

WikiCausal: Corpus and Evaluation Framework for Causal Knowledge Graph Construction

Okkie Hassanzadeh

IBM Research

Abstract

Recently, there has been an increasing interest in the construction of general-domain and domain-specific causal knowledge graphs. Such knowledge graphs enable reasoning for causal analysis and event prediction, and so have a range of applications across different domains. While great progress has been made toward automated construction of causal knowledge graphs, the evaluation of such solutions has either focused on low-level tasks (e.g., cause-effect phrase extraction) or on ad hoc evaluation data and small manual evaluations. In this work, we present a corpus, task, and evaluation framework for causal knowledge graph construction. Our corpus consists of Wikipedia articles for a collection of event-related concepts in Wikidata. The task is to extract causal relations between event concepts from the corpus. The evaluation is performed in part using existing causal relations in Wikidata to measure recall, and in part using Large Language Models to avoid the need for manual or crowd-sourced evaluation. We evaluate a pipeline for causal knowledge graph construction that relies on neural models for question answering and concept linking, and show how the corpus and the evaluation framework allow us to effectively find the right model for each task.

Corpus: <https://doi.org/10.5281/zenodo.7897996>

Evaluation Framework: <https://github.com/IBM/wikicausal>

Keywords

Causal Knowledge, Knowledge Graph Construction, Knowledge Extraction from Text


1. Introduction


Extracting and representing causal knowledge has been a topic of extensive research, with applications in decision support and event forecasting in a variety of domains such as sociopolitical event forecasting [1, 2, 3, 4], enterprise risk management and finance [5, 6, 7], and healthcare [8, 9, 10]. One way to derive causal knowledge is by using observations in the form of structured data, and performing causal inference [11, 12]. An alternative is to extract causal knowledge stated explicitly or implicitly in text documents. Such statements are abundant across domains and applications in various forms, such as analyst reports, news articles, financial reports, medical documents, books, and scientific literature. As a result, there is a body of research on extracting causal knowledge from text documents with the goal of turning the knowledge into structured form for various retrieval, analysis, and reasoning tasks.

In this paper, we present a dataset and an evaluation framework for assessing the quality of causal knowledge graphs extracted automatically from text documents. To the best of our knowledge, this is the first evaluation framework that allows for measuring the quality of end-to-end causal extraction solutions. Our target solutions are those that take textual corpora

Posters, Demos, and Industry Tracks at ISWC 2024, November 13–15, 2024, Baltimore, USA

 hassanzadeh@us.ibm.com (O. Hassanzadeh)

 0000-0001-5307-9857 (O. Hassanzadeh)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

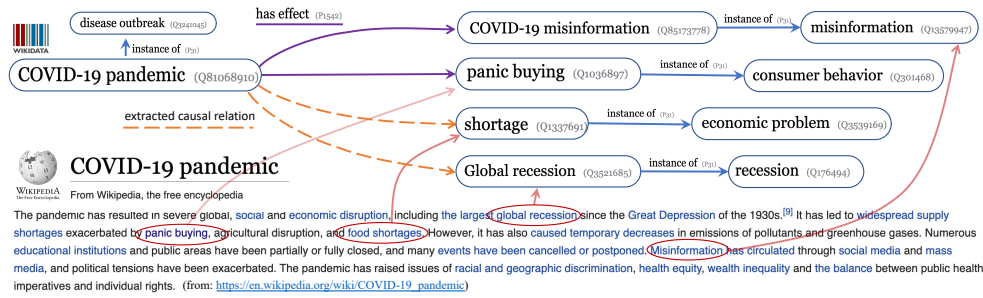


Figure 1: Examples of Event-Related Causal Knowledge in Wikidata and Wikipedia

as input, and produce a KG of causal relations among a set of concepts. Our dataset is curated from event-related Wikipedia articles. The evaluation of recall is performed by measuring the coverage of causal relations that are already in Wikidata, as the majority of such relations are described in text in the associated Wikipedia articles. For the evaluation of precision, inspired by a recent trend in the use of large language models (LLMs) in lieu of crowdsourcing [13, 14], we devise a mechanism for automatically creating prompts and probing LLMs to measure the accuracy of the cause-effect concept pairs in the output that is being evaluated. To show the effectiveness of the evaluation framework, we use a modular causal knowledge extraction pipeline to generate four versions of a Wikidata-based causal KG.

