"There will be consequences!"

Datasets and Benchmarks towards Causal Relation Extraction and Societal Event Forecasting

Gaurav Dass
Cognitive and Immersive Systems Lab
Rensselaer Polytechnic Institute
Troy, NY 12180
dassg2@rpi.edu

Abstract

In this paper, we discuss the creation of datasets (from which future benchmarks and other derived datasets can be created), geared towards causal relation extraction from a given context (obtained from the Google Natural Questions dataset) and event forecasting, especially that of making predictions of effects of societal events based on past iterations of similar events and more, obtained from Wikipedia and WikiNews. We introduce multiple novel datasets in this work:- 1.) Four named versions of benchmarks towards extracting verbatim phrasal causes and effects based on contextual than commonsense reasoning from Google’s Natural Questions dataset; two being denoted as the final set, called the Final Evaluation versions set of the Cause-Effect-Context from Natural Questions (NQ-CE) 2.) Three datasets of impacts and benchmarks from Wikipedia, so that, given some query event X, we try to answer the question, "what will follow?"; namely for the COVID-19, 9/11 and the Hurricane Sandy usecases, collectively referred to as the Dataset of Significant Societal Events and their Impacts from Wikipedia (SSIW) and two collections of significant societal events from Wikipedia year pages, called the Significant Societal Events from Wikipedia Year Pages (SSWY). 3.) One dataset of events from WikiNews, called the Event Consequence Collections from WikiNews (ECCW); extracting source events, their consequences and relevant and best matching semantically related non consequences for thirty-nine category pages in WikiNews after having filtered them using a combination of automated and manual filtering approaches from over 10k category pages. Future work, including evaluating benchmarks and creating datasets with more categories and types of events, along with creating derived benchmarks from any created datasets is planned.

1 Introduction

Causal reasoning is at the forefronts of research in artificial intelligence. Even though the idea of learning patterns from data to create an “intelligent system” has already gained a lot of attention and progress, it is thought that true intelligent agents of the future would utilize "causal reasoning" or be able to formulate understanding over chains of cause-consequence pairs to model real world phenomenon in some way or form. For building these maps of causality, many different kinds of cause effect extraction / causal relation extraction methods may exist and to this end, we hope to create

benchmarks and other supporting datasets which would cater to the causal relation extraction and
event forecasting systems of tomorrow, building not only on past ideas but also introducing new ones.
Hoping to curate, not only clean and accurate benchmarks but also datasets with diversity of domain
specific topical coverage, which can be used to train better machine learning models. To actually
realize these datasets we use methods ranging from pattern matching to question answering using
neural language models to even web data mining; utilizing sources such as WikiNews, Wikipedia
and Google Natural Questions dataset (derived from Wikipedia). So far we’ve already published
multiple datasets on zenodo, one concerning the extraction of cause and effect pairs from a given
context, two relating to mining significant societal events from Wikipedia and finally one containing
event consequence pairs from WikiNews.

