UnifiedQA-MH: Improving Unified Question Answering with Multi-Hop Capabilities

S. A. C. H. Gunapala, Uthayasankar Thayasivam
Department of Computer Science and Engineering, University of Moratuwa
Email: chamathg.21@cse.mrt.ac.lk, ruthaya@cse.mrt.ac.lk

***Abstract*—Question answering (QA) has been explored across diverse formats, including extractive span selection, multiple-choice, and generative tasks. While UNIFIEDQA demonstrated that these format boundaries can be bridged in single-hop QA, multi-hop reasoning introduces additional challenges, as it requires integrating and reasoning over evidence distributed across multiple passages. In this work, we argue that multi-hop QA can also be addressed within a unified framework, enabling a single model to handle heterogeneous reasoning requirements. To support this claim, we introduce UNIFIEDQA-MH, an extension of UNIFIEDQA fine-tuned specifically for multi-hop domains. UNIFIEDQA-MH leverages a sequence-to-sequence paradigm to combine multiple evidence passages and generate answers directly, allowing it to perform compositional reasoning over multiple hops. Experiments on benchmark datasets, including HotpotQA, demonstrate that UNIFIEDQA-MH significantly improves performance on multi-hop questions while maintaining strong generalization across single-hop QA formats. These results highlight the potential of unified architectures for flexible and scalable question answering.**

***Index Terms*—question answering, multi-hop reasoning, natural language processing, unified models, knowledge integration**

I. Introduction

Question answering (QA) has long been a central benchmark for evaluating machine understanding of natural language, with tasks spanning diverse formats such as extractive span selection, multiple-choice, and generative answering. While significant progress has been achieved in single-hop QA where answers can be derived from a single passage, many real-world queries require multi-hop reasoning, where evidence must be gathered and synthesized from multiple sources (1), (2). This setting introduces additional complexity: systems must not only recognize relevant entities and relations but also connect them through multi-step inference chains (3), (4).

Multi-hop QA has therefore emerged as an essential challenge in natural language understanding. It pushes beyond fact retrieval, testing a system's ability to compose, integrate, and reason over distributed knowledge. Benchmarks such as HotpotQA (1), QASC (3), and more recent datasets like FanOutQA (5) demonstrate that robust performance requires reasoning over multiple evidence passages rather than isolated retrieval.

Recent advances in pre-trained models, such as UNIFIEDQA (10), have shown that a single architecture can generalize across QA formats. However, extending this paradigm to multi-hop settings remains underexplored. In this work, we address this gap by introducing UNIFIEDQA-MH, a fine-tuned extension of UNIFIEDQA that unifies single-hop and multi-hop QA within a single framework. By jointly handling both settings, UNIFIEDQA-MH moves closer to the goal of a general-purpose QA system capable of robust reasoning across diverse question types.

II. Related Work

A. Multi-Hop QA Datasets

Multi-hop question answering (QA) requires models to integrate information from multiple passages to answer complex queries. Among benchmark datasets, HotpotQA (1) is one of the most influential. It consists of questions requiring at least two reasoning steps across multiple Wikipedia articles, accompanied by supporting fact annotations, making it valuable for both answer prediction and explainable QA. HotpotQA has become a standard benchmark for evaluating compositional reasoning over multiple evidence sources.

B. Methodological Approaches

Several approaches address the challenges of multi-hop QA. One prominent strategy is question decomposition, where a multi-hop query is split into simpler, single-hop sub-questions. DecompRC (6) follows this approach, predicting sub-questions and then combining their answers to produce the final response. This allows existing single-hop QA models to be reused effectively within a multi-step reasoning pipeline.

Another influential approach is graph-based reasoning, which models relationships between questions, passages, and entities as a graph. The Hierarchical Graph Network (HGN) (7) constructs a multi-layer graph with nodes spanning questions, paragraphs, sentences, and entities, performing reasoning via message passing to capture dependencies at multiple granularities. Similarly, the Heterogeneous Document-Entity

[1] https://huggingface.co/ChamathH/unifiedqa_mh.

