A Survey on Advances in Retrieval-Augmented Generation over Tabular Data and Table QA

Hassan Soliman and Iryna Gurevych

Ubiquitous Knowledge Processing Lab (UKP Lab)
Department of Computer Science and Hessian Center for AI (hessian.AI)
Technical University of Darmstadt
www.ukp.tu-darmstadt.de

Abstract. Recent developments in retrieval-augmented generation (RAG) and question answering (QA) for tabular data have shown considerable promise in tackling issues related to data retrieval, semantic comprehension, and intricate reasoning. This paper examines significant trends, insights, and constraints across multiple domains, highlighting progress in TableQA, multi-table retrieval, multimodal table retrieval, and generative information retrieval (GenIR). These innovations are essential for enhancing machine interactions with structured datasets, setting the stage for scalable and precise decision-making solutions in practical applications.

Keywords: TableQA · Retrieval-Augmented Generation · Multi-Table Retrieval · Multimodal Table Retrieval · Generative Information Retrieval.

1 Introduction

The swift expansion of data-driven systems demands effective techniques for retrieving and reasoning with structured data. This review centers on four interrelated domains: TableQA, multi-table retrieval, multimodal table retrieval, and GenIR. These fields were chosen to encompass the range of recent advancements in managing structured and hybrid datasets, which enhance scalability, reasoning capabilities, and data interaction.

2 Key Trends and Insights

TableQA: Strategies such as TableRAG [1] improve table comprehension for extensive datasets through schema and cell retrieval, effectively overcoming restrictions related to token and context length. Multi-hop retrieval methods like MURRE [2] boost query accuracy by tackling the shortcomings of single-hop retrieval with a removal strategy. Furthermore, augment-before-you-try techniques utilize external sources to enrich tables, thereby enhancing SQL generation and reasoning [3]. Nonetheless, challenges remain due to the reliance on query design and the precision of SQL generation by large language models (LLMs).

Multi-Table Retrieval: Approaches prioritize understanding join relationships among tables, as illustrated in research such as [4]. This improves relevance between tables, while benchmarks like TQA-Bench [5] provide scalable context management and symbolic extensions that capture the complexities of real-world QA involving tables. MultiTabQA [6] is geared towards generating tabular answers from several tables for intricate inquiries, adapting single-table QA models to multi-table scenarios. It excels in enhancing interpretability, scalability across various domains, and facilitating multi-step aggregation. However, challenges in this area include reliance on precise retrieval and the propagation of errors throughout the process. Important aspects include modeling implicit relationships between tables (such as non-key joins) and bridging gaps in existing benchmarks with new datasets or scenarios. Improved accuracy can be achieved through join-aware retrieval and query decomposition, yet obstacles persist due to the strong dependence on accurate table joins and difficulties in modeling relationships that extend beyond single-column keys.

Multimodal Table Retrieval: Datasets like MMTABQA and MMQA combine text, tables, and images for cross-modal reasoning [7] [8]. Model improvements such as ImplicitDecomp [8] and modality selection networks [9] target multi-hop reasoning and the resolution of ambiguities. However, the computational demands of multimodal table retrieval present a significant challenge, highlighting the need for advancements focused on efficiency.

GenIR: Techniques like RIPOR [10] transition from matching-based retrieval of textual data to generating document identifiers, thus improving scalability. The existing literature in this field predominantly centers on GenIR for text, revealing issues such as resource-heavy autoregressive generation and continual text updates [11][12]. Nevertheless, GenIR for tabular data remains largely uncharted, presenting specific challenges. These include effectively encoding or defining identifiers (IDs) for structured tabular data, managing scalability for real-time inference and dynamic data modifications, and establishing suitable evaluation metrics such as BLEU or cell recall.

3 Conclusion

Current systems face limitations such as syntax issues in generating complex SQL, dependence on certain datasets like Wikipedia, challenges in keeping knowledge updated for generative models, and high computational costs in retrieving multimodal tables. The retrieval of multiple tables encounters the propagation of errors throughout the process because it relies on precise table joins and struggles with implicit relationship modeling. Generative information retrieval techniques for tabular data are still largely uncharted, presenting considerable challenges in terms of scalability, representation, and assessment.

Future investigations should aim at creating streamlined, adaptable GenIR models specifically designed for structured data, broadening the datasets for multi-table retrieval to improve real-world relevance, and setting up benchmarks that are independent of domain. Furthermore, tackling the computational demands in multimodal table retrieval and refining approaches for modeling implicit relationships within tables will foster the development of more resilient and efficient question-answering systems for various applications.

Acknowledgment

This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

1. Chen, Si-An, et al. "TableRAG: Million-Token Table Understanding with Language Models." arXiv, 2024. https://arxiv.org/abs/2410.04739.
2. Zhang, Xuanliang, et al. "MURRE: Multi-Hop Table Retrieval with Removal for Open-Domain Text-to-SQL." In Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025), 2025. https://aclanthology.org/2025.coling-main.386/.
3. Liu, Yujian, et al. "Augment before You Try: Knowledge-Enhanced Table Question Answering via Table Expansion." arXiv, 2024. https://arXiv.org/abs/2401.15555.
4. Chen, Peter Baile, et al. "Is Table Retrieval a Solved Problem? Exploring Join-Aware Multi-Table Retrieval." In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024), 2024. https://aclanthology.org/2024.acl-long.148/.
5. Qiu, Zipeng, et al. "TQA-Bench: Evaluating LLMs for Multi-Table Question Answering with Scalable Context and Symbolic Extension." arXiv, 2024. https://arXiv.org/abs/2411.19504.
6. Pal, Vaishali, et al. "MultiTabQA: Generating Tabular Answers for Multi-Table Question Answering." In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023), 2023. https://aclanthology.org/2023.acl-long.348/.
7. Mathur, Suyash Vardhan, et al. "Knowledge-Aware Reasoning over Multimodal Semi-Structured Tables." In Findings of the Association for Computational Linguistics: EMNLP 2024, 2024. https://aclanthology.org/2024.findings-emnlp.822/.
8. Talmor, Alon, et al. "MultiModalQA: Complex Question Answering over Text, Tables and Images." arXiv, 2021. https://arxiv.org/abs/2104.06039.
9. Hannan, Darryl, et al. "ManyModalQA: Modality Disambiguation and QA over Diverse Inputs." In Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI 2020), 2020. https://ojs.aaai.org/index.php/AAAI/article/view/6294.
10. Zeng, Hansi, et al. "Scalable and Effective Generative Information Retrieval." In Proceedings of the ACM Web Conference 2024 (WWW '24), 2024. https://dl.acm.org/doi/10.1145/3589334.3645477.
11. Li, Xiaoxi, et al. "From Matching to Generation: A Survey on Generative Information Retrieval." arXiv, 2024. https://arXiv.org/abs/2404.14851.
12. Wang, Ye, et al. "Generative Retrieval with Large Language Models." arXiv, 2024. https://arxiv.org/abs/2402.17010.