A Time-Sensitive Temporal Knowledge Editing Benchmark

Jae-Wook Lee[O], Ki-Nam Park[*]
Kora University, HIAI Research
{Jaewook133, spknn}@korea.ac.kr

Abstract

Temporal knowledge editing refers to the task of modifying a model's knowledge to reflect information that includes temporal aspects. Existing temporal knowledge editing benchmarks typically evaluate whether the knowledge editing techniques successfully update the model's knowledge and preserve related historical knowledge. However, they do not focus on the impact of temporal factors embedded within the knowledge on the editing process. In this study, we propose TTKEB, a benchmark for knowledge editing that incorporates various types of temporal reasoning. Unlike previous benchmarks, TTKEB highlights the influence of temporal factors on knowledge editing, providing new directions for future research on temporal knowledge editing involving temporal reasoning.

Keywords : Temporal knowledge editing, Benchmark

1. Introduction

Existing knowledge editing techniques primarily focus on injecting new knowledge, which often results in insufficient preservation of prior knowledge, especially when dealing with time-sensitive information such as historical events or accumulating data. This limitation highlights the need for "temporal knowledge editing," where related previous knowledge is maintained while new information is edited, and the necessity of benchmarks to evaluate this process.

Previous studies on temporal knowledge editing have attempted to optimize the model's predictions by simultaneously editing both past and new knowledge, balancing the acquisition of new information with the retention of prior data [1]. Other research has utilized temporal knowledge graphs (TKG) to perform multi-hop question answering involving temporal data [2]. These studies aim to satisfy both the injection of new knowledge and the retention of existing knowledge.

Existing works on editing knowledge with temporal information generally rely on fixed temporal expressions like \textit{"from (year) to (year)"}. However, temporal information in knowledge can be expressed in various ways, and these different representations can influence the range of time that can be inferred and the form of knowledge implied [3]. Therefore, datasets that use a fixed temporal representation do not reflect this diversity and are limited in their ability to infer information outside of a specified time range, making it difficult to make reasoning about change or continuity over time.

To address these issues, we propose TTKEB (Time-sensitive Temporal Knowledge Editing Benchmark), a benchmark designed to account for various types of temporal reasoning in knowledge editing. Based on the typs of temporal reasoning, TTKEB generates questions by modifying temporal expressions within the time ranges indicated by the knowledge possessed by the model.

2. Related Works

Knowledge editing aims to update the model with new information [4,5]. Existing methods include hypernetworks, meta-learning, and layer-specific parameter updates to selectively edit knowledge [6,7,8]. Other approaches modify outputs through external memory or in-context learning without updating parameters [9,10,11].

Temporal reasoning is key to human cognition, involving concepts like event order and duration, and is crucial for complex reasoning such as causality. There is growing interest in LLMs′ temporal reasoning abilities [12,13,14].

However, research on temporal knowledge editing in LLMs is limited. Current datasets do not present significant challenges for temporal reasoning, making detailed analysis difficult. To address this, we propose a benchmark that evaluates knowledge editing by incorporating diverse types of temporal reasoning.

3. TTKEB: Time-sensitive TKE Benchmark

In this study, we developed a benchmark to evaluate the impact of knowledge editing techniques when modifying knowledge involving various types of temporal reasoning. For this, we utilized passages, questions, and answers from the Time-Sensitive QA dataset [15], which is based on WikiText, to create data that allows for the assessment of existing knowledge editing methods.

3.1 Temporal Knowledge Correction

[*] corresponding author

Temporal Knowledge Correction is a scenario where the model′s existing temporal knowledge is updated with new information. This setting evaluates not only whether the knowledge editing technique successfully modified the model′s knowledge but also assesses the model′s question-answering performance regarding the modified temporal range after the correction.

The Temporal Knowledge Correction dataset consists of questions about specific entities, target answers related to the model’s stored knowledge, paraphrased versions of the original questions that preserve the core content, questions with modified temporal expressions based on different reasoning types within the original time range, and unrelated questions along with their respective answers that are not affected by the knowledge being edited.

4. Dataset Statistics

Table 1 shows the number of questions categorized by reasoning type included in the benchmark we developed.

Table 1. The number of questions by reasoning type in TTKEB

Reasoning Type	# of samples
in (explicit)	283
in (abstract)	265
after	275
before	270
between (explicit)	271
total	1419

4.1. Experiments settings

Model In this study, we utilized the following large language models for our experiments.

- **GPT-J (6B)**
- **LLaMA-2 (7B)**

Methods we utilize the following knowledge editing methods for out experiments.

- **FT-L** [16]: A method that directly fine-tunes a single layer’s feed-forward network (FFN).
- **ROME** [8]: A method that identifies the critical neuron activations responsible for the model's factual predictions and updates the feed-forward weights to modify the corresponding facts.
- **MEMIT** [9]: A method that performs a scalable multi-layer update by explicitly calculating parameter updates to insert new memory into the existing model.

Metrics In this study, we evaluate the performance of each method on the benchmark using the following metrics.

- **Efficacy** (E): Evaluates whether the targeted knowledge was correctly modified by the model.
- **Generality** (G): Assesses the model's ability to correctly answer generalized questions related to the modified knowledge.
- **Specificity** (S): Measures whether the knowledge editing technique has left unrelated knowledge unaffected.

Additionally, to assess the impact of knowledge editing techniques on the model's performance when modifying temporal knowledge, we define new metrics for each setting of the benchmark in this study.

Temporal Robustness In the Temporal Knowledge Correction setting, we evaluate whether the modified knowledge is accurately reflected in questions that have been converted into various temporal expressions within the time range indicated by the original question. To do this, each question is transformed based on different reasoning types, and the model’s ability to reflect the modified knowledge in these transformed questions is assessed.

