Predicting the Future with Wikidata and Wikipedia

Oktie Hassanzadeh[0000−0001−5307−9857]

IBM Research
hassanzadeh@us.ibm.com

Abstract. In this demonstration, we present a prototype knowledge-based event forecasting system powered by Wikidata and Wikipedia. The system relies on existing event-related concepts and relations in Wikidata to build a base knowledge graph of events and consequences. It then uses a combination of state-of-the-art knowledge extraction methods to augment the base knowledge graph using natural language descriptions of events and their consequences that can be found in Wikipedia articles. Using a number of use case scenarios, we show how the resulting knowledge graph can be used as a part of a human-in-the-loop explainable solution for event forecasting and analysis.

1 Introduction

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. Figure 1 shows a few examples of such sources of *causal knowledge* around COVID-19 related 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 [10] 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. We will show in our demonstration how these existing links can be used for an analysis of potential effects of a given type of event. More importantly, we show how we can turn the existing links into a base knowledge graph of events and consequences, and then use the textual descriptions of events in Wikipedia articles to augment the base knowledge graph. In what follows, we describe the architecture of our prototype forecasting solution. We then present a brief sketch of our demonstration plan.

Distribution Statement "A" (Approved for Public Release, Distribution Unlimited).
Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-2980/paper322.pdf

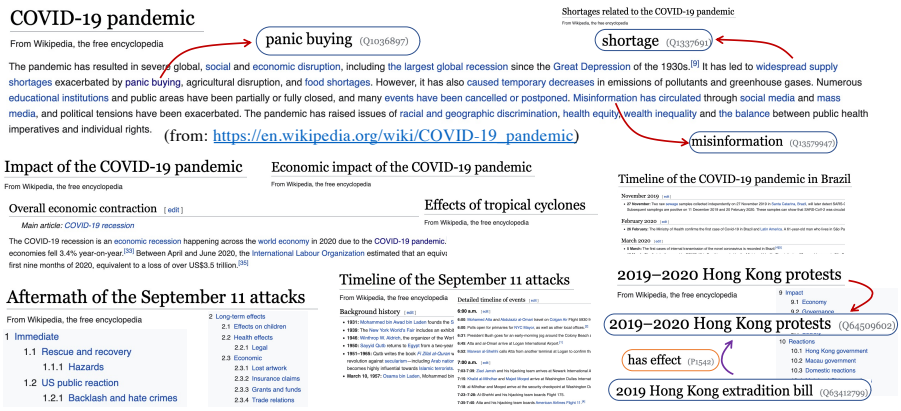

Fig. 1. Examples of Sources of Causal Knowledge in Wikipedia & Wikidata

2 Event Forecasting System

Figure 2 shows the overall architecture of our system. At the core of the system is a knowledge graph of events and consequences that is curated from existing event-related concepts and relations in Wikidata. The knowledge graph is then augmented with causal knowledge extracted from Wikipedia articles. The user interacts with the system through a dashboard that allows performing various analysis tasks over ongoing events and their potential consequences primarily through matching to similar events and event types in the knowledge graph.

Knowledge Graph of Events and Consequences A base knowledge graph is curated from existing concepts and links in Wikidata. Since our goal in this work is analysis of major newsworthy events, 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. Each link between event types is 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 cause-effect pair and scores that can be used for ranking of potential consequences for a given event.

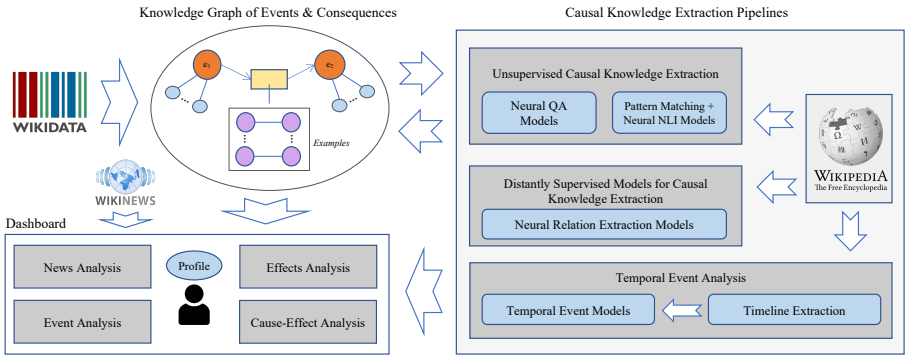

Fig. 2. System Architecture

Causal Knowledge Extraction Pipelines The base knowledge graph is augmented with causal knowledge extracted from Wikipedia articles using a number of causal knowledge extraction pipelines. This augmentation can be in the form of a) finding new consequences for a given event of interest, b) finding new example cause-effect pairs of instances for a pair of event types, and c) calculating scores reflecting the likelihood or significance for a causal relation between two events. Given the variety of ways that causal knowledge can be captured in Wikipedia documents as depicted in Figure 1, we need a number of different knowledge extraction approaches. In this demonstration, we show examples of three different kinds of pipelines we have implemented in our prototype:

Unsupervised Causal Knowledge Extraction: 1) An approach relying on pattern matching and neural Natural Language Inference (NLI) models. Briefly, the approach we show is an adaptation of the approach of Bhandari et al. [2] which is a fully unsupervised pipeline with a high precision of nearly 80% in manual evaluations. We link the output phrases to Wikidata concepts using BLINK [11], keeping only high-confidence links. 2) An approach relying on neural Question Answering (QA) models that a) generates questions 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) uses pre-trained neural QA models and Wikipedia articles associated with the target event to retrieve an answer for the generated questions, and c) performs entity linking to link the answer to Wikidata.

