Answering Binary Causal Questions Through Large-Scale Text Mining: An Evaluation Using Cause-Effect Pairs from Human Experts

Oktie Hassanzadeh, **Debarun Bhattacharjya**, **Mark Feblowitz**, **Kavitha Srinivas**, **Michael Perrone**, **Shirin Sohrabi** and **Michael Katz**
IBM Research
hassanzadeh@us.ibm.com, debarunb@us.ibm.com, mfeb@us.ibm.com, kavitha.srinivas@ibm.com, mpp@us.ibm.com, ssohrab@us.ibm.com, michael.katz1@ibm.com

Abstract

In this paper, we study the problem of answering questions of type "*Could X cause Y?*" where X and Y are general phrases *without any constraints*. Answering such questions will assist with various decision support tasks such as verifying and extending presumed causal associations used for decision making. Our goal is to analyze the ability of an AI agent built using state-of-the-art unsupervised methods in answering causal questions derived from collections of cause-effect pairs from human experts. We focus only on unsupervised and weakly supervised methods due to the difficulty of creating a large enough training set with a reasonable quality and coverage. The methods we examine rely on a large corpus of text derived from news articles, and include methods ranging from large-scale application of classic NLP techniques and statistical analysis to the use of neural network based phrase embeddings and state-of-the-art neural language models.

1 Introduction

Despite recent advancements in various areas of artificial intelligence and machine learning, it is still a common view that machines do not match human knowledge that is often essential for complex analysis and decision making tasks. For instance, when humans are tasked with answering questions regarding future sociopolitical and economic events, Tetlock and Gardner [2015] claim that "...a few hundred ordinary people and some simple math can not only compete with professionals supported by a multibillion-dollar apparatus but also beat them". A key part of answering such questions lies in not only gaining knowledge about current conditions and events, but also understanding "causal" relations between events and conditions. That is, knowing that an event or condition X has happened, is it likely that event or condition Y will follow? In other words, we need the ability to answer questions of type "Could X cause Y?"; we refer to these questions as *binary causal questions*.

Causal discovery is widely studied in artificial intelligence [Pearl, 2009] and growingly recognized as highly desirable for practical AI systems. While defining and identifying "actual" causality remains a topic of great interest [Halpern, 2016], we focus on a basic problem that humans can easily tackle. Our goal is to understand whether a causal relation could exist, regardless of whether it holds true in a particular scenario. For example, it should be clear to a human that "explosion" can cause "fire inside a factory building" or "several injuries", while "several injuries" is not a reasonable cause for "explosion". Our focus is on such binary causal questions that humans could answer either from prior knowledge or by searching through a source of knowledge (e.g., reading a book or performing a Web search), combined with some basic reasoning. While answering such questions could have a number of applications in various domains, our primary target application is risk management and situation awareness, where the "causal" knowledge can help analysts and decision makers better understand the impact of past and current events and conditions on relevant outcomes [Sohrabi *et al.*, 2018b; Sohrabi *et al.*, 2019]. A key requirement in this application is the ability to support causes (Xs) and effects (Ys) that are general phrases, *without any semantic constraints*.

The main contributions of this paper are as follows: 1) a set of novel benchmark data sets of cause-effect pairs from real-world risk management and decision support applications (Section 3) 2) a set of unsupervised methods of answering binary causal questions that do not impose any constraints on cause and effect phrases (Section 4) 3) Detailed experimental evaluation of the methods along with an analysis of strengths and shortcomings of each approach (Section 5).

2 Related Work

Our work is related to a number of active research areas in natural language processing, knowledge management, and data mining. Given our research focus, we only mention closely related work in terms of methods used and/or the application. To the best of our knowledge, no prior work has directly addressed the problem of answering binary causal questions of the form "Could X cause Y?" where X and Y do not necessarily subscribe to any semantic constraints.

In terms of the end application and the source of knowledge, our work is most related to that of Radinsky et al. [2012]. The Pundit algorithm in this work predicts the possible future effects of an "event" that has occurred. Here an event is defined as a set of concepts with semantic types

such as Action, Actor, Object, Instrument, and Location. For example, an event derived from the news article title "The U.S Army destroyed a warehouse in Iraq with explosives, which occurred on October 2004" is represented as: Destroy (Action); U.S Army (Actor); warehouse (Object); explosives (Instrument); Iraq (Location); October 2004 (Time). While this is a powerful event representation for our target application, the approach is limited to when proper concepts can be extracted and more importantly can be matched to concepts in an existing knowledge base. For example, the algorithm can predict "Tsunami-warning will be issued in the Pacific Ocean" given an event "Magnitude 6.5 earthquake rocks the Solomon Islands", based on causal statements the system is trained on such as "Tsunami warning issued for Indian Ocean after 7.6 earthquake strikes island near India". This is done based on mapping extracted concepts to an existing knowledge base that contains knowledge about islands and their nearest ocean. Follow-up work [Radinsky and Horvitz, 2013] describes an approach capable of generating general-purpose predictions, relying on "storylines" (collections of related news stories) instead of causal statements, but similarly uses a mapping to existing knowledge bases.