4.2 Experiment Results

Table 2 shows the performance of models after modifying temporal knowledge. MEMIT achieves the highest score across both backbone models used in this study, demonstrating robust performance across various evaluation metrics. Similarly, ROME also efficiently modified temporal knowledge. These results indicate that the gradient-based strategy of identifying and editing knowledge neurons in the FFN layer allows the model to clearly recognize and modify time-sensitive knowledge. On the other hand, FT-L proved to be ineffective for knowledge editing, as it made minimal changes to the original model.

Table 2. Performance of different KE methods on TTKEB.

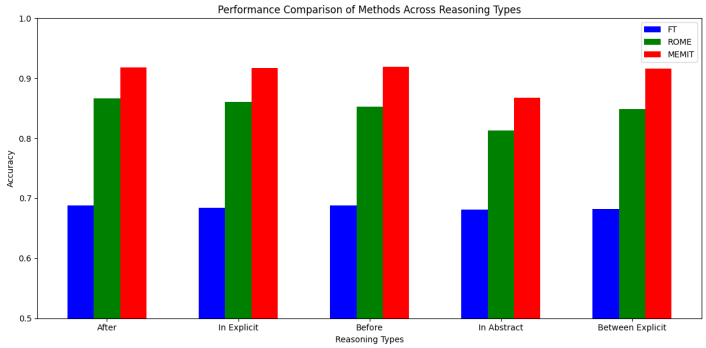
Method	Score	E	G	S	TR
GPT-J	0.4517	0.3795	0.3839	1.0000	0.3823
FT-L	0.4516	0.3795	0.3838	**0.9995**	0.3823
ROME	0.9426	0.9936	**0.9493**	0.8642	**0.9739**
MEMIT	**0.9434**	**0.9972**	0.8527	0.9942	0.9448
Llama-2	0.5393	0.4614	0.4736	1.0000	0.4677
FT-L	0.5394	0.4616	0.4736	**0.9999**	0.4679
ROME	0.9682	**0.9941**	0.9390	0.9786	0.9627
MEMIT	**0.9747**	0.9906	**0.9498**	0.9903	**0.9693**

4.3 Perfornace Analysis by Reasoning Type

Figure 1 illustrates the performance trends of knowledge editing techniques across different reasoning types. FT showed minimal

variation in performance across reasoning types, likely because this method did not make significant modifications to the model. Both ROME and MEMIT exhibited some performance differences across types, with the "In-Abstract" type showing lower performance compared to others. This is likely due to the abstract representation of temporal information in the questions, making it harder for the model to capture the time-related context compared to more explicit temporal expressions.

Figure 1. Performance of different Reasoning Types.



5. Conclusion

In this study, we extended existing temporal editing benchmarks, which use limited temporal expressions, by developing TTKEB, a benchmark that includes editing samples based on various types of temporal reasoning. Using TTKEB, we analyzed the performance of existing knowledge editing techniques when modifying temporal knowledge and examined how different reasoning types affect editing performance. Future work will focus on developing robust temporal editing techniques that are independent of specific reasoning types.

Acknowledgements

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI), This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425), Following are results of a study on the "Leaders in INdustry-university Cooperation 3.0" Project, supported by the Ministry of Education and National Research Foundation of Korea.

Reference

[1] Yin, Xunjian, et al. "History matters: Temporal knowledge editing in large language model." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 17. 2024.

[2] Cheng, Keyuan, et al. "Multi-hop question answering under temporal knowledge editing." arXiv preprint arXiv:2404.00492 (2024).

[3] Cai, Li, et al. "A Survey on Temporal Knowledge Graph: Representation Learning and Applications." arXiv preprint arXiv:2403.04782 (2024).

[4] Zhang, Ningyu, Yunzhi Yao, and Shumin Deng. "Knowledge Editing for Large Language Models." Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): Tutorial Summaries. 2024.

[5] Zhang, Ningyu, et al. "A comprehensive study of knowledge editing for large language models." arXiv preprint arXiv:2401.01286 (2024).

[6] De Cao, N., W. Aziz, and I. Titov. "Editing Factual Knowledge in Language Models." EMNLP 2021-2021 Conference on Empirical Methods in Natural Language Processing, Proceedings. 2021.

[7] Mitchell, Eric, et al. "Fast Model Editing at Scale." International Conference on Learning Representations, 2022.

[8] Meng, Kevin, et al. "Locating and editing factual associations in GPT." Advances in Neural Information Processing Systems 35 (2022): 17359-17372.

[9] Meng, Kevin, et al. "Mass Editing Memory in a Transformer." Proceedings of the International Conference on Learning Representations, 2023.

[10] Mitchell, Eric, et al. "Memory-based model editing at scale." International Conference on Machine Learning. PMLR, 2022.

[11] Hartvigsen, Tom, et al. "Aging with grace: Lifelong model editing with discrete key-value adaptors." Advances in Neural Information Processing Systems 36 (2024).

[12] Xia, Yuwei, et al. "Chain-of-History Reasoning for Temporal Knowledge Graph Forecasting." Findings of the Association for Computational Linguistics ACL 2024. 2024.

[13] Jiayang, Cheng, et al. "EventGround: Narrative Reasoning by Grounding to Eventuality-centric Knowledge Graphs." arXiv preprint arXiv:2404.00209 (2024).

[14] Chen, Jianhao, et al. "Timeline-based Sentence Decomposition with In-Context Learning for Temporal Fact Extraction." arXiv preprint arXiv:2405.10288 (2024).

[15] Mousavi, Seyed Mahed, Simone Alghisi, and Giuseppe Riccardi. "Is Your LLM Outdated? Benchmarking LLMs & Alignment Algorithms for Time-Sensitive Knowledge." arXiv preprint arXiv:2404.08700 (2024).