2. Task Definition and Use Cases

Our target task is as follows: given a corpus of text documents and a select set of concepts (e.g., from an existing KG), automatically generate a causal KG in which nodes are the given concepts, and an edge between two concepts indicates a causal relation between the concepts. The select concepts in the KGs could be either event-related classes, or instances of such classes. We assume that no annotations or training data are available. That is, while we know what concept each document is associated with, we do not have annotations of concepts or relations in the corpus. Figure 1 shows a snippet of a document from our Wikipedia-based corpus, along with a set of concepts from Wikidata. For this example, an application of the task defined above is to augment and/or validate the available causal knowledge. This case can also arise in applications such as healthcare or enterprise risk management, where part of the causal knowledge has already been captured in a structured form. Another use case for this task is construction of a domain-specific causal knowledge graph from a given corpus (e.g., analyst reports), with the goal of facilitating automated reasoning and planning [7, 15].

3. Corpus Creation

Given our task definition, we curate a collection of text documents, each associated with an event-related concept. We use Wikipedia as the source of our text documents and Wikidata as our source of event-related concepts. The first step in curating our corpus is identifying a set of event-related concepts in Wikidata. We do so by querying Wikidata for concepts that

have associated Wikinews articles. An associated Wikinews article implies that the article’s topic is on a newsworthy event instance. We then find the set of all the classes of the retrieved instances that are subclasses of class *occurrence* (*Q1190554*) to ensure that the chosen class is an event class as some non-event classes also have links to Wikinews. We then further manually verify each of the concepts and drop those that are not event-related. The next step is to retrieve all the instances of the identified event-related classes in Wikidata. We then use the Wikipedia “sitelinks” to collect the URL of all the associated English Wikipedia documents. We use the list of URLs over a dump of English Wikipedia to retrieve the associated Wikipedia articles, and process the contents of each article into plain text in addition to some meta-data about the page such as section headlines, categories, and infoboxes. We store the outcome in the form of a jsonl file, with each line being a JSON object containing the page contents, meta-data, and associated event concept(s). The first version of the dataset contains 68,391 articles, associated with a select set of 50 top-level event-related concepts in Wikidata.

4. Evaluation Framework

As with any automated knowledge graph construction task, we need to measure the quality of the output both in terms of the number of causal relations expressed in text that have been extracted (recall) and the number of extracted causal relations that are accurate (precision). Given that manually extracting all the expressed causal relations over the corpus is not feasible, our automated recall evaluation relies on existing causal relations in Wikidata. In the absence of a complete knowledge graph for a given corpus, the standard way to evaluate the precision of the extracted knowledge is manual evaluation. Manual evaluation, however, is tedious and time-consuming, which limits the possibility of experimenting on a large scale with a wide range of methods and parameters. Inspired by a recent trend in the use of large language models (LLMs) as an alternative to crowd-sourcing and manual annotation [13, 14, 16], we devise a mechanism to automatically create prompts for generative LLMs to evaluate the precision of the extracted causal relations. This approach works well for our corpus and task since LLMs have been exposed to the knowledge that is available on Wikipedia and Wikidata and are therefore likely to perform very well in the verification of the extracted relations.

5. Experiments & Results

We have used the corpus and our evaluation framework to evaluate a causal knowledge extraction pipeline that relies on the extraction of cause-effect phrases and linking the outcome to event concepts. As a part of our evaluation framework, in addition to the evaluation scripts and the corpus, we have made the extracted knowledge graph outputs publicly available: <https://github.com/IBM/wikicausal/tree/main/data/extracted-kg> as well as the results of our evaluation: <https://github.com/IBM/wikicausal/tree/main/results>. Our goal is to engage the community to extend the framework and perform a thorough evaluation of state-of-the-art KG extraction solutions, particularly those that rely on Retrieval Augmented Generation (RAG) [17]. Further details regarding the framework, results, and some interesting lessons learned can be found in the extended version of this work [18].

