Framework for Weakly Supervised Causal Knowledge Extraction from Text

Anonymous ACL submission

Abstract

In this paper, we address the problem of extracting causal knowledge from text documents in a weakly supervised manner. We target use cases in decision support and risk management, where causes and effects are general phrases without any constraints. We present a unified framework that supports three classes of tasks with varying degrees of available information. We provide approaches for each of the tasks using pre-trained, Natural Language Inference (NLI) and Question Answering (QA) models. We present a novel evaluation scheme and use existing and new benchmark data sets to measure the relative performance of each of the approaches.

1 Introduction

Extracting causal knowledge from natural language descriptions of such knowledge in text documents is a challenging problem with a wide range of applications in AI systems. There is a relatively large body of work in the literature addressing different flavors of this problem. One major application area has been event forecasting (Radinsky et al., 2012a), as well as decision support and risk management (Dasgupta et al., 2018; Hassanzadeh et al., 2019, 2020). Our work targets the latter application area, where causes and effects are general phrases which may or may not be describing actions or events.

A major challenge in applying state-of-the-art supervised knowledge extraction methods is the need for a large manually-annotated corpus, which is not feasible for large-scale generic causal knowledge extraction. Our focus in this paper is on weakly supervised methods where the input is a corpus of text documents that contain descriptions of causal knowledge required in the target application, and the output is a high-quality collection of cause-effect pairs, which can then be further processed, represented as a causal knowledge graph, and used as input for decision support or predictive analytics. Table 1 shows an example of a few cause-effect pairs extracted by one of our methods where the only input is a collection of Wikipedia articles about COVID-19.

Cause	Effect
COVID-19 pandemic	wave of solidarity
COVID-19 pandemic	sharp increase in the use of telemedical services
COVID-19 outbreak	fear of a potential economic breakdown
COVID-19	reductions in bus route frequencies
fears of supply shortages	panic buying
panic buying	shortages of some products

Table 1: Examples of Cause-Effect pairs extracted by one of our proposed methods where the only input is a collection of Wikipedia articles on COVID-19.

Our contributions in this paper are as follows:

1. We present a framework for weakly supervised causal knowledge extraction from text, depicted in Figure 1, with three classes of solutions based on whether the input is only a corpus of text documents or consists of a set of candidate causes and/or effects.

2. For each class of solutions, we present a method using state-of-the-art natural language understanding techniques including methods that rely on neural models for Natural Language Inference (NLI) or Question Answering (QA). To our knowledge, we are the first to use NLI and open-ended QA for causal knowledge extraction.

3. We present a novel scheme for evaluation of weakly supervised causal knowledge extraction techniques and present the results of our experiments on existing and new benchmarks. We will make our benchmark data sets publicly available.

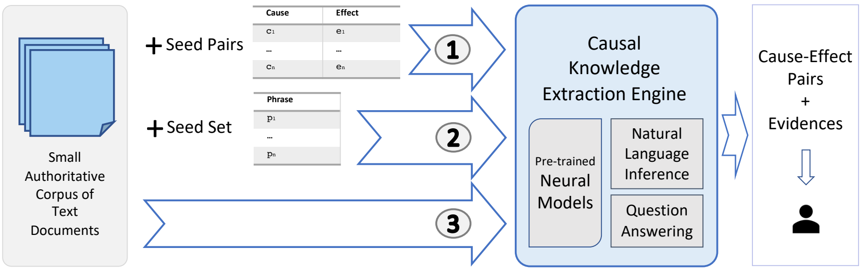


Figure 1: Causal Knowledge Extraction Framework. The three approaches labeled ①, ② and ③ are presented in Sec 3.1, 3.2, and 3.3 respectively.

2 Related Work

There is a large body of work on causal knowledge extraction from text. Section 2 of (Hassanzadeh et al., 2019), Section 5 of (Li et al., 2020), and Xu et al. 2020 provide excellent summaries of related work in this area. Here, we discuss a few key approaches and their main characteristics of solutions as compared to our approach. Table 2 lists several prior works, along with their main characteristics based on these dimensions: 1) the end application; whether the goal is primarily commonsense reasoning, or decision support and risk management 2) whether the approach is supervised or unsupervised 3) if causes and effects are simply words/phrases/text spans, or have a specific semantic representation 4) whether patterns or discourse cues are used or not, and 5) if the approach relies on a very large corpus or not. Note that these dimensions are not entirely independent. For example, work primarily focused on commonsense reasoning can take advantage of the vast volume of textual descriptions of such knowledge available on the Web, whereas in other domains such large corpora may not be available or may result in an too much noise for the end application.

