CS4681 - Advanced Machine Learning

Progress Evaluation Report

Project

UnifiedQA - NLP005

Improving Multi-Hop Reasoning

210190R - Gunapala S.A.C.H.

Table Of Contents

Abstract

Multi-hop reasoning question answering (MHQA) represents a significant advancement in natural language processing, requiring systems to aggregate information from multiple sources and perform complex reasoning across interconnected facts. This literature review examines the current state of multi-hop QA systems, exploring key methodologies, architectural innovations, evaluation metrics, and recent developments in the field. Based on these insights, this project plan and methodology is proposed to enhance UnifiedQA with multi-hop reasoning capabilities through dataset integration, retrieval augmentation, and reasoning supervision.

1. Introduction

Multi-hop question answering has emerged as a critical challenge in natural language understanding, extending beyond simple fact retrieval to require complex reasoning across multiple pieces of evidence [1]. Unlike single-hop QA, where answers can be extracted from a single passage, multi-hop QA demands that systems identify, connect, and synthesize information from diverse sources [2,3]. This ability is essential for real-world applications, where complex queries often require multi-step reasoning.

The complexity arises from two requirements: (1) possessing prerequisite knowledge of entities and relations, and (2) the ability to compose and chain knowledge [4,5]. While large language models (LLMs) such as T5 and GPT excel at memorizing facts, they frequently struggle with compositional reasoning [10].

2. Literature Review

Multi-Hop QA Datasets

A number of benchmark datasets have been developed to push progress in multi-hop reasoning. HotpotQA (Yang et al., EMNLP 2018) is one of the most influential, requiring models to combine information across multiple Wikipedia articles. Each question involves at least two reasoning steps and is accompanied by *supporting fact* annotations, making it valuable not only for answer prediction but also for training explainable systems. WikiHop (Welbl et al., TACL 2018), released earlier as part of the QAngaroo project, presented a similar challenge where models must connect information spread across multiple passages for example, linking entities and their properties across documents. QASC (Khot et al., AAAI 2020) focused specifically on science questions, with nearly 10,000 multiple-choice items designed so that the correct answer can only be reached by retrieving and composing two supporting facts from a large corpus. WikiMultiHopQA (Ho et al., COLING 2020) extended this idea further by creating synthetic questions that come with explicit reasoning chains, ensuring that each query truly requires multi-step inference.

More recently, newer datasets have emerged to test modern retrieval-augmented systems. FanOutQA (Zhu et al., ACL 2024) emphasizes “fan-out” reasoning, where a single question requires aggregating information about multiple related entities, and provides annotated decompositions of reasoning steps. MultiHop-RAG (Tang & Yang, 2024) is

specifically tailored for retrieval-augmented generation, supplying gold-standard evidence to evaluate how well models handle multi-step queries in an open-domain setting.

Together, these datasets span diverse reasoning types bridging, comparison, and synthesis and evaluation formats, from multiple-choice to open-ended generation. They have now become the standard benchmarks for testing how effectively models can perform compositional, multi-hop reasoning.

Methodological Approaches

Several approaches have been proposed to address the challenges of multi-hop QA, with one of the most prominent being question decomposition. The idea is to break down a complex query into simpler, single-hop sub-questions that can be handled effectively by existing QA models. For instance, Min et al. (2019) introduced *DecompRC*, which automatically predicts sub-questions through span detection and then combines their answers via a rescoring step to arrive at the final response. This decomposition strategy enables researchers to reuse powerful single-hop reading comprehension models within a structured, multi-step reasoning pipeline effectively reducing the difficulty of the task while still preserving the multi-hop requirement.

Another influential direction in multi-hop QA is graph-based reasoning, where models explicitly represent the relationships between questions, passages, and entities as a graph structure. For example, Fang et al. (2020) introduced the *Hierarchical Graph Network (HGN)*, which builds a multi-layer graph with nodes spanning different levels questions, paragraphs, sentences, and entities. Reasoning is then performed through message passing across these layers, enabling the model to capture dependencies at multiple granularities and achieving state-of-the-art performance on HotpotQA. Similarly, Tu et al. (2019) proposed the *Heterogeneous Document-Entity (HDE) graph*, which links candidate answers, documents, and entities into a unified graph structure. A graph neural network (GNN) is then applied to aggregate evidence across documents, making it possible to combine scattered pieces of information into a coherent reasoning chain. Overall, graph-based methods excel at modeling the complex interconnections (such as entity co-references and contextual overlaps) that are often necessary to answer multi-hop questions correctly.