There is a rich literature on extraction of causal relations from text. To our knowledge, all previous work in this area falls under one or both of the following two categories: 1) requires extensive training data; 2) semantically restricts X and Y, in most cases by defining them as events with a strict representation that requires the use of NLP techniques for effective extraction of events.

An example of research that relies on extensive training data is a body of work by a group based in Japan who have applied supervised learning techniques using a benchmark training data set with over 100K human-annotated cause-effect pairs in Japanese [Hashimoto *et al.*, 2014; Kruengkrai *et al.*, 2017]. This relies heavily on a significant amount of labeled data and only considers events that involve a predicate with a single argument noun, such as "exacerbate desertification" (translation to English). A recent approach that can perform general causal extraction but also requires training data used annotations involving around 5,000 sentences by three annotators [Dasgupta *et al.*, 2018].

The work of Do et al. [2011] is an example of a minimally supervised approach, but puts restrictions on X and Y. Here an event is defined as $p(a_1, a_2, \ldots, a_n)$, where p is the word that triggers the presence of event in text (e.g., verbs such as "attacked" or nouns such as "explosion"), and $a_1, a_2, \ldots, a_n$ are the associated arguments (e.g., subject and object nouns for "attacked" event). As we will illustrate in the next section using examples from the cause-effect pairs from human experts, such definitions do not always support our target application. Another recent example of causal extraction over restricted event types involves commonsense reasoning, where X and Y are events involving agents, e.g. "PersonX wanted to be nice" could be a cause for "PersonX pays PersonY a compliment" [Sap *et al.*, 2019]. Asghar [2016], Radinsky et al. [2012, Sec. 5] and Dasgupta et al. [2018, Sec. 2] provide good reviews of causal relation extraction techniques.

We note that much of the early work on causal extraction from text uses lexical co-occurrence as a proxy or at least supporting evidence for causal relationships. The work of Church and Hank [1990] is a classic work in this domain that proposed the use of pointwise mutual information (PMI) for word association in text, computed by identifying co-occurrence of words in a corpus. Since then, mutual information and its variations have been widely used to measure causality between phrases or fragments of text [Chambers and Jurafsky, 2008; Riaz and Girju, 2010; Gordon *et al.*, 2011; Do *et al.*, 2011; Luo *et al.*, 2016].

Another popular approach to causal extraction uses discourse cues, i.e. lexical patterns in the form of 'A led to B', 'if A then B', etc., which provide semantic knowledge about how phrases relate to each other [Khoo *et al.*, 2000; Girju, 2003; Radinsky *et al.*, 2012]. For instance, in the sentence 'the police arrested him because he killed someone', the connective 'because' evokes a contingency relation between the adjacent text spans 'The police arrested him' and 'he killed someone'. It is common for causal extraction techniques to combine notions of co-occurrence together with discourse cues [Do *et al.*, 2011; Luo *et al.*, 2016].

In theory, it is also possible to apply a question answering solution to address our problem of answering binary causal questions. In practice, however, such an approach would need a high-quality corpus of text and training data, which could be too hard to collect for a generic and scalable solution. We consider such a solution out of scope for this paper but a potential avenue for future research. We note that some work on causal relation extraction has been framed as causal question answering solutions [Girju, 2003; Sharp *et al.*, 2016].

3 Cause-Effect Pairs Benchmark Data Sets

In this section, we describe a group of benchmark data sets of cause-effect pairs. Note that we do not use these data sets as training sets. Our causal extraction methods rely solely on a large corpus of text (news articles) that is used as an external source, i.e., it has not been used in any way for the creation of the cause-effect pairs data sets. Apart from the first collection which we include for the sake of comparison with state-of-the-art methods, the other three collections have not been used in the past and target our use case in risk management. We have made these data sets publicly available [Hassanzadeh *et al.*, 2019].

