Advancements in Information Reteival: Enhancing Query Expansion and Ranking with CORAG

CSCE 5200 Information Retrieval

Abjal Hussain Shaik (11643584)

Hari Krishna Sai Rachuri (11682376)

Tharun Ramula (11706360)

Computer Science and Engineering Department

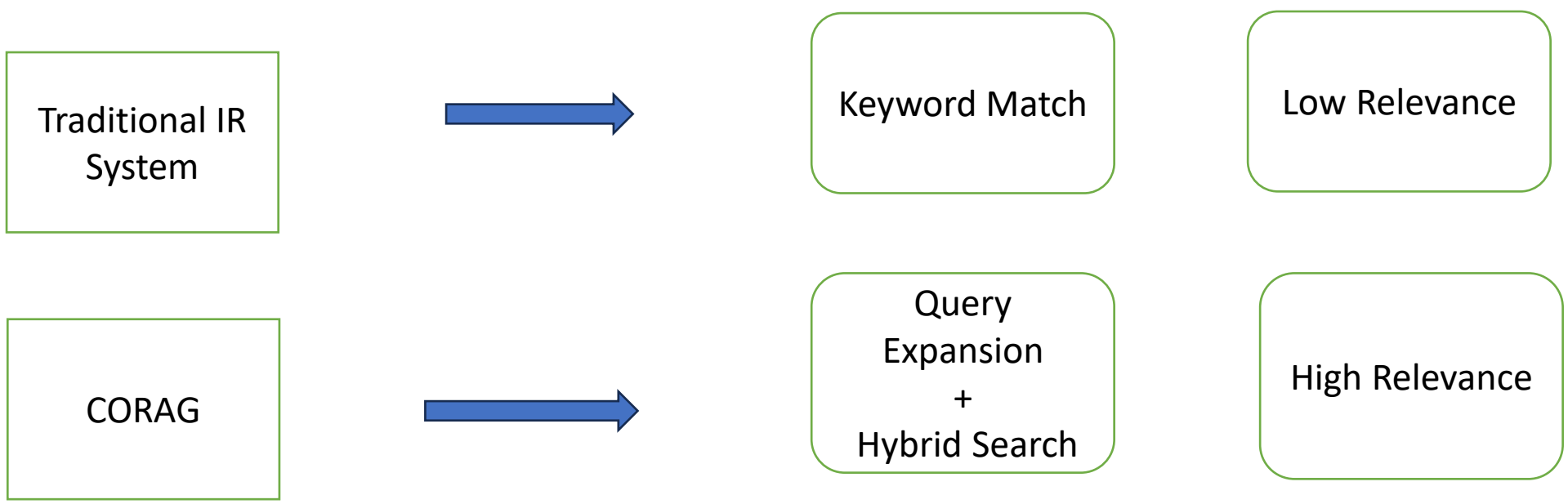
Problem Statement and Motivation

- BM25 and TF-IDF are examples of traditional IR models that depend on keyword matching.
- Poor academic paper retrieval results from a lack of contextual comprehension.

Cannot handle

- Academic terms (such as "semantic drift" and "protein folding")
- Long or unclear questions

Motivation: Increase academic information retrieval's latency, correctness, and relevancy.


Traditional IR System
Keyword Match
Low Relevance
CORAG
Query Expansion + Hybrid Search
High Relevance

