Toward Fine-grained Causality Reasoning and CausalQA

Anonymous ACL submission

Abstract

Understanding causality is key to the success of NLP applications, especially in high-stakes domains. Causality comes in various perspectives such as *enable* and *prevent* that, despite their importance, have been largely ignored in the literature. This paper introduces a first-of-its-kind, fine-grained causal reasoning dataset that contains seven causal relations and defines a series of NLP tasks, from causality detection to event causality extraction and causal reasoning. Our dataset contains human annotations of 25K cause-effect event pairs and 24K question-answering pairs within multi-sentence samples, where each can contain multiple causal relationships. Through extensive experiments and analysis, we show that the complex relations in our dataset bring unique challenges to state-of-the-art methods across all three tasks and highlight potential research opportunities, especially in developing "causal-thinking" methods.

1 Introduction

Causality has received much research attention in recent years (Gao et al., 2019a; Schölkopf et al., 2021; Feder et al., 2021; Scherrer et al., 2021). It has been shown that causal reasoning entails a new goal of building more powerful AI systems beyond making predictions using statistical correlations (Kaushik et al., 2021; Srivastava et al., 2020; Li et al., 2021). In particular, understanding fine-grained causal relations between events in a document is an important step in language understanding and is beneficial to various NLP applications – information extraction, question answering, and machine reading comprehension, especially in high-stakes domains such as medicine and finance.

Much work has been done on detecting a shallow "cause" relationship automatically for text (Khoo et al., 1998; Mirza et al., 2014; Chang and Chen, 2019; Mariko et al., 2020b). However, a single "cause" relationship cannot cover a plethora of causal concepts in the real-world scenarios reported by the previous psychology research (Talmy, 1988; Wolff et al., 2005). For example, the spread of COVID-19 has *led to* the boom in online shopping – i.e., (cause) – but it also has *deterred* – i.e. (prevent) – people from going shopping-centres. According to classical psychology (Wolff and Song, 2003), it is important to understand possible fine-grained relationships between two events from three different causal perspectives, including *cause*, *enable*, and *prevent*.

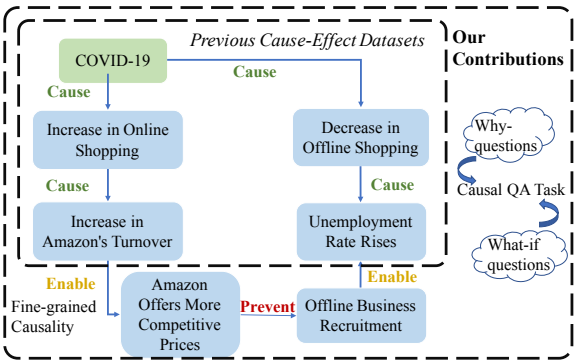


Figure 1: Comparisons between existing datasets and our dataset in the task of event causality analysis.

In this paper, we construct a large-scale, hand-labeled, fine-grained causal reasoning (FineCR) dataset in the financial domain. A contrast between our dataset and the previous causality detection dataset is shown in Fig. 1. As can be seen, given the same passage "*COVID-19 has accelerated change in online shopping, and given Amazon's ... it will result in economic returns for years to come and offering more competitive prices compared to an offline business that brings pressures for the offline business recruitment.*", previous work can extract facts such as "COVID-19 causes an increase in online shopping", yet cannot detect the subsequence for Amazon to "offer more competitive prices", and further the negative influence on offline business recruitment, both of which can be valuable for predicting the future events.

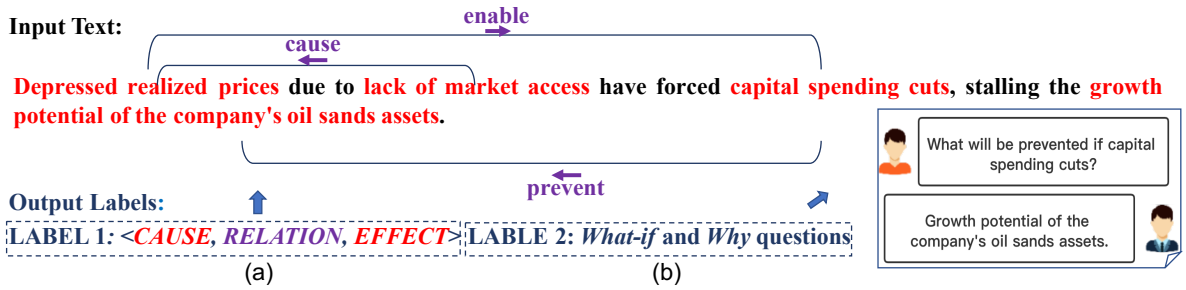


Figure 2: Illustration of our crowdsourcing tasks using an example that contains all three types of causal relationship.

Our dataset can also potentially benefit downstream applications such as financial analysis (Ding et al., 2015) and BioNLP (Demner-Fushman et al., 2021). As one practical application, we investigate the utility of our dataset for the causal question reasoning (CausalQA) task (Oh et al., 2016) in this work. The resulting dataset, FineCR, consists of 25, 193 cause-effect pairs and 24, 486 question-answering pairs, in terms of almost all questions involving "why" and "what-if" scenarios belonging to three fine-grained causalities. To establish the benchmark performance on FineCR, which consists of causality detection, fine-grained causality extraction, and CausalQA tasks, we explore several state-of-the-art neural models. Experimental results show a significant gap between machine and human ceiling performance (74.1% vs. 90.53% accuracy in fine-grained classification). To the best of our knowledge, FineCR is the first human-labeled fine-grained event causality dataset, and we define a novel CausalQA task based on that.

