Leveraging Isomorphisms to facilitate Zero-Shot KBQA Generalization

Ritam Dutt[1], Dongfang Ling[1], Yu Gu [2], Carolyn Penstein Rosé[1]

[1]Carnegie Mellon University, [2]Ohio State University
{rdutt,dling,cprose}@cs.cmu.edu,gu.826@osu.edu

Abstract

Designing systems to perform question answering over knowledge bases (KBQA) on unseen schema items in a zero-shot setting remains a challenge. Isomorphisms help characterize the complexity of questions and serve as a lens to assess the capabilities of KBQA systems. We investigate the role of isomorphisms as scaffolds on the zero-shot generalization performance of pre-existing models without any retraining. We note the utility of incorporating isomorphism information across two models and propose several ways of predicting isomorphism categories in an automated manner.

1 Introduction

The task of retrieving information from structured knowledge sources such as knowledge graphs, tables, or databases has received considerable interest from the research community. (Liu et al., 2022; Gu et al., 2021; Zhang et al., 2023; Li et al., 2024). Of recent, there has been tremendous progress towards devising systems that can generalize beyond the i.i.d setting and can operate over unseen knowledge sources with minimal supervision. (Xie et al., 2022; Zhuang et al., 2024; Agarwal et al.)

In the context of question answering over knowledge bases or KBQA, "zero-shot generalization" refers to the ability of KBQA systems to operate over schema items (such as relations and classes) that were unobserved during training. The most salient work in this space is that of Gu et al. (2021) where the authors construct the GrailQA dataset to benchmarks the zero-shot generalization capabilities of KBQA systems (Ye et al., 2021; Yu et al., 2022; Gu and Su, 2022; Shu et al., 2022; Gu et al., 2023; Liu et al., 2022; Luo et al., 2023).

However, in a recent study, Dutt et al. (2023) demonstrated that the original GrailQA dataset was biased towards simpler questions leading to an inaccurate assessment of KBQA systems on zero-shot generalization. The authors leveraged the idea of graphical isomorphism akin to semantic structures (Li and Ji, 2022), reasoning paths (Das et al., 2022) or query graphs (Xu et al., 2023) to characterize the complexity of a given question. Isomorphisms served as a lens to identify which input populations were better serviced by KBQA systems, and revealed that most models were biased towards simpler isomorphism categories, due to the prevalence of such categories in the training distribution.

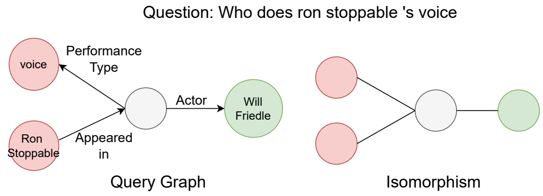


Figure 1: An example of a KBQA question and its corresponding query graph and isomorphism category.

In this extended abstract, we conduct preliminary experiments to assess the role of isomorphisms, as an auxiliary source of information to mitigate the distribution shift during inference. Incorporating this isomorphism information could eliminate the need to retrain KBQA systems on unseen distributions and be readily applied to off-the-shelf systems. Our preliminary experiments highlight the utility of gold isomorphisms for two state-of-the-art KBQA systems on the challenging GrailQA++ dataset (Dutt et al., 2023). We also propose possible ways of predicting the isomorphism category in an automated manner to facilitate zero-shot generalization of KBQA systems during inference.

2 Methodology

Inference Dataset: The GrailQA++ dataset of Dutt et al. (2023) was constructed to evaluate the zero-shot generalizability of KBQA systems trained on the original GrailQA (Gu et al., 2021); it was designed to have an equal proportion of simple and complex isomorphism categories. We show distri-

Table 1: EM/F1 scores of KBQA systems in presence and absence of gold isomorphism (Gold Iso) across different categories and the dataset (ALL). Performance gains are highlighted in bold and drops are colored in red.

ISO Group	T-0	T-1	T-2	T-3	T-4	T-5	ALL
RNG-KBQA	45.8/ 55.2	50.8/ 55.1	41.6/ 59.3	16.5/ 35.1	28.4/ 37.4	1.0/ 10.1	33.2/ 44.4
+ Gold Iso	45.8/ 55.2	50.8/ 55.1	41.8/ 56.3	**29.7**/ 35.7	**31.9/ 39.2**	1.0/ 10.1	**36.6**/ 44.3
PANGU-BERT	47.9/ 60.2	47.0/ 54.8	39.5/ 61.8	25.2/ 48.0	13.5/ 37.6	19.5/ 27.3	35.2/ 50.5
+ Gold Iso	**49.2/ 61.6**	**57.4/ 65.2**	**41.1**/ 61.8	22.3/ 44.4	**24.1**/ 37.7	**25.3/ 31.6**	**39.2/ 53.2**

bution of these simple and complex isomorphism types in Table 2 in the Appendix.