A third line of work leverages memory-augmented models, which incorporate external memory structures to support multi-step reasoning. These models are designed to store and recall intermediate information as the reasoning process unfolds, thereby improving the ability to chain evidence across multiple hops. For instance, Li et al. (2022) proposed *QA2MN*, a framework that integrates knowledge-graph embeddings with a question-aware memory network. By dynamically updating attention over question tokens during multi-hop reasoning, QA2MN can better track relevant entities and relations across steps. The model demonstrated strong performance on challenging multi-hop benchmarks such as *PathQuestion* and *WorldCup2014*, achieving high hits@1 accuracy. Memory-augmented designs such as this highlight the importance of retaining intermediate reasoning states, which ensures that the model can effectively build upon earlier steps rather than treating each inference in isolation.

Retrieval-Augmented Generation (RAG) frameworks integrate dense retrieval with generation, allowing models to ground answers in external evidence. Lewis et al. (2020) introduced RAG, combining neural retrievers with LLMs, while Tang & Yang (2024) benchmarked multi-hop RAG and showed that standard retrievers often fail to capture full reasoning chains. As a result, modern multi-hop QA systems typically pair retrievers (e.g., DPR) with answer models, concatenating the top-k retrieved passages for reasoning.

Other methods include explicit training with intermediate supervision (e.g. using annotated sub-questions or supporting facts during training) and curriculum learning (starting from simpler reasoning to harder queries). Knowledge-graph integration is another avenue (e.g. models that ground parts of the question to a KG), though most

recent multi-hop QA works focus on text corpora. Overall, the trend is towards combining powerful pre-trained LMs with structured retrieval and reasoning mechanisms that enforce multi-step inference.

Evaluation

Performance in multi-hop QA is measured not only by final answer correctness but also by reasoning quality. Standard metrics include Exact Match (EM) and F1, while reasoning-aware benchmarks add process-level evaluations. For example, HotpotQA reports a supporting-facts F1, requiring models to identify the evidence sentences used in inference. Other proposals include sub-question accuracy (Tang et al., 2021), which checks if models can solve the implied simpler steps. Such metrics highlight gaps where models reach the right answer without true reasoning. Increasingly, evaluations consider rationale quality and chain-of-thought coherence, though EM/F1 remain the primary benchmarks.

Role of UnifiedQA

UnifiedQA (Khashabi et al., 2020) is a T5-based, format-agnostic QA system trained on a wide range of QA tasks by casting them into a unified text-to-text framework. It achieves strong results across extractive, multiple-choice, and yes/no settings, but was not explicitly optimized for multi-hop reasoning since most of its pretraining corpora involve single-hop or commonsense QA. Nevertheless, its flexibility makes it an ideal foundation for multi-hop adaptation. In this project, we will fine-tune UnifiedQA on multi-hop datasets such as HotpotQA, QASC, and DROP, while augmenting it with retrieval mechanisms and reasoning supervision (e.g., supporting facts, sub-questions). Leveraging its robust pretraining, we hypothesize that UnifiedQA can integrate these additional signals to better handle compositional queries, moving toward a unified multi-hop reasoning system.

3. Project Planning

3.1 Objectives

- Extend UnifiedQA to handle multi-hop reasoning effectively.
- Evaluate improvements on HotpotQA, QASC, and DROP.
- Explore retrieval augmentation, reasoning supervision, and curriculum learning.

3.2 Timeline

Aug 18 – Aug 20: Baseline setup – fine-tune UnifiedQA on single-hop datasets.

Aug 21 – Aug 24: Baseline evaluation – assess performance and document results.

Aug 25 – Sep 08: Multi-hop integration – train on QASC and DROP.

Sep 06 – Sep 09: Multi-hop integration – train on HotpotQA with supporting facts.