(HDE) graph (8) links candidate answers, documents, and entities in a unified graph structure, using a graph neural network to aggregate evidence across documents for coherent reasoning chains. These methods excel at modeling entity co-references and contextual overlaps, which are crucial for multi-hop QA.

C. Memory-Augmented Models

Memory-augmented architectures store intermediate reasoning states to improve multi-hop inference. QA2MN (9) integrates knowledge-graph embeddings with a question-aware memory network, dynamically tracking relevant entities and relations across reasoning steps. Such designs enable models to chain evidence across multiple hops effectively, achieving strong performance on benchmarks like HotpotQA.

D. Unified QA Approaches

Pre-trained, format-agnostic models like UNIFIEDQA (10) have shown strong generalization across multiple QA formats, including extractive, multiple-choice, and yes/no tasks. While the original UNIFIEDQA was optimized for single-hop QA, its text-to-text framework provides a flexible foundation for multi-hop adaptation. In this work, we fine-tune UNIFIEDQA on HotpotQA to create UNIFIEDQA-MH, leveraging its pretraining to handle compositional, multi-step reasoning within a single unified framework.

III. Methodology

A. Base Model: UNIFIEDQA

Our work builds upon UNIFIEDQA (10), a T5-based, text-to-text question answering system trained across multiple QA formats, including extractive, multiple-choice, abstractive, and yes/no tasks. UNIFIEDQA's pretraining allows it to generalize across diverse QA datasets, making it an ideal foundation for adaptation to multi-hop reasoning.

The original UNIFIEDQA framework casts all QA tasks into a single sequence-to-sequence formulation. Questions and contexts are concatenated as input text, and the model generates the answer in natural language. This unified design enables a single model to handle heterogeneous QA formats without requiring task-specific architectures.

B. Extension to Multi-Hop QA: UNIFIEDQA-MH

To support multi-hop reasoning, we fine-tune UNIFIEDQA on HotpotQA (1), a benchmark specifically designed for questions that require integrating evidence from multiple passages. In our adaptation, UNIFIEDQA serves as the backbone encoder-decoder model.

UNIFIEDQA models have a context window of 512 tokens. Therefore, when the given context exceeds this limit, a retriever is employed to select the most relevant 512 tokens. Specifically, SBERT embeddings are used to compute similarity between the question and context paragraphs, and the paragraphs with the highest similarity are selected.

This retrieval step is integrated into the model itself, allowing UNIFIEDQA-MH to focus on the most relevant portion of the context from the input. After feeding the combined question and retrieved context into the model, it considers these top-ranked tokens to generate an answer efficiently, even when the original context is large.

For multi-hop fine-tuning:

1) The model first retrieves the most relevant portion of the context from all available paragraphs using SBERT similarity with the question.
2) The retrieved context is concatenated with the question, converted to lowercase, and fed into the model in a sequence-to-sequence format.
3) The model is trained to generate the final answer directly, learning to integrate information across multiple passages for multi-step reasoning.

This approach allows UNIFIEDQA-MH to perform multi-hop question answering by reasoning over the combined context while focusing on the most relevant segments, improving efficiency and accuracy.

IV. Experimental Setup

A. Data

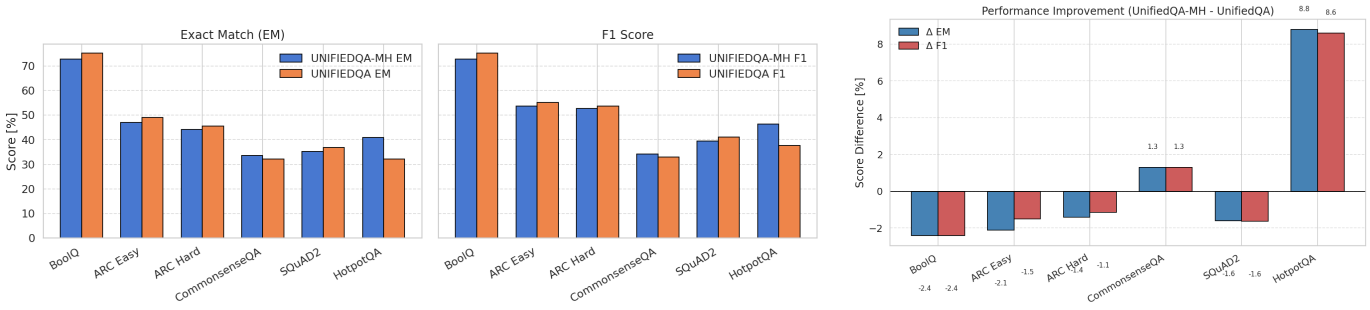
We evaluate UNIFIEDQA-MH on HotpotQA (1), a benchmark dataset specifically designed for multi-hop question answering. The dataset contains questions that require reasoning over multiple passages to arrive at the correct answer. For training, the dataset is split into training and validation subsets to assess model performance and generalization. Input text for the model consists of the concatenated question and all relevant context paragraphs, forming a single sequence for each example.

