Building a Knowledge Graph of Events and Consequences Using Wikipedia and Wikidata

Oktie Hassanzadeh
hassanzadeh@us.ibm.com
IBM Research
Yorktown Heights, New York, USA

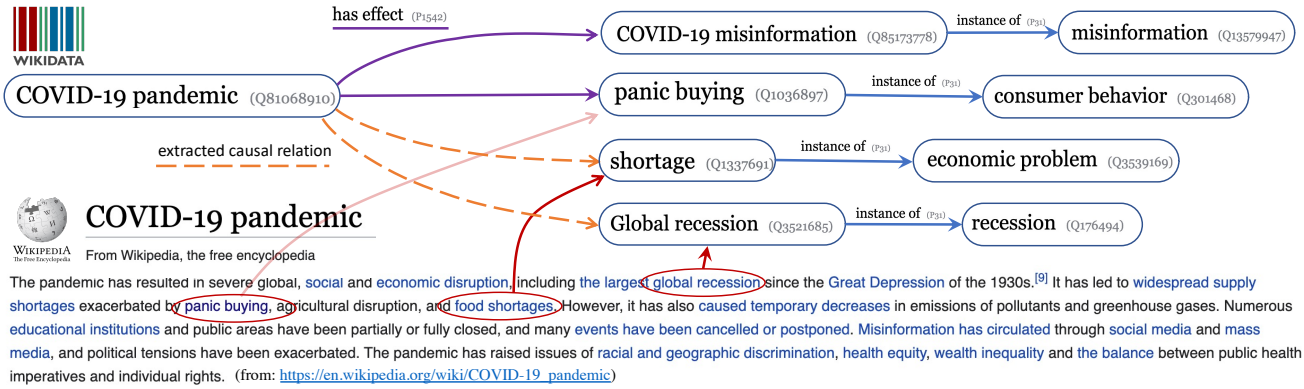

Figure 1: Examples of Event-Related Causal Knowledge in Wikidata and Wikipedia

ABSTRACT

In this short paper, we present our preliminary results on building a Knowledge Graph (KG) of events and consequences with application to event forecasting and analysis. A base KG is first constructed using existing concepts and relations in Wikidata. Using an automated unsupervised knowledge extraction pipeline, causal knowledge is extracted from Wikipedia articles to augment the base KG. We show examples from the base and the augmented KG, and discuss a few challenges in building a high-quality KG. We also discuss a few potential directions that the Wikimedia community can work on to improve the representation of event-related knowledge in Wikipedia and Wikidata.[1]

KEYWORDS

knowledge graphs, causal knowledge, knowledge extraction from text, natural language understanding, Wikipedia, Wikidata

[1]This paper has been presented at the Wikidata workshop at ISWC 2021.

Wiki Workshop 2022, April 25, 2022, Virtual

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Oktie Hassanzadeh. 2022. Building a Knowledge Graph of Events and Consequences Using Wikipedia and Wikidata. In *Proceedings of The Web Conference 2022 (Wiki Workshop 2022).* ACM, New York, NY, USA, 4 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

While prior work has considered knowledge-driven forecasting of future events [9, 10], curating large collections of causes and effects [4, 8], and event-based knowledge graphs (e.g., GDELT GKG [7]), there are no rich structured sources of knowledge around major societal events that can be queried directly to reason about the potential consequences of ongoing events. In this paper, we report on our initial results on curating such a source of knowledge from Wikidata and Wikipedia.

Wikipedia is a rich source of knowledge about major events and their consequences. Major newsworthy events often result in many additions and new pages describing various aspects of the events in detail. In particular, there are often descriptions of causes and effects of events, either explicitly in text, or implicitly in statements, sections, or descriptions of timelines of events. An effective representation of this knowledge in the form of a rich knowledge graph can enable a deep analysis of past events and their consequences. This can in turn be used as a mechanism of predicting the potential consequences of ongoing events by mapping them to past similar events in the knowledge graph.

