Knowledge-Based News Event Analysis & Forecasting Toolkit

Oktie Hassanzadeh , Parul Awasthy , Ken Barker , Onkar Bhardwaj , Debarun Bhattacharjya , Mark Feblowitz , Lee Martie , Jian Ni , Kavitha Srinivas and **Lucy Yip**
IBM Research
hassanzadeh@us.ibm.com, awasthyp@us.ibm.com, kjbarker@us.ibm.com, onkarbhardwaj@ibm.com, debarunb@us.ibm.com, mfeb@us.ibm.com, Lee.Martie@ibm.com, nij@us.ibm.com, kavitha.srinivas@ibm.com, Lucy.Yip@ibm.com

Abstract

We present a toolkit for knowledge-based news event analysis and forecasting. The toolkit is powered by a Knowledge Graph (KG) of events curated from structured and unstructured sources of event-related knowledge. The toolkit provides functions for 1) mapping ongoing news headlines to concepts in the KG, 2) retrieval, reasoning, and visualization for causal analysis and forecasting, and 3) extraction of causal knowledge from text documents to augment the KG with additional domain knowledge. Each function has a number of implementations using state-of-the-art neuro-symbolic techniques. We show how the toolkit enables building a human-in-the-loop explainable solution for event analysis and forecasting.

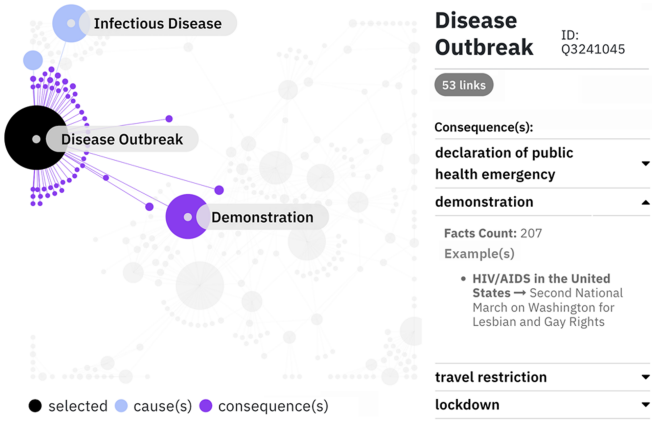

Figure 1: Visualization Function Showing A Portion of the Toolkit KG around Disease Outbreak Events

1 Introduction

Monitoring and analyzing ongoing news events and predicting their consequences has been a long-standing challenge in political science, intelligence, and finance communities [Elliott and Timmermann, 2016; Muthiah *et al.*, 2016; Sohrabi *et al.*, 2018]. Most prior work on event analysis with the goal of forecasting has been based on methods that require structured data (e.g., time series data, event databases) as input [Gmati *et al.*, 2019; Zhao, 2021]. In this demonstration, we present a toolkit for event analysis and forecasting that is primarily based on understanding natural language descriptions of events from text documents. Our goal is to demonstrate that a knowledge graph curated using generic knowledge extraction methods over event descriptions in publicly available text documents can enable a simple yet powerful AI agent for event analysis and forecasting. While our toolkit includes generic functions and can be used in applications ranging from scientific discovery to finance and risk management, our demonstration is focused on analyzing significant societal events such as disease outbreaks, natural disasters, and protests. Figure 1 depicts a portion of our toolkit KG.

2 News Event Analysis & Forecasting Toolkit

Figure 2 shows the components of our toolkit. At the core is a knowledge graph of events and consequences. The nodes in the KG are event types, which are connected to their potential consequences. Each edge is associated with evidence in the form of example instances of past events and their consequences, or text passages describing how the source event could result in the target event. Our toolkit comes with a rich KG of causal knowledge around significant societal events, curated from publicly available sources using state-of-the-art knowledge extraction methods. The KG can be augmented or replaced with domain-specific knowledge using the knowledge extraction methods that our toolkit provides.

The toolkit provides several functions for performing various analysis tasks over events that could be identified from ongoing and trending news headlines. The analysis is done primarily through matching to similar events in the KG. Our toolkit functions can be used to build an end-to-end event forecasting solution by monitoring ongoing news events, mapping them to past similar events, and reasoning about their potential consequences. They can also be used in a number of applications in knowledge-based decision support, risk management, and event analysis.