Our primary motivation in this paper is application in generic decision support systems and risk management, where the system needs to be capable of extracting causal relations between a wide variety of causes and effects, and so specifying a specific semantic representation for causes and effects (e.g. an event representation) and a large enough annotated corpus could be unfeasible. As a result, we focus on weakly supervised approaches that perform causal relation extraction over text spans with very little training data. The output of our solution can then be used to build models for risk management (Chapman, 2013; Sohrabi et al., 2018), or be further refined into a knowledge base for e.g. forecasting future events (Radinsky and Horvitz, 2013; Muthiah et al., 2016).

	Commonsense Goal	Unsupervised	Text Span Based	Uses Patterns (Cues)	Not Requires Large Corpus
Our Work		✓	✓	✓	✓
(Li et al., 2020)	✓		✓	✓	
(Hassanzadeh et al., 2019)		✓	✓	✓	
(Dasgupta et al., 2018)			✓	✓	✓
(Kruengkrai et al., 2017)					
(Hashimoto et al., 2014)				✓	
(Dunietz et al., 2017b)		✓	✓		
(Luo et al., 2016)	✓	✓	✓	✓	
(Sap et al., 2018)	✓			✓	
(Radinsky et al., 2012b)					
(Soares et al., 2019)			✓		
(Li and Tian, 2020)			✓		✓

Table 2: Characteristics Prior Work & Our Work

Our approach in using patterns to extract candidate cause-effect pairs follows the approach used in prior work (Girju, 2003; Luo et al., 2016; Hassanzadeh et al., 2019; Li et al., 2020). Most recently, such patterns are used to create very large collections of cause-effect pairs given a large corpus of documents. Li et al. (Li et al., 2020) use such an approach over a large corpus of Web documents to create a large collection of cause-effect pairs, referred to as CausalBank, which is then used to generate a "Cause Effect Graph" with application to training a BERT-based model that significantly outperforms similar methods in the Choice Of Plausible Alternatives (COPA) evaluation task which is geared towards commonsense reasoning. Hassanzadeh et al. (2019) extract cause-effect pairs from a large collection of news articles and use the outcome for a binary classification task to answer binary causal questions. Our work has a different goal: creating a high-quality collection of cause-effect pairs from a smaller authoritative source of text documents in a particular domain, similar to

1. X causes Y	5. If X then Y
2. X is the reason for Y	6. Effect of X is Y
3. Because of X, Y	7. Y as a result of X
4. X leads to Y	

Table 3: Causal patterns used for NLI

the pairs shown in Table 1.

3 Causal Knowledge Extraction Framework

Our framework for causal knowledge extraction is depicted in Figure 1. This framework allows us to approach Below, we describe the three categories of tasks and propose weakly supervised causal extraction methods in each category.

3.1 Cause, Effect and Context

Given a potential cause-effect pair and the context in which it appears, the task is a binary classification problem - to label the pair as causal or non-causal. We approach this task using Natural Language Inference. Let S_1 be the original sentence and (X, Y) be a candidate cause-effect pair. We construct a new causal sentence $S_2^i, i \in 1 \dots k$ in k different ways based on $k = 7$ syntactically different causal patterns shown in Table 3. For instance, S_2^1 = *"X causes Y"*, S_2^2 = *"X is the reason for Y"*. We then use a pre-trained NLI model to get the probability P_i of inferring the causal sentence S_2^i from the original sentence S_1. We use the mean of the k probabilities as the probability of (X, Y) being causal.

3.2 Cause/Effect and Context

In this category, methods have access to the text and either the cause or the effect but not both. The task is to discover the corresponding effect or cause. This is a common scenario in practice where a user might be interested in the causes of a major set of events such as "covid-19". We approach this task using Question Answering. Let S be the given sentence and X be the candidate cause. We create a causal question q = *"What does X cause?"*. We then use a pre-trained QA model to extract answers (Y_i) to q from S along with their confidence scores. We retain the one with the highest score and pair it with X to form the causal pair (X, Y). Correspondingly, we could treat X as the candidate effect and change the causal question to q = *"What causes X"* and follow the same procedure above to extract the cause.