2 Related Work

If we were to focus on named textual benchmarks and datasets in English which are prominent
towards causality (and solely on work where named datasets or corpuses have been released, rather
than just cause effect pairs or causal sentences extracted as training or testing data towards a novel
method for a causality related problem, e.g. causal relation extraction), through the lens of what is
relevant here i.e. style of curation, how causes and effects are discerned from text in these benchmarks
and the boundary of the causes and effects themselves, we find that there are benchmarks with word
based causal spans [Hendrickx et al. [2019]], benchmarks and datasets with phrase based causal
spans [Prasad et al. [2008] Hidey and McKeown [2016] Dunietz et al. [2017] Zhang et al. [2017],
those focusing on commonsense reasoning Mostafazadeh et al. [2020] Mostafazadeh et al. [2016a]
Mostafazadeh et al. [2016b] Roemmele et al. [2011] Ponti et al. [2020] and both commonsense and
contextual reasoning Rogers et al. [2020], those focusing only on explicit Dunietz et al. [2017] Mirza
et al. [2014] and implicit causality Mostafazadeh et al. [2020] Mostafazadeh et al. [2016a], on both
Prasad et al. [2008], capturing causality between events Caselli and Vossen [2017] Mostafazadeh et al.
[2016b] Mirza et al. [2014] Zhang et al. [2017] , capturing many other relations other than causality
Zhang et al. [2017] Hendrickx et al. [2019] and even some in the biomedical domain Pyysalo et al.
[2007] Gurulingappa et al. [2012]. Now, in trying to build benchmarks and datasets towards causal
relation extraction, one must always hold the benchmark itself against high standards, e.g. clean and
accurate sets of cause and effect pairs, otherwise one risks evaluating methods against a poor set of
criteria which would always lead to inconsistent analysis. Typically people leverage human curation
or even crowdsourcing to solve this problem (and create good benchmark). In fact in most of these
causality related datasets that we just discussed, domain expertise or some form of crowdsourcing is
used directly (albeit mixed with hybrid approaches, i.e. curating some automated results by human
beings), since that human labor cannot be and should not be avoided. In light of that, some other work
has done interesting by taking it one step further and utilising pre-existing and pre-curated
causal knowledge in already existing resources (thereby leveraging past curations). Hassanazadeh et
al. release 4 such datasets towards a new kind of causal reasoning task, called Binary Causal Question
Answering (BCQA) Hassanazadeh et al. [2019]: namely, SemEval Hassanazadeh et al. [2019] which
is nothing but causes and effects from Semeval 2010 task 8 [Hendrickx et al. [2019], an existing
resource to create a word based causality benchmark towards BCQA; NATO-SFA Hassanazadeh et
al. 2019] which is a set of causes and effects found in the document Strategic Foresight Analysis (SFA)
2017 Report Hassanazadeh et al. [2019], leveraging domain knowledge of past experts to obtain a
precured set of cause effect pairs; Risk Models Hassanazadeh et al. [2019], which is obtained from
domain oriented causal mindmaps created by human experts; and finally CE Pairs Hassanazadeh et
al. [2019] which again uses human curators to obtain more causes and effects from the risk models
benchmark by using some of those already in there as seeds to manually obtain more. Hassanazadeh
et al’s work was properly rooted in precision and expertise, i.e. the benchmarks they curated were
based on curations that were already plausible (e.g. SemEval) or based on strong domain expertise
(e.g. NatoSFA). Therefore the end result was a set of very clean and very accurate cause effect pairs.
Although point to note here, that such preexisting resources of good causal knowledge are hard to
come by, especially those vet by domain experts. In fact, not all preexisting sources are even good
ones. To look at this, if we were to consider knowledge bases as such a potential source (but there
can be many more as well e.g. authoritative source documents like the Strategic Foresight Analysis
Document from NATO Hassanazadeh et al. [2019] or other pre-existing datasets) they have their own
limitations, e.g. they might be too noisy Liu and Singh [2004] Vrandečić and Krötzsch [2014] and
even if they do have an automatically generated score associated with causal pairs, it does not help to
mitigate noise [Liu and Singh, 2004] or they could contain causal knowledge that might be limited to a certain domain [Mitchell et al., 2018]. None of these knowledge bases discussed just now seem to have any expert curation and one of them actually introduces noise due to cause affect pairs added by random users [Liu and Singh, 2004]. Hence, only curating datasets by non-experts can actually have potential to add noise to the set of causal pairs (than mitigate it) and that such crowdsourcing and human labor would be valuable only when actual human experts annotate the causal pairs or even if non-expert or crowdsourced non-expert curation is used, someone checks the curations for accuracy, as that ensures the goodness of the vetting. But from the above, we can see that if one wanted to build a good benchmark towards causality with limited human curation (or starting off with a rich set of causal pairs which have been curated by domain experts) then what is required here is not just the presence of causal knowledge but clean sources containing past or preexisting domain or human expert or validity checked and curated causal knowledge (just like SemEval and NatoSFA seen above in Hassanazadeh et al.’s work). That way, one can leverage past curation by domain experts or validity in past curations as a step in the current dataset creation process. This can be a giant leap to ensure quality of cause effect pairs. In terms of actual methods for performing cause effect extraction from text, although there exists a large variety of work, state of the art neural language models, one such instance of which is BERT [Devlin et al., 2018], (that is itself a language model that can learn contextual representations [Liu et al., 2019], e.g. the word counting in “I am counting on you” is treated differently than that in “I am counting sheep”) have been shown to have capability to extract cause effect spans from text after being fine tuned over the Stanford Question Answering Dataset (SQUAD) [Rajpurkar et al., 2016]. Here, cause and effect spans are extracted by asking BERT-SQUAD [Devlin et al., 2018] leading questions of the form “What are the effects of X?” given a passage, where X is some candidate cause, whose effects we’re looking for.

Towards event forecasting, Zhao [2021] has done a very thorough review on various event forecasting methods. From Zhao [2021] we can infer that event forecasting can be divided into 4 types, time prediction, location prediction, semantic prediction and multi-faceted prediction. As the name suggests, time and location prediction deal with predicting time and location of upcoming events; Zhao [2021], these seem to have a few datasets or collections of event datasets associated with them already e.g. GDELT [Leetaru and Schrodt, 2013] and and MITRE Gold Standard Report (GSR) [Ramakrishnan et al., 2014] to name a few (which are the most prominently named datasets). Although prominent, these datasets are either not openly available without explicit permission of the owners or due to their large scale nature (or other attributes such as requiring querying to even access small subsets) not freely available. The latter aspect, that of a large scale nature which requires querying to analyze or utilize is not a negative aspect and actually might present a boon and a strong necessity at other times (and has it’s own function in doing so), i.e. say when an overarching coverage of the events needs to be well detailed and from all over the world; but currently we’re only looking at this through the lens of problems which require datasets to immediately train models (or for other quick tasks like researchers needing to benchmark their models) and although such datasets and benchmarks could easily be curated from the queried subsets (in case of large scale data), we note that there should definitely be a freely (no requirement to seek permissions of owners) and easily (no need to query datasets or other additional measures to obtain the data) accessible datasets for this endeavor.