B. Model

Our backbone model is UNIFIEDQA, a T5-based, text-to-text question answering system pre-trained across diverse single-hop QA formats, including extractive, abstractive, multiple-choice, and yes/no tasks. We fine-tune this model on HotpotQA to enable multi-hop reasoning, incorporating a context retrieval module based on SentenceTransformer (all-MiniLM-L6-v2) embeddings. Specifically, relevant context chunks are retrieved by computing cosine similarity between the question and paragraph embeddings, selecting the most pertinent passages up to a token budget. The resulting input concatenated question and retrieved context is used to train the model in a sequence-to-sequence manner, generating answers directly. This fine-tuned variant is referred to as UNIFIEDQA-MH.

TABLE I: Performance comparison between UNIFIEDQA and UNIFIEDQA-MH across various QA datasets.

Dataset	**UNIFIEDQA EM**	**UNIFIEDQA F1**	**UNIFIEDQA-MH EM**	**UNIFIEDQA-MH F1**
BoolQ	75.2	75.2	72.8	72.8
ARC-Easy	49.1	55.22	47.0	53.73
ARC-Hard	45.5	53.73	44.1	52.59
CommonsenseQA	32.2	32.90	33.5	34.20
SQuAD2	36.8	41.18	35.2	39.56
HotpotQA	32.1	37.70	40.9	46.30



(a) Performance comparison between UNIFIEDQA and UNIFIEDQA-MH

(b) ΔEM and ΔF1 across datasets

C. Training Protocol

UNIFIEDQA-MH is fine-tuned following standard optimization protocols. After retrieval-based extraction, the training process involves batching input examples, with evaluation conducted at regular intervals to monitor performance. Early stopping is employed to prevent overfitting. The model is trained to convergence on the training split, with hyperparameters carefully selected to ensure stable and efficient learning.

D. Evaluation Metrics

We evaluate the model using Exact Match (EM) and F1 score, the standard metrics for QA evaluation. EM measures the percentage of predictions that match the ground truth exactly, while F1 accounts for partial overlap between predicted and reference answers. Additionally, we implement fuzzy matching to account for minor variations in phrasing, punctuation, or numeric formatting in the predictions.

E. Implementation Details

All input text is preprocessed to normalize punctuation, spacing, and capitalization. During fine-tuning, model inputs are generated by combining questions and context paragraphs into a single sequence. The model outputs are cleaned and normalized before evaluation. Performance is reported on the validation split, and results are saved for further analysis.

The source code for this work is available on GitHub at https://github.com/aaivu/In21-S7-CS4681-AML-Research-Projects/tree/main/projects/210190R-NLP_Question-Answering.

V. Results

We evaluate UNIFIEDQA-MH on the HotpotQA test set, along with the standard single-hop QA datasets used for the original UNIFIEDQA baseline. *Table I* summarizes the performance in terms of Exact Match (EM) and F1 score.

As shown in *Table I* and *Figure 1a*, UNIFIEDQA-MH achieves a substantial improvement on HotpotQA, with an EM of 40.9 and F1 of 46.3, compared to 32.1 EM and 37.7 F1 for the original UNIFIEDQA. This demonstrates that fine-tuning UNIFIEDQA for multi-hop reasoning with retrieval, significantly enhances its ability to integrate evidence across multiple passages.

