GraphRAG with Knowledge Graphs for Question Answering on Administrative Meeting Records

Kumi Ushio[1,*], Daichi Tsuji[2] and Yohei Kobashi[1]

[1] *The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan*
[2] *Kyoto Prefectural government, Yabunouchi-cho, Nishiiru, Shinmachi, Shimodachiuri-dori, Kamigyo-ku, Kyoto-shi, Kyoto 602-8570, Japan*

Abstract

This study introduces a GraphRAG-based question-answering system for Japanese administrative meeting minutes, addressing challenges in accessing policy-related information. Using the minutes from Japan's Financial Services Agency, we constructed a lightweight ontology-aware knowledge graph capturing participants, meetings, and utterances, and integrated it with LLMs. The system applies a GraphRAG approach, leveraging graph-based context expansion to integrate related nodes and enrich contextual understanding, combined with dynamic tool selection to support multi-step reasoning. Evaluation with questions showed high accuracy for both simple retrieval and relation-exploration queries. Future work includes improving retrieval accuracy, developing domain-specific ontologies, automating tool generation, and deploying the system as an interactive application.

Keywords

Graph RAG, knowledge graph, administrative text, meeting minutes, question answering, sustainable finance

1. Introduction

Councils of the Japanese government and study groups involve diverse stakeholders, and their meeting minutes contain essential information for understanding how policy decisions are formulated. While these minutes are disclosed to enhance transparency in administrative activities and decision-making processes, their lack of standardized formats hinders effective search and utilization. To understand the context and intent of policy decisions, advanced methods for information extraction, retrieval, and analytical support are required.

Recent years show growing interest in integrating large language models (LLMs) with external knowledge for search and generation [1][2]. Traditional Retrieval-Augmented Generation (RAG) mainly uses vector-based retrieval to identify semantically similar text fragments and feed them into generative models, but faces limitations in capturing complex inter-textual relationships and structural context. Against this backdrop, Graph Retrieval-Augmented Generation (GraphRAG) emerged as a promising approach [3]. GraphRAG structures entities and relations as a knowledge graph and operates through three stages: indexing, retrieval, and generation [4]. This approach improves contextual coherence and supports multi-hop reasoning and relational understanding. Recent surveys reviewed GraphRAG's architecture and highlighted its applications across domains like question answering, report generation, legal analysis, and scientific research, while identifying challenges including scalability and multimodal integration [5]. However, most existing studies have focused on English encyclopedic texts and domain-specific documents such as legal, medical, and scientific texts, with few reports on the application of GraphRAG to non-English policy documents or administrative meeting records [6][7]. These documents include meeting-level context, inter-speaker relationships, and temporal structures, which make simple keyword or vector-based retrieval struggle to achieve sufficient accuracy.

ISWC 2025 Companion Volume, November 2–6, 2025, Nara, Japan
* Corresponding author.
ushio-kumi@g.ecc.u-tokyo.ac.jp (K. Ushio)

In this study, we target the meeting minutes of the "Expert Panel on Sustainable Finance," disclosed by Japan's Financial Services Agency (FSA), which oversees financial administration in Japan. Based on metadata attached to individual utterances, we construct a knowledge graph and integrate it with a large language model to design and prototype a question-answering system. Unlike conventional semantic retrieval approaches, the proposed system employs graph-based context expansion, leveraging relationships among nodes in the knowledge graph, and aims to generate coherent responses to natural language queries that account for the connections among utterances, meetings, and participants.

2. Construction of the Minutes Knowledge Graph

This study builds a knowledge graph from transcripts of the Japanese FSA's "Sustainable Finance Expert Panel" for integration with a GraphRAG-based Question Answering (QA) system, while also supporting analyses such as discourse structure and speaker networks. The graph is designed to preserve the semantic structure of the data while ensuring extensibility for future reasoning and interoperability.