1. X causes Y	6. X is responsible for Y
2. Y because X	7. whenever X, Y
3. X triggers Y	8. Y arises from X
4. Y results from	9. X contributes to Y
5. attribute Y to X	10. following X, Y

Table 4: Examples of causal patterns used for matching

3.3 Only Context

This category consists of methods which have access to only the text and the task is to extract cause-effect pairs from the text. This is the most difficult task among the three since it assumes access to the least amount of information. We approach this task by first using pattern matching (`PM`) to construct candidate cause-effect pairs from the given text and then classifying them as causal or non-causal using the NLI method described in Sec 3.1. We use a list of nearly 200 causal patterns created by Dunietz et al. (2017a) as a guide to annotate linguistic evidences of causality. Table 4 shows a sample of these patterns. We lemmatize all the patterns and the sentences to enable matching verbs in their root form and convert the patterns to regexes e.g. *"(.*) cause (.*)"*. Finally we match them against the sentence obtaining the parts of the sentence corresponding to the candidate cause-effect pair.

In the cases where the given text is long, the patterns lead to long candidate causes and effects which may provide details that are irrelevant to the causal pair. In such cases, we extract phrases from the candidates and pair them with each other to form candidate cause-effect pairs. To extract phrases, we experiment with two phrase extraction techniques - `NPFST` and `CP`. `NPFST` (Handler et al., 2016) extracts noun phrases using Finite State Transducers while `CP` extracts all constituent phrases from a constituency parse of the sentence.

4 Evaluation

4.1 Datasets

We benchmark the performance of the proposed methods on three datasets described below. Full datasets are included in supplementary material and will be released publicly.

The BECauSE 2.0 corpus (Dunietz et al., 2017a) consists of general phrases as causes and effects, tagged by annotators from within a sentence. Overall, there are 2150 pairs in the dataset out of which 1472 are causal. Table 5 shows a sample of cause-effect pairs from this dataset.

The SemEval dataset has been constructed

Cause	Effect	Context
The regulatory regime we establish and follow	market discipline	**The regulatory regime we establish and follow** must accomplish three things: ensure **market discipline**; provide a shock absorber against systemic risk; and, first and foremost, protect the taxpayer.
This bill	regulation more efficient	**This bill** seeks to make **regulation more efficient** by closing gaps in our regulatory structure and by promoting consolidation and cooperation among regulatory agencies.
The federal reserve's actions	preserve confidence and bring stability to our financial markets and institutions	And Chairman Bernanke, **the Federal Reserve's actions** continue to help **preserve confidence and bring stability to our financial markets and institutions**.

Table 5: Causal pairs from the BECauSE 2.0 corpus.

from SemEval 2010 Task 8 (Hendrickx et al., 2010) The dataset consists of 2662 pairs of words instead of phrases with equal number of causal and non-causal pairs. Table 6 shows some examples from this dataset.

Cause	Effect	Context
disease	blindness	a rare and incurable congenital **disease** which causes **blindness** has been successfully treated for the first time using gene therapy.
vaccine	fever	convulsions that occur after dtap are usually not caused directly by the **vaccine**, but by a **fever**, which in turn was triggered by the vaccine.
explosion	damage	iraqi soldiers inspect the **damage** after the **explosion** in a school in baghdad.

Table 6: Causal pairs from the SemEval dataset.

The MultiCause dataset (Anonymous, 2020) is a new dataset created using the Natural Questions dataset (Kwiatkowski et al., 2019). The dataset is created by first finding causal questions by filtering questions that have a causal verb (e.g., "causes", "leads to") and start with "What" or "Would". There are several questions that result in more than one cause for an effect or more than one effect for a cause. The final set consists of 140 cause-effect pairs from 112 causal questions. We expand a pair with multiple causes (or effects) into multiple pairs, each having the same cause (or effect). Table 7 shows a sample of pairs from this dataset.

4.2 Evaluation Metrics

We match an extracted pair with the ground truth pair if both the phrases - cause and the effect match. To match a phrase in the BECAUSE dataset we

Cause	Effect	Context
(1) excessive nutrient pollution from human activities coupled with other factors that deplete the oxygen	(1) a dead zone in the ocean	**Dead zones** are hypoxic (low-oxygen) areas in the world's oceans and large lakes, caused by **excessive nutrient pollution from human activities coupled with other factors that deplete the oxygen** required to support most marine life in bottom and near-bottom water. (NOAA)
(1) cold weather (2) anticyclone and windless conditions (3) collected airborne pollutants	the deadly smog in london in 1952	The **Great Smog of London** , or Great Smog of 1952 , was a severe air - pollution event that affected the British capital of London in early December 1952 . A period of cold weather , combined with an anticyclone and windless conditions , **collected airborne pollutants** – mostly arising from the use of coal – to form a thick layer of smog over the city .
(1) bacterium treponema pallidum	(1) syphilis (2) bejel (3) pinta (4) yawns	**Treponema pallidum** is a spirochaete bacterium with subspecies that cause treponemal diseases such as **syphilis** , **bejel** , **pinta** , and **yaws**. The treponemes have a cytoplasmic and an outer membrane. . . .