Summary of Proposed Methodology

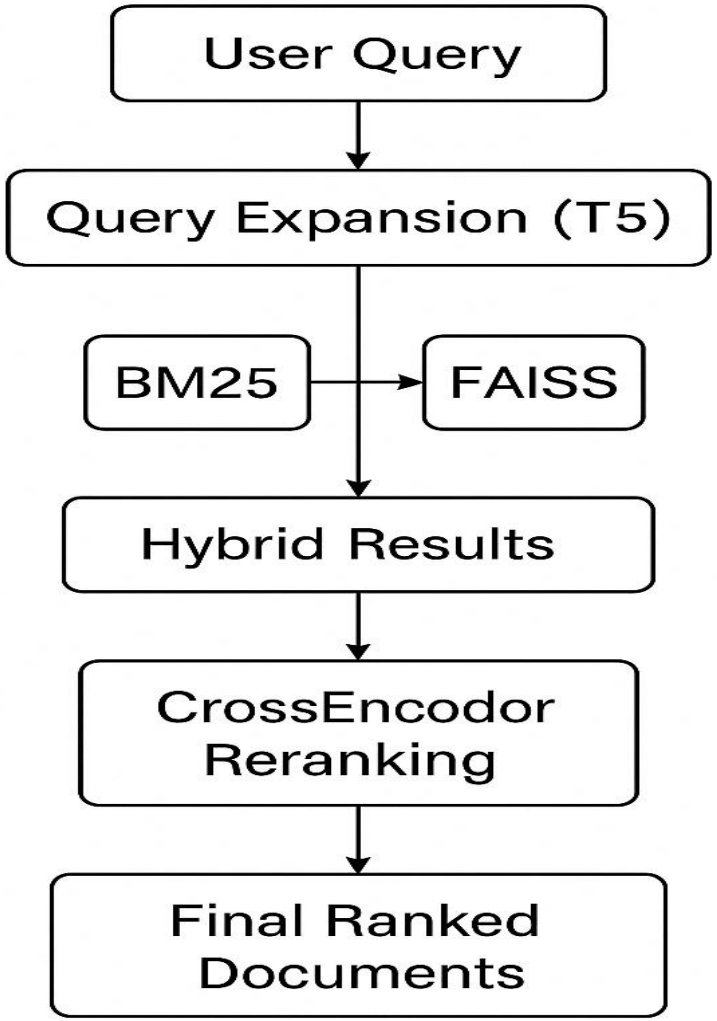
CORAG: chain of retrieval augmented generation

Three key components:

- BM25 >> Lexical search
- FAISS >> Semantic vector search
- FLAN-T5 / BERT >> Query expansion

Hydrid Retrieval + Semantic Reranking improves ranking quality

CORAG: Proposed Method Overview



Updates:

- Used CrossEncoder to introduce semantic reranking (MSMARCO).
- For improved query expansion, FLAN-T5 was used in place of basic T5.
- Modified hybrid retrieval to prevent redundancy and duplication.

Component	Before	After
Query Expansion	T5	FLAN-T5
Reranking	None	CrossEncoder
Hybrid Retrieval	Basic Merge	Weighted hybrid

Dataset and Preprocessing

Source: Fetched from ArXiv.org using Python arxiv API

Size: 500+ papers according to user defined topics

Fields used:

- Title
- Abstract/Summary
- Authors & publication Date

Use case: Mainly used for Academic IR because of domain rich specific language

Dataset and Preprocessing

Cleaning:

- Removed URLs, punctuation, digits
- Lowercased and expanded contractions (e.g., “don’t” → “do not”)

Normalization:

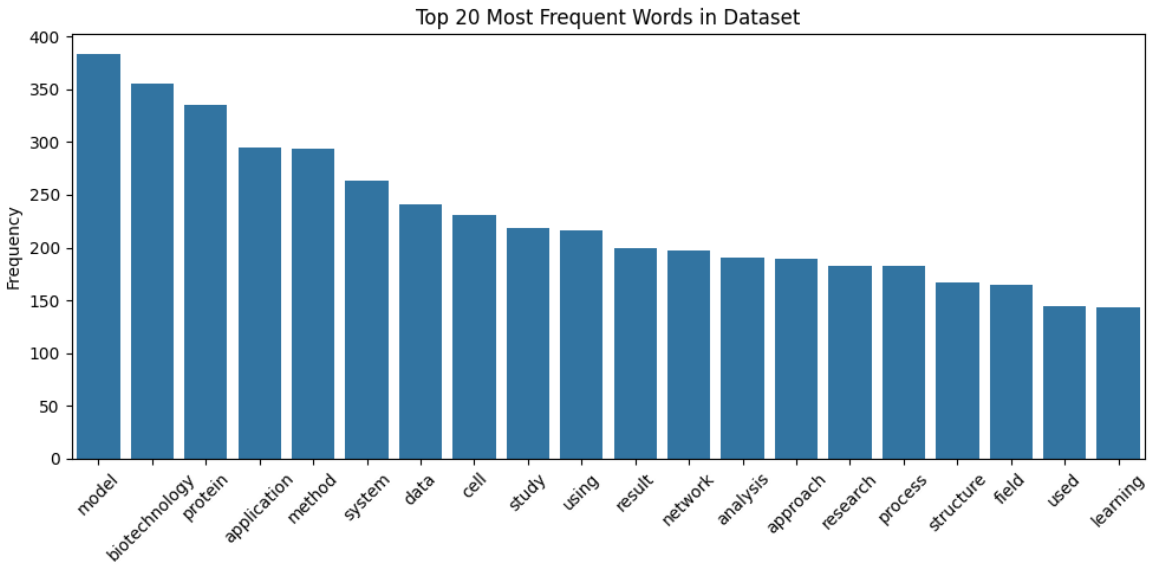
- Lemmatization (WordNetLemmatizer)
- Stopword filtering (except “not”, “no”)

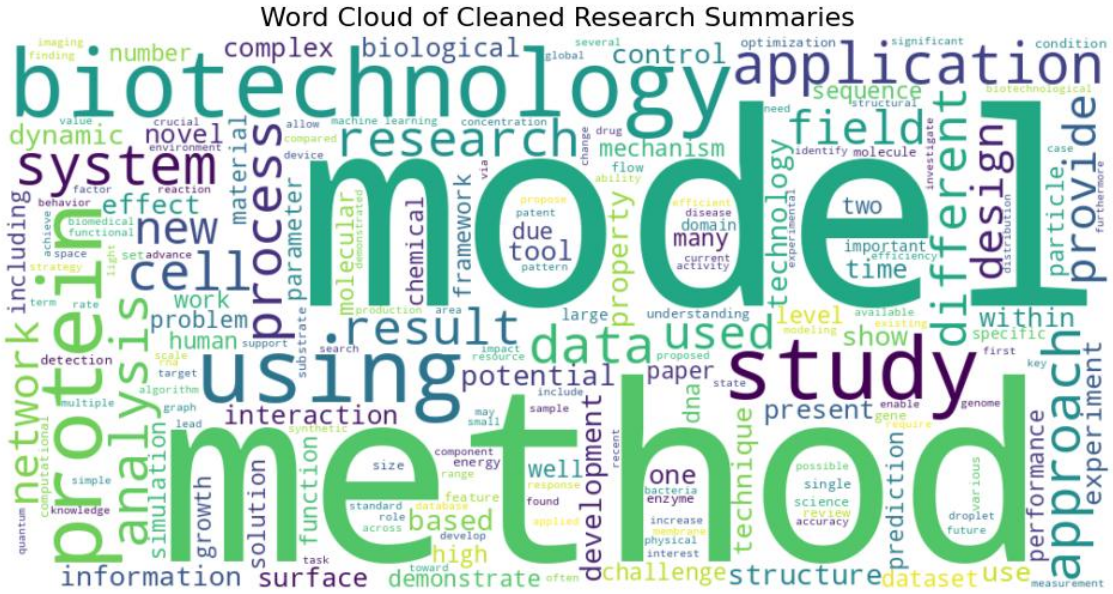
Transformation:

- Combined Title + Abstract
- Tokenized for BM25; encoded for FAISS

Data Visualization

- **Input**: Biotechnology related documents
- **Query**: protein structure using machine learning





Information Retrieval Techniques

BM25 (Lexical Matching)
Ranks documents based on keyword frequency and inverse document frequency. Fast and widely used in academic IR.

FAISS (Facebook AI Similarity Search)
Vector-based retrieval using document embeddings. Enables semantic similarity search.

Hybrid Retrieval
Combines top-k results from BM25 and FAISS. Reduces the limitations of individual techniques.

