Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/71/20241636

Application of large language models based on knowledge graphs in question-answering systems: A review

Yani Wang

School of Mathematics and Computer Science, Northwest Minzu University, Chengguan District, Lanzhou, 730030, China

210136075@qq.com

Abstract. The integration of Knowledge Graphs (KGs) and Large Language Models (LLMs) is emerging as a transformative advancement in AI, particularly within Question Answering (QA) systems. Traditional QA systems, constrained by static knowledge bases, have struggled with multimodal queries and personalized responses. The deep integration of KGs and LLMs offers a novel approach, combining the structured, contextual understanding of KGs with the semantic parsing capabilities of LLMs. This review explores the methodologies, algorithms, datasets, and applications of KG-LLM integration in QA systems, highlighting key research that enhances model performance, reasoning, and domain-specific applications. It addresses technical challenges such as data consistency, semantic understanding, and system optimization. Future directions focus on improving intrinsic model performance and achieving deeper integration of KGs and LLMs. This comprehensive overview underscores the potential of KG-LLM synergy to significantly improve service quality across diverse industries, including education and healthcare.

Keywords: Knowledge Graphs, KGs Large Language Models, LLMs Question Answering.

1. Introduction

The integration of Knowledge Graphs (KGs) and Large Language Models (LLMs) has become a transformative force in AI technology, particularly in Question Answering (QA) systems. Traditional QA systems relied on static knowledge bases, often limiting questions and answers to a single modality of natural language. These systems struggled with multimodal questions involving images, audio, and other formats, and their accuracy was constrained by the finite scope of the knowledge base, lacking intelligent, personalized experiences [1].

The deep integration of KGs and LLMs offers a novel approach to developing high-precision, intelligent QA systems. KGs provide the ability to understand context and integrate fragmented knowledge in a precise, coherent, and structured manner, while LLMs excel at discovering knowledge and parsing semantic complexities. For example, transformer-based models like the GPT series have been integrated into KGs for training and optimization, ensuring efficient collaboration between KGs and LLMs. This synergy enhances query accuracy, enables handling diverse data types, and effectively answers multimodal questions. Additionally, LLMs can leverage the structured information in KGs to understand entities and relationships, thereby solving more complex problems. Application of Large Language Models Based on Knowledge Graphs in Question-Answering Systems has been studied by

DOI: 10.54254/2755-2721/71/20241636

many researchers. Yasunaga et al. through both quantitative and qualitative analyses, we showed QA-GNN's improvements over existing LM and LM+KG models on question answering tasks, as well as its capability to perform interpretable and structured reasoning, e.g., correctly handling negation in questions[2]. Hu et al. presented a investigate the application of PLMs to solve a knowledge-intensive task, namely knowledge graph question answering[3]. Conduct comprehensive experiments to explore the accuracy and efficiency performance of PLMs on KGQA, as well as the scalability of PLMs as KG size increases. Wu and Wang proposed to use large language models and knowledge graph technology to construct an intelligent question answering system for specific fields[4]. Through systematic training and optimization, efficient domain specific knowledge Q&A has been achieved, improving the satisfaction rate of domain specific knowledge Q&A. The above reviews highlight the application of large language models integrated with knowledge graphs in Q&A systems, emphasizing the importance of model performance, reasoning capabilities, data processing capabilities, and domain-specific applications.

This review explores research on the integration of large language models and knowledge graphs in QA systems, examining various algorithms, datasets, and applications in this field. By evaluating the current state of research, identifying challenges, and highlighting future directions, this review aims to provide a comprehensive overview of enhancing AI systems through the combined analysis of KGs and LLMs. It seeks to serve as a reliable auxiliary tool for QA systems, improving service quality across various industries. For example, it can offer personalized teaching plans for students in education, and assist in diagnosis and clinical decision-making in the medical field.

2. Research Methodology

2.1. Construction and Maintenance of High-Quality Knowledge Graphs

In the process of data collection, web crawlers can be used for automated internet traversal, APIs can be leveraged to access databases, and human annotators can conduct manual reviews. Data preprocessing involves cleaning, transformation, and organization. For entity recognition, linking, and relation extraction, we can adopt rule-based methods, machine learning, and deep learning technologies, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and autoencoders, to achieve automation and enhance accuracy[5,6].

2.2. How to Selection of Large Language Models

Existing language models are mainly built on Transformer, with various architectures and routes. For example, BERT (Bidirectional Encoder Representations from Transformers) is a bidirectional model built on the Transformer encoder module, GPT (Generative Pre-trained Transformer) is a unidirectional model built on the Transformer decoder module, and T5 (Text-to-Text Transfer Transformer) is a model built on both the Transformer encoder and decoder. The introduction of BERT and GPT series models has gradually converged the large language model framework into two different technical paradigms: "pre-training + fine-tuning" paradigm and "pre-training + prompting" paradigm. Among them, the "pre-training + fine-tuning" paradigm is represented by the BERT model, while the "pre-training + prompting" paradigm is represented by the GPT series models 2.3 Integration of Knowledge Graphs and Large Language Models[7].

