Q&A SYSTEM
INDIGO

GAURAV SARDAR

INTRODUCTION

In the era of digital communication, question-answering systems have become a pivotal tool in various applications ranging from customer service to personal assistants. This project aims to develop a state-of-the-art question-answering model using the Quora Question Answer Dataset. By harnessing advanced natural language processing (NLP) techniques and leveraging powerful models such as BERT, T5, and GPT, aim to create an AI system that can understand and generate accurate responses to a diverse range of user queries, providing human-like interactions.

LITERATURE SURVEY

Question-answering (QA) systems are a subset of information retrieval and extraction that focus on providing precise answers to user queries. Traditional QA systems relied heavily on predefined rules and databases, but with the advent of machine learning and NLP, modern QA systems have significantly improved in their ability to understand and process natural language.

MODELS

BERT

BERT is a transformer-based model designed to pre-train deep bidirectional representations by joint conditioning on both left and right contexts in all layers. This allows it to understand the context of a word based on its surrounding words, making it highly effective for various NLP tasks including question answering.

T5

T5 is an innovative model that frames all NLP tasks as a text-to-text problem. By converting input data into a text format, T5 can be fine-tuned to perform a wide range of tasks, including QA, translation, and summarization.

GPT2

GPT, particularly in its latest iterations, has set new benchmarks in generating human-like text. Its autoregressive nature allows it to generate coherent and contextually relevant responses, making it suitable for conversational AI and QA systems.

EVALUATION METRICS

ROUGE

ROUGE focuses on recall, assessing how well the generated answer covers the reference answer by measuring the overlap of n-grams between the two. Higher overlap indicates better recall. It includes variants like ROUGE-N (specific n-grams), ROUGE-L (longest common subsequence), ROUGE-S (skip-bigrams), and ROUGE-SU (unigrams and skip-bigrams). This metric is particularly effective for evaluating summaries and longer answers, where capturing all relevant information is crucial.

BLEU

BLEU measures the precision of n-grams in the generated text compared to the reference text, focusing on how many of the generated n-grams appear in the reference. It includes precision-based metrics and typically uses a brevity penalty to avoid overly short answers. BLEU is effective for evaluating machine translation and shorter answers, emphasizing exact matches over recall.

F1 SCORE

The F1 score balances precision and recall, providing a harmonic mean of both. It measures the overlap between the predicted and reference tokens, considering both false positives and false negatives. This metric is useful for evaluating classification tasks, ensuring that the model performs well in identifying relevant information and minimizing incorrect predictions.

RESULTS

Accuracy Score	BERT	T-5	GPT – 2
ROUGE-1 Score	0.29275887427351016	0.2664656218095823	0.2367169576991198
ROUGE-2 Score	0.22658508279133857	0.19582659978780728	0.13765840923439
ROUGE-L Score	0.29275887427351016	0.26646535439076274	0.2367169576991198
BLEU Score	0.15501534531770816	0.1346996297472278	0.0777791024144392
F1 Score	0.3131977970899778	0.2849548102818311	0.25391952899405495

·BERT's Superior Performance: BERT's bidirectional training approach allows it to capture more context from the input data, leading to better performance in the QA task.

·FLAN T5's Versatility: While FLAN T5's performance is slightly lower than BERT, its text-to-text framework makes it highly versatile for various NLP tasks.

·GPT-2's Limitations: GPT-2, despite being a powerful generative model, shows lower performance in the QA task, possibly due to its autoregressive nature which might not capture the full context as effectively as BERT.

NOVEL RECOMMENDATIONS

Enhanced Fine-Tuning:

- BERT: Further fine-tuning BERT on domain-specific data or using more recent versions of BERT (like RoBERTa or DistilBERT) could improve performance. Training on a diverse set of question-answer pairs from various domains can help the model generalize better.

- FLAN-T5: Experiment with different configurations of T5, such as T5-large or T5-3B, to assess if larger models provide better performance. Fine-tuning on additional datasets related to specific question-answering tasks may also be beneficial.

- GPT-2: Explore other GPT-2 variants, such as GPT-3 or GPT-4, which have more advanced architectures and larger training datasets. Implementing more sophisticated prompt engineering and few-shot learning techniques could also enhance performance.

Model Ensembling:

- Ensemble Methods: Combine predictions from BERT, FLAN-T5, and GPT-2 using ensemble methods. Techniques like voting, stacking, or averaging the output probabilities of the models could leverage the strengths of each model to improve overall accuracy.

- Weighted Averaging: Assign weights to each model based on its performance metrics (e.g., ROUGE, BLEU, F1) to create a weighted average of the answers. This approach can integrate the strengths of different models effectively.

c. Advanced Architectures:

- Hybrid Models: Develop hybrid models that combine the strengths of transformer-based architectures with other neural network types, such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs). This could improve the model's ability to capture both local and global contextual information.

THANK YOU