2 Dataset

We collected a financial analyst report dataset from Yahoo Finance[1], which contains 6,786 well-processed articles between December 2020 and July 2021. Each instance corresponds to a specific financial analyst report on a U.S. listed company, which highlights the financial strengths and weaknesses of the company business.

2.1 Crowdsourcing

The original FineCR dataset consists of 6,786 articles in 54, 289 sentences. We employ editors from a crowd-sourcing company to complete several human annotation tasks. Several pre-processing steps required crowd-sourcing efforts were carried out to prepare the raw dataset, including (1) A binary classification task for the causality detection; (2) Mark the cause and effect formatted as text chunks (a given instance may contain multiple causal relations), and give event pairs a fine-grained causality label, including *cause*, *cause_by*), *enable*, *enable_by*), *prevent*, *prevent_by*), and *irrelevant* relation, where the suffix *"_by"* means the effect comes before its cause; (3) Generating the following question-answering dataset by using the labeled event triples.

[1]We have received the written consent from the Yahoo Finance.

Metric	Counts
Causality Sentence Classification (**Task1**)	
#Positive Instances	21,046
#Negative Instances	29,979
#Multi-sentence Samples	846
#Average Token Length of POS Samples	42.8
#Average Token Length of NEG Samples	41.3
Cause-Effect Event Pairs (**Task2**)	
#Causal Text Chunks	45, 710
#Uni-causal Text Spans	18, 457
#Multi-causal Text Spans	3, 017
#Average Token Length of Cause Spans	16.0
#Average Token Length of Effect Spans	15.2
CausalQA Pairs (**Task3**)	
#Total Number of QA pairs	24, 486
#Average Token Length of Context	191.7
#Average Token Length of Questions	20.1
#Average Token Length of Answers	15.4
#Variance of Answer Length	69.15

Table 1: Statistics for causality detection, cause-effect pairs and QA pairs.

Causality Detection. We first focus on a binary classification task of the causality detection, as such, removed sentences with outcome types of non-causal relationships, leaving only those text sequences (one or two sentences) that are considered containing at least a causal relation.

Fine-grained Event Causality. Given the sentences each containing at least one event causality, human annotators are required to highlight all the event causalities and give each instance a fine-grained label. As shown in Fig. 2(a), a single sentence can have more than one event causality,

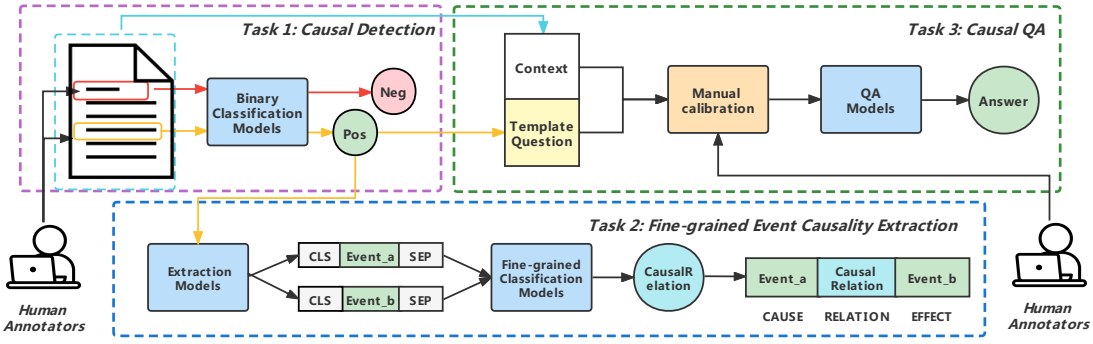


Figure 3: The pipeline of experiments based on the FineCR dataset.

which will be stored as triples containing ⟨ *cause, relation, effect* ⟩.

CausalQA. As shown in Fig. 2(b), we design a novel and challenging causal reasoning QA task based on the fine-grained causality labels. We expand each $< cause, relation, effect >$ triple for generating a plausible question-answer pair. Different templates have been designed for different types of questions. For example, the active causal relations – *CAUSE*, *ENABLE*, and *PREVENT* – could usually be used for generating why-questions while the corresponding passive causal relations could be used for generating what-if questions.

Quality Control. To ensure high quality, we restricted the participants to experienced human labelers with relevant records. For each task, we conducted pilot tests before the crowd-sourcing work officially began, receiving feedback from quality inspectors and revising instructions accordingly. We filter out the sentences regarding the estimation of the stock price movement due to the naturally high-sensitive features and uncertainty of the complex financial market. After the first-round annotation (half of the data), we manually organized spot checks for 10% samples in the dataset and revised the incorrect labels. After review, we revised roughly 3% of instances and refused the labelers with above 10% error rate from participating in the second-round data annotation. Finally, the inter-annotators agreement ratio is 91% for fine-grained causality labels, and the F1 score of the inter-annotators agreement ratio is 0.94 for causal question-answer pairs.

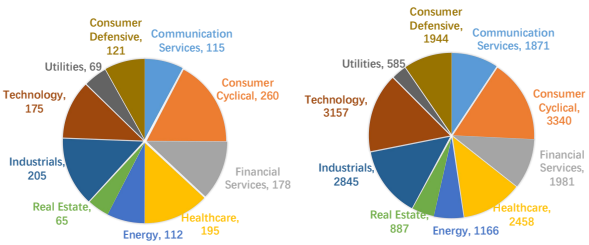
Finally, we obtained a dataset of 51,025 instances (21,046 contain at least one causal relation) with fine-grained labels of cause-effect relations that were subsequently divided into training, validation, and testing sets for the following experiments. It may be worth noting that we sort the dataset in chronological order because the future data is not expected to be used for predictions.