2.3. System Integration

The combination of LLM and KG has become a powerful tool in the field of artificial intelligence, especially in question answering systems. By leveraging fine-tuning techniques or directly applying large models to relevant operations[8], researchers have made significant progress in enhancing question answering models.

Taffa and Usbeck developed the Knowledge Graph-based Thought (KGT) framework, an innovative solution that integrates LLMs with Knowledge Graphs (KGs) to improve their initial responses by utilizing verifiable information from KGs, thus significantly reducing factual errors in reasoning[9].

DOI: 10.54254/2755-2721/71/20241636

Zhang et al. proposed a new method called Knowledge Preference Alignment (KnowPAT), aiming to improve the performance of large language models (LLMs) in domain-specific question answering (QA). By constructing style preference sets and knowledge preference sets, KnowPAT can tackle the challenges of user-friendliness and domain knowledge utilization simultaneously[10].

Agarwal et al. introduced BYOKG, a universal KGQA system that can work with any target KG and does not require any human-annotated training data. Inspired by human curiosity-driven learning, this system first adopts unsupervised learning to explore unknown knowledge graphs through graph traversal. It then leverages large models to generate natural language questions to supplement the exploration corpus, and finally uses retrieval-enhanced reasoning to achieve question-answering prediction[11].

These algorithms not only improve the accuracy of LLMs but also enhance their domain adaptability, promoting the deep integration of LLMs and KGs. This overview sets the stage for exploring the methodologies and results achieved through various Q&A systems approaches in the subsequent sections.

3. Interdisciplinary Collaboration

In the medical field, the Graph-Care project leverages prompt engineering to extract knowledge from extensive clinical databases, constructing personalized medical knowledge graphs for patients. It then utilizes a Bi-Attention enhanced (BAT) Graph Neural Network (GNN) model to predict downstream tasks[12]. For more complex entity relation extraction, GPT-RE adopts a task-aware retrieval and gold label-induced reasoning approach, enabling contextual learning for relation extraction[13]. Additionally, to mitigate hallucination phenomena, the REALM model innovatively proposes a Retrieval Augmented Generation (RAG)-driven framework. This framework can extract entities from various unstructured data sources (such as clinical notes and electronic health records) and match them with external specialized knowledge graphs, ensuring the consistency and accuracy of the model's output[14].

Li et al. conducted an evaluation of the application of different pre-trained embeddings in an educational content recommendation system based on textual semantic similarity. These embeddings were derived from three unique model types: firstly, contextual embeddings extracted from a pre-trained language model known as SBERT; secondly, contextual embeddings derived from a specific Large Language Model (LLM), namely ADA-002; and thirdly, static embeddings trained using a concept-based knowledge graph, particularly ConceptNet Numberbatch. The experimental results revealed that, across the three academic domains of biology, social science, and physics, and for both quizzes and study questions, leveraging the power of contextual embeddings from a Large Language Model, particularly ADA-002, resulted in the most accurate recommendations, as evidenced by the Recall@3 and Mean Reciprocal Rank (MRR) metrics [15].

4. Technical Challenges and Solutions

4.1. Semantic Understanding and Reasoning

Semantic understanding and reasoning are core to NLP & AI. Interpreting nuanced semantic relationships in human language is challenging. Approaches include rule-based (relying on expertise & domain knowledge), statistical (using ML algorithms like HMMs, CRFs), and neural network-based (RNNs, CNNs, Attention) methods. Neural nets provide end-to-end training, learning semantic representations from raw text [16].

4.2. System Performance Optimization

System performance optimization is vital for efficient operation, especially with soaring data and complex apps. To tackle bottlenecks like slow response and resource inefficiency, we must adopt diverse strategies: using buffers to bridge performance gaps, caching to shorten data access, load balancing via tools like Apache to distribute loads, indexing for faster queries, compression for storage & transmission efficiency, and parallel processing for higher throughput, all requiring robust concurrency control and resource scheduling.

DOI: 10.54254/2755-2721/71/20241636

5. Future Development Directions

5.1. Enhancing the intrinsic performance of large models

Large language models have surpassed expectations, necessitating a focus on integrating structured, high-quality knowledge systems. Research should explore how to enhance these models' understanding of structured data, particularly through knowledge graphs. Teaching models to internalize knowledge from these graphs is crucial. Utilizing technologies like deep learning to extract valuable information from knowledge graphs is essential[17]. Advanced knowledge encoding strategies, such as Graph Neural Networks (GNNs), are important for capturing relational networks and deep semantic information, providing comprehensive and accurate support for language models.