Table 7: Causal pairs from the MultiCause dataset.

check if the Jaccard similarity between the tokens is more than 0.5. Since the SemEval dataset consists of words, we check if the true word is contained within the extracted phrase. On the other hand, in the MultiCause dataset, the cause-effect pairs do not occur inside the sentence verbatim. Hence, we calculate the cosine similarity between the mean of the Siamese BERT (Reimers and Gurevych, 2019) word vectors of the two phrases and use a threshold of 0.5 to declare a match. Finally, we report the Precision, Recall, and F1-score of the extracted pairs as well as the causes and effects.

5 Experiments

Settings For QA, we use the ALBERT-xxlarge model (Lan et al., 2020) fine-tuned on the SQuAD v2.0 dataset (Rajpurkar et al., 2018) while for NLI we use the RoBERTa model (Liu et al., 2020) fine-tuned on MNLI (A. Williams and Bowman, 2018). For every dataset, we first split it into dev and test sets with 20% and 80% points respectively and search for a threshold confidence on the dev set from the range $[0, 1]$ with steps of 0.01. An extracted pair is marked causal if its confidence is more than the selected threshold. All our experiments were conducted using PyTorch framework on one Tesla P100 GPU with 16GB memory.

DS	Input	Method	Thresh	Pairs			Causes			Effects		
				P	R	F	P	R	F	P	R	F
BECauSE	Context only	PM	-	26.3	36.5	30.6	28.5	38.5	32.7	27.4	37.0	31.5
	Context only	PM + NLI	0.9	41.6	27.7	33.3	40.9	28.0	33.3	39.5	27.1	32.1
	Context only	PM + CP + NLI	0.89	34.0	23.2	27.6	39.6	27.2	32.2	35.9	24.2	29.6
	Context only	PM + NP + NLI	0.28	7.0	3.5	4.6	15.5	7.6	10.2	9.1	4.4	5.9
	Context + Cause/Effect	QA	0.23	56.6	55.4	56.0	46.2	45.0	45.6	51.8	51.6	51.7
	Context + Cause + Effect	NLI	0.6	74	81.0	77.3	74.3	80.7	77.4	74.6	80.9	77.6
MultiCause	Context only	PM	-	7.5	19.1	10.8	22.3	51.2	31.1	14.8	33.3	20.5
	Context only	PM + NLI	0.9	9.0	19.1	12.2	22.8	47.6	30.8	16.1	31.0	21.2
	Context only	PM + CP + NLI	0.9	9.0	10.7	9.8	23.1	26.2	24.5	16.7	20.2	18.3
	Context only	PM + NP + NLI	0.9	15.0	7.1	9.7	25.0	15.5	19.1	22.5	10.7	14.5
	Context + Cause/Effect	QA	0.3	54.7	53.6	54.1	41.3	41.7	41.5	34.7	34.5	34.6
	Context + Cause + Effect	NLI	0.5	58.4	75.0	65.7	68.1	76.2	71.9	72.6	78.6	75.5
SemEval	Context only	PM	-	36.1	66.2	46.7	45.7	72.6	56.1	46.3	72.9	56.6
	Context only	PM + NLI	0.95	45.9	57.7	51.1	55.2	63.0	58.8	56.1	62.8	59.3
	Context only	PM + CP + NLI	0.97	43.3	44.2	43.7	56.3	49.8	52.9	55.6	49.5	52.4
	Context only	PM + NP + NLI	0.92	26.6	14.1	18.4	46.5	18.6	26.6	43.8	17.8	25.3
	Context + Cause/Effect	QA	0.33	76.0	78.6	77.3	76.0	78.6	77.3	77.4	80.3	78.8
	Context + Cause + Effect	NLI	0.58	81.9	87.6	84.7	81.9	87.6	84.7	81.9	87.7	84.7

Table 8: The performance of different classes of models based on their input, across three diverse datasets. P, R and F refer to the Precision, Recall and F-score of the different methods and Thresh refers to the threshold picked on a small dev set. The standard deviation across 5 random runs for all the methods is smaller than 0.6

5.1 Overall Results

In Table 8 we show the performance of our models. We can observe that as we add more information to the methods, their performance improves i.e. `NLI` performs better than `QA` which performs better than `PM` based methods.