Multifaceted systems [Yang et al., 2002] on the other hand, seem to use a mix of domain or problem specific datasets [Li et al., 2017] and user collected [Yang et al., 2002] or other synthetic data [Brandt et al., 2011]. Semantic prediction systems [Zhao, 2021], specifically using causality based semantic prediction [Radinsky et al., 2012] deal with the “problem of forecasting topics, descriptions, or other meta-attributes in addition to future events’ times and locations” [Zhao, 2021] by using causality, i.e. future events are predicted by making causal inferences amongst past iterations of similar events or conversely, cause and consequence (effect) pairs of historical events are used to make predictions of effects for similar future events. This is what we focus on in our work. For this kind of a problem, previous work seems to use their own problem specific datasets (to benchmark their methods or for other training data), which authors curate from whatever holdout sets they segregated from their original data or whatever collection they performed, if any [Radinsky et al., 2012] [Zhao et al., 2017] [Dami et al., 2018] or domain specific datasets [Acharya et al., 2017]. From looking at past work we also note that there’s no named datasets (that have been released in the fashion we are doing now) towards event forecasting (in general) or causality based event semantics prediction directly from Wikipedia [Card, 2003], a collaborative and crowdsourced online encyclopedia, and WikiNews [Dube 2004] (which is quite rich in events, especially event chains and useful event metadata), a similar platform but for news.
Towards creating a benchmark for causal relation extraction, our work takes on from the above in that it deals not necessarily with explicit causality but implicit causal relationships between two phrases or words in a passage (also taking on from what was done before to utilize hand curated boundaries of cause and effect arguments than automated extractions of entities), regardless of the fact if there’s an explicit connector between them or not (e.g. the two phrases could be from different sentences in a given passage) with contextual causality, i.e. detection or extraction of cause and effect within the passage (with the passage being the context for the extraction) itself rather than some sort of a commonsense resolution task (e.g. if a given candidate causal pair is causal or not). Although, rather than curate from scratch, we scrape the crowdsourced curation already done by the domain experts who answered queries related to different Wikipedia Card [2003] pages in the Google Natural dataset [Kwiatkowski et al., 2019] and use that as our data source as it provides an (almost) "pre-curated" source of causal knowledge. Towards event forecasting, we create datasets relating to textual event data from Wikipedia that can be used for causality based forecasting, as past work shows that there’s definitely a lack in terms of such datasets for the given problem. We also create open source and easily accessible datasets of events and their consequences from WikiNews, (modeled using news articles and curated for temporal ordering and other attributes by human experts and further cleaned by us) which can be used to generate other datasets or benchmarks or can be used as a corpus for training and testing models towards an event forecasting task. In the following sections we describe creation process for each dataset, current progress and future work to be done.

3 Benchmark Curation towards Causal Relation Extraction

Collections of very clean and accurate cause effect pairs curated by domain experts are hard to come by and as discussed in section 2, curations by non experts can actually introduce noise than mitigate it. Liu and Singh [2004]. To create a good collection independently, one would need to pull in domain experts, but their time and availability is typically limited. Hence in our work, we did something different, we consider the Google Natural Questions (NQ) dataset [Kwiatkowski et al., 2019], which consists of a question and a response to the question found in Wikipedia Card [2003], annotated by domain experts (who annotate long and short answers in articles). This is because in answering queries within Wikipedia, we theorized that if domain experts were to ever have answered open questions (here open means that one of cause or effect is missing) of the form, "What would X result in?" or "What is the cause of X?" and marked short answers within Wikipedia, they would essentially be curating causal knowledge for us within Wikipedia and the task for us would simply be finding those causal questions. Now, there might be a margin of error in finding those causal questions but even then the causal knowledge present within the NQ dataset (in the fashion described above, since they are curated by domain experts) is not affected by errors in the process of obtaining causal questions and worst case, one engages in some form of limited manual cleaning. We might end up missing some causal pairs (depending on how we do our causal question extraction) but what we do obtain in the end is clean and accurate. In doing this, we leverage past crowdsourcing which was already performed and make it part of our own curation process. Using these cause effect pairs we create a benchmark for causal reasoning where the task is that of phrasal extraction of causes and effects from a given text, whether they might be explicit or implicit, followed by matching the extractions against annotated causes and effects such that the extraction is not necessarily based on world knowledge or commonsense reasoning but contextual information present solely within the text that signals causality. In the next section we discuss the creation and curation process for such a benchmark.