Wikidata [12] aims at representing the rich knowledge available in Wikipedia in structured form. As shown in Figure 1, there are existing causal relations such as `has_cause` and `has_effect` between many event-related concepts. However, there are many explicit and implicit causal relations described in Wikipedia articles

that are missing from Wikidata. In what follows, we first show how we can turn the existing event-related concepts and causal relations in Wikidata into a base knowledge graph of events and consequences geared towards future event prediction and analysis. We then describe an unsupervised knowledge extraction pipeline that uses the textual descriptions of events in Wikipedia articles to augment the base knowledge graph. Using a few examples, we discuss the strengths and weaknesses of the approach. Finally, we discuss a few directions for future work.

2 KNOWLEDGE GRAPH OF EVENTS AND CONSEQUENCES

We first construct a base knowledge graph of events and consequences from existing concepts and links in Wikidata. Since our goal is analyzing major newsworthy events and their consequences, we only include in the base knowledge graph those event types that at least one of their instances have an existing link to a Wikinews article. This way, we ensure that out of the thousands of subclasses of type `occurrence (Q1190554)` and their instances, we only include events that are likely to receive news coverage. We then query for all the existing causal relations in Wikidata using properties such as `has effect (P1542)`, `contributing factor of (P1537)`, `immediate cause of (P1536)` and their inverse properties. We then group the event types that are linked directly or through their instances. The result is a collection of `event` objects that are event types (classes in Wikidata), each associated with a list of `consequences` which are also event types. Each consequence for an event has a list of `examples` with each example having a `cause` event instance and an `effect` event instance. Events and consequences are also annotated with a set of base scores derived from simple frequency analysis, e.g. the number of example pairs of instances, the number of triples for the event type and its instances, and the number of Wikipedia pages linked to instances of the type. The result is a collection of event types and their consequences, along with examples for each consequence, and scores that can be used for ranking of potential consequences for a given event.

Our current version of the base KG contains 50 source events (classes), 427 consequences, and 563 examples (instances). This output is a result of running 2,762 SPARQL queries to retrieve all the concepts and relations as well as their included properties and statistics. Figure 2 shows a few event types linked to `coup d'état (Q45382)`, instances (examples) for one of the consequences, and their JSON representations. This Wikidata-based representation of events and consequences not only enables retrieval of potential consequences for a given type of event, it also enables a deeper analysis of potential consequences using the rich structured knowledge around the Wikidata concepts. As a simple example, one can group potential consequences by geographic locations associated with the cause and effect events.

3 CAUSAL KNOWLEDGE EXTRACTION

As mentioned earlier, there are many causal relations expressed explicitly or implicitly in Wikipedia articles that cannot be found on Wikidata. We use an automated unsupervised causal knowledge extraction pipeline to augment the base KG using natural language understanding. The pipeline, shown in Figure 3, relies on pre-trained neural Question Answering (QA) and Entity Linking (EL) models. It consists of the following steps: a) A collection of causal questions are generated using a set of templates, such as "What could X cause?" or "What was a major consequence of X?" where X is a label of an event type or instance, b) a pre-trained neural QA model is used to find the answer from Wikipedia articles associated with the target event, and c) the answers are linked to Wikidata using pre-trained neural entity linking models based on BLINK [13].

At the time of this writing, we have applied the causal knowledge extraction pipeline only to the opening paragraphs of a collection of Wikipedia articles that describe instances of events that can be found in the base KG. Figure 4 shows examples of the extracted causes and consequences for the same `coup d'état (Q45382)` event used in the base KG example in Figure 2. Out of the six extracted consequences, one was also in the base KG (`conflict (Q180684)`), one is a superclass of of an event in the base KG `murder (Q132821)` which is a superclass of `political murder (Q1139665)` and the other four extractions were not in the base KG. The figure also shows an example for the discovered consequence `bomb attack (Q891854)`. The `examples` in the output KG from this pipeline also include a list of `mentions` that are answers from the QA model and come with a confidence score `answer_score`, and a `linking_score` that is the confidence score of linking the mention `text` to the Wikidata entity.