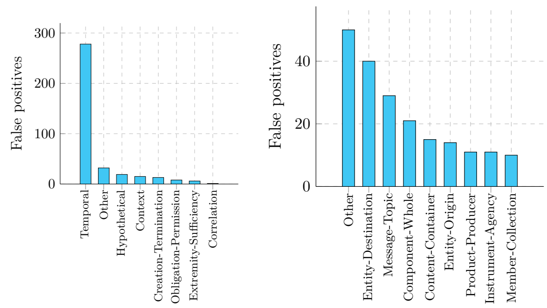
We also observe that `CP` performs better than `NPFST`, likely due to the fact that NPFST focuses on extracting only the noun phrases while `CP` has no such restriction. However, the `PM+NLI` approach which does not perform any phrase extraction outperforms both. This is likely due to the fact that for short, well formed sentences, extracting phrases might remove critical context e.g. in the sentence *“Failure to comply with the new regulations could result in denying entry or a fine of AU $62,800 ."* `NPFST` extracts the phrase *“new regulations"* as the cause whereas the precise cause is *“Failure to comply with the new regulations"*.

5.2 Error Analysis

Here we analyze the errors made by the `NLI` model on BECauSE and SemEval datasets. We focus on the false positives as these errors are more critical to our target application in risk management.

In the BECauSE 2.0 dataset, all pairs are labeled with eight relations like *temporal*, *hypothetical* etc. in addition to the causal/non-causal relation. Figure 2a shows the distribution of false positives of the `NLI` model. We find that most of the false positives are actually only temporal relations between the phrases. We find many instances in which even though liguistically there is little evidence of causality, the `NLI` model gives a reasonable output. For example in the sentence *“In Iraq violence, three american soldiers died over the weekend, the military said in a statement"* it is reasonable to assume that the Iraq violence caused the death of three American soldiers. We also find some cases in which the context implies that the cause *prevents* the effect from occurring but the `NLI` model mistakes it for a causal relationship.

In the SemEval dataset, all the non-causal pairs are labeled with one of nine relations like *Component-Whole*, *Entity-Destination*, *Other* etc. Fig 2b shows the distribution of false positives made by the `NLI` model on the SemEval dataset. We find that most errors belong to the *Other* category followed by the *Entity-Destination* category. Here also, we find some instances where the model makes a reasonable assumption about causality even though there is little linguistic evidence in the sentence e.g. in the sentence *“The typical flu infection start with fever, muscular pains, headache and general fatigue"*, infection is the cause of fever.



(a) BECauSE (b) SemEval

Figure 2: Distriution of false positives of the NLI model on the BECauSE and SemEval datasets.

5.3 Comparison with supervised baselines

To show the efficacy of our pattern matching and NLI based methods, we compare them against a strong, supervised baseline based on BERT (Devlin et al., 2018; Soares et al., 2019) on the BECAUSE dataset. BERT is finetuned as a tagging model for comparison against pattern matching based approaches and as a relation classification model (Soares et al., 2019) for comparison against the NLI method. We follow the same procedure as Devlin et al. (2018); Soares et al. (2019) for these two scenarios. We also compare the methods in the more real-world setting where very little training data is available by randomly sampling 20% datapoints as the training set for BERT and development set for our methods. These comparisons are shown in Table 9. We find that in the presence of small amounts of training data, the semi-supervised approaches perform much better. However, given a large amount of training data, the supervised method outperforms the semi-supervised methods.

Method	Full Data			20% Data		
	P	R	F	P	R	F
PM+NLI	**35.1**	34.5	34.8	**40.8**	26.8	**32.3**
BERT	33.1	**49.6**	**39.7**	19.3	**38.9**	25.8
NLI	71.8	**95.9**	82.1	**71.7**	**95.4**	**81.9**
BERT	**83.4**	83.4	**83.4**	69.0	69.0	69.0

Table 9: Comparison of our best performing semi-supervised models (PM+NLI and NLI) against a strong supervised baseline based on BERT.