3.1 Cause-Effect-Context from Natural Questions (NQ-CE)

To obtain the causal knowledge within the expert curated NQ dataset, we perform causal relation extraction on the questions of the NQ dataset. Then we find causal questions which start with what or would, followed by finding only those questions who have short responses in the Wikipedia text, discarding all the questions which don’t have short responses at all, e.g. only yes/no response or are noisy responses e.g. Wikipedia tables having been picked up in responses. This was done so because to truly leverage the curation done by domain experts in NQ dataset, we wish to find short responses which could either be causes or effects for certain arguments in the query. We do causal relation extraction as seen in Hassanzadeh et al. [2019] by looking for the presence of certain causal signals (girju’s list [Girju et al., 2002] of such signals and their derived forms) in questions and then checking
which ones are forward or backward, i.e. are cause-signal-effect or effect-signal-cause; here cause
and effect are the parts of the sentences to the left and the right of the causal signal found. Depending
on the order, we populate a new set of fields, cause and effect which are nothing but the short query
and the part of the query which has been found to be cause or effect. All of these fields, along with
short responses and the long response containing the short response (called here as the passage),
along with the URL and question text are present per query in a JSONlines file, with each line being a
query and relevant fields attached to it. Then we clean all queries by hand and after multiple iterations
of hand cleaning we’re able to obtain a set of 106 causal queries which are near perfect in terms of
having phrasal causes and effects obtained as present verbatim within the context passage (e.g. if the
cause is "military coup" and the effect is "The Nigerian Civil War" then they would be present exactly
as is in the corpus) and require reading the passage to answer and need not always be resolved
through world knowledge, e.g. in a causal pair with the cause being "Letters of Marque" and effect
being "plundering of Spanish ships", such a causal pair cannot be resolved through world knowledge
but a historical one, learnt only after reading the passage. In fact to enforce this verbatim nature,
causes and effects can even carry the same "typos", capitalizations and other syntactic cues as in the
original text, e.g. for a cause, "Economic policy", the effect is "Shays ' Rebellion", the text "Shays '
Rebellion" is present as is in the passage and is left untouched (in hopes that an automated extraction
algorithm which extracts a correct span of text does not fail to make even any simple matches due to
typecase). Sometimes it was found that there were repeating questions in the NQ dataset which had
the same answers but were duplicates simply because they were phrased differently, e.g. "what factors
led to the war of 1812 quizlet?", "what were the 4 causes of the war of 1812?" and "what factors led
to the war with britain", these were all merged into one in the final version. In the 106 causal queries
obtained, we sample random short phrases from the contextual passage and randomly replace either
the cause or the effect, these form our non causal queries. Finally we put them together and compose
them into a single set of 212 causal and non causal queries. In the 106 causal queries only version,
we have 166 causal pairs (as some queries can have multiple causes or effects) and in the 212 version,
297 candidate causal pairs. We release the final curated set which is in fact the third iteration of such
a dataset, called the Final Evaluation versions set due to having verbatim causes and effects which
makes matching predictions to true annotations when evaluating a model. Each iteration has two
versions, all versions of family 1 (1, 1.5 and 1.5b) contain only causal queries, while those that of
family 2 (2, 2.5 and 2.5b) contain both causal and non causal queries. Older versions may not have
verbatim consistency enforced (instead have manual annotations based on human understanding) and
hence may have some slightly different queries as they are simply semi-intermediate towards the final
benchmarks seen in the Final Evaluation versions set (versions 1.5b and 2.5b). We release along with
this paper, the Final Evaluation versions set (and older ones are accessible as well, although people
looking to use this benchmark are encouraged to use the final evaluation versions as that has been
made targeting models which extract span based cause and effect phrases from the contextual text).

4 Dataset and Benchmark Curation for Significant Societal Events from
Wikipedia

In wanting to build a benchmark or even datasets for this problem, one would seek to collect significant
events, compose them into cause effect pairs, or cause consequence pairs (for the independent events
and their impacts) and then finally find some way to generalize them (and later on being able to
match them to query events and then extrapolate to impacts based on past most similar event found
in the collection, although this step might not be required for a benchmark). Hence the first step is
to obviously collect significant societal events. To do that, a clean data source (or one that could
be reasonably hand or automatically cleaned) is sought of significant societal events that contained
human annotations or were originally curated by human beings (to decide the impacts of the event
or even in some cases decide significance). That is, the impacts of selected significant events or
some contribution to the significance of new events have been done through the sensemaking and
organization of information done by experts or people reasonably knowledgeable about the domain.
Wikipedia seemed to fit these criteria (similar to the criteria for data sources as seen in section 3.1) and
hence we utilised it for this purpose. In doing so, we sought to build datasets to address the following
problems and tasks listed here:- 1) Creation of a dataset of impacts for significant societal events
so that, given some query event X, we try to answer the question, "what will follow?", i.e. given
some Y event in the past, and a similar X event happening in present, what can we infer about X
by information we have on Y? 2) Creation of a benchmark to see how well a method performs
in obtaining impacts of a societal event, given a Wikipedia page about the same, i.e. how well
does any method benchmark in collecting impacts of that event (as per the Wikipedia page) and
finally 3) Building a dataset of societal events from Wikipedia year pages, whose significance is
born of being selected by Wikipedia editors (that they are placed in there) and manual curation done
later on by us. The following sections describe each problem in order.