SemEval. We use the same data set used by Sharp et al. [2016] for the sake of comparison with state-of-the-art methods and pointing out some of the shortcomings of prior methods. The data set is derived from the SemEval 2010 Task 8 [Hendrickx *et al.*, 2010], originally a classification of semantic relations between nominals (words). Creating the collection involved a relatively complex annotation procedure involving humans following a set of guidelines to annotate around 1,200 sentences manually collected through pattern-based Web searches. The collection of cause-effect pairs created from this collection consists of 1,730 word pairs, out of which 865 (half) are from the cause-effect relations in the original data and so marked as causal, and the rest are a random subset of non-causal relations in the original data. Table 1 shows several examples of cause-effect pairs found in the

causal		non_causal	
word1	**word2**	**word1**	**word2**
disease	blindness	method	mathematician
vaccine	fever	aliens	space
protests	revolution	horses	stable
flight	crisis	blade	saw
explosion	damage	review	paper
officers	suffering	report	panel

Table 1: Examples from the SemEval collection

Cause	Effect
Rising regional tensions	Increased defence spending
Climate Change	New opportunities
Fractured and/or polarized societies	Instability and civil war
Ageing population and Youth bulges	Straining resources
Public discontent/disaffection and polarization	Lack of trust in governments and institutions
Globalization of financial resources	Lack of visibility on transactions supporting criminal and terrorist activities

Table 2: Examples from the NATO-SFA collection

collection. As these examples show, most pairs can turn into questions that are relatively easy to answer by humans (e.g., "Could vaccine cause fever?"), while some can be more challenging with no context (e.g., "Could flight cause crisis?").

NATO-SFA. Strategic Foresight Analysis (SFA) 2017 Report is a publication of NATO (the North Atlantic Treaty Organization) that "examines the main trends of global change and the resultant defence and security implications for NATO" [NATO-SFA, 2017]. This report is a result of a deep understanding of various trends and conditions throughout the world by a large number of human experts involved either directly in producing this report or indirectly by performing studies, meetings, and reviews, as acknowledged under "Sources and Acknowledgments" section of the report. The Appendix of the document contains a summary table of 20 "Trends" and 59 "Implications". We use the title text of each trend as a cause and the text of the implication (a word or a phrase) as the effect. We use the text as-is, with only slight modifications of a few effects/implications as in some cases the text repeated the cause/trend. Examples of the cause-effect pairs are shown in Table 2. Since there are no negative (non-causal) relations in the document, we use the set of unique cause or effect phrases to create an equal number of random pairs that do not appear in the table (i.e., there is no stated causal relationship between them) and mark them as non-causal. We note that while lack of a causal relation between randomly chosen cause and effect phrases in the non-causal portion is not guaranteed, the chances of having such a pair is very low.

Risk Models. As another source of causal knowledge by human experts, we take advantage of models designed by expert analysts for setting up a decision support system at a large organization [Sohrabi *et al.*, 2018a; Sohrabi *et al.*, 2018b; Sohrabi *et al.*, 2019]. The experts have created these models using a so-called "mind-mapping software" [Wikipedia, 2019]. The models can be seen as graphs where nodes are short descriptions of conditions or events (e.g. "High Inflation Rate" or "Increase in Corruption") and edges

Cause	Effect
currency appreciation against US dollar	low inflation
decreased local protectionism	decreased tariffs on foreign firms
decreasing government and political stability	decrease in government spending (infrastructure, education, public benefits)
increase in corruption	unfair allocation of government budget
rising unemployment rates	growing social tension
weakening economic environment	rising unemployment rates

Table 3: Examples from the Risk Models collection

Cause	Effect
wealth inequalities	social fissures
OPEC's agreement to raise production quota	Low oil prices
Expansionary fiscal policy	increased government spending
social programs	higher quality of life
improvement in global demand	boost commodity exports
reduction of the broad money supply	inflation

Table 4: Examples from the CE Pairs collection

imply a causal relation. These models are in part based on the experts' domain knowledge in enterprise risk management, and in part based on studying review literature and reports. We create a collection of cause-effect pairs by turning each edge in the graph to a pair with the label of the nodes as the phrase for cause and effect. The result is a set of 368 causal pairs with 223 unique cause/effect phrases. As in the NATO-SFA data, we extend the data with 368 randomly chosen (and so most likely non-causal) pairs and mark them as non-causal. Table 3 shows examples of these pairs.

CE Pairs. As another collection of cause-effect pairs targeting our primary use case in risk management, we manually extracted a set of cause-effect pairs where either cause or effect is related to one of the node labels in the above risk models, but the phrase comes from an external source. For this, we asked each of 7 people to take 30 unique labels from the Risk Models, and for each label, find sentences (from online news or other documents) that state a causal relationship between the node label (or its paraphrase) and another concept/phrase. For example, for a node label "increased tariffs on foreign firms", the cause could be "higher tariffs" with the effect being "lower consumer consumption" which is derived from this text found through Web search on the Investopedia website (investopedia.com): "The effect of tariffs and trade barriers on businesses, consumers and the government shifts over time. In the short run, higher prices for goods can reduce consumption by individual consumers and by businesses." Our goal in creating this new collection is twofold: 1) expanding the set of cause and effect phrases beyond the limited number of original node labels 2) extracting phrases that are used within one or a small number of sentences written by humans and so turn into more natural "Could X cause Y?" questions. Table 4 shows examples from this collection, that currently consists of 160 causal and 160 non-causal pairs.

