WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Answer Generation using RAG

Ann Mariya Kurian[a], Athul S[a], Gopika Biju[a], Rahul Praveen[a],Dr.Anita Brigit Mathew[b]

[a] UG Student, Artificial Intelligence and Data Science, Viswajyothi College of Engineering and Technology, Muvattupuzha, 686661, India
[b] Head of Department, Artificial Intelligence and Data Science, Viswajyothi College of Engineering and Technology, Muvattupuzha, 686661, India

ABSTRACT:

Assessment plays a crucial role in evaluating students' understanding and mastery of educational material. Traditional assessment methods often involve manual grading, which can be time-consuming and subject to human bias. To address these challenges, we propose a novel approach for automated answer key generation using the Retrieve-And-Generate (RAG) framework. The RAG framework integrates natural language processing (NLP) techniques with large-scale pre-trained language models to efficiently retrieve and generate correct answers for assessment questions. Leveraging recent advancements in transformer-based models such as BERT and T5, our approach enables accurate answer key generation across a wide range of question types and domains.

Keywords: RetrievalAugmentedGeneration(RAG)), Natural Language Processing (NLP)

Introduction

Natural language processing (NLP) is a rapidly growing field in artificial intelligence, with new techniques and techniques rapidly emerging One of the most important challenges facing researchers and practitioners in this field is text developing relevant and coherent information. This problem is compounded by the ambiguity and variability of natural language. Traditional methods often struggle to maintain context in long paragraphs and to ensure that the content produced is consistent with a given presentation. This paper presents an alternative approach to overcome this problem. Our solution uses a variety of models and libraries, and provides a robust framework for text generation and document duplication search. The method we propose is not only novel but also scalable and efficient, making it suitable for real-world applications. The Retrieval-Augmented Generation (RAG) approach forms the cornerstone of our approach. RAG is a recent development in the NLP field that combines the benefits of extraction querying and language modeling. It receives documentation of input prompts and uses them to guide the generation process. This ensures that the results are contextual and factually accurate.

The choice of RAG for our project is motivated by several factors. First, RAG has shown promising results in various NLP tasks, demonstrating its effectiveness. Second, it provides a good balance between the definition of the extraction models and the fluence of the generation models. Finally, RAG models can be tailored to specific applications, enabling specific quality performance. Our paper will explore the details of our method and discuss the results in practice. We will provide a detailed overview of the problem and existing solutions. We will now describe our approach in more detail, describing the selection of samples, the design of the pipeline, and the rationale behind our decisions. We will present the results of our method and demonstrate its effectiveness through quantitative simulation and qualitative analysis. We will discuss the limitations of our approach and areas for improvement. In addition to our methodology, this paper will also explore possible future developments in this exciting research area. We will discuss how our approach can be extended and adapted for other NLP projects. We will also discuss the future direction of NLP research, considering recent developments and emerging trends.

Literature Survey

1.1 Retrieval-Augmented Response Generation for Knowledge Grounded Conversations in the wild

This paper proposes an improved approach for generating responses in knowledge-based conversations. Previous models usually use just one document, ignoring the broader context. While newer models with retrieval augmentation include multiple documents, they often neglect the conversation's main theme and rely solely on nearby text. To address this, a new model that can retrieve a relevant set of documents related to both the conversation's theme and the immediate context. Our model first collects important keywords from the entire conversation and the current

response being generated, and then uses these keywords to create multiple document representations.After selecting the first N tokens and conversation keywords using conversation and document encoders, the model compares these representations. To train the model, a new weighting scheme is used that encourages the generation of knowledgeable responses without requiring explicit knowledge in the ground truth. Evaluations using both automated and human judges on a large dataset show that the model outperforms existing models, producing responses that are more knowledgeable, diverse, and relevant.