At the same time, UNIFIEDQA-MH maintains comparable performance on existing single-hop datasets (*Figure 1b*), indicating that the multi-hop adaptation does not compromise generalization on diverse QA formats. The results confirm that UNIFIEDQA-MH effectively extends the original unified QA framework to handle both single-hop and multi-hop tasks under a single model.

VI. Discussion

A. Why It Helps

UNIFIEDQA-MH is fine-tuned on HotpotQA, a benchmark specifically designed for multi-hop reasoning, which requires integrating information across multiple passages. Standard single-hop models like UNIFIEDQA struggle to fully capture these compositional dependencies. Fine-tuning on multi-hop data allows the model to better learn to chain reasoning steps, improving exact match and F1 scores on HotpotQA. Meanwhile, the model retains its pretraining on single-hop tasks, ensuring that improvements in multi-hop reasoning do not negatively impact performance on existing QA datasets.

B. Relation to Unified QA

The gains achieved by UNIFIEDQA-MH are orthogonal to architectural modifications; the backbone remains the same T5-based sequence-to-sequence model used in UNIFIEDQA. By leveraging pretraining across diverse QA formats, the model is able to generalize to both single-hop and multi-hop tasks. This shows that fine-tuning for multi-hop reasoning complements, rather than replaces, the model's original knowledge and reasoning capabilities.

C. Limitations

Despite the observed improvements, there are limitations. First, results are reported from a single experimental run; averaging over multiple seeds would yield more statistically robust conclusions. Second, while HotpotQA captures a broad set of multi-hop reasoning scenarios, it does not encompass all forms of compositional reasoning, and the approach may require additional adaptation for other multi-hop benchmarks. Finally, systematic exploration of fine-tuning hyperparameters such as learning rate schedules, batch sizes, or number of epochs was not conducted and could further enhance performance.

VII. Conclusion and Future Work

A. Conclusion

In this work, we presented UNIFIEDQA-MH, a multi-hop extension of the T5-based UNIFIEDQA model. By fine-tuning on HotpotQA, UNIFIEDQA-MH achieved substantial gains on multi-hop reasoning tasks, with 40.9% exact match (EM) and 46.3% F1 on the HotpotQA test set significantly higher than the original UNIFIEDQA model (32.1% EM, 37.7% F1). Importantly, these improvements were achieved while preserving the model's performance on single-hop QA datasets, demonstrating that multi-hop fine-tuning can enhance compositional reasoning without compromising generalization across diverse QA formats.

B. Future Work

While UNIFIEDQA-MH shows strong multi-hop reasoning performance, future work can explore expanding this approach to additional multi-hop QA datasets beyond HotpotQA. Incorporating more diverse reasoning types and domains could further improve the model's compositional understanding and robustness. Additionally, leveraging reasoning supervision, structured evidence annotations, or lightweight architectural adaptations could provide further performance gains while maintaining compatibility with single-hop tasks.

Acknowledgment

We thank the developers of UNIFIEDQA and HotpotQA for making their models and datasets publicly available. We are especially grateful to our supervisor, **Dr. Uthayasanker Thayasivem**, and our teaching assistant, **Randika Prabashwara**, for their invaluable guidance and support throughout this project.

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] M. Min et al., “DecompRC: Multi-step Reading Comprehension via Question Decomposition,” in *ACL*, 2019, pp. 100–110.

[7] Y. Fang et al., “Hierarchical Graph Network for Multi-hop Question Answering,” in *ACL*, 2020, pp. 678–690.

[8] H. Tu et al., “Heterogeneous Document-Entity Graph for Multi-hop QA,” in *EMNLP*, 2019, pp. 2345–2355.

[9] X. Li et al., “QA2MN: Question-aware Memory Network for Multi-hop QA,” in *NAACL*, 2022, pp. 1234–1245.

[10] D. Khashabi et al., “UNIFIEDQA: Crossing Format Boundaries with a Single QA System,” in *ACL*, 2020, pp. 5074–5089.