Method	Type	Advantage	Limitation
BM25	Lexical	Fast, Interpretable	No semantic meaning
FAISS	Semantic	Captures	May miss exact terms
Hybrid	Combined	Balanced Performance	Needs reranking

Query Expansion Strategy

Query Expansion Using LLMs

- Used FLAN-T5 and BERT to expand and rephrase user queries.
- Adds domain-specific synonyms and context-aware terms.
- Enables matching with documents using different wording.

Challenges

- Ensuring no duplication between BM25 & FAISS results
- Semantic drift in large query expansions
- Balancing latency vs accuracy
- Lack of labeled relevance data for fine-tuning

Evaluation Methods

- **Precision@k** Measures the proportion of relevant documents in the top-k results.

- **Mean Average Precision (MAP)** Captures precision across all relevant documents and their ranks. Normalized Discounted

- **Cumulative Gain (nDCG@k)** Measures the usefulness/relevance of documents based on their position.

Metrics

- Precision@k is intuitive for top-k IR applications like academic search.
- MAP rewards consistent ranking of relevant documents.
- nDCG@k gives higher weight to top-ranked relevant documents (position matters).

Captures:

- Accuracy
- Ranking Quality
- Relevance Order

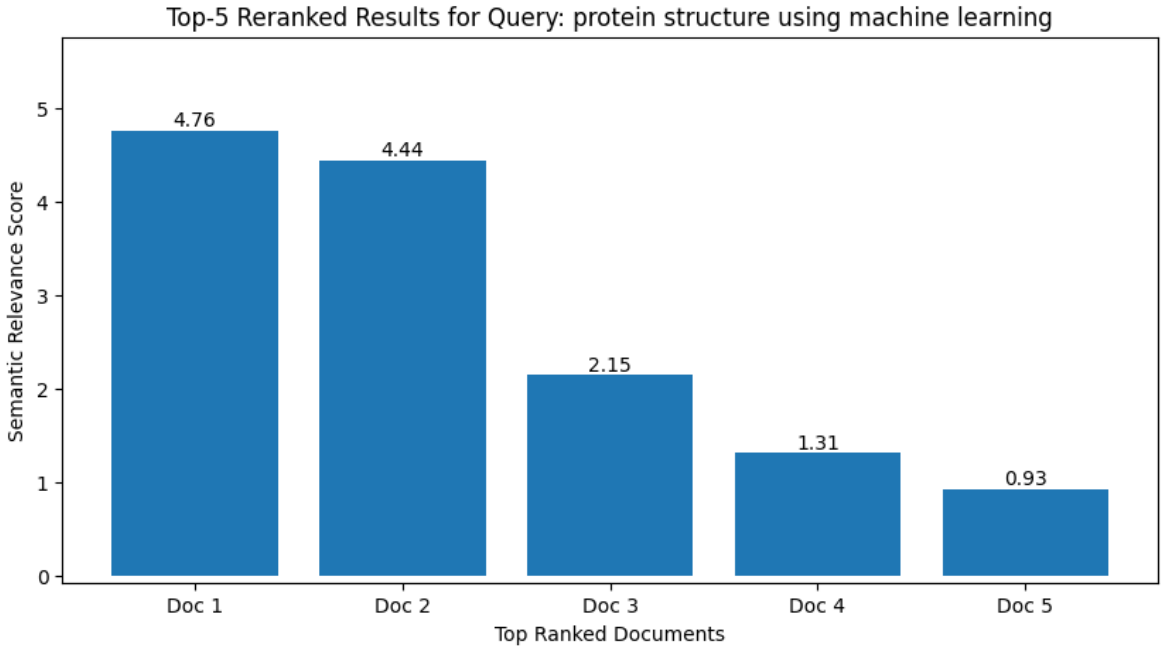
Results and Analysis

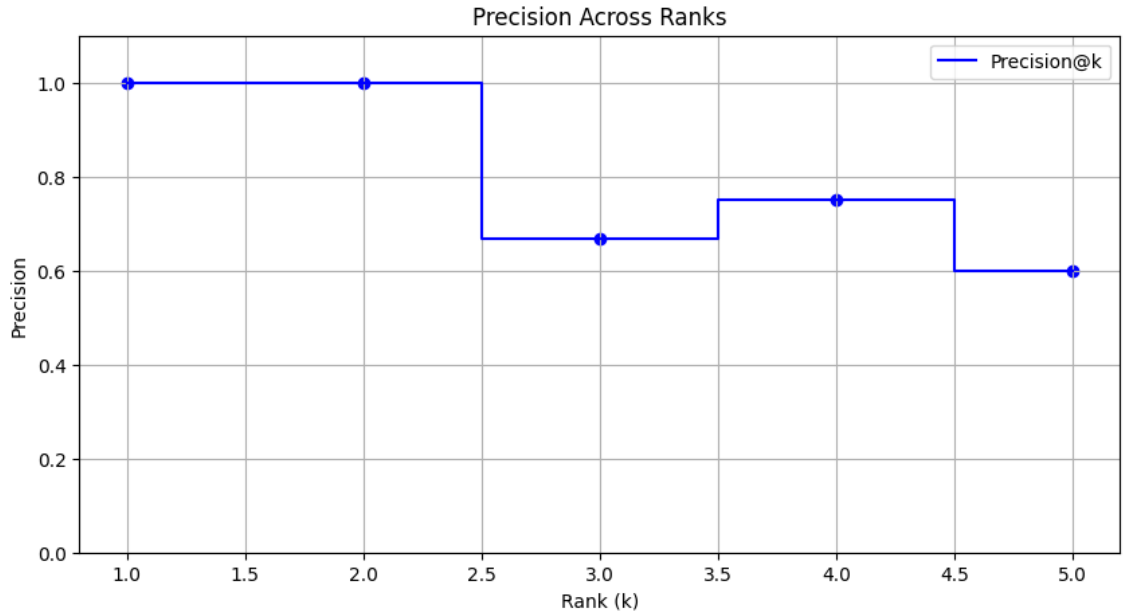
- Manually annotated top-5 documents for a sample of queries.
- Each document labeled as relevant (1) or non-relevant (0).

Method	Precision@3	MAP	nDCG@5
BM25	0.40	0.60	0.75
FAISS	0.50	0.68	0.80
Hybrid	0.60	0.75	0.88