References

- [1] A. Hürriyetoglu, E. Yörük, O. Mutlu, F. Durusan, Ç. Yoltar, D. Yüret, B. Gürel, Cross-context news corpus for protest event-related knowledge base construction, *Data Intell.* 3 (2021) 308–335. URL: https://doi.org/10.1162/dint_a_00092. doi:10.1162/dint_a_00092.
- [2] F. Morstatter, A. Galstyan, G. Satyukov, D. Benjamin, A. Abeliuk, M. Mirtaheri, K. S. M. T. Hossain, P. A. Szekely, E. Ferrara, A. Matsui, M. Steyvers, S. Bennett, D. V. Budescu, M. Himmelstein, M. D. Ward, A. Beger, M. Catasta, R. Sosic, J. Leskovec, P. Atanasov, R. Joseph, R. Sethi, A. E. Abbas, SAGE: A hybrid geopolitical event forecasting system, in: S. Kraus (Ed.), *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, ijcai.org, 2019, pp. 6557–6559. URL: <https://doi.org/10.24963/ijcai.2019/955>. doi:10.24963/ijcai.2019/955.
- [3] S. Muthiah, et al., Embers at 4 years: Experiences operating an open source indicators forecasting system, in: *KDD*, 2016, pp. 205–214. doi:10.1145/2939672.2939709.
- [4] K. Radinsky, S. Davidovich, S. Markovitch, Learning causality for news events prediction, in: *WWW*, 2012, p. 909–918.
- [5] P. Bromiley, M. McShane, A. Nair, E. Rustambekov, Enterprise risk management: Review, critique, and research directions, *Long range planning* 48 (2015) 265–276.
- [6] F. Iwama, M. Enoki, S. Yoshihama, HOPE-Graph: A Hypothesis Evaluation Service considering News and Causality Knowledge, in: *2021 IEEE International Conference on Smart Data Services (SMDS)*, 2021, pp. 198–209. doi:10.1109/SMDS53860.2021.00034.
- [7] S. Sohrabi, M. Katz, O. Hassanzadeh, O. Udrea, M. D. Feblowitz, A. Riabov, IBM scenario planning advisor: Plan recognition as AI planning in practice, *AI Commun.* 32 (2019) 1–13. URL: <https://doi.org/10.3233/AIC-180602>.
- [8] R. Barnard-Mayers, E. Childs, L. Corlin, E. C. Caniglia, M. P. Fox, J. P. Donnelly, E. J. Murray, Assessing knowledge, attitudes, and practices towards causal directed acyclic graphs: A qualitative research project, *European Journal of Epidemiology* 36 (2021) 659–667. URL: <https://doi.org/10.1007/s10654-021-00771-3>.
- [9] M. Prospero, Y. Guo, M. Sperrin, J. S. Koopman, J. S. Min, X. He, S. Rich, M. Wang, I. E. Buchan, J. Bian, Causal inference and counterfactual prediction in machine learning for actionable healthcare, *Nature Machine Intelligence* 2 (2020) 369–375.
- [10] H. Q. Yu, S. Reiff-Marganiec, Learning disease causality knowledge from the web of health data, *International Journal on Semantic Web and Information Systems (IJSWIS)* 18 (2022) 1–19. URL: <https://doi.org/10.4018/IJSWIS.297145>. doi:10.4018/IJSWIS.297145.
- [11] J. Pearl, *Causality*, Cambridge University Press, 2009.
- [12] J. Pearl, Causal Inference, in: *Proceedings of Workshop on Causality: Objectives and Assessment at NIPS 2008*, PMLR, 2010, pp. 39–58. URL: <https://proceedings.mlr.press/v6/pearl10a.html>.
- [13] X. He, Z. Lin, Y. Gong, A.-L. Jin, H. Zhang, C. Lin, J. Jiao, S. M. Yiu, N. Duan, W. Chen, Annollm: Making large language models to be better crowdsourced annotators, 2023. arXiv:2303.16854.
- [14] M. Zhao, F. Mi, Y. Wang, M. Li, X. Jiang, Q. Liu, H. Schuetze, LMTurk: Few-shot learners as crowdsourcing workers in a language-model-as-a-service framework, in: *Findings of the Association for Computational Linguistics: NAACL 2022*, Association for Computational

- Linguistics, Seattle, United States, 2022, pp. 675–692. URL: <https://aclanthology.org/2022.findings-naacl.51>. doi:10.18653/v1/2022.findings-naacl.51.
- [15] O. Hassanzadeh, P. Awasthy, K. Barker, O. Bhardwaj, D. Bhattacharjya, M. Feblowitz, L. Martie, J. Ni, K. Srinivas, L. Yip, Knowledge-based news event analysis and forecasting toolkit, in: L. D. Raedt (Ed.), Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, International Joint Conferences on Artificial Intelligence Organization, 2022, pp. 5904–5907. URL: <https://doi.org/10.24963/ijcai.2022/850>. doi:10.24963/ijcai.2022/850, demo Track.
- [16] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, I. Stoica, Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. arXiv:2306.05685.
- [17] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, *Advances in Neural Information Processing Systems* 33 (2020) 9459–9474.
- [18] O. Hassanzadeh, WikiCausal: Corpus and evaluation framework for causal knowledge graph construction, 2024. Preprint available at <http://purl.org/wikicausalpaper>.