2.1 News Retrieval

Our toolkit provides interfaces for retrieving news headlines from third-party sources. Our current options include an interface to retrieve news from Wikinews and EventRegistry [Leban *et al.*, 2014]. Wikinews provides a source of

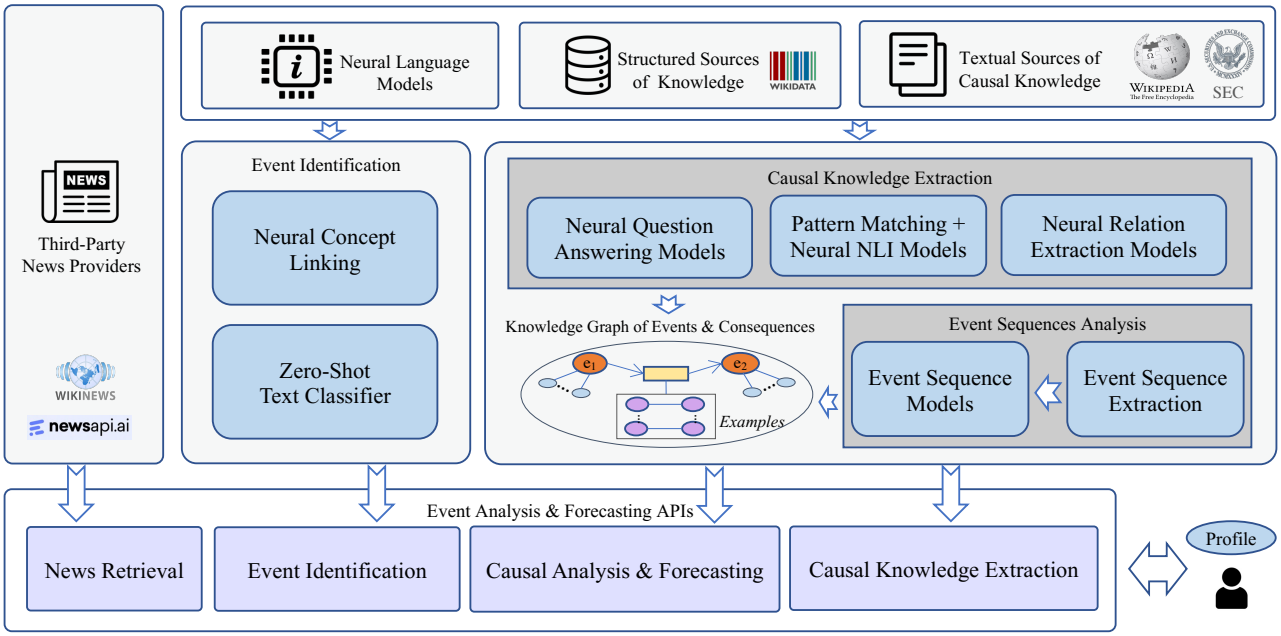

Figure 2: Toolkit Components and APIs

news that may not be as comprehensive and timely as other sources, but has no restriction for research and commercial use. It also comes with Wikimedia categories that makes integration with knowledge derived from other Wikimedia based sources easier. EventRegistry provides a high-quality source of text about ongoing news events along with rich meta-data, and also includes a free tier plan.

2.2 Event Identification

Event identification functions allow mapping a news headline to an event type in the KG. They also capture the context of the event, e.g. locations, people, or organizations involved in the event. The toolkit currently includes two implementations of this function, including a baseline method that relies on semantic parsing and concept linking, and a method using a zero-shot text classifier model. The baseline method performs semantic parsing to identify verbs and named entities and then links them to the KG using a neural concept linking model based on BLINK [Wu *et al.*, 2020]. Our zero-shot text classifier model is based the work of [Barker *et al.*, 2021] and is built around a Natural Language Inference (NLI) model. The NLI model scores the strength with which a news headline implies textual descriptions of each of the target event types, without requiring training data for the types.

2.3 Causal Analysis & Forecasting.

Given an event of interest, e.g. a news event from the output of the event identification functions, causal analysis functions allow for analysis of its potential causes and consequences. They also allow exploring the relationship between two or more events.