4 Question Answering Methods

Following the approach of Radinsky et al. [2012], and with the same motivation, the methods we consider in this paper rely on a large collection of text. We seek generic methods ca-

pable of handling causes and effects that are general phrases, without any restriction on the type of the phrase (e.g., representing an event with certain arguments) or the ability to map the phrase to an existing source of knowledge or dictionary. All the methods considered in this paper are unsupervised, i.e., they do not require an existing collection of cause-effect pairs for training. Instead, our methods use the knowledge extracted from a large external text corpus to assign scores to the input questions (or cause-effect pairs). The scores can be used along with a threshold to provide a binary Yes/No answer for a given question, but also can be used as a confidence measure for a given answer. A primary feature of these methods is that the answers can be explained using the external corpus, e.g. by providing example similar causal relations from the external corpus. This explainability feature is often a requirement in risk management applications.

While our approach can work on any large corpus of text, we used a corpus of around 180 million titles and snippets of news articles covering around three years of news. Clearly, a larger corpus can result in a larger number of extracted cause-effect pairs but will also result in higher noise and increased scalability requirements for the methods. A smaller but higher quality source such as the New York Times corpus used by Radinsky et al. [2012] could result in less noise but also more limited representation of causes and effects. As some of the examples from our cause-effect pairs data sets show, the language used to express a particular causal relation can be very complex, and there could be numerous ways for humans to express the same relation. As a result, regardless of the size of the input corpus, an explicit mention of the cause-effect pairs may never be found in the corpus especially when no restrictions are imposed on cause/effect phrases.

4.1 Co-occurrence Based Methods

Since temporal co-occurrence methods are prevalent in the literature on causal extraction from text, we use a couple of these approaches as baselines.

Method 1 (`PMI`)
As mentioned earlier, the point-wise mutual information (PMI) between occurrences of words in a corpus has been particularly popular. For any pair of words x and y, we compute $PMI = \log \frac{p(y,x)}{p(x)p(y)}$, where $p(x)$ and $p(y)$ are the probabilities that x and y will be observed at least once in a document in the corpus, respectively, and $p(x, y)$ is the probability that the words co-occur in the same sentence at least once in a document in the corpus. To generalize this notion to a pair of text spans X, Y, we first use a phrase extractor to process the spans into bags of phrases, and then compute the average PMI between all possible combinations of cause-effect phrase pairs. We refer to this average as the PMI score for X, Y. We use the PMI score along with a threshold value to provide a Yes/No answer to the input question.

Method 2 (`CEA`)
The second temporal co-occurrence baseline is denoted CEA (cause-effect association) and is based on the measure proposed in Do et al. [2011]. It is a modification of the PMI that multiplies other factors, such as the joint inverse document frequency to penalize phrases that occur frequently, as well as measures for how phrases co-occur relative to other causes and effects. Hashimoto et al. [2014] used a similar co-occurrence measure as a baseline in their work. We refer the reader to Do et al. [2011] for details about the CEA score. We use the CEA score along with a threshold value to return a Yes/No answer to the input question.

4.2 Discourse Cue Based Methods

Our second class of methods is based on extraction of cause-effect pairs explicitly mentioned in sentences in the corpus. The extraction process consists of the following steps:

1. Filtering sentences with explicit causal verbs. For this, we use a dictionary of verbs derived from Girju and Moldovan's list of causal verbs with low ambiguity [Girju and Moldovan, 2002, Table 1].
2. Extraction of causal mentions following the approach of Sharp et al. [2016], i.e., turning each sentence into one or more ordered (X, Y) pair(s) where X and Y are phrases.
3. Indexing the extracted cause-effect pairs using a standard information retrieval engine, supporting substring and keyword searches.

Method 3 (`DCC`)
The simplest method of using the above index to answer a "Could X cause Y?" question is to do a search for (X, Y) and provide the number of hits (`count`) as the score and the hits as evidence/explanation for the answer. When X and/or Y contain more than one phrase, we first perform a phrase extraction and then perform a boolean OR search for all the phrase combinations.

