INDIGO PROJECT REPORT

(Hack to Hire 2024)

Question Answer Model

SUBMITTED BY-

Rewaa Mital

INDEX

INTRODUCTION

In the realm of natural language processing (NLP), developing sophisticated question-answering (QA) systems has been a paramount goal due to their wide-ranging applications in customer service, information retrieval, and virtual assistants. This project aims to create a state-of-the-art question-answering model utilizing the Quora Question Answer Dataset, which is renowned for its diverse and complex question-answer pairs. The primary objective is to develop an AI system that can understand and generate accurate responses to various user queries, thus mimicking human-like interaction. By leveraging advanced models such as BART and FLAN-T5, this project endeavors to push the boundaries of current QA systems and deliver a highly efficient, reliable, and scalable solution.

PROBLEM STATEMENT

Develop a state-of-the-art question-answering model leveraging the Quora
Question Answer Dataset. The objective is to create an AI system capable of understanding and generating accurate responses to a variety of user queries, mimicking a human-like interaction.

LITERATURE REVIEW

SR NO.	TITLE	AUTHOR	YEAR	METHODOLOGY	LIMITATIONS
1.	Enhancing Question Prediction with Flan T5 -A Context-AwareLanguage Model Approach	Jay Oza, Hrishikesh Yadav	2023	The technique incorporates a memory mechanism to maintain the created question inside the given parameters. Because of the model's effectiveness in gathering context and formulating relevant questions, user engagement is improved, which promotes better results in a variety of applications. A methodical approach to building the machine learning model is covered in the research, which includes phases for data collecting, preprocessing, tokenisation, model implementation, and fine-tuning.	The research has limitations, including the dependency on dataset size and diversity, which impacts model generalizability. The significant computational resources required for larger models pose practical constraints.

2.	BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension	Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer	2019	This paper uses a two-step approach to train a denoising autoencoder called BART: first, a noising function is used to damage text, and then a model is trained to recover the original text. BART combines a left-to-right decoder and a bidirectional encoder in a Transformer-based sequence-to-sequence architecture. The study assesses several noising techniques and determines that employing an in-filling strategy and randomly rearranging the sequence of sentences yields the best results. Subsequently, BART is optimised for text production and assessed on comprehension assignments, attaining cutting-edge outcomes across various NLP benchmarks.	The limitations of the paper include the reliance on specific noising functions for text corruption, which may not be optimal for all tasks. Additionally, while BART performs well on both discriminative and generative tasks, the study suggests that future work should explore more tailored corruption methods to potentially enhance performance for specific end tasks.

SR NO .	TITLE	AUTHOR	YEAR	METHODOLOGY	LIMITATIONS
3.	Question and Answer Test-Train Overlap in Open-Domain Question Answering Datasets	Patrick Lewis, Pontus Stenetorp , Sebastian Riedel	2020	The methodology involves analyzing three open-domain benchmark test sets, revealing that 60-70% of test answers and 30% of test questions are also found in the training sets. The study evaluates various models, showing a 63% performance drop on non-memorized questions. It also finds that simple nearest-neighbor models outperform a BART closed-book QA model, highlighting the impact of training set memorization.	The limitations of the paper include a high overlap between training and test sets, which questions the true generalization capability of the evaluated models. The findings suggest that current evaluation methods relying on overall QA accuracy are insufficient and should shift towards behavior-driven evaluation to better assess model performance.

4.	Can Generative Pre-trained Language Models Serve as Knowledge Bases for Closed-book QA?	Cunxiang Wang, Pai Liu and Yue Zhang	2021	The paper creates a new closed-book QA dataset using SQuAD to evaluate BART's performance as a knowledge base for answering open questions. The study examines BART's ability to remember training facts and answer questions with retained knowledge. It explores the challenges faced by BART and identifies promising directions, such as decoupling the knowledge memorization process from the QA fine-tuning process and forcing the model to recall relevant knowledge during question answering.	The limitations of the paper include the difficulty generative pre-trained models like BART face in accurately remembering and utilizing detailed knowledge for closed-book QA. The study highlights the challenge of answering questions even with retained knowledge and suggests the need for improved methods, such as decoupling the LM-finetuning and QA-finetuning processes and explicitly prompting models to recall relevant knowledge.