The model consists of three node types—Person, Utterance, and Meeting—and three edge types—made_statement, participated_in, and occurred_in. Person nodes represent meeting participants and include attributes such as name, organization, and position. Utterance nodes represent individual statements made during meetings, retaining speech content as their main attribute. Meeting nodes represent individual meetings, including attributes such as date and fiscal year. Edges capture the semantic relationships between entities: made_statement links a person to their utterances, participated_in links a person to meetings, and occurred_in links utterances to meetings.

The design follows a lightweight ontology-aware approach that ensures semantic consistency and aligns well with established vocabularies such as FOAF and Schema.org [8][9]. For example, Person nodes can map to foaf:Person, Meetings to Event, and Utterances to Speech or CreativeWork. Relationship types can correspond to existing terms such as attendee or partOfEvent, enabling future integration into Linked Open Data ecosystems. This structure allows for future extensions, including hierarchical categorization of meetings, temporal reasoning over utterances, and interoperability with other meeting datasets. Beyond network representation, the approach provides a semantically grounded model, supporting advanced analysis and knowledge-based reasoning.

3. Design of the Question Answering System

The proposed QA system is designed to analyze structured meeting minutes of expert panels on sustainable finance, combining flexibility for natural language queries with precision in retrieving relevant information. Its graph-based architecture represents entities such as Persons, Meetings, and Utterances as nodes and edges, capturing semantic relationships. This graph-centric foundation supports both reasoning and advanced analytics.

The system begins by loading node and edge data and building a MultiDiGraph using the NetworkX library. Each node retains metadata, such as names, organizational affiliations, roles, dates, and content [10]. The graph structure captures relations like 'spoke at' or 'participated in,' linking utterances to speakers and meetings for reasoning. To support question answering, the system integrates GPT-4o via LangChain's agent-based framework [11]. The agent uses a custom prompt to plan queries and select tools instead of relying solely on text generation. This enables the system to execute analytical tasks such as ranking, filtering, and keyword-based search that require precise computation. A comprehensive set of tools is implemented to support various functions, from retrieving meetings and participants to performing statistical and network-based analyses. The system can identify highly active participants by analyzing utterance frequency, detect thematic engagement through keyword filtering, and uncover co-occurrence patterns to map collaboration.

One distinctive feature is its ability to construct co-attendance networks and compute centrality measures—such as betweenness, eigenvector, and degree—to highlight influential participants. Additionally, it allows temporal and semantic filtering for targeted content extraction, supporting trend analysis and thematic exploration.

This pipeline processes user queries by dynamically invoking tools, aggregating results, and producing responses in natural language. By combining symbolic graph reasoning with LLM interpretation, the system ensures responses are accurate, interpretable, and grounded in verifiable data.

4. Response Experiments and Discussion

To evaluate the proposed QA system, we created 50 questions based on the meeting minutes of the "Expert Panel on Sustainable Finance" published by Japan's FSA and applied them to the system for performance assessment. The questions reflected diverse user needs for meeting minutes, ranging from simple retrieval-oriented queries to those requiring multi-step reasoning that integrates information about participants, meetings, and utterances (Table 1). Answer correctness was assessed manually.

Table 2 summarizes the results. In addition to achieving high accuracy for simple retrieval questions, the system also performed well on relation-exploration questions. Because accuracy and F1 score were identical in this evaluation, accuracy was adopted as the primary metric. The proposed system demonstrated an advantage in handling complex queries by leveraging tool selection and step-by-step reasoning via a LangChain agent, enabling deductive responses utilizing structured data. However, for queries requiring exhaustive answers, response time tended to increase, indicating that optimizing interaction costs remains an open issue. Furthermore, the current system relies on manual addition and design of tools; therefore, future work should focus on enhancing generalizability through automated tool generation and integration with SPARQL.