We use an additional score which we refer to as the `c-score`, calculated by counting the number of hits for (X, Y) divided by the number of hits for (Y, X). This score is based on the intuition that the majority of causal relations are directional, i.e., when X is highly likely to cause Y, then in general it is unlikely for Y to cause X as well. As an example, our index returns 212 hits for (`explosion`, `injuries`) and 37 hits for (`injuries`, `explosion`)[1] and so the `c-score` is 5.73. A score larger than 1.0 is very likely a causal relation, while a score much lower than 1.0 shows lack of a causal relation or inability to provide a confident answer to the question.

Our first method which we refer to as `DCC`, uses both `count` and `c-score` along with two threshold values to provide a Yes/No answer to the input question.

Method 4 (`DCC-embed`)
A basic problem with the `DCC` method is its inability to capture various ways X and Y can be represented in natural language. As an example, a question with cause-effect pair (`High Inflation`,`Interest Rate Hike`) can be answered based on a pair like (`Increased Consumer Price Index`,`Higher Borrowing Costs`) although the phrases have no lexical similarity. Unlike previous work

[1] These hits are a result of noise in our causal extraction method. For example, "Another one of the three workers injured in Sunday's explosion has died as a result of injuries" results in a hit for (`injuries`, `explosion`).

Synonyms / Variations		Phrases capturing the same "concept"	
phrase	score	phrase	score
the_inflation_rate	0.896	consumer_price_index	0.750
an_inflation_rate	0.794	consumer_prices	0.745
annual_inflation_rate	0.792	the_consumer_price_index	0.742
annual_inflation	0.771	the_consumer_prices_index	0.716
inflation	0.767	…	
headline_inflation	0.757	gdp_deflator	0.708
inflation_rates	0.751	…	
…		gross_domestic_product	0.702
core_inflation	.727	consumer_price_growth	0.695
…		consumer_prices_index	0.694
overall_inflation	0.719	the_consumer-price_index	0.693
…		the_wholesale_price_index	0.688
retail_inflation	0.680	overall_consumer_prices	0.687

Figure 1: Similarity query results for inflation_rate

that relies on dictionaries and ontologies to perform semantic mapping, we build neural network based semantic embeddings of phrases using our external corpus to effectively capture semantic relatedness across phrases. Figure 1 shows the results of a nearest neighbor query for phrase Inflation Rate, and how top-k query results could contain synonyms/variations as well as related concepts. Our phrase embeddings are trained over the full corpus (not just the causal sentences) using a modified version of the classic word2vec [Mikolov *et al.*, 2013] with skip-gram architecture [Hassanzadeh *et al.*, 2018], where sentences are first turned into a collection of phrases, and the context for each phrase is all the other phrases in the same sentence. We believe our custom phrase embeddings better address our goal of discovering related phrases in the same corpus, although we have not performed a comparison with other methods including those that use pre-built models on another corpus.

Using the embeddings, our DCC-embed method extends X and Y with k phrases using a top-k nearest neighbor search query, and performs a Boolean OR query over the index for all the possible combinations. The method then uses count (the number of hits for the query) and c-score (the ratio of the number of hits for (X, Y) over the number of hits for (Y, X)) along with two threshold values to provide a Yes/No answer to the input question.

4.3 Neural Language Model Based Method

Recently, there has been an increasing interest around the use of neural language models to improve various NLP tasks. These language models are designed to capture sentence structures much more effectively compared to classic statistical language models [Bengio *et al.*, 2003]. Classic models, even if trained on large amounts of text, are very likely to face a sequence not seen in training data. As a complementary approach to our DCC-embed method which aims at capturing various representations of phrases, our goal is to use a neural language model that can capture the semantics of sentences involving both the cause and the effect. For this, we use BERT (Bidirectional Encoder Representations from Transformers) [Devlin *et al.*, 2018] which unlike previous neural language models, "is designed to pre-train deep bidirectional representations by jointly conditioning on both left and right context in all layers" and has been shown to improve the state of the art in eleven NLP tasks.

Method 5 (NLM-BERT)

Our method based on BERT also relies on extraction of causal sentences using discourse cues as explained in the previous section. However, instead of performing causal mention extraction and using phrases for lookup, we encode each causal sentence into a vector using BERT, and index all the vectors for efficient nearest neighbor search queries. We then compute two scores. Given a pair (X, Y), we first perform a search for top k causal sentences most similar to "X may cause Y" and compute the average cosine similarity score returned by the search. We refer to this score as bert-sim-score. For example, the average score for top 10 hits for "Explosion may cause death" is 0.916. We then perform a second search for the reverse relation, retrieving top k causal sentences most similar to "Y may cause X" and calculate the average cosine similarity scores. We divide bert-sim-score by this average similarity score of the reverse relation to compute our score for this method which we refer to as the bert-c-score. For example, the average cosine similarity score for top 10 hits for "Death may cause explosion" is 0.847, and so the bert-c-score is $0.916/0.847 = 1.081$.