4.1 Dataset of Significant Societal Events and their Impacts from Wikipedia (SSIW)

To put together a collection of the impacts of significant societal events, we selected certain categories
(i.e. the classes of past events Y) of events (to mine for) manually and then looked at Wikipedia
pages for different instances of the same. The following classes of events have been considered so
far, Disease Outbreaks, Terrorist attacks and Natural Disasters; with the instances of them being,
COVID-19, 9/11 and Hurricane Sandy respectively. On looking at their Wikipedia pages, we found
that the impacts were always listed at the bottom of the page, in form of a footer table (whose title
was always the event name or the name of the Wikipedia page) in a row having the Impacts row
name, sometimes the Impacts row was itself also a link to a Wikipedia article. Other times it might
even be called something else. Entries in this row typically featured other Wikipedia articles which
were impacts of this event. The names and links for all these pages were extracted and stored in a
JSONlines file. These links (for Wikipedia pages) are then curated by hand and then used to obtain the
Wikipedia text for each page, this is stored as another JSONlines file. BERT-SQUAD is used to obtain
impacts from every Wikipedia page by repeatedly forming and asking questions of the form, "What
are the impacts of page-name?". Where page-name is the name of the main Wikipedia page (e.g.
COVID-19) being processed. Top 200 responses are saved per impact page. It was found that these
responses were quite repetitive in nature, due to which some processing would be required before
they can be manually vetted. Hence they are first preprocessed to remove substrings (strings which
are completely present in another string) and deduplications (strict deduplications, i.e. complete and
perfect matches of same sized strings) amongst response text. To account for string matches which
had overlaps but not perfect duplicates or substrings all responses are clustered, for every cluster,
the longest cluster element was selected as the response to be saved. It was found that there were
still candidates which remained to be filtered hence on this set, hence overlap merging was done, i.e.
overlaps are detected amongst responses and longest merge amongst all overlaps found, this is done
till convergence so that length of the resultant response list remains the same. It was seen that certain
responses were still colliding since their provenances had an overlap but the responses did not, hence
the final filtering step comprised of finding provenances and then merging responses on basis of same
provenances (end response chosen from one group having same provenance would be that having
longest length). The provenance here means the original sentence or span of text from where the
response was picked by the BERT-SQUAD model. These extractions are saved as JSONlines and then
manual vetting done on them for cleaning and correction purposes. Each line contains the a) corrected
text, b) URL, c) original extraction (that it was corrected from) d) provenance and a probability field
which (is an extraction artifact) and can be ignored. While manually vetting, some of the extractions
are not made verbatim to the provenance present but rather based on how a human would understand
the impact (or the annotator paraphrases the same), e.g. the impact "closure of commercial airports"
came from provenance, "This was the first total closure of a United States commercial airport for
demand-related reasons.". This was an instance where the extractions were made generic, i.e. US
commercial airports became commercial airports, i.e. generalizability was enforced where ever
possible. In some cases though the specificity of the impacts were maintained, e.g. "disc golf saw
an increase in participation". Other times significant changes were made to the impacts in case it
was found that the provenance was valid but the extraction not so much, e.g. "censorship in media
coverage in some countries" vs original one of "2019-20 coronavirus pandemic has varied". Due
to imperfections in finding provenance sometimes there can be more than one provenance, in these
cases the correct impact is still listed. Using this method, we obtained 249 impacts for the 9/11, 49
for Hurricane Sandy and 768 for the COVID-19 usecase. Since, this dataset is hand curated, hence
beyond being used as collection of cause effect pairs to do lookups for similar current events, it can
also be used as a benchmark to see how well methods perform in extracting impacts from textual
description of societal events, (in this case the text being from the Wikipedia pages for the events).
4.2 Significant Societal Events from Wikipedia Year Pages (SSWY)

As per section 4.1 we pick certain classes of events and then choose events to do extraction for from Wikipedia. But to build a more varied collection of events with a greater number of impacts so that anyone looking for impacts of similar societal events obtains a comprehensive set of results with more diversity than them being restricted to only single instances (only having one instance of a given class, e.g. only having COVID-19 for disease outbreaks) or subtypes (having only one kind of a class, e.g. having only hurricanes in natural disasters), hand picked categories and classes would be relevant (and although succinctly picked categories like these are useful in when there’s something specific that needs to be done) but not suffice to a complete solution. To resolve this issue, some sort of source of bulk human curated significant events have to be picked out. Towards this end, Wikipedia was looked at and it was found that there were certain Year pages which contained lists of events that happened in a given calendar year as per the human editors of Wikipedia. Since they were picked by human beings, one can say they are at least more significant and a better starting point than just a million point corpus of news articles to process and curate. Combined with manual hand curation we do, significant events can be picked out more numerable than those that were selected in 4.1.