Models: We employ two popular off-the-shelf KBQA models: RNG-KBQA (Ye et al., 2021) and PANGU (Gu et al., 2023). We choose the former due to its popularity and the latter due to its SOTA performance on GrailQA dataset [1]. We use the model's saved checkpoints for inference on GrailQA++, which serves as the baseline. The two models also highlight the different ways in which isomorphism information can be integrated.

For ranking-based models like RNG-KBQA, we filter out candidates during the enumeration phase whose logical form does not correspond with the gold isomorphism category. This effectively prunes the search space of candidates and thus reduces the overhead on the ranker.

For exploration-based models like PANGU, which iteratively builds up the logical form by searching the KB, we constrain the generation of the logical form by ensuring that exploration only considers paths that produce candidates corresponding to the given isomorphism category.

Metrics: We evaluate the performance of the models (in presence and absence of gold isomorphisms) in terms of EM (exact match) and F1 scores (between the predicted and gold answers).

3 Discussion

3.1 Present Findings:

Table 1 presents an overview of the results for the two baseline models on the GrailQA++ dataset in presence and absence of gold isomorphisms. While, we observe that adding isomorphism information improves the overall EM score by $\approx 10\%$ for both baselines (ALL column), the improvements are not consistent across different isomorphism types or even across models.

For example, we see that adding isomorphism information improves performance across all isomorphism categories except T-3 for the PANGU model. On the other hand, for RNG-KBQA, while adding isomorphism information brought about no performance gains for the simple isomorphism categories (T-0, T-1, and T-2), instances corresponding to T-3 observed the greatest jump in performance ($\approx 80\%$ relative gain) followed by T-4. A deeper dive reveals that the poor-performance of PANGU on T-3 results from the inability of systems to handle queries that involve superlative comparisons, e.g. "What war did the US lose the most soldiers?" Overall, we observe substantial improvements in performance on the GrailQA++ dataset by incorporating the isomorphism information.

3.2 Proposed Solutions:

A drawback of our exploratory approach is that it requires the gold isomorphism to be available during inference. We thus propose a few techniques to predict the isomorphism category. While this does require us to be aware of possible isomorphism classes during inference, we make no assumption about their distribution. We restrict ourselves to only those isomorphism classes that were seen during training. Some of our proposed solutions in this space are the following.

(i) Fine-tune language models like BERT (Devlin et al., 2019) or LLama (Touvron et al., 2023) for isomorphism prediction.

(ii) Learn representations of the KB schema and train a Graph Neural Network (GNN) like RGCN (Schlichtkrull et al., 2018) to predict the isomorphism category.

(iii) Perform data augmentation by sampling query graphs from the KB corresponding to the infrequent categories of isomorphism. The sampled query graphs, which consist of schema items present in the GrailQA training data, is then converted to natural language using LLMs (Agarwal et al.; Shu and Yu, 2024). This augmented data can then be used in addition to the original GrailQA training dataset to finetune LMs or train GNNs.

[1]https://dki-lab.github.io/GrailQA/

References

Dhruv Agarwal, Rajarshi Das, Sopan Khosla, and Rashmi Gangadharaiah. Bring your own kg: Self-supervised program synthesis for zero-shot kgqa. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Rajarshi Das, Ameya Godbole, Ankita Naik, Elliot Tower, Manzil Zaheer, Hannaneh Hajishirzi, Robin Jia, and Andrew Mccallum. 2022. Knowledge base question answering by case-based reasoning over subgraphs. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 4777–4793. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ritam Dutt, Sopan Khosla, Vinayshekhar Bannihatti Kumar, and Rashmi Gangadharaiah. 2023. GrailQA++: A challenging zero-shot benchmark for knowledge base question answering. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 897–909, Nusa Dua, Bali. Association for Computational Linguistics.

Yu Gu, Xiang Deng, and Yu Su. 2023. Don't generate, discriminate: A proposal for grounding language models to real-world environments. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4928–4949, Toronto, Canada. Association for Computational Linguistics.