2.2 Discussion

The primary data statistic of the FineCR dataset is shown in Table 1 for three different tasks. We observe that there is no significant difference in the average token numbers between positive and negative examples for Task 1 and 2, which shows that predictive models are difficult to learn from shortcut features (Sugawara et al., 2018, 2020; Lai et al., 2021) (e.g., the instance length) during the training process. Furthermore, our dataset contains 846 multi-sentence samples, and 3,017 text chunks contain more than one causal relation in one instance, which requires a complex reasoning process to get the correct answer, even for a human. Most importantly, unlike other QA datasets (Sugawara et al., 2018, 2020) that can easily benefit from the test-train overlap as revealed by (Lewis et al., 2021a; Liu et al., 2021; Wang et al., 2021), our dataset is sorted in chronological order so that the future test data could be theoretically difficult to coincide with the training set. This allows us to obtain greater insight into what extent models can actually generalize.

2.3 Meta-information

Our dataset contains multi-sentence instances with fine-grained causality labels and the meta-information (company names and published dates). As shown in Fig. 4, we list the number of financial documents from different sectors, where the top three largest sectors belong to Consumer Cyclical, Industrial, and Technology. In contrast, companies from the Utilities are the smallest group in our dataset. The use of the meta-information is two-fold. First, we choose the top three largest domains to perform out-of-domain evaluations (see



(a) Companies Count (b) Reports Count

Figure 4: Sector distributions on companies and reports.

Appendix A). Second, company names would be used for generating question templates. Besides, the meta-information is crucial for benefiting the potential applications in NLP related to the domain of Finance (Tang et al., 2021; Chen et al., 2021).

3 Tasks and Methods

The pipeline of our experiments is shown in Fig. 3. We define three tasks on our FineCR dataset and build strong benchmark results for each task. First, as a prerequisite, models are evaluated on a binary classification task to predict whether a given text sequence contains a causal relation (**Task 1**). Second, we set up a joint event extraction and fine-grained causality task for identifying text chunks describing the cause and effect, respectively, and which fine-grained causality category it belongs to (**Task 2**). Finally, we design cause-effect question answering (**Task 3**).

3.1 Data Settings

For hold-out evaluation, we split our dataset into mutually exclusive training/validation/testing sets in the same ratio of 8:1:1 for all tasks. Predictive models and data splitting strategies have been kept the same among these tasks for building the benchmark results of each task. In line with the best practice, model hyper-parameters are tuned using the validation set. Both validation results and testing results will be reported in experiments.

3.2 Models

We consider using both classical deep learning models – CNN-Test (Kim, 2014) and HAN (Yang et al., 2016) – and Transformer-based models downloaded from Huggingface[2] – BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and SpanBERT (Joshi et al., 2020) – as predictive models.

In addition, we perform a causal reasoning QA task by leveraging six Transformer-based pre-trained models provided by Huggingface (Wolf et al., 2020) on our dataset, including BERT-base, BERT-large, RoBERTa-base, RoBERTa-large, RoBERTa-base-with-squad, and RoBERTa-large-with-squad[3]. Furthermore, pre-trained seq2seq models such as T5 (Raffel et al., 2020), or BART (Lewis et al., 2020) are fine-tuned on QA-pairs as the benchmark methods of the generative QA tasks. In particular, we consider T5-small, T5-base, T5-large, BART-base, and BART-large models for building the benchmark results.

We use Adam as the optimizer and adopt the trick of decay learning-rate with the steps increase to train our model until converging for all models.

3.3 Methods

Our methods are built on the recently advanced Transformer architectures (Vaswani et al., 2017a) with the framework provided by Huggingface[4]. As follows, we introduce the detailed implementation of deep neural methods on three tasks completed on the FineCR dataset.

3.3.1 Causal Detection

The dataset consists of instances labeled with positive for the binary classification task if a given instance contains one causal relation and negative non-causal instances. The input data is extracted from the raw dataset directly, which contains 846 multi-sentence samples. We include multi-sentence samples besides a single sentence because causality could be found in multi-sentence contexts.

3.3.2 Fine-grained Event Causality Extraction

We use [CLS] and [SEP] to mark the event's begin and end positions, respectively. For example, we have "[CLS] Better card analytics, increased capital markets and M&A offerings, and bolt-on acquisitions [SEP] should help drive [CLS] growth in fee income [SEP]", in which "Better card analytics, increased capital markets and M&A offerings, and bolt-on acquisitions" is a "cause" event and "growth in fee come" is an "effect" event. In total, there are 33, 634 event triples with seven-class labels used for our experiments.

Then, we conduct the cause-effect extraction between samples. We consider cause-effect extrac-

[2] https://github.com/huggingface/models

[3] https://huggingface.co/navteca/roberta-large-squad2

[4] https://github.com/huggingface/transformers

Methods	Dev.		Test	
	F1	Acc	F1	Acc
CNN-Text	81.35	81.59	80.03	81.01
HAN	81.18	81.23	80.60	81.26
BERT-base	83.72	84.23	84.02	84.43
BERT-large	84.03	84.41	84.63	84.90
SpanBERT-base	84.09	84.38	84.51	84.72
SpanBERT-large	84.43	84.80	84.55	84.82
RoBERTa-base	**84.59**	**85.16**	84.31	84.76
RoBERTa-large	84.39	84.75	**84.64**	**84.89**
Human	-	-	94.32	95.94
Best Results @FinCausal	-	-	97.75	97.76

Table 2: The results of the causal sentence classification. 'F1' refers to the Macro F1. 'ACC.' is short for the accuracy.

tion a multi-span event extraction task as complex causal scenarios containing multiple causes or effects within a single instance is under consideration. We set the label for the first token of a cause or effect to "B", the rest of the tokens within the detected text chunks are given the label "I", and the other words in a given instance are set to "O". The results of event causality extraction are reserved for generating causal questions.