Our BERT-based method uses bert-sim-score and bert-c-score along with two threshold values to provide a Yes/No answer to the input question.

5 Experiments

Implementation Details. We omit a detailed report on running times due to space constraints, but note that a primary requirement for all our methods has been scalability and near real-time response at query time. The pre-processing of our co-occurrence based solutions and the causal extraction part of our discourse cue based methods are implemented on Apache Spark, with jobs that take a few minutes to a few hours on a cluster with 256 executors. We tried both spaCy[2] and NLTK[3] libraries for phrase extraction but found only a simple extraction method based on regular expressions on top of POS tagging of NLTK to scale well, i.e., finish preprocessing within at most a few hours. Our corpus of over 180 million news article titles and snippets comes from Event Registry [Leban *et al.*, 2014]. We extracted over 320K causal sentences from titles and 16.7 million causal sentences from article snippets. For nearest neighbor search for DCC-embed and NLM methods, we used gensim[4] and faiss[5] libraries.

Accuracy Measures. Let tp be the number of true positives, fp the number of false positives, tn the number of true negatives and fn the number of false negatives. We measure precision ($pr = \frac{tp}{tp+fp}$), recall ($re = \frac{tp}{tp+fn}$), accuracy ($acc = \frac{tp+tn}{tp+tn+fp+fn}$), and F1 ($f1 = \frac{2*pr*re}{pr+re}$) for each method, data set, and varying threshold values.

We first report on overall accuracy results across the four data sets for various methods, and then provide a deeper analysis of the results for each data set.

[2]https://spacy.io/

[3]https://www.nltk.org/

[4]https://radimrehurek.com/gensim/

[5]https://github.com/facebookresearch/faiss

ds	method	mxv	thr1	thr2	tp	fp	pr	re	f1	acc
SemEval	PMI	F1	-	-∞	865	865	0.500	1.000	0.666	0.500
		Acc	-	-1.55	442	392	0.530	0.511	0.520	0.529
	CEA	F1	-	-∞	865	865	0.500	1.000	0.666	0.500
		Acc	-	-0.06	455	385	0.542	0.526	0.534	0.540
	DCC	F1	0	1.10	589	228	0.721	0.681	0.700	0.709
		Acc	0	1.90	516	135	0.793	0.597	0.681	0.720
	DCC-embed	F1	10	0.00	674	339	0.665	0.779	**0.718**	0.694
		Acc	10	1.60	501	97	0.838	0.579	0.685	**0.734**
	NLM-BERT	F1	0.82	0.62	863	855	0.502	0.998	0.668	0.505
		Acc	0.90	0.63	513	307	0.626	0.593	0.609	0.619
NATO-SFA	PMI	F1	-	-∞	59	59	0.500	1.000	0.667	0.500
		Acc	-	-1.84	23	11	0.676	0.390	0.495	0.602
	CEA	F1	-	-∞	59	59	0.500	1.000	0.667	0.500
		Acc	-	-0.06	13	7	0.650	0.220	0.329	0.551
	DCC	F1	0	0.00	59	59	0.500	1.000	0.667	0.500
		Acc	3	0.63	33	14	0.702	0.559	0.623	0.661
	DCC-embed	F1	0	0.00	59	59	0.500	1.000	0.667	0.500
		Acc	8	0.90	29	9	0.763	0.492	0.598	**0.669**
	NLM-BERT	F1	0.94	0.82	58	53	0.523	0.983	**0.682**	0.542
		Acc	0.93	0.90	50	43	0.538	0.847	0.658	0.559
Risk Models	PMI	F1	-	-∞	368	368	0.500	1.000	0.667	0.500
		Acc	-	-0.05	60	37	0.619	0.163	0.258	0.531
	CEA	F1	-	-∞	368	368	0.500	1.000	0.667	0.500
		Acc	-	-0.92	185	153	0.547	0.503	0.524	0.543
	DCC	F1	0	0.00	368	368	0.500	1.000	0.667	0.500
		Acc	0	1.60	47	41	0.534	0.128	0.206	0.508
	DCC-embed	F1	0	0.00	368	368	0.500	1.000	0.667	0.500
		Acc	0	1.56	67	52	0.563	0.182	0.275	0.520
	NLM-BERT	F1	0	0.90	345	318	0.520	0.938	**0.669**	0.537
		Acc	0	1.00	226	184	0.551	0.614	0.581	**0.557**
CE Pairs	PMI	F1	-	-∞	160	160	0.500	1.000	0.666	0.500
		Acc	-	-0.10	7	4	0.636	0.044	0.082	0.509
	CEA	F1	-	-∞	160	160	0.500	1.000	0.667	0.500
		Acc	-	2.55	36	23	0.610	0.225	0.329	0.541
	DCC	F1	0	0.00	160	160	0.500	1.000	0.667	0.500
		Acc	2	1.38	54	35	0.607	0.338	0.434	0.559
	DCC-embed	F1	0	0.00	160	160	0.500	1.000	0.667	0.500
		Acc	2	1.20	66	42	0.611	0.413	0.493	**0.575**
	NLM-BERT	F1	0.91	0.65	160	157	0.505	1.000	**0.671**	0.509
		Acc	0.96	1.02	44	24	0.647	0.275	0.386	0.563