Table 1
Example Questions and Answers

Question Type	Example Questions	Example Answers
Simple Retrieval	"Please list all the meeting dates on which carbon pricing was discussed."	"The meetings where carbon pricing was discussed were held on the following dates: 1. March 2, 2021 2. March 25, 2021 3. April 22, 2021 4. May 28, 2021 5. June 10, 2022 6. December 21, 2023 7. March 1, 2024 8. June 17, 2024 9. October 18, 2024."
Aggregation	"Please identify the three individuals who made the most comments on "TCFD" between April 2021 and March 2022."	"The three individuals who made the most comments on TCFD between April 2021 and March 2022 are: 1. Joji Iguchi (11 comments) 2. Hiroyuki Tezuka (11 comments) 3. Tsuyoshi Mizuguchi (8 comments)."
Relation Exploration	"Please provide the three individuals with the highest betweenness centrality in the network analysis."	"The three individuals with the highest betweenness centrality in the network analysis are: 1. Tsuyoshi Mizuguchi (0.121508) 2. Reiko Hayashi (0.121508) 3. Eiichiro Adachi (0.075881)."
Complex Analysis	"Please provide the number of comments on "TCFD" and "TNFD", along with up to three representative examples and illustrate their trends."	"The numbers of comments and representative examples for TCFD and TNFD are as follows: TCFD — Joji Iguchi and Hiroyuki Tezuka (11 comments each). Remarks emphasized sustainable finance, climate change, and Paris Agreement goals. [...] TNFD — Tsuyoshi Mizuguchi (7 comments). Other remarks highlighted biodiversity disclosure and business impacts. [...] Trends: TCFD focused on climate finance, while TNFD emphasized biodiversity."

Table 2

Evaluation of System Performance by Question Type

Question Type	Questions	Answers	Correct Answers	Accuracy
Simple Retrieval	12	12	9	0.750
Aggregation	19	19	11	0.579
Relation Exploration	16	16	13	0.813
Complex Analysis	3	3	1	0.333

Table 3

Aggregated Retrieval, Generation, and Judgment Metrics for Compared Methods

Method Group	Recall@10	Precision@10	Hit rate@10
BaselineNgram/BaselineNgram+LLM	0.114	0.372	0.88
BM25/BM25+LLM	0.318	0.624	0.960
BM25+Graph/BM25+Graph+LLM	0.223	0.484	0.92
VectorEmbed/VectorEmbed+LLM	0.316	0.624	0.960
Vector+Graph/Vector+Graph+LLM	0.255	0.568	0.94
RRF(BM25+TFIDF+Ngram)/ RRF(BM25+TFIDF+Ngram)+LLM	0.271	0.57	0.960
GraphRAG-Hybrid/ GraphRAG-Hybrid+LLM	0.111	0.382	0.88
GraphRAG-HardenedHybrid/ GraphRAG-HardenedHybrid+LLM	0.128	0.398	0.92
GraphRAG-Seed+Neighbor/ GraphRAG-Seed+Neighbor+LLM	0.113	0.378	0.8
Ontology-aware GraphRAG (adopted approach)	0.311	0.606	0.62

Method Group	Faithfulness	Groundedness	Coherence
BaselineNgram/BaselineNgram+LLM	0.713	0.352	0.535
BM25/BM25+LLM	0.886	0.606	0.757
BM25+Graph/BM25+Graph+LLM	0.756	0.33	0.555
VectorEmbed/VectorEmbed+LLM	0.778	0.392	0.569
Vector+Graph/Vector+Graph+LLM	0.765	0.326	0.541
RRF(BM25+TFIDF+Ngram)/ RRF(BM25+TFIDF+Ngram)+LLM	0.827	0.484	0.653
GraphRAG-Hybrid/ GraphRAG-Hybrid+LLM	0.711	0.348	0.541
GraphRAG-HardenedHybrid/ GraphRAG-HardenedHybrid+LLM	0.721	0.338	0.556
GraphRAG-Seed+Neighbor/ GraphRAG-Seed+Neighbor+LLM	0.7	0.332	0.526
Ontology-aware GraphRAG (adopted approach)	0.933	0.684	0.874