3.3.3 CausalQA

Both extractive methods and generative methods have been evaluated. For the extractive QA task, we adopt the same methods as the previous Transfomer-based QA works (Kayesh et al., 2020). In particular, we first convert the context $C = (c_1, c_2, ..., c_l)$ and question $Q = (q_1, q_2, ..., q_{l'})$ into a single sequence $X = [CLS]\ c_1 c_2 \dots c_l\ [SEP]\ q_1 q_2 \dots q_{l'}$[SEP], passing it to the pre-trained Transformer encoders for predicting the answer span boundary (start and end).

3.4 Metrics

The F1-score and accuracy are used for evaluating the event causality analysis task, and the exact match and F1-score are used for CausalQA.

The Macro F1-score is defined as the mean of label-wise F1-scores:

$$\text{Macro F1-score} = \frac{1}{N}\sum_{i=0}^{N} \text{F1-score}_i \quad (1)$$

where i is the label index and N is the number of classes.

4 Results and Discussion

We present the results of **Tasks 1-3** based on our dataset in this section.

4.1 Causality Detection

The causal detection result is shown in Table 2. We find that although Transformer-based methods achieve much better results than other methods – CNN and HAN using ELMO embeddings – on judging whether an instance contains at least a causal relationship (RoBERTa-Large can get the highest F1 Score – 84.64), it is still significantly below the human performance (84.64 vs. 94.32). The results of human performance are reported by quality inspectors from the crowdsourcing company. It is worth noting that the best results on the FinCausal (Mariko et al., 2020b) dataset can reach the human-level result (F1 = 97.75), providing indirect evidence that our dataset is more challenging caused by more complex causality instances.

4.2 Fine-grained Event Causality Extraction

The results of the fine-grained event causality extraction task are shown in Table 3. We find that SpanBERT and RoBERTa model can achieve the best performance for event causality extraction (F1 = 86.82 and EM = 60.26) and fine-grained classification (F1 = 68.99 and EM = 74.09), respectively. Nevertheless, all methods perform dramatically worse on the more challenging joint task, where the prediction is judged true only if event extraction and classification results exactly match the ground truth. Although the SpanBERT-large model can achieve the highest 21.78 EM on the test set, there is still much room for improvement.

We find that the large Transformer-based models (Vaswani et al., 2017b) with larger parameter sizes could not improve the performance on these tasks based on the FineCR dataset by comparing the test performance of BERT-base (**63.72** in F1, **71.72** in ACC) with BERT-large (60.24 in F1, 69.85 in ACC) on the task of the fine-grained classification. It sheds new light that increasing the parameter size could not be helpful for causal reasoning tasks.

A more detailed error analysis by using the best-performed RoBERTa-Large model is given in Table 4. The model performs well in terms of the F1 score when predicting simple causal relations – Irrelevant (84.40), Cause (74.00), Cause_by (79.62), and Prevent (76.34). In contrast, complex relations – Enable (62.61) and Enable_by (41.49) – and the category with few examples – Prevent_by (64.46) – are not well predicted.

Model	Event Causality Extraction				Fine-grained Classification				Joint Evaluation	
	Dev.		Test		Dev.		Test		Dev.	Test
	F1	EM	F1	EM	F1	ACC	F1	ACC	EM	EM
BERT-base	84.37	51.48	85.30	53.53	71.74	70.43	63.72	71.72	21.21	20.15
BERT-large	85.13	50.34	86.93	52.88	70.90	64.16	60.24	69.85	17.54	21.73
RoBERTa-base	85.67	53.41	86.32	56.04	73.09	68.37	65.99	71.63	20.45	19.08
RoBERTa-large	85.12	54.70	85.95	56.77	**74.54**	**71.99**	**68.99**	**74.09**	20.46	19.77
SpanBERT-base	**85.84**	55.40	**86.82**	57.26	71.18	68.40	63.73	70.52	21.17	21.09
SpanBERT-large	85.50	**57.40**	86.33	**60.26**	73.65	68.15	64.43	72.93	**23.01**	**21.78**
Human Performance	-	-	94.32	81.34	-	-	88.61	90.53	-	-

Table 3: The results of the joint event causality detection (task2), 'F1' refers to the Macro F1. 'ACC.' is short for the accuracy, 'EM' refers to exact match and spe.

Category	Counts	Dev.		Test	
		F1	Acc	F1	Acc
Irrelevant	8,441	84.17	86.49	**84.40**	**85.55**
Cause	8,428	73.60	73.93	74.00	76.61
Cause_By	7,437	80.21	84.94	79.62	83.69
Enable	5,506	63.42	60.91	62.61	58.70
Enable_By	2,367	47.66	41.06	*41.49*	*35.68*
Prevent	1,086	79.79	71.43	76.34	67.87
Prevent_By	369	55.88	52.78	64.46	65.00

Table 4: Error analysis for fine-grained classifications.