Analyzing Causes & Consequences. A primary mechanism of causal analysis that our toolkit enables is based on understanding causes and consequences of past similar events, as represented in the KG that is curated through knowledge extraction from text. Two functions `get_causes` and `get_consequences` take as input an event of interest (e.g., from the output of event identification functions) along with (optional) parameters that define the context of the event and the user profile, and provide as output a ranked list of causes and consequences. The ranking is based on one or more scores that are calculated from input event characteristics as well as the user profile. Our current implementation of these functions includes two scores: 1) **impact** as a measure of the potential impact of an event, and 2) **surprise** as a measure of how surprising the occurrence of the event might be. These scores are calculated using a combination of statistical analysis of past causes and consequences of events present in the KG, as well as reasoning about their context and the user profile (if available).

Sequence-Based Analysis. Another mechanism of understanding the causes of an event and its potential consequences is through an analysis of historical sequences of events involving the same type of event. This function of our toolkit is implemented by analyzing pre-built event models that are learned using sequences of event types extracted from textual descriptions of events. The temporal event models are based on a variation of the work of [Bhattacharjya *et al.*, 2018] around graphical event models, and allow for retrieval of a set of *influencers* for a given event type, i.e., those event types that most affect its probability of occurrence. Our current APIs include capabilities for identifying sets of event types that are amplifiers or inhibitors for a user-specifed event type, i.e. make it more or less likely to happen, respectively.

Visualization. We also provide a visualization function for the KG to support effective causal analysis. This function displays an interactive directed node-edge network graph, as

shown in Figure 1, with options for adjusting node and edge sizes based on impact and other frequency-based measures.

2.4 Causal Knowledge Extraction

As mentioned earlier, our toolkit comes with a rich KG curated from publicly available sources using state-of-the-art causal knowledge extraction methods. The extraction methods are also provided as functions in the toolkit to assist with augmenting or replacing the available KG using user-provided sources of knowledge in a particular domain. The 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 text documents, we need a number of different knowledge extraction approaches. Our toolkit currently includes three classes of knowledge extraction solutions.

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 [Bhandari *et al.*, 2021] which is a fully unsupervised pipeline with a high precision of nearly 80% in manual evaluations. We link the output phrases to KG concepts using our concept linker based on BLINK [Wu *et al.*, 2020], 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 pretrained neural QA models and articles associated with the target event to retrieve an answer for the generated questions, and *c)* links the answer to the KG.

Supervised Models for Causal Knowledge Extraction. In this pipeline, we formulate causal knowledge extraction as a sequence labeling problem: for an input sequence, each token is assigned one of the following labels: {B-Cause, I-Cause, B-Effect, I-Effect, O}, where "B" stands for "Beginning", "I" for "Inside", and "O" for "Outside". We apply a state-of-the-art Transformer-based sequence labeling model [Awasthy *et al.*, 2021] to extract causal phrases from a corpus of event-related Wikipedia articles. The model uses XLM-RoBERTa [Conneau *et al.*, 2020] as the input sequence encoder and is fine-tuned with the BECAUSE (Bank of Effects and Causes Stated Explicitly) dataset [Dunietz *et al.*, 2017]. We then link the output phrases to KG concepts.

Event Sequences Analysis. This pipeline first extracts event sequences from descriptions of sequences or timelines of events in text, then maps the extracted sequences to event types in the KG, and then applies temporal event models [Bhattacharjya *et al.*, 2018; Bhattacharjya *et al.*, 2020; Bhattacharjya *et al.*, 2022] 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.

2.5 Toolkit Knowledge Graph

We apply the above pipelines to publicly available textual sources of event-related knowledge to curate a KG of events and consequences that is used to power the causal analysis and forecasting functions in our toolkit. Currently, our primary source of knowledge is Wikipedia and Wikidata [Vrandecic and Krötzsch, 2014]. 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. Wikidata provides a structured representation of the majority of the events described in Wikipedia, often along with a rich collection of event-related facts.

The KG is constructed first by curating a base KG from existing concepts and links in Wikidata. For our current focus application, which is the analysis of major newsworthy events, we only include in the base KG those event types having at least one instance with an existing link to a Wikinews article. 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 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.

The KG is then augmented through knowledge extraction from Wikipedia articles using the functions described in Section 2.4. Our event sequence models are constructed from the timeline sections of event-related Wikipedia articles.

