CompRAG: Retrieval for Multiple Hops

Authors: Indraneel Paranjape, Nayanthara Prathap, Nura Tamton, Vangmay Sachan

Advisor: Prof. Kan Min Yen

CompRAG (COMPrehension and COMPosition RAG) explores how relations in a text can be expressed as composite to enhance multi-hop answering outcomes in RAG. We extract entity-relation-entity triplets from text, but innovate by leveraging Holographic Reduced Representations (HRR) [1] to compose individual triplet vectors together, preserving the direction of relations in vector embeddings. At query time, processed query vectors are matched to the most similar triplets in the index. The associated chunks form the context for the LM. By focusing retrieval on the relations present in the text rather than raw similarity or pure entity-linked graphs, CompRAG aims to bridge semantic and graph methods to improve multi-hop retrieval outcomes.

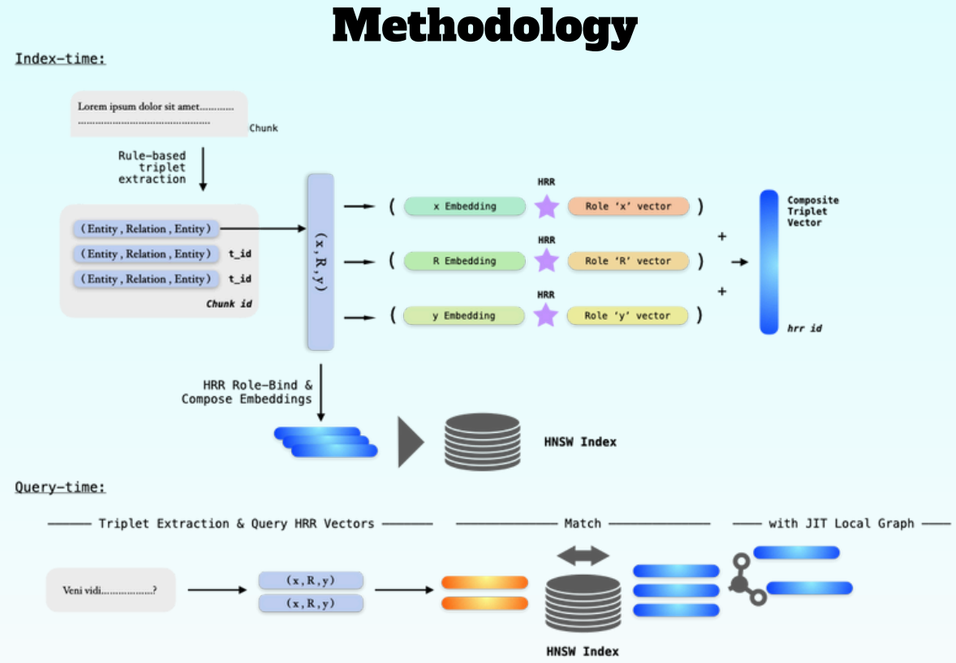
Research Question

Does compositional triplet representation with graph-based reranking provide better retrieval quality than standard embedding approaches?

Background

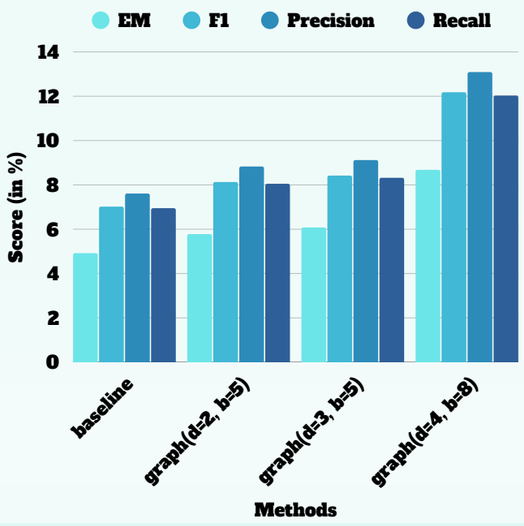
Recent works like GraphRAG (Microsoft, 2024) [2] and LightRAG [3] explore graph-based retrieval for reasoning across document connections. HotpotQA [4] establish a benchmark for multi-hop QA that motivate structured retrieval approaches. HRR [1] provides a method to bind symbolic structures into fixed-size vectors, while PageRank [5] inspired techniques suggest ways to rank relational importance in graphs. CompRAG integrates these ideas by combining triple extraction and HRR encoding, allowing compositional, structure-aware retrieval that supports reasoning without scaling model size.

Methodology



CompRag transforms text into structured graph-like knowledge units to enable compositional retrieval. Each document is first divided into chunks, and every chunk is parsed using SpaCy to extract triplets of the form (entity–relation–entity). Each of the three elements is encoded into a vector and combined using HRR to produce a composite vector representing the triplet. These triplets and their associated chunks are stored in a vector database. At query time, triples are extracted from the question and encoded using HRR. The system retrieves the top N most similar triplets from the database using vector similarity. The chunks associated with these triples are then added to the prompt context, forming a RAG input that captures relationships across multiple sources.

Results



- Graph retrieval outperforms baseline: Best graph configuration (depth=4, k=8) achieved 12.2% EM and 13.2% F1, representing ~2.5× improvement over baseline (4.9% EM, 7.0% F1)
- Depth matters: Performance increases with graph search depth (d=2 → d=4), suggesting multi-hop reasoning benefits from exploring more distant semantic connections.
- The answering style of the LLM could have a significant effect on the EM score so even if the model was able to get the answer right, it may not be counted.
- Eg: "predicted_answer": "NCAA Division I Men's College Soccer".
  "gold_answer": "the North Atlantic Conference"
- This requires manual checks to be done to understand whether the model is actually getting the correct answer and if our metrics just does not capture that.

Analysis

- Lemmatisation does not improve retrieval outcomes in our use case. Since triplets are already minimal, lemmatisation reduces their natural meaningfulness, and this effect carries over to the composite embeddings. E.g (he, argue, confident) vs (he, argued, confidently). The latter preserves meaning better, both to the reader and in the embedding.
- HRR does not natively support directionality since it is commutative (output of A HRR B = B HRR A). To resolve this, we introduce the role-bind and vector composition.
- When scaling to large amounts of data, retrieval becomes difficult with so many relationships to capture. We used Just-in-time (JIT) local graphs to capture common topics by finding other similar triplets to retrieved ones and performance increased significantly

Conclusion

We hypothesised that if we could retrieve most of the relations present in a text, it could be sufficient to help retrieve necessary chunks associated with the query. We believe the biggest bottleneck to effective graphs for multi-hop outcomes is the rule-based triplet extraction, which is brittle. We believe the overall method shows promise in modelling the relationships present between chunks.

Related literature

[1] Tony A. Plate. Holographic Reduced Representations: Distributed Representation for Cognitive Structures. CSLI Publications, 2003.

[2] Jonathan Larson, Steven Truitt et al. GraphRAG: Unlocking LLM Discovery on Narrative Private Data. Microsoft Research, 2024.

[3] Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, Chao Huang. LightRAG: Simple and Fast Retrieval-Augmented Generation.arXiv preprint arXiv:2410.05779, 2024.

[4] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, Christopher D. Manning. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. Proceedings of EMNLP, 2018.

[5] Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford InfoLab, 1998.

NUS National University of Singapore

WING Web Information Retrieval / Natural Language Processing Group