CausalQA	Dev.		Test	
	F1	EM	F1	EM
BERT-base	79.90	55.52	79.33	55.70
BERT-large	82.48	59.24	82.37	58.71
RoBERTa-base	82.96	60.13	83.11	60.33
SQuAD2.0-only	*64.87*	*26.71*	*65.20*	*27.36*
SQuAD2.0-enhanced	84.39	61.22	84.34	61.17
RoBERTa-large	84.28	**61.69**	84.35	**61.76**
SQuAD2.0-only	*63.99*	*26.02*	*63.82*	*25.26*
SQuAD2.0-enhanced	**84.65**	61.63	**84.65**	61.58
Generative Methods				
BART-base	74.34	35.81	74.35	36.16
BART-large	65.52	27.24	65.70	26.48
T5-small	75.98	42.31	76.40	41.61
T5-Large	**81.95**	**48.17**	**81.77**	**47.43**

Table 5: The results of causal reasoning QA using both extractive methods and generative methods. 'SQuAD2.0' refers to the evaluation results using the model trained with the training set of SQuAD2.0[5] only.

4.3 CausalQA

We provide both quantitative analysis and qualitative analysis for CausalQA. In addition, we compare the best performance on our dataset and other popular QA datasets.

4.3.1 Quantitative Analysis

The results of CausalQA are given in Table 5, where the bold values indicate the best performance while the italic values show the results of transfer learning methods trained by the SQuAD2.0 training data only. We find that the best-performing generative model – T5-Large – can achieve comparable results with the RoBERTa-large in terms of the F1 (81.77 vs. **84.35**). Meanwhile, the average EM of generative methods is largely below the extractive methods using the same training data. Second, the results of models trained with SQuAD2.0 data are much worse than those models trained with the original FineCR training set in terms of the F1 score (65.20 vs. **83.11** for RoBERTa-base and 63.82 vs. **84.35** for RoBERTa-large). On the other hand, we note a distinct improvement of using SQuAD2.0 data for initially training for both RoBERT-base (from 83.11 to **84.34**) and RoBERTa-large (from 84.35 to **84.65**), which indicates that the training with additional well-labelled data could bring significant benefits for CausalQA. This may hint that the current QA data sources are still helpful for improving the performance of the causal reasoning QA task, although further research is required, as to what extent models can actually benefit from the additional data for the generalization is hard to be evaluated.

4.3.2 Qualitative Analysis for Answers

Table 6 presents a qualitative analysis for CausalQA, where we highlight the question and answer parts extracted from the raw context. Human labelers label the gold answers while the BERT-based model generates the output answers. The first three questions are answered correctly by the model, while the last two instances show two typical patterns prone to errors. In the first incorrect example, the model outputs "*targeted marketing*" using the keyword "*through*" but fails to give the gold answer "*analyzing the data and applying artificial intelligence*". This could be because the model fails to identify the difference between the same word appearing in two different positions. The last example shows that the model tends to output the answer closer to the question in the context instead of observing the whole sentence. The real reasons – "*equity and credit markets*" and "*Brexit*"

Context	Question	Gold Answer	Output Answer
(Relation: Cause) Amazon's 2017 purchase of Whole Foods remains a threat ... The COVID-19 outbreak has lifted near-term revenue as shoppers spend more time at home.	Why **the COVID-19 outbreak has lifted near-term revenue** for **Amazon**?	**Shoppers spend more time at home**	Shoppers spend more time at home
(Relation: Enable) As a first mover in the local-market daily deals space, Groupon has captured a leadership position, but not robust profitability.	What enable **Groupon capture a leadership position**?	**A first mover in the local-market daily deals space**	A first mover in the local-market daily deals space
(Relation: Prevent_By) In neurology, RNA therapies can reach their intended targets via intrathecal administration into spinal fluid, directly preventing the production of toxic proteins	What will be prevented if **intrathecal administration into spinal fluid**?	**The production of toxic proteins**	The production of toxic proteins
Examples of Incorrect Predictions			
(Relation: Enable_By) ... Through analyzing the data and applying artificial intelligence, the advertisers can improve the efficiency of advertisements through targeted marketing for Tencent ...	What can help **advertisers to improve the efficiency of advertisements**?	**Analyzing the data and applying artificial intelligence**	Targeted marketing
(Relation: Cause_By) Given expectations for more volatile equity and credit markets, as well as some disruption as Brexit moves forward, it remain doubtful that flows will improve too dramatically, a negative 3%-5% annual organic growth...	Why **a negative 3%-5% annual organic growth** happened?	**Given expectations ... as well as some disruption as Brexit moves forward**	It remains doubtful that flows will improve too dramatically.

Table 6: Qualitative analysis of "Why" and "What-if" questions answering tasks based on the best-performed RoBERTa-Large model. The **company name** can be found in the meta-information of our dataset. **Cause** and **Effect** are extracted from the original context. The inputs of models consist with the context and question.

Dataset	Method	F1	ACC	EM
SQuAD1.1 (Rajpurkar et al., 2016)	LUKE (Yamada et al., 2020)	95.7	-	90.6
SQuAD2.0 (Rajpurkar et al., 2018)	IE-Net (Gao et al., 2019b)	93.2	-	90.9
DROP (Dua et al., 2019)	QDGAT (Chen et al., 2020)	88.4	-	-
HotpotQA (Yang et al., 2018)	BigBird-etc (Zaheer et al., 2020)	95.7	-	90.6
Reasoning Based Datasets				
LogiQA(Liu et al., 2020)	DAGAN (Huang et al., 2021)	-	39.3	-
CausalQA (Ours)	RoBERTa-SQuAD	84.7	85.6	61.6

Table 7: The comparison of best performance between our dataset and other popular QA datasets.

– are ignored as it is relatively away for the question position.