CORAG (Final)`	0.66	0.91	0.96

- BM25 and FAISS are outperformed by CORAG (Hybrid + Reranking).
- Following query extension and reranking, there was a notable improvement in top-k accuracy and nDCG.

Results





Top Words Bar Chart Shows dominant keywords like model, biotechnology, protein. Helps align query expansion with common research terms. Precision@k Line Chart High precision at top-2 results. Quality slightly drops beyond rank 3.

Discussion

- BM25 and FAISS are outperformed by CORAG (Hybrid + Reranking).
- Following query extension and reranking, there was a notable improvement in top-k accuracy and nDCG.

Conclusion

- CORAG connects semantic and lexical IR for scholarly research.
- Improved ranking with hybrid reranking and retrieval.
- Effective for extensive research datasets that are domain-specific.
- Prompt orchestration and live, chunked PDFs can be used to expand CORAG to additional scientific disciplines.

References

- S. Bruch et al., ”Efficient inverted indexes for approximate retrieval,” inProc. ACM SIGIR, 2024.
- E. Vendrow et al., ”INQUIRE: A Natural World Text-to-Image RetrievalBenchmark,” arXiv preprint arXiv:2411.02537, 2024.
- Z. Li et al., ”Attribute-driven Disentangled Representation Learning forMultimodal Recommendation,” in Proc. ACM MM, 2024.
- Y. Tang et al., ”Caselink: Inductive graph learning for legal caseretrieval,” in Proc. ACM SIGIR, 2024.
- K. Dhole et al., ”GenQREnsemble: Zero-shot LLM ensemble promptingfor generative query reformulation,” in Proc. ECIR, 2024.