Table 5: Accuracy results

Table 5 shows the overall accuracy results, reporting the maximum accuracy and F1 scores each method can achieve on each of the data sets along with the score threshold(s) used for the result achieved. To better understand what the accuracy numbers mean in practice, we also report the number of questions answered positively correctly (tp) and incorrectly (fp). A higher precision is a desired feature in use cases where a large number of phrases (e.g., events and conditions) of interest are known and the system is queried with all the possible pairs to build a causal graph. Since the number of causal relations is usually a small fraction of all the possible pairs, lower precision will result in an unacceptably large number of false positive pairs. Also, since the input corpus may or may not be able to support answering a question, negative answers can be viewed as inability to answer a question if the method has a high precision.

On SemEval data, DCC-embed achieves the highest accuracy and F1 scores, outperforming the state-of-the-art method by Sharp et al. [2016] which is based on a "Causal Convolutional Neural Network Model" that achieves a precision of under 60% at recall levels over 70%. DCC-embed achieves a precision of 66.5% at a recall of 77.9% (71.8% F1 score). Figure 2 of Sharp et al. [2016] shows that no method can achieve a precision of over 70% at recall levels over 45% whereas DCC and DCC-embed methods achieve this precision at over 70% recall.

The results over the other data sets show a different overall story. While DCC-embed outperforms the other methods in terms of maximum accuracy in the NATO-SFA and CE Pairs data sets as well, the margin comparing with baselines is much smaller. The NLM-BERT method performs better than the other methods in terms of F1 score in NATO-SFA and CE Pairs data and in both accuracy and F1 score in the Risk Models data, again with a small margin. Overall, we observe that some methods fail at distinguishing between causal and non-causal pairs in the question, while other methods do well only on a small number of the input questions

One surprising result is the poor performance of NLM-BERT on SemEval. Upon inspection of the results, we observe two limitations of this method: 1) the method fails at understanding "unusual" sentences that are unlikely to be written by humans. For example, the pair (eyelids, blinking) from SemEval pairs turns into a query sentence "*eyelids may cause blinking*" which even if correct, is not a statement a human would make; 2) the value of bert-sim-score on its own is meaningless and is only useful for ranking. For example, the average score for highly similar sentences retrieved for a given query can be the same or even lower than the average score for another query with sentences retrieved that have little similarity to the query.

Choice of Threshold Value. We note that while we measured the maximum accuracy scores with varying threshold values for the score(s) of each method, in practice the choice of the right threshold value may not be clear, and may require supervision. However, our DCC and DCC-embed methods have the additional advantage that given the intuition behind the c-score value, a threshold value of around 1.0 results in accuracy scores close to the maximum value in all cases.

6 Conclusion & Future Work

Our primary goal in this paper was to show the ability of an AI engine that uses state-of-the-art methods of large-scale knowledge extraction from text, to answer general binary causal questions of the form "Could X cause Y?". Although our results are promising and show in part the power of the implemented methods, the achieved accuracy is still far from the level of human intelligence and what human experts can achieve. Our methods however show the promise of being able to provide assistance to human experts, e.g., by pruning a large number of potential causal pairs. We intend to use the outcome of this work as a part of the IBM Scenario Planning Advisor [Sohrabi *et al.*, 2019] to enable users to prune a large space of potential pairs, and to provide hints for each question posed to the user. We have made our benchmark data sets publicly available [Hassanzadeh *et al.*, 2019] and will continue to extend and refine them. We hope that these data sets will promote further research on causal knowledge extraction and binary causal question answering methods.

References

[Asghar, 2016] N. Asghar. Automatic extraction of causal relations from natural language texts: A comprehensive survey. *CoRR*, abs/1605.07895, 2016.

[Bengio *et al.*, 2003] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003.