5.4 Manual Evaluation

We also applied the three promising pattern matching based methods (1) `PM+NLI`, (2) `PM+CP+NLI` and (3) `PM+NP+NLI` on articles about COVID-19 from Wikipedia. The collection consists of 236 articles under the COVID-19 Pandemic category and its subcategories, crawled on May 6th 2020. We evaluated the top 50 outputs from each of the three methods (total 150 outputs) using three annotators experienced in this field. They followed a variety of "tests for causality" (Grivaz, 2010; Dunietz et al., 2017a) to annotate the ambiguous cases. Table 10 shows the precision of the three methods. Overall, we observed 82.2% agreement between annotators with Fleiss's Kappa (Fleiss, 1971) of 0.6. We observe that for Wikipedia articles which often have long and complex sentence structures, `PM+NLI` method often gives non-precise extractions while both `PM+CP+NLI` and `PM+NP+NLI` methods have a high precision. Table 1 shows some examples from the `PM+NP+NLI` method. The high precision of the `PM+CP+NLI` and `PM+NP+NLI` methods shows the usefulness of these weakly supervised approaches for generating high-quality collections of cause-effect pairs that are directly usable in decision support and risk management applications. We believe the lower precision of these methods over the SemEval and BECauSE datasets in our automated evaluation results in Table 8 is due to the use of shorter cause and effect phrases and sentences in these datasets, and show the need for new and more diverse datasets for evaluation. The MultiCause dataset is a step towards this direction. All outputs from the three methods along with human judgments can be found in the supplementary material.

Method	Precision
PM + NLI	44.7
PM + CP + NLI	76.7
PM + NP + NLI	80.7

Table 10: Precision of the pattern matching and NLI based methods over COVID-19 Wikipedia articles.

6 Future Work

In the future we would like to explicitly handle cases (1) in which a cause *prevents* the effect from occurring and (2) where multiple causes may lead to multiple effects. Another possible future direction is to use our pattern matching based approaches which only require text as input, to create a seed causal graph and use it to create a distantly supervised causal extractor. Finally, we are planning to explore the application of our framework in decision support and event forecasting. All our datasets and experimental results will be made publicly available.

References

N. Nangia A. Williams and S. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *NAACL-HLT*. ACL.

Anonymous. 2020. Publicly available dataset. details omitted due to double-blind reviewing.

R.J. Chapman. 2013. *Simple Tools and Techniques for Enterprise Risk Management*. Wiley finance series. Wiley.

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

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

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

Jesse Dunietz, Lori S. Levin, and Jaime G. Carbonell. 2017b. Automatically tagging constructions of causation and their slot-fillers. *TACL*, 5:117–133.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

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

Cécile Grivaz. 2010. Human judgements on causation in French texts. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Abram Handler, Matthew Denny, Hanna Wallach, and Brendan O'Connor. 2016. Bag of what? simple noun phrase extraction for text analysis. In *Proceedings of the First Workshop on NLP and Computational Social Science*. ACL.

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

Oktie Hassanzadeh, Debarun Bhattacharjya, Mark Feblowitz, Kavitha Srinivas, Michael Perrone, Shirin Sohrabi, and Michael Katz. 2019. Answering binary causal questions through large-scale text mining: An evaluation using cause-effect pairs from human experts. In *IJCAI*, pages 5003–5009.

Oktie Hassanzadeh, Debarun Bhattacharjya, Mark Feblowitz, Kavitha Srinivas, Michael Perrone, Shirin Sohrabi, and Michael Katz. 2020. Causal knowledge extraction through large-scale text mining. In *AAAI*, pages 13610–13611.

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

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

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *ICLR*.

Cheng Li and Ye Tian. 2020. Downstream model design of pre-trained language model for relation extraction task. *CoRR*, abs/2004.03786.

Zhongyang Li, Xiao Ding, Ting Liu, J. Edward Hu, and Benjamin Van Durme. 2020. Guided generation of cause and effect. In *IJCAI*, pages 3629–3636.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

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

S. Muthiah et al. 2016. Embers at 4 years: Experiences operating an open source indicators forecasting system. In *KDD*.

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

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

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

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *EMNLP-IJCNLP*. ACL.

M. Sap, R. LeBras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, and Y. Choi. 2018. ATOMIC: An atlas of machine commonsense for if-then reasoning. *CoRR*, abs/1811.00146.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *ACL*, pages 2895–2905.

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

Jinghang Xu, Wanli Zuo, Shining Liang, and Xianglin Zuo. 2020. A review of dataset and labeling methods for causality extraction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1519–1531, Barcelona, Spain (Online). International Committee on Computational Linguistics.