SR NO.	TITLE	AUTHOR	YEAR	METHODOLOGY	LIMITATIONS
5.	DQ-BART: Efficient Sequence-to-Sequence Model via Joint Distillation and Quantization	Zheng Li, Zijian Wang, Ming Tan, Ramesh Nallapati, Parminder Bhatia, Andrew Arnold, Bing Xiang, Dan Roth	2022	The methodology involves jointly distilling and quantizing large-scale models like BART and T5 to reduce memory and latency. Knowledge is transferred from a full-precision teacher to a low-precision student model, achieving up to 27.7x compression with minimal performance loss. The work is the first to effectively apply these techniques to pre-trained sequence-to-sequence models for language generation.	The limitations of the paper include the challenge of joint low-bit quantization and distillation for deeper models, particularly in cross-lingual tasks, and the omission of additional compression techniques like attention head pruning and sequence-level distillation. The study also leaves the measurement of latency improvements in various settings for future work.

METHODOLOGY

The methodology of this project is structured to systematically develop and evaluate a state-of-the-art question-answering (QA) model leveraging the Quora Question Answer Dataset. The key steps in the methodology include data preprocessing, model selection, training, evaluation, and real-time interaction implementation.

Data Preprocessing

Data preprocessing is a crucial step in preparing the dataset for training machine learning models. The Quora Question Answer Dataset, containing pairs of questions and corresponding answers, is first loaded and split into training and testing sets. This ensures that the model can be evaluated on unseen data to gauge its generalization capability.

1. Dataset Loading and Splitting: The dataset is loaded using the datasets library, and split into training (80%) and testing (20%) subsets. This split ensures a sufficient amount of data for both training the model and evaluating its performance.
2. Tokenization: Tokenization is the process of converting text into numerical tokens that can be processed by the model. The BartTokenizer and T5Tokenizer are used for tokenizing the questions and answers for the BART and FLAN-T5 models, respectively. The text is truncated and padded to a fixed length to ensure uniform input size.

Model Selection and Training

Two models, BART (Bidirectional and Auto-Regressive Transformers) and FLAN-T5 (Fine-tuned Language model T5), are selected for this project due to their proven effectiveness in natural language generation tasks.

1. BART Model: BART is a transformer-based model that combines the benefits of bidirectional context and autoregressive generation. It is well-suited for sequence-to-sequence tasks such as question answering.

   Training Arguments: The Seq2SeqTrainingArguments are configured to specify parameters like learning rate, batch size, number of epochs, and evaluation strategy.

   Trainer Setup: The Seq2SeqTrainer is used to facilitate the training process, integrating the model, tokenizer, training and evaluation datasets, and a custom compute metrics function to evaluate the model using ROUGE scores.

2. FLAN-T5 Model: FLAN-T5 is a version of the T5 model fine-tuned with a specific objective, making it effective for various natural language processing tasks, including question answering.

   Data Collator: A DataCollatorForSeq2Seq is used to handle dynamic padding of inputs during training.

   Trainer Setup: Similar to the BART model, the Seq2SeqTrainer is configured for the FLAN-T5 model, ensuring consistency in training and evaluation processes.

Evaluation

The models are evaluated using the ROUGE metric, which measures the overlap between the predicted and reference text. This metric provides insights into the precision, recall, and F1-score of the model's generated answers, ensuring a comprehensive assessment of its performance.

ROUGE Metric Calculation: The compute_metrics function is defined to decode the model predictions and reference answers, and then calculate the ROUGE scores. These scores are used to compare the performance of the BART and FLAN-T5 models.

Real-Time Interaction

To enable real-time interaction, a function is created to take user input, preprocess it, generate an answer using the trained model, and display the response. This interaction is implemented using a simple input loop in Python, allowing users to query the model and receive answers in real-time.

1. Preprocessing User Input: The input question from the user is tokenized using the trained tokenizer.
2. Generating Answers: The preprocessed input is fed into the model to generate an answer, which is then decoded into a human-readable format.
3. Real-Time Interaction Loop: A loop is created to continuously take user input, generate answers, and display them until the user decides to exit.