4.3.3 Challenges by CausalQA

We are interested in better understanding the difficulty of the CausalQA task compared to other popular datasets regarding prediction performance. We list the best-performing model of several popular datasets in Table 7. In general, we find that reasoning-based tasks are more complex than other tasks in terms of the relatively low accuracy achieved by the state-of-the-art method. LogiQA is more challenging than our dataset (39.3 vs. 85.6 in accuracy) because it requires heavy logical reasoning rather than identifying causal relations from text. Moreover, we find that the state-of-the-art result on our dataset (RoBERTa-SQuAD) is dramatically worse than the best performance on other datasets (EM = 90.9 on SQuAD2.0 while EM = 61.6 on CausalQA). This may suggest that the model tends to output the partially right answer but fails to output the utterly correct answer, although further research is required, as the model still could be easily perturbed by the length of an event. Meanwhile, the human performance is still ahead of the best-performing model's result in the causal reasoning QA task. Thus, we argue that CausalQA is worth investigating by using more "*causal-thinking*" methods in the future.

5 Related Work

This paper brings together two interesting ideas – event causality and causal question answering –and in what follows, we briefly introduce the existing relevant works and datasets to the present work.

Event Causality. There is a deep literature on causal inference techniques using non-text datasets (Pearl, 2009; Morgan and Winship, 2015; Keith et al., 2020; Feder et al., 2021), and a line of work focusing on discovering the causal relationship between events from textual data (Gordon et al., 2012; Mirza and Tonelli, 2016; Du et al., 2021).Previous efforts lie on the graph-based event causality detec-

Datasets	Event Extraction	Causal Reasoning	Fine-grained Causality	Span-based QA	Sources
FinCausal (Mariko et al., 2020b)	✔	✔	✗	✗	Finance
COPA (Roemmele et al., 2011)	✗	✔	✗	✗	Open
SQuAD (Rajpurkar et al., 2018)	✗	✗	✗	✔	Wikipedia
LogiQA (Liu et al., 2020)	✗	✔	✗	✗	Examination
HotpotQA (Yang et al., 2018)	✗	✗	✗	✔	Wikipedia
DROP (Dua et al., 2019)	✗	✗	✗	✔	Wikipedia
DREAM (Sun et al., 2019)	✗	✔	✗	✗	Examination
RACE (Lai et al., 2017)	✗	✔	✗	✗	Examination
FineCR (Ours)	✔	✔	✔	✔	Finance

Table 8: Comparisons of our fin-grained causal reasoning dataset and related public datasets.

tion tasks (Tanon et al., 2017; Li et al., 2020; Du et al., 2021) and the event-level causality detection tasks (Mariko et al., 2020a; El-Haj et al., 2021; Gusev and Tikhonov, 2021). However, causal reasoning for text data with a special focus on fine-grained causality between events has been relatively little considered. For this reason, we build a fine-grained causality dataset in the financial domain and expect to see whether the state-of-the-art models can achieve human-like accuracy on several causal reasoning tasks, and if not, to what extent.

Datasets. Table 8 compares our dataset with datasets in the domain of both event causality and question answering (QA). FinCausal (Mariko et al., 2020a) dataset is the most relevant to ours, which developed a relatively small dataset from the Edgar Database[6] focusing on the simple "cause" relation only and do not contain QA tasks. In addition, existing popular question answering datasets (Sun et al., 2019; Liu et al., 2020; Cui et al., 2020) mainly focus on *what*, *who*, *where* and *when* questions, making their usage scenarios somewhat limited. SQuAD (Rajpurkar et al., 2016, 2018) consists of factual questions concerning Wikipedia articles, and some unanswerable questions are involved in SQuAD2.0. Although there are some datasets contain the causal reasoning tasks (Lai et al., 2017; Sun et al., 2019; Cui et al., 2020), none of them consider answering questions by text span. Span-based question answering problems have gained wide interest in recent years (Yang et al., 2018; Huang et al., 2019; Lewis et al., 2021b). HotpotQA (Yang et al., 2018) focuses on multi-hop QA where the question can only be answered through analyzing multiple documents. The answers in the (Dua et al., 2019) may come from different spans of a passage and require some combination technologies to get the correct answer. Compared with these datasets, that none of them have features of causal reasoning and span-based QA simultaneously, our dataset is the first to leverage fine-grained human-labeled causality for designing the CausalQA task consisting with "Why" and "What-if" questions.

Our task is similar to the machine reading comprehension setting (Huang et al., 2019) where the algorithms make a multiple-choice selection given a passage and a question. Nevertheless, we focus on causal questions, which turn out to be more challenging. To the best of our knowledge, we are the first to evaluate models on the event causality analysis and causal question answering (CausalQA) tasks based on the fine-grained causality dataset. We will release our code and dataset on Github.

6 Conclusion

We explored the efficacy of current state-of-the-art methods for causal reasoning tasks by considering a novel fine-grained reasoning setting and developing a dataset with rich human labels. Experimental results using the state-of-the-art neural language models provide the evidence that there is still much room for improvement on causal reasoning tasks, where there is a need for designing better solutions to correlation discovery related to event causality analysis and Why/What-if QA tasks.

7 Ethical Statement

This paper honors the ACL Code of Ethics. Public available financial analysis reports are used to extract fine-grained event relationships. No private data or non-public information was used. We obtain permission from Yahoo Finance for non-profit research. All annotators have received labor fees corresponding to their amount of annotated corpus.

[6] https://www.sec.gov/edgar/searchedgar/

References

Ting-Yun Chang and Yun-Nung Chen. 2019. What does this word mean? explaining contextualized embeddings with natural language definition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6064–6070, Hong Kong, China. Association for Computational Linguistics.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. *From Opinion Mining to Financial Argument Mining*. Springer Briefs in Computer Science. Springer.