5.2. Deep Integration of LLS and KG

The integration of knowledge graphs with large models has progressed significantly. Studies like Joint LK and QA-GNN have used Graph Neural Networks (GNN) to combine knowledge graphs with large models, achieving deep integration[18,19]. The project advanced this further by incorporating self-supervised learning, enhancing the comprehension of complex structural information and improving reasoning accuracy and efficiency[20]. These advancements demonstrate the potential of integrating knowledge graphs with large models and highlight promising directions for future intelligent technologies.

6. Conclusions

This review examines the methodologies, algorithms, datasets, and applications in KG-LLM integration. It addresses the construction and maintenance of high-quality KGs, selection of LLMs, and their integration. Technical challenges like data consistency, semantic understanding, and system optimization are discussed. Future directions emphasize enhancing model performance and deeper KG-LLM integration. The review underscores the importance of these advancements in improving service quality across various industries.

References

[1] Jovanović, M., & Campbell, M. (2023). Connecting AI: Merging large language models and knowledge graph. Computer, 56(11), 103-108.

[2] Yasunaga, M., Bosselut, A., Ren, H. Y., et al. (2021). QA-GNN: Reasoning with language models and knowledge graphs for question answering. arXiv:2104.06378.

[3] Hu, N., Wu, Y., Qi, G., Min, D., Chen, J., Pan, J. Z., & Ali, Z. (2023). An empirical study of pre-trained language models in simple knowledge graph question answering. World Wide Web, 26(5), 2855-2886.

[4] Wu, Q., & Wang, Y. (2023). Research on intelligent question-answering systems based on large language models and knowledge graphs. 2023 16th International Symposium on Computational Intelligence and Design (ISCID). IEEE.

[5] Shannon, C. E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27(3), 379-423.

[6] Page, L., & Brin, S. (1998). The PageRank citation ranking: Bringing order to the web. Stanford Information Sciences and Engineering Report, 98-03.

[7] Devlin, J., Chang, M., Lee, K., et al. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171-4186). Stroudsburg, PA: Association for Computational Linguistics.

[8] Li, X. X., Zhao, R. C., Chia, Y. K., et al. (2023). Chain of knowledge: A framework for grounding large language models with structured knowledge bases. arXiv:2305.13269.

[9] Taffa, T. A., & Usbeck, R. (2023). Leveraging LLMs in scholarly knowledge graph question answering. arXiv:2311.09841.

DOI: 10.54254/2755-2721/71/20241636

[10] Zhang, Y. C., Chen, Z., Fang, Y., et al. (2023). Knowledgeable preference alignment for LLMs in domain-specific question answering. arXiv:2311.06503.
[11] Agarwal, D., Das, R., Khosla, S., et al. (2023). Bring your own KG: Self-supervised program synthesis for zero-shot KGQA. arXiv:2311.12788.
[12] Jiang, P. C., Xiao, C., Cross, A., et al. (2023). GraphCare: Enhancing healthcare predictions with personalized knowledge graphs. arXiv:2305.12788.
[13] Wan, Z., Cheng, F., Mao, Z. Y., et al. (2023). GPT-RE: In-context learning for relation extraction using large language models. arXiv:2305.02105.
[14] Zhu, Y. H., Ren, C. Y., Xie, S. Y., et al. (2023). REALM: RAG driven enhancement of multimodal.
[15] Li, X., Henriksson, A., & Duneld, J. W. Y. (2024). Evaluating embeddings from pre-trained language models and knowledge graphs for educational content recommendation. Future Internet, 16(1).
[16] Yang, X., Yumer, E., Asente, P., Kraley, M., Kifer, D., & Lee Giles, C. (2017). Learning to extract semantic structure from documents using multimodal fully convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5315-5324).
[17] Zhang, Q. G., Dong, J. N., Chen, H., et al. (2024). KnowGPT: Black-box knowledge injection for large language models. arXiv:2312.06185.
[18] Sun, Y. Q., Shi, Q., Qi, L., et al. (2021). JointLK: Joint reasoning with language models and knowledge graphs for commonsense question answering. arXiv:2112.15027.
[19] Yasunaga, M., Ren, H., Bosselut, A., Liang, P., & Leskovec, J. (2021). QA-GNN: Reasoning with language models and knowledge graphs for question answering. arXiv:2104.06378.
[20] Yasunaga, M., Bosselut, A., Ren, H. Y., et al. (2022). Deep bidirectional language-knowledge graph pretraining. Advances in Neural Information Processing Systems, 35, 37309-37323.