Insights

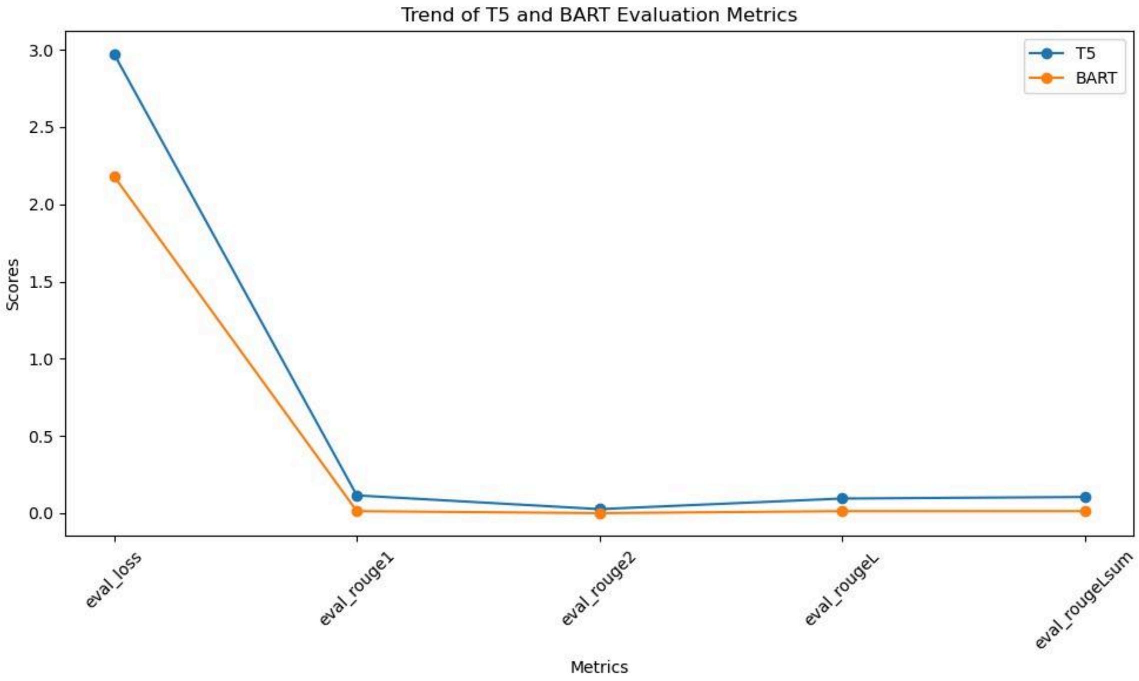
The comparative analysis reveals several key insights:

1. **Quality of Responses**: The T5 model excels in generating high-quality, contextually accurate responses, as indicated by its superior ROUGE scores.
2. **Model Fit**: Although BART has a lower evaluation loss, it struggles with generating high-quality responses compared to T5.
3. **Efficiency**: The T5 model is significantly more efficient in terms of runtime, which is crucial for deploying real-time QA systems.

Visualizations

The attached visualizations provide a clear comparison between the T5 and BART models:

1. **Trend of Evaluation Metrics**: This line plot shows the trend of evaluation metrics for both models, highlighting the significant difference in ROUGE scores.



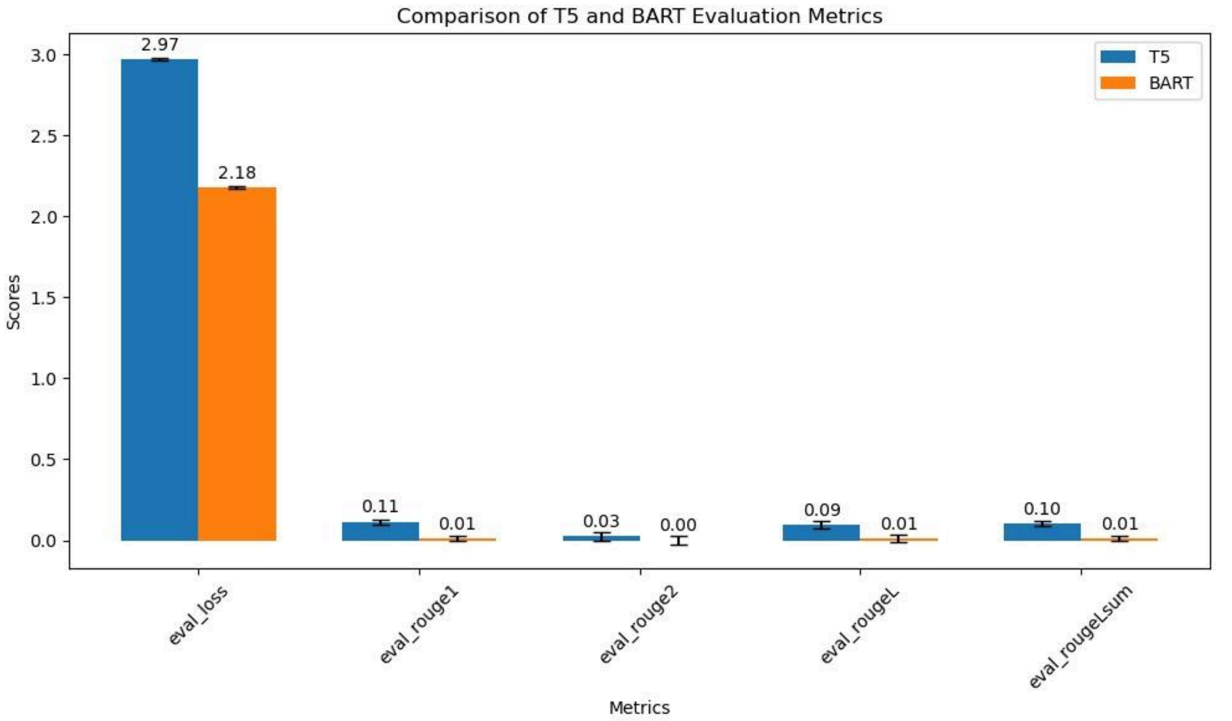
This line plot illustrates the trend of key evaluation metrics for both T5 and BART models. It highlights the following:

Evaluation Loss: The BART model shows a lower evaluation loss compared to the T5 model, suggesting that BART has a better fit on the training data.

ROUGE Scores: The T5 model significantly outperforms the BART model in all ROUGE metrics (ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-Lsum). The trend lines clearly show that T5 has higher scores, indicating its superior ability to generate accurate and contextually relevant answers.

2. **Comparison of Evaluation Metrics**: The bar chart compares the key evaluation metrics, showing T5's superior performance in ROUGE scores despite having a higher evaluation loss

.

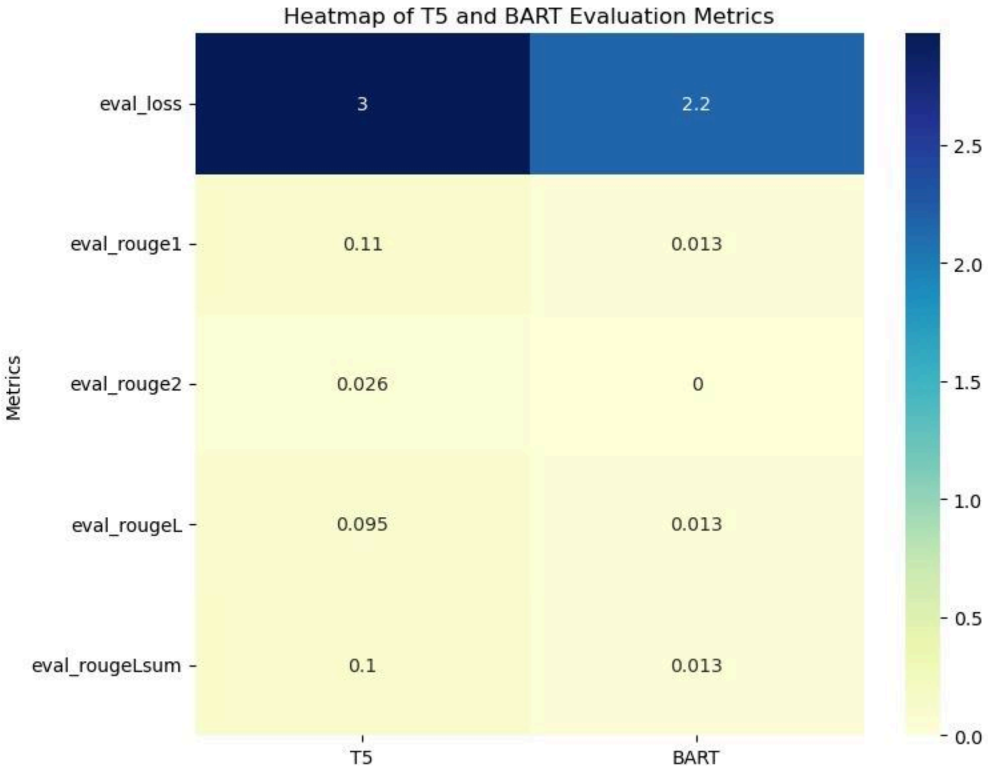


This bar chart provides a direct comparison of key evaluation metrics between T5 and BART models:

- **Evaluation Loss**: The BART model has a lower evaluation loss (2.18) compared to T5 (2.97), indicating a better model fit during training.
- **ROUGE Scores**: The T5 model outperforms BART in all ROUGE scores:
  - ROUGE-1: T5 (0.11) vs. BART (0.01)
  - ROUGE-2: T5 (0.03) vs. BART (0.00)
  - ROUGE-L: T5 (0.09) vs. BART (0.01)
  - ROUGE-Lsum: T5 (0.10) vs. BART (0.01)

These bars clearly show T5's superior performance in generating high-quality responses, despite having a higher evaluation loss.