Method Group	LLM-judged Faithfulness	LLM-judged Groundedness	LLM-judged Coherence	Exact Match Accuracy
BaselineNgram/BaselineNgram+LLM	0.428	0.352	0.535	0.02
BM25/BM25+LLM	0.674	0.606	0.757	0.02
BM25+Graph/BM25+Graph+LLM	0.416	0.33	0.555	0.02
VectorEmbed/VectorEmbed+LLM	0.454	0.392	0.569	0.02
Vector+Graph/Vector+Graph+LLM	0.418	0.326	0.541	0.02
RRF(BM25+TFIDF+Ngram)/ RRF(BM25+TFIDF+Ngram)+LLM	0.568	0.484	0.653	0.02
GraphRAG-Hybrid/ GraphRAG-Hybrid+LLM	0.434	0.348	0.541	0.02
GraphRAG-HardenedHybrid/ GraphRAG-HardenedHybrid+LLM	0.418	0.338	0.556	0.02
GraphRAG-Seed+Neighbor/ GraphRAG-Seed+Neighbor+LLM	0.42	0.332	0.526	0.02
Ontology-aware GraphRAG (adopted approach)	0.756	0.684	0.874	0.480

To strengthen the evaluation, we additionally conducted comparative experiments across ten groups of retrieval methods covering both classical and graph-based approaches. These include: BM25 (with/without Graph and LLM integration) [12], n-gram baselines [13], vector-based methods (VectorEmbed, Vector+Graph) [14], hybrid retrieval with reciprocal rank fusion (RRF) [15], multiple GraphRAG variants (Hybrid, HardenedHybrid, Seed+Neighbor) [16], and the ontology-aware GraphRAG approach we adopted [17].

To carry out these comparative experiments, we prepared an evaluation tool that computes both retrieval and generation metrics for all method groups. Retrieval performance was measured at the utterance level using Recall@10, Precision@10, and Hit Rate@10. Generation quality was evaluated with Faithfulness, Groundedness, and Coherence, while LLM-based judgment functions were additionally applied to automatically score outputs. Finally, overall correctness was captured with Exact Match Accuracy.

The Ontology-aware GraphRAG (adopted approach) achieved the highest performance, with exact match accuracy at 0.48 and superior generation quality (faithfulness 0.933, coherence 0.874, answer score 0.968), clearly outperforming all baselines despite weaker recall and precision (Table 3). BM25 (with/without LLM) provided the most stable retrieval strength, leading in Recall@10 and Precision@10, though its exact match accuracy remained low (0.02), confirming its role as a reliable baseline. RRF (BM25+TFIDF+Ngram) further enhanced recall and hit rate relative to BM25, validating hybrid retrieval, but still exhibited low correctness and mid-level generation. Vector-based and GraphRAG variants showed limited gains, and BaselineNgram performed the worst. Overall, structured ontology-aware graphs proved decisive for trustworthy QA.

5. Conclusion and Future Work

This paper presents ongoing work on constructing a knowledge graph from meeting minutes and applying GraphRAG to Japanese administrative texts, with the aim of improving comprehensibility in discussions and policy-making at national-level meetings. We introduced a graph-based QA system that integrates this knowledge graph with LLMs to support policy-related information retrieval and analysis. Focusing on the “Expert Panel on Sustainable Finance” minutes published by Japan’s Financial Services Agency, we developed a lightweight ontology-oriented graph that captures relationships among participants, utterances, and meetings. The proposed system employs graph-based context expansion combined with dynamic tool selection and stepwise reasoning, enabling coherent responses that leverage both structured and unstructured information.

Evaluation using a diverse set of questions demonstrated that the system achieves high accuracy for both simple retrieval and relation-exploration queries, highlighting the benefits of incorporating graph reasoning into QA workflows. However, the results also revealed limitations, including increased response time for exhaustive queries and the reliance on manually configured tools, which constrain scalability and adaptability.