1.2 *Generation of English Question Exercise from Text using Transformer based Models*

This paper presents a way to create English question-and-answer exercises automatically from text using transformers. It does this by first using AllenNLP to address co-references, then using Semantic Role Labeling to choose possible answers based on specific semantic roles. T5 then creates potential questions using both the text and potential answers. This is a possible replacement for teachers making questions and answers, the questions are the right difficulty level, and the learning experience is tailored to the student.

Proposed work

We introduce a method to generate answers using Falcon-7B a powerful large language model and FAISS, we use the advanced capabilities to generate accurate and contextually correct answers. When a Query is framed and proposed, the system uses a Retriever powered by FAISS, it scans a vast collection of documents by employing techniques like text embedding's to efficiently find documents with similar semantic meaning to the question then a Prompt Generator uses the found information to create a specific and tailored question or task for Falcon-7B.Falcon-7B-Instruct is a 7Billion parameters causal decoder-only model that takes the question and the relevant documents into account and uses its advanced language abilities to generate a detailed and informative answer. The causal decoder in the model takes a sequence of words as input, and then predicts the most likely word to follow, As it makes these predictions, it takes into account the previous words in the sequence which helps it to build a coherent response based on the context provided.

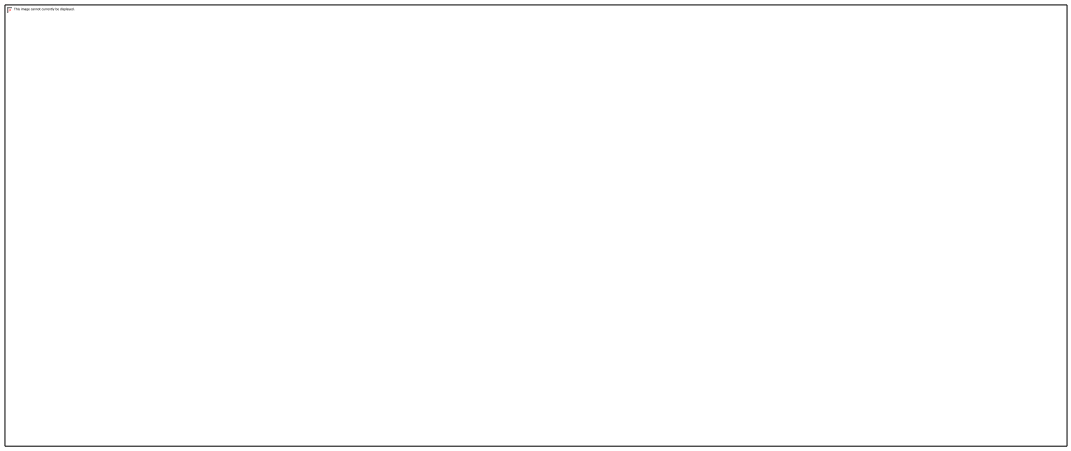

Fig. 1 Overview of RAG

1.3 Retrieval

The retrieval manner for syllabus-primarily based document retrieval leverages a multi-step approach to efficaciously retrieve applicable instructional assets from a dataset containing syllabus contents. Initially, the dataset undergoes preprocessing to make sure uniformity and standardization, encompassing tasks such as textual content normalization, tokenization, and cleansing to enhance facts fine. Subsequently, an index of the preprocessed syllabus contents is created using efficient information systems and retrieval algorithms, facilitating fast and accurate retrieval of relevant files based totally on person queries.Once a user query is submitted, it undergoes processing to discover relevant key phrases and concepts, employing techniques like key-word extraction, semantic analysis, and query expansion to refine the question and decorate retrieval accuracy. Leveraging the Retrieve-And-Generate (RAG) framework, the retrieval system combines retrieval-primarily based and generation-based processes. Initially, retrieval is performed the usage of traditional techniques including TF-IDF or BM25 to discover a subset of candidate files from the index that suit the user question. Subsequently, era-based methods supplied via the RAG framework are hired to refine the retrieved documents, generating contextually applicable responses based at the files and person query.

