Advancements in Infromation Retrieval: Enhancing Query Expansion and Ranking with CORAG

Abjal Hussain Shaik
11643584
abjalhussainshaik@my.unt.edu

Hari Krishna Sai Rachuri
11682376
harikrishnasairachuri@my.unt.edu

Tharun Ramula
11706360
tharunramula@my.unt.edu

***Abstract*—Traditional Key word based search is difficult for researchers to access relevant information from the document or publications. Traditional IR methods like BM25, TD-IDF these techniques were not able to do semantic search or awareness which genreally leds to poor accuracy or performance when there are domain related or context sensitive quries. To overcome this issue we have implemented a CORAG (Chain of Retrieval Augmented Generation) it is hybrid based model where it combines semantic search, large language model (LLM) for query expansion based on the provided query and lexical retrieval all these three techniques are used to provide the accurate and mostly contextually relevant search results.**

Firstly this CORAG pipeline starts with a user input topic or area of intrest to the user so that it can be used to dynamically fetch academic documents from the ArXiv repository using ArXiv API. Then the actual process starts where the data preporessing is done like cleaning, tokenization, stop word removal, lemmatization, and contraction correction to standarize the dataset. Now the main task document corpus is provided user provides the query and this is query is expanded by using an LLM in our case we are using FLAN-T5 where it transform the short query to more expressive relevant query.

This expanded query is then used for dual retreival BM25 where it main task is to perform keyword based matching and then FAISS is used to retrieve semantically similar documents by using dense embedding from the sentencetransformer model MiniLm-L6-v2. Now the results with the top from the both strategies are merged and then a tansformer based crossEncoder is used to rerank and it evaluates relevance in between the documents and the query which is expanded.

Evaluation is done on more than 450+ documents from the ArXiv abstracts in the particular field of technology which the user provided as input. while coming to the outputs of the model Precision is 0.67 while MAP is 0.92 and nDCG@5 is 0.97 it shows the model ability to perform well on the input provided by the user by providing the contextually appropriate documents. CORAG thus reprents a scalable and modular solution for enhancing information retreival in academic related search environments.

I. Introduction

With the sudden growth in academic publications researchers need a complex system where it can show releavant document based on the query provided while using complex vocabulary but from the quite years there is use of traditional Information retrieval mehtods like TF IDF and BM25 etc which are essential or foundations of document retreival Even these models are quite effective these struggles with domain specific literature where the query intent is complex and vocabulary is highly specialized and semantic matching of keywords is essentials.

In academic related documents there are abstract terms abbreviations or jargon which are related to dicipline which do not match the surface text of documents these lead to many issues like:

(1) Semmantically irrelavant documents retrieval

(2) omission of documents that use different terminology

To overcome the issues we use CORAG (chain of retreival augmented generation) a mutli stage hybrid model or IR pipeline that uses the power of traditional keyword search and semantic vector search and LLM driven query expansion

This desire to implement a model or IR system which can

(a) better comprehends the purpose of the query

(b) obtains relevant information beyond extact key word matches

(c) Contextual relevane is used to rank the results

This is the main motive of the CORAG system this frame work main purpose is to serve the academic research its like a personal assistant to a person who is reaserching on a particular topic this system is used based on the high information density as well as diverge range of terms.

Three essential modules that are in CORAG:

(1) FLAN T5 which is used to extent the queries like making short query into informative phrases

(2) BM25 and FAISS which is used to carry out semantic and lexical retrieval.

(3) CrossEncoder is a deep contextual alignment which is generally used to rerank the results

with all of these mix or combination scalability is maintained while precision and recall are enhanced

This project is build up on using realworld academic data like ArXiv and the evaluation of the CORAG system is measured using Precsion, MAP, nDCG adn visual tools such as word clouds and rankign curves. Finally it can be said that this CORAG system can help to improve domain specific academic discovery.