Now to actually do this, one would need to mass mine the Wikipedia year pages specifically only or mostly for instances of real world occurrences or wikipages of such as opposed to generic wikipages e.g. a Wikipedia page about a specific event like the 2021 Apure clashes vs. the Wikipedia page for Vegetables. The problem here being how to filter out event links or urls from every link on the page. Obviously, anything that is within the bulleted list of the Year page is a candidate, but due to Wikipedia being reference heavy, there are multiple links in almost every sentence. So further inspection was done, and it was found that event pages differed from generic ones in two ways: a) sometimes whenever the sentence right after the date starts with a link, it is an event page, e.g. March 21 – Clashes in Apure, where Clashes in Apure is a hyperlink to the page and more importantly b) since the event is expected to be some sort of particular occurrence at a given time, hence it will be more likely to have the same date inside it’s linked article textual content as listed right next to it’s listing in the year page. Any stray cases or uncertainties that arise outside of these two rules (there can be more) can be mitigated by limited hand curation. While hand curating datasets in this case, we omit entities and organizations that the year pages link to (even if these organizations contain an instance of an "event" in their wikipage, e.g. concert cancellations, postponements, rise in COVID cases which contained many duplicate events etc.) and long drawn events which have been happening for a while and have had minor hallmarks in a year (unless they present some form of significance). Sports and Entertainment events have also not been considered. Although in some select cases, entities have been considered (e.g. someone starting a presidency or a notable in relation to Thomas Cook etc. ). Each event is stored as a JSON object with the following fields: a) text (text within the year page that is linked to the Wikipedia page), b) url, c) event_title (title of the Wikipedia event page), d) date and e) categories (of the Wikipedia page). The text is the We did this for Wikipedia pages for the years 2019 and 2020 and obtained 131 significant societal events for 2019 and 79 for the year 2020. They can be useful towards creation of an independent corpus of significance, perhaps some future work on the definition of significance and of course for obtaining impacts as in section 4.1 amongst many other things (including more year pages).

5 Dataset Curation towards Significant Societal Events from WikiNews

In attempting to design event forecasting tasks for news events, first one obviously needs a clean source of events curated by humans and second, capability to collect various causal (and difficult to discern and semantically similar non causal) information about the news events from that source of data. The latter is so because one would want capability of having both things which could be caused by events and also negative examples which are events which are easy for a human being to discern as being non causal but harder for a machine to do so. Such a source of news events (or organized articles of such events) is WikiNews, an online collaborative and crowdsourced platform (much like Wikipedia) for news, written by anyone and run (and moderated) by citizen journalists all over the world. We focus solely on the English WikiNews for this work. News articles in WikiNews are organized categories, which might themselves be organized further into subcategories (or categories which are contained in other categories). Every category page is a collection of articles and if present, other subcategories (they might have an overlap of articles with original categories). Category pages can be topics (e.g. Transport), entities (e.g. Europe), dates (e.g. May 21st 2021), events (e.g. 2011 England riots) or even many other things (e.g. interviews, original reporting etc.).
To create a collection of significant societal events from WikiNews, we needed to understand what significance could mean and how we pick the societal events from WikiNews, for this we needed an analysis which would require some criteria. With respect to WikiNews, we defined three such criteria for significance:- a) Diversity of events and b) Significance on basis of some numerical criteria within WikiNews c) Human significance (whatever we deem as important on human examination).

From this, we derive an intermediary dataset which can be used as a training (and testing) corpus or as a means to create other datasets and benchmarks. The next section discusses the collection and curation process for such a dataset.

5.1 Event Consequence Collection from WikiNews (ECCW)

To perform an initial analysis on WikiNews, we first manually picked events from different categories (protests, disease outbreaks, natural disasters, terrorist attacks, political events and military conflicts) and then did an initial manual collection of event consequence (EC) pairs. Here events and consequences were articles in WikiNews. We found 83 EC pairs in this manner and this exercise gave us a better idea on how to scope our automated collection. We saw that EC pairs could broadly be classified as micro (occurring over short temporal spans, i.e. smaller sub-events in a larger event, e.g. "US military to carry out review following Wikileaks release of classified 2007 video leads to US intelligence analyst arrested over Wikileaks video") or macro (occurring over longer temporal spans, i.e. can be deemed as impacts of an event many years after it has occurred, e.g. "Iraq War leads to Thousands protesting Iraq war in Washington D.C."); beyond these we also found events as belonging to a traceable chain (e.g. "Bush unveils America’s new Iraq plan" leads to "Republican Senators oppose more troops in Iraq" which leads to "Protesters demonstrate at US Coast Guard Academy" and so on) and also being small independent micro unchainable events which don’t necessarily fit into such chains, e.g. "Leaked cables from wikileaks leads to Standard Operating Procedure changes at Camp Delta, Guantanamo Bay." On this basis, we decided that any event in our dataset would contain:- a) The source event or the first or the best candidate which is the earliest sensible temporally consistent match in a given category page, this would name the event. b) Everything following the source article would be the consequences of that given event, as we found that this was the simplest way to go to cover the largest criteria of candidates, i.e. both the macro and micro and other chained or unchained events. c) Non-consequences or negative examples which would be a pool of candidates.

