better suit itself to perform on that dataset. Transfer learning is also applied in this method for achieving task specific results and data.

Reinforcement Learning approach

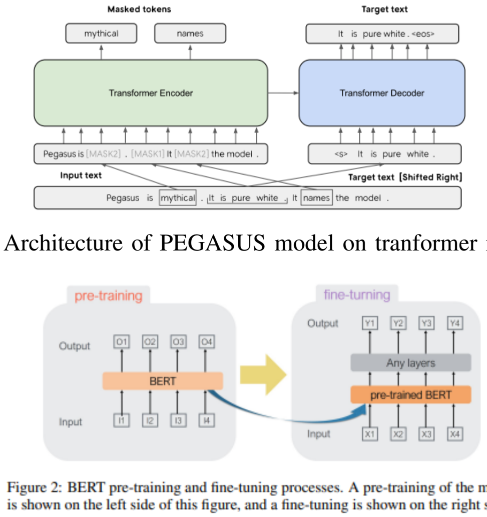
This method points out the pros and cons of extractive and abstractive summarization by supervised approach and alternatively proposes a deep reinforcement learning based approach [2] utilizing semantic similarity [3] measured by humans as a criteria for reward based system. This approach uses human evaluation metric to train a log loss likelihood teacher force algorithm to learn better for this purpose

Among all forms of communication language and dialogue is the most prevalent by a country mile. The amount of information exchange through dialog dominates the current information exchange. This is the motivation which has encouraged us to take an attempt at analyzing the Dialogue based Abstractive summarization which can allow greater understanding of human conversational trends and improve and establish newer models. Much research work has been done in this regard as some mentioned above and definitely will be in the future and these act as the background to this attempt which hopes to make some contribution in current existing methods and approaches.

II. Methodology

Models

For our purpose we have used the two different LLMs suited to this purpose.One of them is the facebook BART [7] modelpretrained on daily cnn-daily-mail dataset and the other is the PEGASUS [8] model also pretrained on the same dataset.



(a) Architecture of PEGASUS model on tranformer model

Figure 2: BERT pre-training and fine-tuning processes. A pre-training of the model is shown on the left side of this figure, and a fine-tuning is shown on the right side.

(b) Architecture of BART model

Fig. 2: The two model architectures

The BART is a sequence to sequence model primarily used for language transduction.Unlike other models this model takes into consideration both left and right words as context for generation.Working on the basis of Masked Language Model(MLM) and Next Sentence Prediction, it generalizes well to task specific datasets for purpose specific results.

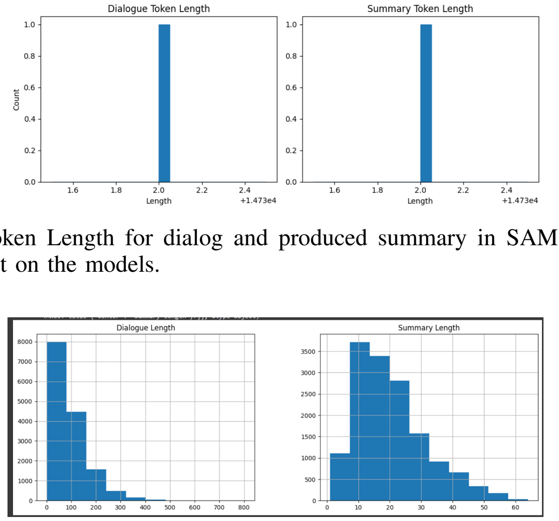
The PEGASUS model is also similar to BART in functionality.However it is shown to be much more specific for summarization purpose particularly abstractive summarization of both text and dialog formats.It is special however in the sense that it is highly efficient in few shot learning.It performs well even with constrained dataset and generalizes well.

Datasets

SAMSum: For our dialog summarization purpose we have used the SAMSum Dataset [4], which is specifically designed for dialog summarization. The SAMSum dataset contains about 16k messenger-like conversations with summaries. Written by linguists fluent in English, this dataset is very enriched in the sense that it contains informal, semi-formal, or formal language and they may contain slang words, emoticons, and typos. Then, the conversations are annotated with summaries. It is prepared by Samsung R&D Institute Poland. This dataset is formed of 16369 conversations distributed uniformly into 4 groups. 75% of data is between two person, the rest have 3 or 4.
The data has 3 fields namely -
dialogue: text of dialogue.
summary: summary of it by humans.
id: unique id of each (not needed for our work).
Also it has a split like this-
train: 14732
val: 818
test: 819

It is used for the fine tuning purpose of the pipelines generated from PEGASUS and BART models in our purpose and the result generated is evaluated to measure performance. It is to be noted that these two datasets are previously pretrained on cnn-daily-mail dataset which allows it to be used for transfer learning for task specific purpose like ours - abstractive dialog summarization.