In the future, we plan to improve retrieval and response generation accuracy through advanced ranking methods and context-aware retrieval strategies. Additionally, we are working on enhancing semantic richness by developing or aligning with domain-specific ontologies to enable deeper reasoning, semantic interoperability, and potential integration into Linked Open Data ecosystems. Other directions include automated tool generation, SPARQL integration for structured querying, and latency reduction strategies to optimize interaction costs. We also aim to extend the system to support multimodal datasets and to deploy it as an interactive application to facilitate usability testing and practical adoption. Ultimately, the goal is for this system to be deployed and utilized not only in various national-level conferences but also across local governments in Japan. Collectively, these enhancements aim to advance knowledge-driven QA for policy documents and administrative records. We hope this work will share key insights and generate valuable feedback from the research community.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, et al., Retrieval-augmented generation for knowledge-intensive NLP tasks, in: Advances in Neural Information Processing Systems (NeurIPS), 2020. URL: https://arxiv.org/abs/2005.11401.

[2] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, et al., On the opportunities and risks of foundation models, arXiv preprint arXiv:2108.07258 (2021). URL: https://arxiv.org/abs/2108.07258.

[3] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, J. Larson, From local to global: A graph RAG approach to query-focused summarization, arXiv preprint arXiv:2404.16130 (2024). URL: https://arxiv.org/abs/2404.16130.

[4] B. Peng, et al., Graph retrieval-augmented generation: A survey, arXiv preprint arXiv:2408.08921 (2024). URL: https://arxiv.org/abs/2408.08921.

[5] H. Han, et al., Retrieval-augmented generation with graphs (GraphRAG), arXiv preprint arXiv:2501.00309 (2024). URL: https://arxiv.org/abs/2501.00309.

[6] T. T. Procko, O. Ochoa, Graph retrieval-augmented generation for large language models: A survey, in: 2024 Conference on AI, Science, Engineering, and Technology (AIxSET), IEEE, 2024, pp. 166–169. doi:10.1109/AIxSET60776.2024.10771030.

[7] Q. Zhang, S. Chen, Y.-Q. Bei, Z. Yuan, H. Zhou, Z. Hong, et al., A survey of graph retrieval-augmented generation for customized large language models, arXiv preprint arXiv:2501.13958 (2025). URL: https://arxiv.org/abs/2501.13958.

[8] FOAF Vocabulary Specification. URL: http://xmlns.com/foaf/spec/.

[9] Schema.org, Schema.org vocabulary. URL: https://schema.org/.

[10] E. Hagberg, D. Schult, P. Swart, Exploring network structure, dynamics, and function using NetworkX, in: Proceedings of the 7th Python in Science Conference (SciPy), 2008, pp. 11–15. URL: https://proceedings.scipy.org/2008.

[11] LangChain, LangChain documentation. URL: https://www.langchain.com/.

[12] S. Robertson, H. Zaragoza, The probabilistic relevance framework: BM25 and beyond, Foundations and Trends in Information Retrieval 3(4) (2009) 333–389. doi:10.1561/1500000019.

[13] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. Della Pietra, J. C. Lai, Class-based n-gram models of natural language, Computational Linguistics 18(4) (1992) 467–479.

[14] V. Karpukhin, B. Oguz, S. Min, L. Wu, S. Edunov, D. Chen, W. Yih, Dense passage retrieval for open-domain question answering, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 6769–6781. URL: https://arxiv.org/abs/2004.04906.

[15] G. V. Cormack, C. R. Clarke, S. Buettcher, Reciprocal rank fusion outperforms Condorcet and individual rank learning methods, in: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2009, pp. 758–759. doi:10.1145/1571941.1572114.

[16] Y. Zhao, X. Li, J. Wang, Ontology-constrained graph retrieval-augmented generation, arXiv preprint arXiv:2412.15235 (2024). URL: https://arxiv.org/abs/2412.15235.

[17] J. Yang, M. Chen, L. Wu, Constrained knowledge graph construction for retrieval-augmented generation, arXiv preprint arXiv:2502.18992 (2025). URL: https://arxiv.org/abs/2502.18992.