Although this small test was good, we wanted to know which topics (and events born of those topics) to do a larger automated extraction to do from, for this we extracted all 10,873 category pages within WikiNews and then sorted them by number of subcategories, idea being that significant category pages would contain a larger number of subcategories. At this level, when we say category, we’re trying to discover significant topics which we could do extractions from. From the 10k+ category pages, we filtered out those which contained far too few subcategories (<11) and manually examined the rest. We identified the following topics of interest:- Politics and Conflicts, Infectious disease, Disasters and Accidents, Aviation accidents and incidents, Economy and business, Crime and law and Human rights and Environment. There were certain derived topics that were subcategories of some of these listed topics which did not have significant subcategory counts but stood out as important on doing manual examination, they are:- Impeachment, Tsunamis, Earthquakes and Stock market. From these 11 topically diverse topics we picked out 39 events (seen in table [2]), where every event is a category page within WikiNews and we take the first or best temporally consistent and earliest article from there as the source and then the consecutive articles as consequences. For non-consequences, all articles under the main category or if the main category is not found then the most semantically similar category to a given event or category page is considered (within the list of categories of source event, omitting itself). The main category is the topic present as topicname within the markup of the source article (sometimes it is not present). The negative examples chosen

### Table 1: List of thirty-nine events selected from within the eleven topical WikiNews categories

<table>
<thead>
<tr>
<th>Category</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>Politics and Conflicts</td>
<td>Impeachment</td>
</tr>
<tr>
<td>Infectious disease</td>
<td>Ebola virus</td>
</tr>
<tr>
<td>Disasters and Accidents</td>
<td>Tsunamis, Earthquakes and Stock market</td>
</tr>
<tr>
<td>Aviation accidents and incidents</td>
<td>Plane crash, Commercial aviation accidents</td>
</tr>
<tr>
<td>Economy and business</td>
<td>Financial crisis, Bank collapse</td>
</tr>
<tr>
<td>Crime and law</td>
<td>Criminal investigation, Law enforcement</td>
</tr>
<tr>
<td>Human rights and Environment</td>
<td>Human rights violation, Environmental disaster</td>
</tr>
<tr>
<td>Impeachment</td>
<td>Impeachment process, Impeachment trial</td>
</tr>
<tr>
<td>Tsunamis, Earthquakes and Stock market</td>
<td>Tsunami, Earthquake, Stock market crash</td>
</tr>
</tbody>
</table>

Note: The table is not directly transcribed due to the nature of the text and the complexity of the content. The table is meant to illustrate the process of selecting events and categorizing them into specific topics.
are done so that they happen around time of the source event. Out of the 39 selected events, every
event is represented as a JSON object which contains the source event, the consequences and the
negative examples. Each consequence, negative example and the source event is itself also a JSON
with the following fields:- a) categories, b) category links, c) category name, d) content, e) date, f)
title and g) URL. This is all packaged into a JSONlines file. This whole process was done for the
English version of WikiNews and took multiple iterations of hand curation and automated cleaning
which required a lot of back and forth and at times the process was complicated due to noise and
errors within WikiNews itself, i.e. WikiNews sometimes seemed not updated in terms of contents e.g.
in terms of 2011 Egypt anti-government protests, the earliest event mentioned in the timeline view of
the category page is different from the earliest one amongst all pulled in alphabetically or WikiNews
has mismatches, e.g. "Greta Thunberg named 2019 Time Person of the Year" is listed under "2019
Hong Kong protests" and in the "Sendai Earthquake" category, a completely unrelated event from 4
years ago has been filed as the first event. These errors were accounted for to some degree along with
that, events which were not really direct impacts of the given source event or were simply things that
curred at the same time as the source events were omitted, e.g. "Strong rain and wind kill one in
Chile" might not have been necessarily caused by "2010 Pichilemu earthquake", but happened at
the same time. In the 39 events that have been released, there are a total of 570 consequences and
780 negative examples, with mean consequences (average number of consequences per event) being
14.6153 and median consequences being 6.