As the examples show, the pipeline is capable of extracting some very interesting causal relations that could not be found on Wikidata. We have found the overall quality of the output to be high even without any effort on tuning the models and parameters. One major quality issue in the current output is the wrong direction of the edges which is also evident in Figure 4. This is mainly a result of the question answering model returning cause instead of effect and vice versa. A potential solution is applying a custom classifier on top of the output, possibly by applying the outcome of our prior work on binary causal question answering [3] and using Natural Language Inference (NLI) for causal relation classification [1].

4 LESSONS LEARNED & FUTURE WORK

Our current results show a number of challenges, some of which could be addressed by the Wikidata community:

- One simple but classic problem we are facing in using event-related Wikidata entities is the inconsistency in `instance of (P31)` statements. One example as shown in Figure 4 is the event `death of Eduardo Frei Montalva (Q5247432)`, which at the time of this writing, is an instance of `murder (Q132821)`, `death (Q4)` and `certain aspects of a person's life (Q20127274)`, whereas the right class consistent with the base KG would be `political murder (Q1139665)` (which is a subclass of class `murder (Q132821)`).
- Some causal relations expressed in text cannot be represented using the existing entities and relations on Wikidata. For example, the Wikipedia article in the example in Figure 1 states that the pandemic has caused "temporary decreases in emissions of pollutants and greenhouse gases". There are currently no events or event types representing a decrease or