Yu Gu, Sue Kase, Michelle Vanni, Brian Sadler, Percy Liang, Xifeng Yan, and Yu Su. 2021. Beyond iid: three levels of generalization for question answering on knowledge bases. In *Proceedings of the Web Conference 2021*, pages 3477–3488.

Yu Gu and Yu Su. 2022. Arcaneqa: Dynamic program induction and contextualized encoding for knowledge base question answering. *arXiv preprint arXiv:2204.08109*.

Chunhui Li, Yifan Wang, Zhen Wu, Zhen Yu, Fei Zhao, Shujian Huang, and Xinyu Dai. 2024. Multisql: A schema-integrated context-dependent text2sql dataset with diverse sql operations. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13857–13867.

Mingchen Li and Jonathan Shihao Ji. 2022. Semantic structure based query graph prediction for question answering over knowledge graph. *arXiv preprint arXiv:2204.10194*.

Ye Liu, Semih Yavuz, Rui Meng, Dragomir Radev, Caiming Xiong, and Yingbo Zhou. 2022. Uniparser: Unified semantic parser for question answering on knowledge base and database. *arXiv preprint arXiv:2211.05165*.

Haoran Luo, Zichen Tang, Shiyao Peng, Yikai Guo, Wentai Zhang, Chenghao Ma, Guanting Dong, Meina Song, Wei Lin, et al. 2023. Chatkbqa: A generate-then-retrieve framework for knowledge base question answering with fine-tuned large language models. *arXiv preprint arXiv:2310.08975*.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pages 593–607. Springer.

Yiheng Shu and Zhiwei Yu. 2024. Distribution shifts are bottlenecks: Extensive evaluation for grounding language models to knowledge bases. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 71–88, St. Julian's, Malta. Association for Computational Linguistics.

Yiheng Shu, Zhiwei Yu, Yuhan Li, Börje F Karlsson, Tingting Ma, Yuzhong Qu, and Chin-Yew Lin. 2022. Tiara: Multi-grained retrieval for robust question answering over large knowledge bases. *arXiv preprint arXiv:2210.12925*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I. Wang, Victor Zhong, Bailin Wang, Chengzu Li, Connor Boyle, Ansong Ni, Ziyu Yao, Dragomir Radev, Caiming Xiong, Lingpeng Kong, Rui Zhang, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2022. UnifiedSKG: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 602–631, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yao Xu, Shizhu He, Cunguang Wang, Li Cai, Kang Liu, and Jun Zhao. 2023. Query2Triple: Unified query encoding for answering diverse complex queries over knowledge graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11369–11382, Singapore. Association for Computational Linguistics.

Xi Ye, Semih Yavuz, Kazuma Hashimoto, Yingbo Zhou, and Caiming Xiong. 2021. Rng-kbqa: Generation augmented iterative ranking for knowledge base question answering. *arXiv preprint arXiv:2109.08678*.

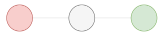
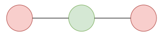
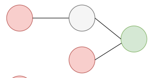
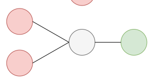
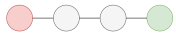
Donghan Yu, Sheng Zhang, Patrick Ng, Henghui Zhu, Alexander Hanbo Li, Jun Wang, Yiqun Hu, William Wang, Zhiguo Wang, and Bing Xiang. 2022. Decaf: Joint decoding of answers and logical forms for question answering over knowledge bases. *arXiv preprint arXiv:2210.00063*.

Zhehao Zhang, Xitao Li, Yan Gao, and Jian-Guang Lou. 2023. CRT-QA: A dataset of complex reasoning question answering over tabular data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2131–2153, Singapore. Association for Computational Linguistics.

Alex Zhuang, Ge Zhang, Tianyu Zheng, Xinrun Du, Junjie Wang, Weiming Ren, Stephen W Huang, Jie Fu, Xiang Yue, and Wenhu Chen. 2024. Structlm: Towards building generalist models for structured knowledge grounding. *arXiv preprint arXiv:2402.16671*.

A Appendix

Table 2: Distribution of different categories of isomorphisms in the GrailQA++ dataset.

Iso-code	Isomorphism	Count	Fraction
T-0		624	18.22
T-1		894	26.11
T-2		428	12.50
T-3		812	23.71
T-4		282	8.24
T-5		384	11.21