[Chambers and Jurafsky, 2008] N. Chambers and D. Jurafsky. Unsupervised learning of narrative event chains. In *ACL*, 2008.

[Church and Hank, 1990] K. W. Church and P. Hank. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

[Dasgupta *et al.*, 2018] T. Dasgupta, R. Saha, L. Dey, and A. Naskar. Automatic extraction of causal relations from text using linguistically informed deep neural networks. In *SIGDIAL*, 2018.

[Devlin *et al.*, 2018] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[Do *et al.*, 2011] Q. Do, Y. Seng Chan, and D. Roth. Minimally supervised event causality identification. In *EMNLP*, 2011.

[Girju and Moldovan, 2002] R. Girju and D. I. Moldovan. Text mining for causal relations. In *FLAIRS*, 2002.

[Girju, 2003] R. Girju. Automatic detection of causal relations for question answering. In *MultiSumQA*, 2003.

[Gordon *et al.*, 2011] A. S. Gordon, C. A. Bejan, and K. Sagae. Commonsense causal reasoning using millions of personal stories. In *AAAI*, 2011.

[Halpern, 2016] J. Y. Halpern. *Actual Causality*. The MIT Press, 2016.

[Hashimoto *et al.*, 2014] C. Hashimoto, K. Torisawa, J. Kloetzer, M. Sano, I. Varga, J.-H. Oh, and Y. Kidawara. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In *ACL*, 2014.

[Hassanzadeh *et al.*, 2018] O. Hassanzadeh, S. Trewin, and A. Gliozzo. Semantic concept discovery over event databases. In *ESWC*, 2018.

[Hassanzadeh *et al.*, 2019] O. Hassanzadeh, D. Bhattacharjya, M. Feblowitz, K. Srinivas, M. Perrone, S. Sohrabi, and M. Katz. Data sets of cause-effect pairs, May 2019. https://doi.org/10.5281/zenodo.3214925.

[Hendrickx *et al.*, 2010] I. Hendrickx, S. Kim, Z. Kozareva, P. Nakov, D. Ó. Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, and S. Szpakowicz. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *SemEval*, 2010.

[Khoo *et al.*, 2000] C. Khoo, S. Chan, and Y. Niu. Extracting causal knowledge from a medical database using graphical patterns. In *ACL*, 2000.

[Kruengkrai *et al.*, 2017] C. Kruengkrai, K. Torisawa, C. Hashimoto, J. Kloetzer, J.-H. Oh, and M. Tanaka. Improving event causality recognition with multiple background knowledge sources using multi-column convolutional neural networks. In *AAAI*, 2017.

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

[Luo *et al.*, 2016] Z. Luo, Y. Sha, K. Q. Zhu, S. Hwang, and Z. Wang. Commonsense causal reasoning between short texts. In *KR*, 2016.

[Mikolov *et al.*, 2013] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed Representations of Words and Phrases and Their Compositionality. In *NIPS*, 2013.

[NATO-SFA, 2017] Strategic Foresight Analysis 2017 Report. https://www.act.nato.int/publications-ffao, 2017. [Online; accessed February 21, 2019].

[Pearl, 2009] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2009.

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

[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.

[Riaz and Girju, 2010] M. Riaz and R. Girju. Another look at causality: Discovering scenario-specific contingency relationships with no supervision. In *ICSC*, 2010.

[Sap *et al.*, 2019] M. Sap, R. LeBras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, and Y. Choi. ATOMIC: An atlas of machine commonsense for if-then reasoning. In *AAAI*, 2019.

[Sharp *et al.*, 2016] R. Sharp, M. Surdeanu, P. Jansen, P. Clark, and M. Hammond. Creating causal embeddings for question answering with minimal supervision. In *EMNLP*, 2016.

[Sohrabi *et al.*, 2018a] 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.

[Sohrabi *et al.*, 2018b] S. Sohrabi, A. V. Riabov, M. Katz, and O. Udrea. An AI planning solution to scenario generation for enterprise risk management. In *AAAI*, 2018.

[Sohrabi *et al.*, 2019] S. Sohrabi, M. Katz, O. Hassanzadeh, O. Udrea, M. D. Feblowitz, and A. Riabov. IBM scenario planning advisor: Plan recognition as AI planning in practice. *AI Commun.*, 32(1):1–13, 2019.

[Tetlock and Gardner, 2015] P. E. Tetlock and D. Gardner. *Superforecasting: The Art and Science of Prediction*. Crown Publishing Group, New York, USA, 2015.

[Wikipedia, 2019] https://en.wikipedia.org/wiki/List_of_concept-_and_mind-mapping_software, 2019. [Online; accessed February 21, 2019].