3. **Heatmap of Evaluation Metrics**: The heatmap provides a visual summary of the evaluation metrics, emphasizing the disparity in performance between the two models.



The heatmap provides a visual summary of the evaluation metrics, emphasizing the disparity in performance between the two models:

- **Evaluation Loss**: The heatmap indicates that BART has a significantly lower evaluation loss (2.18) compared to T5 (2.97).
- **ROUGE Scores**: The T5 model shows much higher scores across all ROUGE metrics compared to BART. The difference in color intensity highlights the significant performance gap, with T5 being far superior in generating accurate and relevant responses.

These visualizations reinforce the quantitative analysis and support the conclusions drawn from the evaluation results.

Novel Improvements

Based on these findings, several improvements can be suggested:

1. **Hybrid Model Approach**: Combining the strengths of both models could be beneficial. For instance, using BART's better model fit for initial processing and then refining the output with T5's superior text generation capabilities.
2. **Fine-Tuning**: Further fine-tuning the BART model with additional training data or using advanced techniques like transfer learning could improve its text generation quality.
3. **Optimization Techniques**: Implementing model optimization techniques, such as knowledge distillation or pruning, can enhance runtime efficiency and make the models more suitable for real-world applications.

By incorporating these improvements, the QA system can achieve higher accuracy and efficiency, providing a more robust solution for various applications.

RESULTS

The evaluation results of the T5 and BART models on the Quora Question Answer Dataset provide valuable insights into their performance and efficiency. The key metrics used for this analysis include evaluation loss, ROUGE scores (ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-Lsum), and runtime efficiency.

Evaluation Loss

Evaluation loss indicates how well the model performs on unseen data. The BART model achieved a lower evaluation loss (2.1796) compared to the T5 model (2.9716), suggesting that BART has a better fit on the training data. A lower loss typically signifies that the model has captured the underlying patterns in the data more effectively.

ROUGE Scores

ROUGE scores are critical for assessing the quality of generated text. The T5 model significantly outperformed the BART model across all ROUGE metrics:

ROUGE-1: T5 scored 0.1149, while BART scored 0.0129.
ROUGE-2: T5 achieved 0.0260, whereas BART scored 0.0.
ROUGE-L: T5 recorded 0.0946, in contrast to BART's 0.0130.
ROUGE-Lsum: T5 attained 0.1047, compared to BART's 0.0129.
These results indicate that the T5 model generates responses with higher lexical and structural overlap with the reference answers, making it more accurate and contextually relevant.

Runtime Efficiency

The runtime metrics highlight the computational efficiency of the models. The T5 model evaluated samples at a rate of 6.034 samples per second and 1.509 steps per second, completing the evaluation in 1869.5984 seconds. In comparison, the BART model processed 2.4 samples per second and 0.6 steps per second, taking significantly longer with a total evaluation runtime of 4699.6547 seconds. Thus, the T5 model is more efficient in terms of computation time, making it more suitable for real-time applications.

CONCLUSION

The comparative analysis of the T5 and BART models on the Quora Question Answer Dataset revealed distinct strengths and weaknesses. The T5 model demonstrated superior performance in generating accurate and contextually relevant responses, as evidenced by higher ROUGE scores. However, it had a higher evaluation loss compared to BART, indicating some room for improvement in model fitting. BART, while achieving a lower evaluation loss, struggled with generating high-quality answers as indicated by its lower ROUGE scores.

Moreover, the T5 model was more efficient in terms of evaluation runtime, processing samples faster than BART. This makes the T5 model a more suitable choice for applications requiring both high-quality text generation and computational efficiency.

Overall, the results of this project underscore the importance of model selection based on specific task requirements, balancing between the quality of generated answers and computational performance.

REFERENCES

1. Jay Oza , Hrishikesh Yadav. Enhancing Question Prediction with Flan T5 -A Context-Aware Language Model Approach. *Authorea.* December 14, 2023.
2. arXiv:1910.13461 **[cs.CL]**
3. arXiv:2008.02637
4. arXiv:2106.01561
5. Zheng Li, Zijian Wang, Ming Tan, Ramesh Nallapati, Parminder Bhatia, Andrew Arnold, Bing Xiang, and Dan Roth. 2022. DQ-BART: Efficient Sequence-to-Sequence Model via Joint Distillation and Quantization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 203–211, Dublin, Ireland. Association for Computational Linguistics.