II. Literature Review

The development of this CORAG system is based on the critical evaluation of the current traditional methodologies especially those that have sought to address understanding of the dynamic query, semantic relevance and domain specific difficulties. In our study we used information from five publications where this CORAG system is build based on the knowledge Our system consists of various retrieval techniques like sparse indexing, dense vector search, graph based retrieval and LLM powered query reformulation which are represented in the following publications.

- **Seismic: Efficient Indexes for Approximate Retrieval (SIGIR 2024)** Contribution: This paper introduces a technique called seismic an inverted index which is optimized for learned sparse representations. It achieves a latency of sub millisecond retrieval. Limitations: It is ineffective for complex search tasks because of lacking query reformulation mechanisms. Furthermore, its sparse indexing design restricts academic search's ability to retrieve semantic information.

  In order to approximate the retrieval over learnt sparse representations this paper proposed effective approach which is inverted indexes it performs extremely efficiently in low latency settings by organizing the documents into compressed vector blocks and also enabling quick top k searches. It is primarily designed for recall without semantic reranking but it supports reformulation or query interpolation. This CORAG system on the other hand has more emphasis on contextual reranking and query extension to improve precision and without compromising speed.

- **Inquire: A Natural World Text to Image Retrieval Benchmark (NeurIPS 2024)** Contribution: A large scale multimodal dataset known as iNaturalist 2024 is developed for text to image retrieval it highlights problems in semantic search across modalities. Limitations: There is no adaptive query expansion, which lowers accuracy because queries are static. It does not have domain-specific retrieval enhancements (such as technical or legal searches).

  In this paper they developed large scale multimodal by using ecological datasets. It illustrates the weakness of the traditional retrieval models while using long tailed or fine grained queries Unlike CORAG it doesn't use query expansion techniques or semantic reranking so the model performance is completely based on the architecture while our CORAG system address these gaps with FLAN T5 and CrossEncoder reranking

- **AD-DRL: Attribute-Driven Disentangled Representation Learning for Multimodal Recommendation (ACM MM 2024)** Contribution: AD-DRL is proposed, which improves retrieval interpretability and robustness by decoupling semantic components in multimodal data. Limitations: Not generic IR, although recommendation systems are the main focus. lacks query reformulation, which reduces its usefulness for changing search requirements.

  This paper is based on the technique for recommendation systems where a modular attribute driven representation learning is used. Even it is not much related to CORAG hybrid design which evaluates and and reranks lexical (BM25) and semantic (FAISS) components independently—was motivated by its concept of deciphering and interpreting latent representations.

- **CaseLink: Inductive Graph Learning for Legal Case Retrieval (SIGIR 2024)** Contributions: Increases the accuracy of legal case retrieval by using graph neural networks (GNNs) to record case-to-case links. Limitations: Limited real-time applications due to high processing expense. Furthermore, its relevance to academic document retrieval is limited because it was created especially for legal case searches.

  In this paper it is about learning the relation ship between document retrieval by using graph neural networks(GNN) where this is very powerful case to case semantics the main purpose of this implementation is to use it on large scale academic search similar to our hydrib model CORAG avoid latency using FAISS and transformer models, which achieves competitive relevance with higher scalability.

- **GenQREnsemble: Zero-Shot LLM-Based Query Reformulation (JIR 2024)** Contribution: Offers a zero-shot large language model (LLM) framework for query reformulation and enlargement called GenQREnsemble. Limitations: The inability to optimize for domain-specific retrieval (e.g., legal, medical) It has trouble with long-tail searches, which reduces the accuracy of retrieval.

  It is a zero shot ensemble prompting technique for LLM based query reformulation in this study it is shown that by using LLM (large language models) we can improve the retrieval accuracy based on the query refinement by using FLAN T5 CORAG expands on this concept in a more straightforward but effectively it provides an advantages without providing the numerous prompts

III. Methodology

Advance NLP and IR techniques are used in the CORAG framework across a carefully designed modular pipeline. In this section we discuss about the implementation, covering data sourcing, query expansion using LLM, retrieval techniques, and text preprocessing reranking and evaluation. Each element I used in order to overcome specific weakness in the traditional IR system mainly focusing on the lack of semantic understanding and adaptive query resolution in academic search engines.