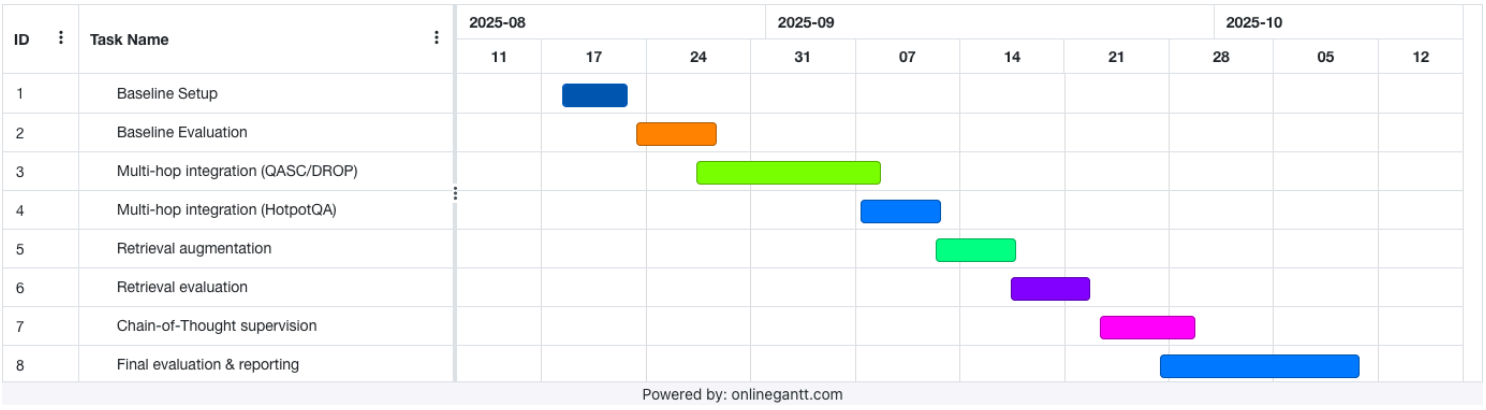
Sep 10 – Sep 16: Retrieval augmentation – implement DPR/BM25 and integrate top-k passages.

Sep 17 – Sep 22: Retrieval evaluation – compare retrieval-augmented model with baseline.

Sep 23 – Sep 29: Chain-of-Thought supervision – train with rationales and evaluate reasoning steps.

Sep 26 – Oct 10: Final evaluation & reporting – comprehensive evaluation and finalize report.

3.3 Gantt chart



4. Methodology

4.1 Baseline Setup

Model Selection - The project will utilize UnifiedQA, specifically the T5-11B checkpoint, as the foundational model. UnifiedQA provides a robust, format-agnostic question answering framework capable of handling multiple QA paradigms.

Implementation Framework - All model fine-tuning and experimentation will be conducted using the HuggingFace Transformers library, complemented by T5X where necessary for large-scale training and optimization. These frameworks ensure scalability, reproducibility, and access to state-of-the-art model utilities.

Baseline Training Datasets - To establish a performance benchmark, the model will first be fine-tuned on widely adopted single-hop QA datasets, namely SQuAD, BoolQ, and OBQA. These datasets serve as standard baselines in QA research and will provide comparative insights before introducing multi-hop reasoning datasets.

4.2 Multi-Hop Data Integration

Multi-Hop Data Integration - The fine-tuning phase will incorporate multi-hop reasoning datasets to extend UnifiedQA's capabilities beyond single-hop question answering. Specifically, the model will be fine-tuned on HotpotQA, QASC, and DROP, which are widely recognized benchmarks for multi-hop and compositional reasoning.

To maximize learning efficiency, a curriculum learning strategy will be employed. The model will first be trained on comparatively simpler reasoning datasets (QASC and DROP) and subsequently scaled to the more complex HotpotQA full wiki setting. This progressive approach is designed to facilitate smoother adaptation of the model to increasingly challenging reasoning tasks.

4.3 Retrieval-Augmented Pipeline

To enhance evidence gathering for multi-hop reasoning, the project will implement a Dense Passage Retriever (DPR) for retrieving the top-k most relevant passages. The retrieved evidence will then be concatenated with the input query and passed to UnifiedQA, ensuring that the model has access to supporting context during answer generation.

A comparative evaluation will be conducted against a baseline configuration without retrieval augmentation, in order to quantify the contribution of the retrieval module to overall system performance.

4.4 Reasoning Supervision

The project will incorporate Chain-of-Thought (CoT) training to explicitly model intermediate reasoning steps. Using supporting facts from HotpotQA, the model will be trained to generate rationales alongside final answers, thereby encouraging stepwise reasoning rather than direct answer prediction.