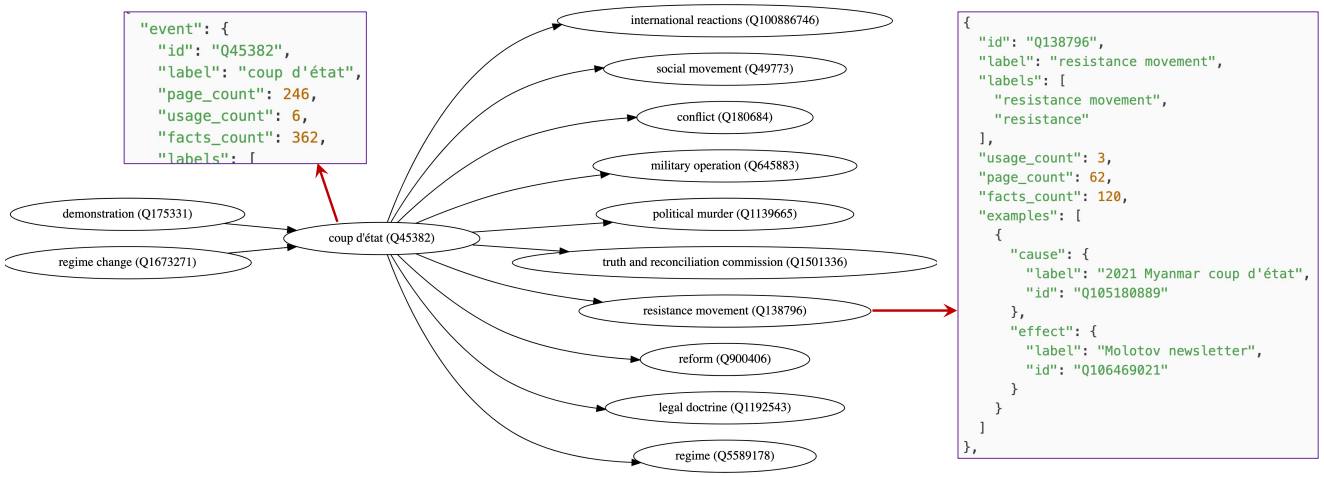

Figure 2: Example Events and Consequences from the Base KG

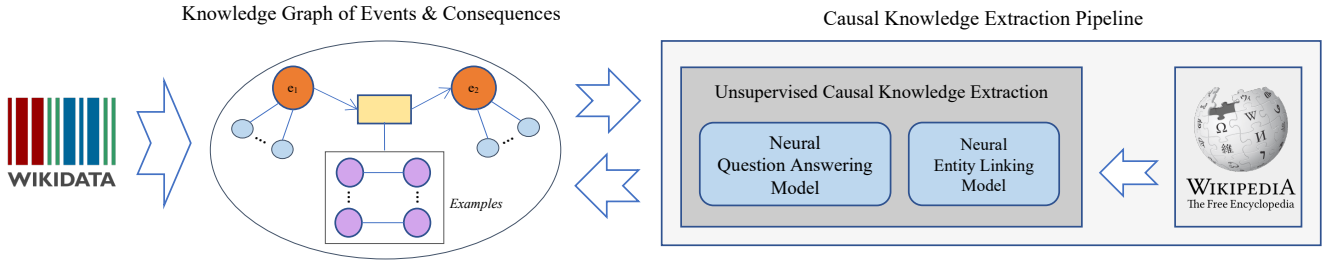

Figure 3: Causal Knowledge Extraction Pipeline

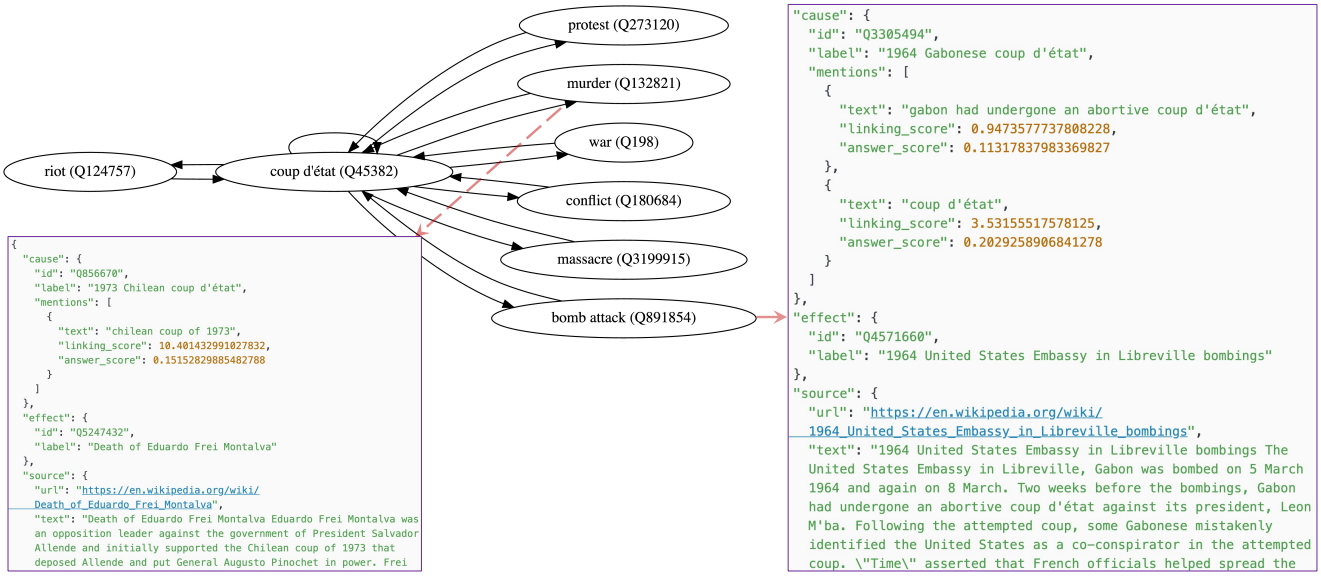

Figure 4: Example Augmentations by the Causal Knowledge Extraction Pipeline

a change in pollutants and greenhouse gases. One potential solution could be a “has effect on" relation that could link the pandemic concept to e.g. `carbon dioxide emissions (Q3588927)` along with attributes that could state whether the effect is temporary and whether it is a decrease or increase.