A. Dataset Collection and Preprocessing

In order to ensure the relevance of the CORAG system in the academic context, over 450 paper abstracts were fetched using the ArXiv API. Initially, the user input the relevant topic which he needs. For example, biotechnology, machine learning, nanotechnology. Based on these topics, the system queries, ArXiv, and extracts metadata like title, summary, authors, published data, for each matched paper.

- **ArXiv Research Papers Dataset:** The datasets used for this research guaranteeThe ArXiv Research Papers Dataset is perfect for assessing academic search performance because it includes scientific research papers with metadata (titles, abstracts, and full-text PDFs).
- It enables testing the efficacy of query expansion, ranking enhancements, and the accuracy of largescale document retrieval. a thorough assessment of CORAG's retrieval capabilities in text-based academic paper searches.

Preprocessing: Raw data undersgoes various processing steps such as

- Text lowercasing
- Removing punctuations, digits, and also special characters.
- contractions expansion (like doesn't to does not)
- Using NLTK to remove stop words but preserving negators like "not"
- Using WordNetLemmatizer to lemmatize all tokens.

Finally these results are normalized and we use preporcessed clean corpous of academic abstracts used for indexing and embedding.

Query Expansion using FLAN T5: In this, query reformulation technique is used in order to address vocabulary mismatch and short query ambiguity. Initially, a user provides query, and this query is passed to a FLAN T5 model using the prompt. By using this prompting, we expand a query. For example, expand this academic search with a query. For example, in our model, we used a query known as protein structure using machine learning. Now, after the query expansion, it becomes the structure of protein using machine learning for prediction and drug discovery. We use this expansion because it perfectly aligns with the technical document phrasing improving recall.

Embedding and Indexing with FAISS: A model known as ALL-MINI-LM-L6V2, which is a sentence transformer. It is used to transform each clean document into a dense vector, which will generally generate 384-dimensional embeddings. And after this, these are indexed using FAISS to support high-speed vector similarity search. So basically, this enables semantic search capabilities, where it is robust to various vocabulary differences.

Keyword based retreival with BM25: In order to tokenize the vectors, we use a rank-bm25 package, where we use bm25 algorithm which tokenize various versions of the same document and these are indexed using the algorithm. So, this algorithm documents based on the frequency and distribution of query terms. Its formula adjusts for document length and term rarity, where it offers a strong lexical baseline in order to complement the FAISS semantic strength.

Hybrid retrieval and Fusion: By using th eexpanded query CORAG retreives the top k results from both the FAISS and BM25 and these are results are then:

- Merged as combined list
- Based on the document ID and title these are deduplicated
- Finally passed to the ranking stage

This CORAG hybrid model ensures both semantic relevance and matching of the extact term

Semantic ranking with CrossEncoder: We use a cross-encoder model, which is known as Cross-Encoder MS-Macro MiniLM L6V2. It is used in order to refine the list of the candidate and it scores each query document paired based on the embedding jointly. By using the Deep Contextual Similarity, it outputs a single relevant score. The top results are sorted by these scores, resulting in optimized semantically final output or ranking.

B. Tools and Technologies

To effectively process and retrieve research articles, CORAG combines a number of machine learning frameworks and information retrieval tools:

- Python: Used to preprocess textual input and integrate retrieval models.
- An effective vector-based retrieval package for quick nearest-neighbor searches in huge academic datasets is called FAISS (Facebook AI Similarity Search).
- Hugging Face Transformers: Large language models (FLAN T5) and CrossEncoder models
- NLTK: A text preprocessing and Stop word handling
- A keyword-based retrieval model called BM25
- ArXiv API: For academic papers retrieval
- Text is transformed into dense embeddings for FAISS semantic search using sentence transformers.