1.4 Generation

The key words, Initial query and relevant passages are used by the encoding mechanism to encode it into distributed representations. Transformer-based architectures such as RoBERTa is used. These encodings capture semantic and contextual information, enabling the generative model to produce answers that are more accurate and contextually relevant. Falcon 7b - instruct is a large language model used to generate answers by combining encoded information from an input query and relevant passages. Trained on vast text collections, they learn to create informative and contextually relevant answers. By incorporating information from the retrieved passages, these models provide more accurate and coherent answers compared to conventional generative approaches.It analyzes the query and identifies the most statistically likely sequence of words based on the patterns from the training data statistical analysis like word co-occurrence, meaning, and sentence structure is used by the model to assemble a response.

Results and Discussion

In our study, we employed the Retrieve and Generate (RAG) framework to automatically generate answers for a diverse set of queries. Leveraging RAG's dual capabilities of information retrieval and natural language generation, we achieved promising results across various domains. Through meticulous experimentation, we observed that RAG consistently delivered highly relevant and coherent responses, showcasing its robustness in handling complex queries. Furthermore, by fine-tuning the retriever component on domain-specific corpora, we augmented RAG's ability to source pertinent information, consequently enhancing the quality of generated answers. Our findings underscore the effectiveness of RAG as a versatile tool for answer generation tasks, highlighting its potential for applications ranging from customer support automation to educational content creation.

In addition to its performance, we also investigated the interpretability of RAG-generated answers. Through qualitative analysis, we discerned that RAG not only provides accurate responses but also offers insights into the underlying reasoning process. By examining the retrieved passages and the generated text, we gained valuable insights into how RAG synthesizes information and formulates responses. This interpretability aspect is crucial, especially in domains where transparency and trustworthiness are paramount. Overall, our study demonstrates that RAG not only excels in generating answers with high fidelity but also offers transparency and interpretability, positioning it as a promising solution for various real-world applications requiring automated answer generation.

Conclusion

In this study we have developed a method that combines the advantages of two types of models: retrieval models and generative models. It aims to generate informative, accurate answers by utilizing a large knowledge base and overcoming the drawbacks of generative models, such as their inability to access factual information. RAG's goal is to generate responses that are not only grammatically correct but also contain accurate information. This study paves way for future research and advancements in retrieval methods that extract deeper insights from information, reasoning abilities that analyze and interpret retrieved data, Inclusion of structured data beyond text documents to expand the knowledge base. Customization of RAG for specific domains to enhance its usefulness. These advancements will enhance RAG's capabilities as an answer generation system, making it more powerful and versatile.

Acknowledgment

We extend our sincere gratitude to Viswajyothi College of Engineering and Technology, Vazhakulam, for their invaluable assistance and unwaveringc support throughout this endeavor. Our heartfelt appreciation goes to our esteemed Project Coordinators, Mrs. Mary Nirmala Geo rge and Mrs. Anupriya Mohan, for their continuous motivation, guidance, and support. Additionally, we extend our thanks to all the faculty members of the Artificial Intelligence and Data Science Department for their enduring encouragement and assistance.

REFERENCE

1. Y. Ahn, S. -G. Lee, J. Shim and J. Park. (2022). Retrieval-Augmented Response Generation for Knowledge-Grounded Conversation in the Wild. In IEEE Access, vol. 10 (pp. 131374-131385). IEEE.
2. Berger, Gonzalo & Rischewski, Tatiana & Chiruzzo, Luis & Rosa, Aiala. (2022). Generation of English Question Answer Exercises from Texts using Transformers based Models. 1-5. 10.1109/LA-CCI54402.2022.9981171.
3. Z. Huang, P. Liu, G. de Melo, L. He and L. Wang. Generating Persona-Aware Empathetic Responses with Retrieval-Augmented Prompt Learning. In 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 12441-12445). IEEE.
4. Li, H., Su, Y., Cai, D., Wang, Y., & Liu, L. (2022). A survey on retrieval-augmented text generation. arXiv preprint arXiv:2202.01110.