Kunlong Chen, Weidi Xu, Xingyi Cheng, Zou Xiaochuan, Yuyu Zhang, Le Song, Taifeng Wang, Yuan Qi, and Wei Chu. 2020. Question directed graph attention network for numerical reasoning over text. *arXiv preprint arXiv:2009.07448*.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. Mutual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Conference of the Association for Computational Linguistics*. Association for Computational Linguistics.

Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii, editors. 2021. *Proceedings of the 20th Workshop on Biomedical Language Processing*. Association for Computational Linguistics, Online.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2015. Deep learning for event-driven stock prediction. In *Proceedings of the 24th International Conference on Artificial Intelligence*, page 2327–2333, Buenos Aires, Argentina.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2021. ExCAR: Event graph knowledge enhanced explainable causal reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2354–2363, Online. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*.

Mahmoud El-Haj, Paul Rayson, and Nadhem Zmandar, editors. 2021. *Proceedings of the 3rd Financial Narrative Processing Workshop*. Association for Computational Linguistics, Lancaster, United Kingdom.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. 2021. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *arXiv preprint arXiv:2109.00725*.

Lei Gao, Prafulla Kumar Choubey, and Ruihong Huang. 2019a. Modeling document-level causal structures for event causal relation identification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1808–1817, Minneapolis, Minnesota. Association for Computational Linguistics.

Yuan Gao, Zixiang Cai, and Lei Yu. 2019b. Intra-ensemble in neural networks. *arXiv preprint arXiv:1904.04466*.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. Semeval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.

Ilya Gusev and Alexey Tikhonov. 2021. Headlinecause: A dataset of news headlines for detecting casualties. *ArXiv*, abs/2108.12626.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos qa: Machine reading comprehension with contextual commonsense reasoning. *arXiv preprint arXiv:1909.00277*.

Yinya Huang, Meng Fang, Yu Cao, Liwei Wang, and Xiaodan Liang. 2021. Dagn: Discourse-aware graph network for logical reasoning. *arXiv preprint arXiv:2103.14349*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Divyansh Kaushik, Amrith Setlur, Eduard Hovy, and Zachary C Lipton. 2021. Explaining the efficacy of counterfactually augmented data. In *International Conference on Learning Representations*.

Humayun Kayesh, Md. Saiful Islam, Junhu Wang, Shikha Anirban, A.S.M. Kayes, and Paul Watters. 2020. Answering binary causal questions: A transfer learning based approach. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9.

Katherine Keith, David Jensen, and Brendan O'Connor. 2020. Text and causal inference: A review of using text to remove confounding from causal estimates. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5332–5344.

Christopher SG Khoo, Jaklin Kornfilt, Robert N Oddy, and Sung Hyon Myaeng. 1998. Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. *Literary and Linguistic Computing*, 13(4):177–186.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Yuxuan Lai, Chen Zhang, Yansong Feng, Quzhe Huang, and Dongyan Zhao. 2021. Why machine reading comprehension models learn shortcuts? *arXiv preprint arXiv:2106.01024*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 2021a. Question and answer test-train overlap in open-domain question answering datasets. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1000–1008.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021b. Paq: 65 million probably-asked questions and what you can do with them. *arXiv*.

Zhongyang Li, Xiao Ding, Kuo Liao, Ting Liu, and Bing Qin. 2021. Causalbert: Injecting causal knowledge into pre-trained models with minimal supervision. *ArXiv*, abs/2107.09852.

Zhongyang Li, Xiao Ding, Ting Liu, J. Edward Hu, and Benjamin Van Durme. 2020. Guided generation of cause and effect. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3629–3636. International Joint Conferences on Artificial Intelligence Organization. Main track.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*.

Linqing Liu, Patrick Lewis, Sebastian Riedel, and Pontus Stenetorp. 2021. Challenges in generalization in open domain question answering. *arXiv preprint arXiv:2109.01156*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv e-prints*, page arXiv:1907.11692.

Dominique Mariko, Hanna Abi-Akl, Estelle Labidurie, Stephane Durfort, Hugues De Mazancourt, and Mahmoud El-Haj. 2020a. The financial document causality detection shared task (FinCausal 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 23–32, Barcelona, Spain (Online). COLING.

Dominique Mariko, Estelle Labidurie, Yagmur Ozturk, Hanna Abi Akl, and Hugues de Mazancourt. 2020b. Data processing and annotation schemes for fincausal shared task. *arXiv preprint arXiv:2012.02498*.

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. Annotating causality in the tempeval-3 corpus. In *EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 10–19. Association for Computational Linguistics.

Paramita Mirza and Sara Tonelli. 2016. Catena: Causal and temporal relation extraction from natural language texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 64–75.

Stephen L Morgan and Christopher Winship. 2015. *Counterfactuals and causal inference*. Cambridge University Press.

Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Ryu Iida, Masahiro Tanaka, and Julien Kloetzer. 2016. A Semi-Supervised Learning Approach to Why-Question Answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1).

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in Neural Information Processing Systems*, 32:13991–14002.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

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.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Nino Scherrer, Olexa Bilaniuk, Yashas Annadani, Anirudh Goyal, Patrick Schwab, Bernhard Schölkopf, Michael C Mozer, Yoshua Bengio, Stefan Bauer, and Nan Rosemary Ke. 2021. Learning neural causal models with active interventions. *arXiv preprint arXiv:2109.02429*.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634.