C. Evaluation Metrics

In this in order to evalute the model we use several evalaution metrics

- Precision@k: Among the top k results it measures the proportion of relevant documents
- Mean Average Precision (MAP): precision average across all the relevant documents
- nDCG@k (Normalized Discounted Cumulative Gain): placing the relevant documents higher in ranking is rewarded

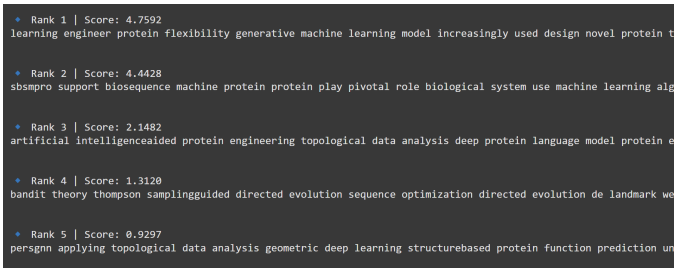

Fig. 1.

IV. Results and Analysis

In this section we discuss about the evaluation of CORAG on real world dataset of around 500+ ArXiv abstracts which are related to various topics but this is extracted based on the user input topic in the model used "Biotechnology ". Where the results tell the performance across query expansion, hybrid retrieval and semantic reranking to validate relevance and ranking query.

A. Query Expansion and Semantic scores:

- Original Query: Protein structure using machine learning
- Expanded query using FLAN T5: the structure of protein using machine learning for prediction and drug discovery. the expansions are used to better align with scientific vocabulary and to also improve the retreival results
- Top 5 Semantic scores (CROSSENCODER):
  - 4.76 - ML model for protein flexibility
  - 4.44 - Classifier for protein function
  - 2.15 - Generative Protein design
  - 1.31 - Sequence optimization
  - 0.93 - Topological protein analysis

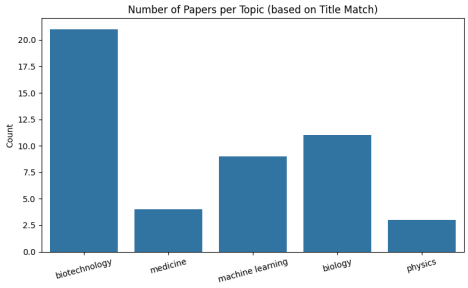

Fig. 2. Top 5 Results

B. Retrieval Metrics

- Precision@3: 0.67
- MAP (Mean Average Precision) 0.92
- nDCG@5: 0.97

These metrics provides the information that our IR System is working well and it retreive and rank documents effectively

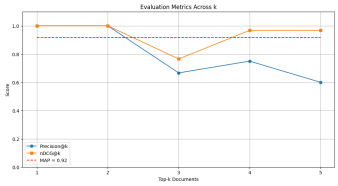

Fig. 3.

C. Visual Analysis

- Precison@k Step Plot: It shows consistent high relevance at top k ranks.

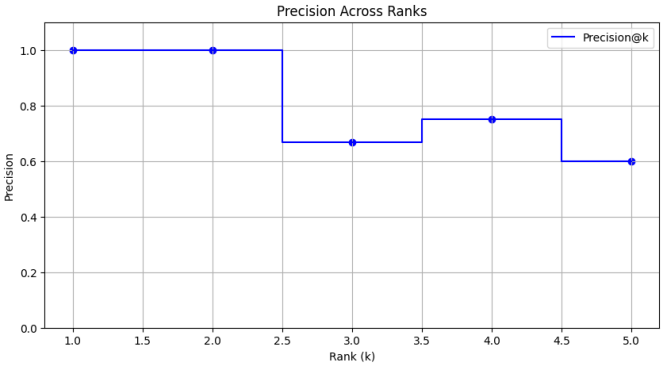

Fig. 4.