Supervised Models for Causal Knowledge Extraction: We use neural models [1, 4] trained on existing annotated data such as the BECauSE Corpus 2.0 [5] for extraction of cause-effect phrases from a corpus of event-related Wikipedia articles, and perform entity linking [6] on the phrases to map them to Wikipedia and then Wikidata concepts in the base knowledge graph. We are also exploring distantly supervised models [7] by constructing a training set through finding

passages containing labels of pairs of events in the base knowledge graph, and using a neural model of relation extraction [9] to extract new causal relations.

Temporal Event Analysis: This pipeline first extracts event timelines from timeline sections and pages (examples shown in Figure 1), then maps the extracted sequences to Wikidata events, and then applies existing and novel temporal event models [3] to the sequences of events that will facilitate more complex analysis of potential temporal and causal relations between event types along with likelihood scores that will better facilitate the ranking of potential consequences for a given event and context.

Dashboard The user dashboard exposes a number of API functions that use the knowledge graph to assist the user with event analysis and forecasting. The APIs allow the user to 1) retrieve the latest news events and their context, 2) retrieve a list of potential consequences for a given event/context, along with explanation in the form of example similar past events and consequences, and 3) rank and re-rank the consequences based on different criteria. The user can optionally define a profile that will be used for ranking the consequences based on how *interesting* or *surprising* the consequence could be for the user.

3 Demonstration Plan

We plan to use a number of use cases involving different recent or ongoing events, and show the ranked list of consequences according to the base knowledge graph as well as different versions of the knowledge graph based on the extraction method used for knowledge augmentation. For this initial prototype demonstration, our primary focus will be on showing the quality and coverage of different versions of the knowledge graph, and how simple major consequences of different types of events are present or missing in different versions. We will use examples of different types of events, including: 1) a "protest" event, e.g., recent protests in Myanmar, highlighting some of the high-ranked extracted causes and consequences from past protest events, which include a coup d'état; 2) a "disease outbreak" event as the cause while excluding COVID-19 related articles from our source, showing how some actual consequences of the COVID-19 outbreak show up in the ranked results of different pipelines; 3) a hypothetical natural disaster event and showing context-specific forecasts, e.g., similar to the work of Radinsky et al. [8] show how an `earthquake (Q7944)` at a location near an ocean would result in a forecast of `tsunami (Q8070)`. We will also highlight a number of challenging examples and wrong forecasts and discuss a number of directions for future work that could turn this simple prototype into a powerful and reliable AI assistant for analysts.

4 Acknowledgements

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. Awasthy, P., Ni, J., Barker, K., Florian, R.: IBM MNLP IE at CASE 2021 task 1: Multigranular and multilingual event detection on protest news. In: Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021). pp. 138–146 (2021)
2. Bhandari, M., Feblowitz, M., Hassanzadeh, O., Srinivas, K., Sohrabi, S.: Unsupervised causal knowledge extraction from text using natural language inference (student abstract). In: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021. pp. 15759–15760 (2021)
3. Bhattacharjya, D., Gao, T., Subramanian, D.: Order-dependent event models for agent interactions. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020. pp. 1977–1983 (2020)
4. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 8440–8451 (2020)
5. Dunietz, J., Levin, L.S., Carbonell, J.G.: The BECauSE corpus 2.0: Annotating causality and overlapping relations. In: Proceedings of the 11th Linguistic Annotation Workshop, LAW@EACL. pp. 95–104 (2017)
6. Li, B.Z., Min, S., Iyer, S., Mehdad, Y., Yih, W.: Efficient one-pass end-to-end entity linking for questions. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, 2020. pp. 6433–6441 (2020)
7. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP. pp. 1003–1011 (2009)
8. Radinsky, K., Davidovich, S., Markovitch, S.: Learning to predict from textual data. J. Artif. Intell. Res. **45**, 641–684 (2012)
9. Soares, L.B., FitzGerald, N., Ling, J., Kwiatkowski, T.: Matching the blanks: Distributional similarity for relation learning. In: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019. pp. 2895–2905 (2019)
10. Vrandecic, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Commun. ACM **57**(10), 78–85 (2014)
11. Wu, L., Petroni, F., Josifoski, M., Riedel, S., Zettlemoyer, L.: Scalable zero-shot entity linking with dense entity retrieval. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020. pp. 6397–6407 (2020)