- Another direction that could have a significant effect on the community would be a tighter integration between Wikinews and Wikidata. For example, the authors of Wikinews articles can be encouraged to create related Wikidata items and specify causes of the events being described. On the Wikidata side, better representation of event classes and event-related concepts along with better alignment with Wikinews categories and sitelinks can provide the community with improved retrieval and news analysis capabilities.

We are currently working on improving our causal knowledge extraction pipeline in several ways, and performing a thorough evaluation of the quality of the extracted knowledge. A major challenge in using state-of-the-art causal relation extraction solutions [14] and benchmarks [5] is their focus on commonsense reasoning as the end application. One direction we are pursuing is publicly releasing our base KG along with a linked corpus of text from Wikipedia, that can be used as a benchmark for causal relation extraction and generic knowledge base completion solutions (e.g., IntKB [6]). We also plan to investigate the application of the knowledge graph in event forecasting [2] and enterprise risk management [11].

ACKNOWLEDGMENTS

This research is based upon work supported in part by U.S. DARPA KAIROS Program No. FA8750-19-C-0206. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

REFERENCES

[1] Manik Bhandari, Mark Feblowitz, Oktie Hassanzadeh, Kavitha Srinivas, and Shirin Sohrabi. 2021. Unsupervised Causal Knowledge Extraction from Text using Natural Language Inference (Student Abstract). In *AAAI*.

[2] Oktie Hassanzadeh. 2021. Predicting the Future with Wikidata and Wikipedia. In *Proceedings of the ISWC 2021 Posters & Demonstrations Tracks co-located with 20th International Semantic Web Conference (ISWC)*.

[3] Oktie Hassanzadeh, Debarun Bhattacharjya, Mark Feblowitz, Kavitha Srinivas, Michael Perrone, Shirin Sohrabi, and Michael Katz. 2019. Answering Binary Causal Questions Through Large-Scale Text Mining: An Evaluation Using Cause-Effect Pairs from Human Experts. In *IJCAI*.

[4] Stefan Heindorf, Yan Scholten, Henning Wachsmuth, Axel-Cyrille Ngonga Ngomo, and Martin Potthast. 2020. CauseNet: Towards a Causality Graph Extracted from the Web. In *CIKM*. ACM.

[5] Pedram Hosseini, David A Broniatowski, and Mona Diab. 2021. Predicting Directionality in Causal Relations in Text. *arXiv preprint arXiv:2103.13606* (2021).

[6] Bernhard Kratzwald, Guo Kunpeng, Stefan Feuerriegel, and Dennis Diefenbach. 2020. IntKB: A Verifiable Interactive Framework for Knowledge Base Completion. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*.

[7] Kalev Leetaru and Philip A Schrodt. 2013. GDELT: Global data on events, location, and tone, 1979–2012. In *ISA Annual Convention*.

[8] Zhiyi Luo, Yuchen Sha, Kenny Q. Zhu, Seung-won Hwang, and Zhongyuan Wang. 2016. Commonsense Causal Reasoning between Short Texts. In *Proceedings of the Fifteenth International Conference on Principles of Knowledge Representation and Reasoning (KR)*.

[9] K. Radinsky, S. Davidovich, and S. Markovitch. 2012. Learning causality for news events prediction. In *WWW*.

[10] K. Radinsky and E. Horvitz. 2013. Mining the web to predict future events. In *WSDM*.

[11] S. Sohrabi, M. Katz, O. Hassanzadeh, O. Udrea, and M. D. Feblowitz. 2018. IBM Scenario Planning Advisor: Plan Recognition as AI Planning in Practice. In *IJCAI*.

[12] Denny Vrandecic and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Commun. ACM* 57, 10 (2014), 78–85.

[13] Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Zero-shot Entity Linking with Dense Entity Retrieval. In *EMNLP*.

[14] Jie Yang, Soyeon Caren Han, and Josiah Poon. 2021. A Survey on Extraction of Causal Relations from Natural Language Text. *CoRR* abs/2101.06426 (2021).