6 Conclusion

In this work, we have curated, created and released open source and freely accessible benchmarks
and datasets towards causality and event forecasting. Even though this is a strong starting endeavor,
there’s a lot of scope for future work. Many other knowledge bases (which could have had some
sort of expert curation or crowd sourcing) could have untapped pre-curated causal knowledge within
them just as the NQ dataset and can be similarly utilized and brought together to curate sets of more
clean cause effect pairs, this time opting to create something large enough to be a dataset for training
machine learning models. Wikipedia has a lot other events and a majority of other topics than the
three considered in this work and work can be expanded onto that. There are also other year pages,
which contain other societal events that can be collected and put together into a similar but larger
collection, albeit with an improved method (perhaps requiring lesser human curation). WikiNews
can have more languages considered than just English (since articles across each language seem not
exactly to be the same) and more events can be considered than the few selected by hand from the
curated topics. More importantly, two new tasks can be created from the WikiNews dataset:- a) A
multiple choice questionnaire much like COPA except instead of being restricted to commonsense
reasoning, the task is to select the right consecutive event following a given event and multiple
choices. b) An "event cloze" task; with the large number of consequences per source event, curate
traceable chains from the same and then the task can be that given a chain of events and a randomly
placed blank at any position of the event chain (could be at the end of the chain, i.e. predict the
last event) along with a set of choices (with one right answer), finding the correct event within that
set of choices that satisfies the chain consistency. Another task can be that given an event chain
and one candidate or a set of candidate option events, find if there’s a match (if any) for the given
event options within the chain at any position within the chain and then place it in the right location
(i.e. find correct position of the event within the chain if there’s a match for a given candidate). It
can be this task or any other task which relies on the narrative attribute of the event chain. Beyond
this, there’s also plans to have a simple baseline for the NQ-CE (causality based) benchmark using
machine reading systems like EIDOS to better get a sense of the numerical evaluations on the same.

References

Saurav Acharya, Byung Suk Lee, and Paul Hines. Causal prediction of top-k event types over

Patrick T Brandt, John R Freeman, and Philip A Schrodt. Real time, time series forecasting of


7 Paper Checklist

7.1 Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope?
Yes we curate the datasets as we described in the abstract.

7.2 Have you read the ethics review guidelines and ensured that your paper conforms to them?
Yes, I’ve read them. I’ve gained permission of the NQ dataset creators and WikiNews and Wikipedia are free to use and derive from. Unless there’s a mad scientist out there who plans on using event forecasting and gaining knowledge of impacts of significant societal events to do evil things, I don’t think there’s any negative impacts at all. In fact I’m collecting impacts of societal events especially because so I can do good. So we can understand something even if very small and minute about low probability high impact events before they happen. The NQ-CE dataset, I cannot fathom how someone can misuse that.

7.3 Did you discuss any potential negative societal impacts of your work?
I don’t see any negative societal impacts of my work but I will include a section for this and limitations in some supplementary materials I attach.

7.4 Did you describe the limitations of your work?
I’m going to include include a special document in the supplementary materials. Will also include that section here as well.

7.5 If you are including theoretical results...
None for this work

7.6 If you ran experiments...
Not in this work

7.7 If your work uses existing assets, did you cite the creators?
Yes, I’ll include more information in the supplementary materials.

7.8 Did you mention the license of the assets?
Please see the supplementary materials.

7.9 Did you include any new assets either in the supplemental material or as a URL?
There will be URLs in the supplementary material

7.10 Did you discuss whether and how consent was obtained from people whose data you’re using/curating?
I obtained consent from the Google NQ creators and Wikipedia and WikiNews are free to use and derive from

7.11 Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content?
Only for major political figures (heads of countries who might be prominent enough to be on world news and that too names), this is generic news data which is public. We don’t even delve into
entertainment or trivia. We just look at events which are significant and news articles in relation to
that. Random names of people might show up but nothing that causes alarm or needs consideration
(the people who’re using the datasets don’t care and those who’re looking to misuse personally
identifiable information here won’t get anything to misuse)

7.12 If you used crowdsourcing or conducted research with human subjects...

No, in fact we try to exercise this idea of curating the best resources with the least amount of people.
As seen in the NQ work, we actually leverage domain experts performing some other task to get

8 Limitations

This is a relatively new and an early version endeavor on many fronts and work is ongoing, hence
there’s a lot of scope for improvement. There can be perhaps much simpler methods to do extraction
of impacts from Wikipedia (better leveraging the human annotated structure of Wikipedia itself) but
such a thing would be restricted to Wikipedia itself and our method is reproducible over mining
impacts over any input text. Web mining WikiNews was hard hence that collection is still noisy but
we gear it towards models that can learn good representations inspite of the noise (although future
versions can be cleaner). We only consider knowledge bases in this paper for past sources of causal
knowledge but there can be many more, including authoritative source documents and even as we
saw other datasets. There could be work released right this very moment which could essentially
contain some hidden and precious curated causal knowledge and is yet another resource to be found
and worked on.

9 Acknowledgements

I would like to thank Dr. Oktie Hassanazadeh of IBM Research for all his support and help in this
work. Without him, this would not be possible and won’t be here. I would also like to thank the IBM
Scenario Planning Advisor team for their help as well.