3 Demonstration Plan

We plan to showcase different functionalities of our toolkit using a Jupyter Notebook environment, which allows calling each of our backend APIs with custom parameters, and navigating through different portions of the KG through an embedded interactive visualization interface. We will use a number of recent or ongoing events at the time of the demonstration, and show an analysis of their potential consequences. We will use different kinds of events such as: 1) a major political event, e.g., the recent coup d'état in Myanmar; 2) a "disease outbreak" event, e.g., a recent Ebola outbreak in Guinea; 3) A natural disaster event, e.g., recent major earthquakes (and similar to the work of [Radinsky *et al.*, 2012] 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 noisy extractions and discuss a number of directions for future work that could turn this simple prototype into a powerful and reliable AI assistant for analysts.

References

[Awasthy *et al.*, 2021] Parul Awasthy, Jian Ni, Ken Barker, and Radu Florian. 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)*, pages 138–146, Online, August 2021. Association for Computational Linguistics.

[Barker *et al.*, 2021] Ken Barker, Parul Awasthy, Jian Ni, and Radu Florian. IBM MNLP IE at CASE 2021 task 2: NLI reranking for zero-shot text classification. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 193–202, Online, August 2021. Association for Computational Linguistics.

[Bhandari *et al.*, 2021] Manik Bhandari, Mark Feblowitz, Oktie Hassanzadeh, Kavitha Srinivas, and Shirin Sohrabi. Unsupervised causal knowledge extraction from text using natural language inference (student abstract). In *AAAI*, pages 15759–15760, 2021.

[Bhattacharjya *et al.*, 2018] Debarun Bhattacharjya, Dharmashankar Subramanian, and Tian Gao. Proximal graphical event models. In *NeurIPS*, pages 8147–8156, 2018.

[Bhattacharjya *et al.*, 2020] Debarun Bhattacharjya, Tian Gao, and Dharmashankar Subramanian. Order-dependent event models for agent interactions. In *IJCAI 2020*, pages 1977–1983, 2020.

[Bhattacharjya *et al.*, 2022] Debarun Bhattacharjya, Saurabh Sihag, Oktie Hassanzadeh, and Liza Bialik. Summary markov models for event sequences. In *IJCAI 2022*, 2022.

[Conneau *et al.*, 2020] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics.

[Dunietz *et al.*, 2017] Jesse Dunietz, Lori Levin, and Jaime Carbonell. The BECauSE corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 95–104, Valencia, Spain, April 2017. Association for Computational Linguistics.

[Elliott and Timmermann, 2016] Graham Elliott and Allan Timmermann. Forecasting in economics and finance. *Annual Review of Economics*, 8(1):81–110, 2016.

[Gmati *et al.*, 2019] Fatma Ezzahra Gmati, Salem Chakhar, Wided Lejouad Chaari, and Mark Xu. A taxonomy of event prediction methods. In Franz Wotawa, Gerhard Friedrich, Ingo Pill, Roxane Koitz-Hristov, and Moonis Ali, editors, *Advances and Trends in Artificial Intelligence. From Theory to Practice*, pages 12–26, Cham, 2019. Springer International Publishing.

[Leban *et al.*, 2014] G. Leban, B. Fortuna, J. Brank, and M. Grobelnik. Event Registry: Learning about world events from news. In *WWW*, 2014.

[Muthiah *et al.*, 2016] Sathappan Muthiah, Patrick Butler, Rupinder Paul Khandpur, Parang Saraf, Nathan Self, Alla Rozovskaya, Liang Zhao, Jose Cadena, Chang-Tien Lu, Anil Vullikanti, Achla Marathe, Kristen Summers, Graham Katz, Andy Doyle, Jaime Arredondo, Dipak K. Gupta, David Mares, and Naren Ramakrishnan. EMBERS at 4 years: Experiences operating an open source indicators forecasting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 205–214, New York, NY, USA, 2016. Association for Computing Machinery.

[Radinsky *et al.*, 2012] K. Radinsky, S. Davidovich, and S. Markovitch. Learning to predict from textual data. *J. Artif. Intell. Res.*, 45:641–684, 2012.

[Sohrabi *et al.*, 2018] S. Sohrabi, M. Katz, O. Hassanzadeh, O. Udrea, and M. D. Feblowitz. IBM scenario planning advisor: Plan recognition as AI planning in practice. In *IJCAI*, 2018.

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

[Wu *et al.*, 2020] Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. Zero-shot entity linking with dense entity retrieval. In *EMNLP*, 2020.

[Zhao, 2021] Liang Zhao. Event prediction in the big data era: A systematic survey. *ACM Comput. Surv.*, 54(5), May 2021.