Evaluation will assess not only the correctness of the final answers but also the quality and coherence of the generated rationales, providing deeper insights into the model's reasoning process.

4.5 Evaluation

The performance of the proposed system will be assessed using the following metrics,

- Exact Match (EM) - Measures the percentage of predictions that exactly match the ground truth answers.
- F1 Score - Captures the overlap between predicted and reference answers, balancing precision and recall.
- Supporting Facts Accuracy - Evaluates the model's ability to correctly identify relevant evidence supporting its answers.

Additionally, ablation studies will be conducted to analyze the contributions of key components:

1. Retrieval Mechanism - Comparing performance with and without document/context retrieval.
2. Chain-of-Thought (CoT) Reasoning - Assessing the impact of CoT prompting on model accuracy.
3. Training Strategy - Evaluating curriculum-based training versus joint training approaches.

5. Expected Outcomes

The anticipated outcomes of this project include:

- An enhanced UnifiedQA model with improved multi-hop reasoning capabilities.
- State-of-the-art or competitive performance on benchmark datasets, including *HotpotQA*, *QASC*, and *DROP*.
- Empirical evidence demonstrating that retrieval augmentation combined with Chain-of-Thought (CoT) prompting enhances compositional reasoning performance

6. Conclusion

This project leverages insights from multi-hop QA research and builds upon the UnifiedQA framework. By combining dataset integration, retrieval-augmented learning, and reasoning supervision, the proposed approach aims to bridge the gap between single-hop QA models and robust, real-world multi-hop reasoning systems.

References

[1] Z. Yang et al., “HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering,” in *EMNLP*, 2018, pp. 2369–2380.

[2] J. Welbl, P. Stenetorp, and S. Riedel, “Constructing Datasets for Multi-hop Reading Comprehension Across Documents,” *Transactions of the ACL*, vol. 6, pp. 287–302, 2018.

[3] T. Khot et al., “QASC: A Dataset for Question Answering via Sentence Composition,” in *AAAI*, 2020.

[4] X. Ho, A.-K. D. Nguyen, S. Sugawara, and A. Aizawa, “Constructing a Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps,” in *COLING*, 2020, pp. 6609–6625.

[5] A. Zhu, A. Hwang, L. Dugan, and C. Callison-Burch, “FanOutQA: A Multi-Hop, Multi-Document Question Answering Benchmark for Large Language Models,” in *ACL (Short Papers)*, 2024, pp. 18–37.

[6] Y. Tang and Y. Yang, “MultiHop-RAG: Benchmarking Retrieval-Augmented Generation for Multi-Hop Queries,” *arXiv preprint arXiv:2401.15391*, 2024.

[7] S. Min, V. Zhong, L. Zettlemoyer, and H. Hajishirzi, “Multi-hop Reading Comprehension through Question Decomposition and Rescoring,” in *ACL*, 2019.

[8] Y. Fang et al., “Hierarchical Graph Network for Multi-hop Question Answering,” in *EMNLP*, 2020.

[9] M. Tu et al., “Multi-hop Reading Comprehension Across Multiple Documents by Reasoning over Heterogeneous Graphs,” in *ACL*, 2019, pp. 2704–2713.

[10] X. Li et al., “Question-aware Memory Network for Multi-hop Question Answering in Human–Robot Interaction,” *Complex & Intelligent Systems*, vol. 8, pp. 851–861, 2022.

[11] C. Zhao et al., “Transformer-XH: Multi-Evidence Reasoning with eXtra Hop Attention,” in *ICLR*, 2020.

[12] Z. Yu, Y. Belinkov, and S. Ananiadou, “Back Attention: Understanding and Enhancing Multi-Hop Reasoning in Large Language Models,” *arXiv preprint arXiv:2502.10835*, 2025.

[13] Y. Tang, H. T. Ng, and A. Tung, “Do Multi-Hop Question Answering Systems Know How to Answer the Single-Hop Sub-Questions?” in *EACL*, 2021, pp. 3244–3249.

[14] D. Khashabi et al., “UNIFIEDQA: Crossing Format Boundaries with a Single QA System,” in *Findings of EMNLP*, 2020, pp. 1896–1907.