- Word cloud: It is used to highlight keywords that are used by indicating the strong topical relevance

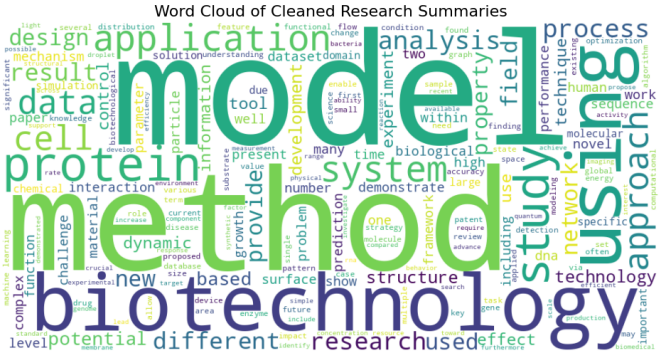

Fig. 5.

D. Summary of Findings

FLAN T5 is used to clarify the vaugue queries which will generally lead to better matches in both semantic (FAISS) and lexical (BM25) retrieval. while CrossEncoder performed well by reranking consistently elevated the most relevant content on top of demonstrating CORAG efficiency and accuracy and academic information retrieval.

V. Discussion and Conclusion

Our CORAG system represents a practical and modular approach to modern academic information retrieval tasks. Our model generally combines semantic depth, lexical precision, and also language modeling In order to overcome the issues generally we found in traditional information retrieval pipelines, for the evaluation we use precision and ranking metrics, which generally shows the power of unifying diverse retrieval techniques into a single coherent framework.

A. Reflections on Results

In our cORAG system, we firstly use FLAN T5, which is an LLM, which is generally used in order to query expansion. It helps close the gap between visual language and domain-specific terms or vocabulary, which by enhancing the expressiveness and contextual relevance of such inputs. While coming to the information retrieval techniques, secondly, we have BM25 plus FAISS, which generally ensures larger coverage. In order to match the exact keywords based on the query, we use BM25, which generally collects and extracts keyword matches. While coming to FAISS, it is used to semantically recover related content that could be overlooked otherwise. And finally, for evaluation tasks, for evaluation, we used Precision and nDCG@5 scores. Coming to cross-encoder, which is a re-ranking, which generally re-ranks based on the effective at giving topologically and conceptually superior document spread.

B. Challenges faced

Eventough our CORAG system is working well we faced quite challenges

- Computational Overhead: In order the run the LLM like FLAN T5 and CrossEncoder which add latency we generally need an GPU for system scalability.
- Manual Labeling: In order to assign binary relevance we need domain level understanding
- Query Generalization: Not all queries works well with the system some expanded queries may use vague expansions with generally dilute the retreieval precision.

C. Future Work

Our system can be upgraded in various ways

- Domain Adaptive Fine tuning: finetuning the CrossEncoder and Large language Model for domain specific corpora.
- Multilingual Retrieval: Use mutli languages to academic search
- Interactive Search: Implement user feedback mechanism

D. Conclusion

Our Corag system sucessfully shows the feasbility anf effectiveness of hybrid retrieval for academic contexts. Precision and relevance are improved by using this dual retreival reranking and LLM based FLAN T5 query expansion and Finally this systems modular design shows that it can be flexible for upcoming IR problems across various studies

References

[1] S. Bruch et al., "Efficient inverted indexes for approximate retrieval," in *Proc. ACM SIGIR*, 2024.

[2] E. Vendrow et al., "INQUIRE: A Natural World Text-to-Image Retrieval Benchmark," *arXiv preprint arXiv:2411.02537*, 2024.

[3] Z. Li et al., "Attribute-driven Disentangled Representation Learning for Multimodal Recommendation," in *Proc. ACM MM*, 2024.

[4] Y. Tang et al., "Caselink: Inductive graph learning for legal case retrieval," in *Proc. ACM SIGIR*, 2024.

[5] K. Dhole et al., "GenQREnsemble: Zero-shot LLM ensemble prompting for generative query reformulation," in *Proc. ECIR*, 2024.