CNN-Daily Mail: The CNN daily mail dataset was used to pretrain the models we have worked with.The CNN / DailyMail Dataset is an English-language dataset containing just over 300k unique news articles as written by journalists at CNN and the Daily Mail. The current version supports both extractive and abstractive summarization, though the original version was created for abstractive QA.
It has three fields - **ID, Article, Highlights**.It has a split of -
Train: 287k rows **Test:** 13.4k rows **Validation:** 11.5k rows



(a) Token Length for dialog and produced summary in SAMSUM dataset on the models.

(b) The dialog length and the produced summary lengths for the dataset training set and so on.

Fig. 3: Token and Dialog Lengths

It can be clearly interpreted from the figure as to the constraints and tokenization to be maintained and suitability of model based on selection criteria.After the model is run the results are analyzed and decision is reached.

III. Result analysis

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is a set of evaluation metrics for automatic summarization and machine translation. It measures the corresponding similarity between the model-generated summaries and reference summaries. ROUGE-1 and ROUGE-2 are two commonly used variants of this metric.

ROUGE-1

ROUGE-1 evaluates the overlap of unigrams (single words) between the generated summary and the reference summary. It calculates the precision, recall, and F1 score based on these unigrams.

Precision: The proportion of unigrams in the generated summary that are also present in the reference summary.

Recall: The proportion of unigrams in the reference summary that are also present in the generated summary.

F1 Score: The harmonic mean of precision and recall.

ROUGE-2

ROUGE-2 evaluates the overlap of bigrams (pairs of consecutive words) between the generated summary and the reference summary. It also calculates the precision, recall, and F1 score based on these bigrams.

Precision: The proportion of bigrams in the generated summary that are also present in the reference summary. Recall: The proportion of bigrams in the reference summary that are also present in the generated summary. F1 Score: The harmonic mean of precision and recall.

For our result analysis of the models we have used four common metrics to calculate the overall score for their comparative analysis of the models.The metrics formulas and their calculation is given in the following tables.

$$\begin{aligned}
\text{Accuracy} &= \frac{TP+TN}{TP+TN+FP+FN} \\
\text{Precision} &= \frac{TP}{TP+FP} \\
\text{Recall} &= \frac{TP}{TP+FN} \\
\text{F1} &= \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}
\end{aligned}$$

Metrics	Precision	Recall	f-measure
rouge1	0.48571	0.56667	0.52308
rouge2	0.20588	0.24138	0.22222
rougeL	0.37143	0.43333	0.40000
rougeLsum			0.40038

TABLE I: Various measures for the facebook bart-large-cnn

Metrics	Scores
rouge1	0.4659
rouge2	0.2345
rougeL	0.3946
rougeLsum	0.3951

TABLE II: Various Scores for PEGASUS Model on various metrics

From the tables it can be observed that pretrained models upon transfer learning perform much better on purpose specific datasets.PEGASUS model here though nearly on par with BART has shown much better performance for this dialogue summarization task than BART. The produced text summaries if taken to human evaluation level, provides verdict in favour of the PEGASUS.

References

[1] S. Gehrmann, Y. Deng, and A. M. Rush, 'Bottom-Up Abstractive Summarization," *arXiv preprint arXiv:1808.10792*, 2018. https://arxiv.org/abs/1808.10792

[2] H. Jang and W. Kim, 'Reinforced Abstractive Text Summarization With Semantic Added Reward," in *IEEE Access*, vol. 9, pp. 103804-103810, 2021, doi: 10.1109/ACCESS.2021.3097087. https://ieeexplore.ieee.org/document/9486438

[3] Beken Fikri F, Oflazer K, Yanıkoğlu B. Abstractive summarization with deep reinforcement learning using semantic similarity rewards. Natural Language Engineering. 2024;30(3):554-576. doi:10.1017/S1351324923000505. Abstractive summarization with deep reinforcement learning using semantic similarity rewards

[4] Gliwa B, Mochol I, Biesek M, Wawer A. SAMSum Corpus: A Human-annotated Dialogue Dataset for Abstractive Summarization. Proceedings of the 2nd Workshop on New Frontiers in Summarization, Association for Computational Linguistics. November 2019. arXiv:1911.12237 [cs.CL]. doi:10.48550/arXiv.1911.12237. https://arxiv.org/abs/1911.12237

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv:1810.04805 [cs.CL], October 2018. https://doi.org/10.48550/arXiv.1810.04805

[6] Alexandra Savelieva, Bryan Au-Yeung, Vasanth Ramani. *Abstractive Summarization of Spoken and Written Instructions with BERT*. arXiv:2008.09676 [cs.CL], August 2020. https://doi.org/10.48550/arXiv.2008.09676

[7] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer, ”BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension”, arXiv:1910.13461 [cs.CL], 2019.

[8] Jingqing Zhang, Yao Zhao, Mohammad Saleh, Peter J. Liu, ”PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization”, arXiv:1912.08777 [cs.CL], 2019.