Megha Srivastava, Tatsunori Hashimoto, and Percy Liang. 2020. Robustness to spurious correlations via human annotations. In *International Conference on Machine Learning*, pages 9109–9119. PMLR.

Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. What makes reading comprehension questions easier? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4208–4219.

Saku Sugawara, Pontus Stenetorp, Kentaro Inui, and Akiko Aizawa. 2020. Assessing the benchmarking capacity of machine reading comprehension datasets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8918–8927.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. Dream: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Leonard Talmy. 1988. Force dynamics in language and cognition. *Cognitive science*, 12(1):49–100.

Tsun-Hsien Tang, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. Retrieving implicit information for stock movement prediction. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2010–2014, New York, NY, USA. Association for Computing Machinery.

Thomas Pellissier Tanon, Daria Stepanova, Simon Razniewski, Paramita Mirza, and Gerhard Weikum. 2017. Completeness-aware rule learning from knowledge graphs. In *International Semantic Web Conference*, pages 507–525. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017a. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017b. Attention is all you need. In *NIPS*.

Cunxiang Wang, Pai Liu, and Yue Zhang. 2021. Can generative pre-trained language models serve as knowledge bases for closed-book QA? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3241–3251, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Phillip Wolff, Bianca Klettke, Tatyana Ventura, and Grace Song. 2005. Expressing causation in english and other languages. *Journal of experimental psychology: General*.

Phillip Wolff and Grace Song. 2003. Models of causation and the semantics of causal verbs. *Cognitive psychology*, 47(3):276–332.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. Luke: deep contextualized entity representations with entity-aware self-attention. *arXiv preprint arXiv:2010.01057*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. In *NeurIPS*.

A Appendix: Out-of-domain Test

Sec	Consumer		Industrial		Technology	
	F1	Acc	F1	Acc	F1	Acc
Con	**59.69**	69.01	50.03	66.47	*47.95*	63.74
Ind	50.60	67.39	**51.14**	64.61	47.49	64.09
Tec	48.65	65.87	47.70	63.56	**50.39**	61.12

Table 9: Out-of-domain test results of the BERT-base model for the fine-grained causality classification task.

Sec	Consumer		Industrial		Technology	
	F1	EM	F1	EM	F1	EM
Con	**86.26**	49.54	85.75	48.37	85.20	47.26
Ind	84.23	49.83	**86.00**	50.68	84.90	47.45
Tec	86.24	47.99	86.03	47.50	**87.09**	61.12

Table 10: Out-of-domain test results of the Span-Large model for the cause-effect extraction task.

It has been shown that sector-relevant features from a given domain could become spurious patterns on the other domains, leading to performance decay under distribution shift (Ovadia et al., 2019). We use instances from three sectors with the largest amounts of samples in our dataset for conducting out-of-domain generalization text. These observe in line with recent works revealing that current deep neural models mostly memorize training instances yet struggle to predict on the out-of-distribution data (Gururangan et al., 2018; Kaushik et al., 2020; Srivastava et al., 2020). To evaluate whether methods can generalize on the out-of-distribution data, and to what extent, the results of the out-of-domain test are shown in Table 9 and Table 10.

In particular, the model achieves the best performance when the training and test sets are extracted from the articles of the same domain companies. In the out-of-domain test, the model shows varying degrees of performance decay for both tasks. For example, in the fine-grained causality classification task, the model trained with the data from the Consumer Cyclical domain achieves 59.69 F1 Score when testing on the Consumer Cyclical data while decreasing to 47.95 when testing on technology companies. Moreover, in the cause-effect extraction task, the model trained with the data from the Consumer Cyclical domain achieves 86.26 F1 Score when testing on itself while decreasing to 85.20 when testing on Technology. This shows that the domain-relevant patterns learned by the model cannot transfer well between domains.

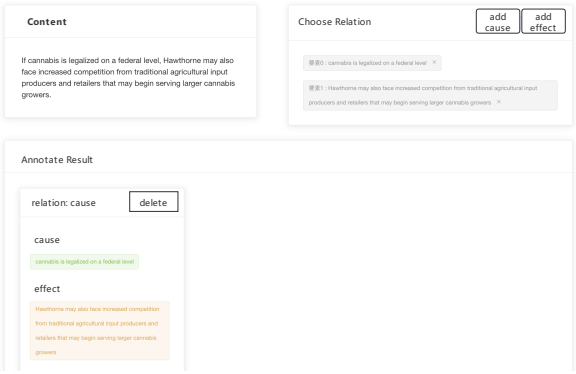


Figure 5: The annotation platform provided the crowdsourcing company for collecting annotations for fine-grained causality reasoning and CausalQA.

B Appendix: Annotation Instructions

The annotation platform used in this work is introduced in Fig. 5. As follows, we provide the detailed annotation instructions used for training the human labelers. Also, we show the annotation of some real examples stored in our dataset.

B.1 General Instruction

This is a annotation task related to event causality. In this task, you are asked to find all the cause-effect pairs and the fine-grained event relationship types from give passages.

B.2 Steps

1. Please read your assigned examples carefully.
2. A sentence is considered contains event causality if at least two events occur in it and the two events are causally related.
3. If a sentence contains causality, mark it as **Positive**; otherwise, mark it as **Negative**.
4. For the **positive** sentence, first find all the events that occur in the sentence, and then pair the events to see if they constitute a causal relationship. The relationship much be one of **Cause**, **Enable** and **Prevent**.
5. "A causes B" means B always happens if A happens. "A enables B" means A is a possible way for B to happen, but not necessarily. "A prevents B" means A and B cannot happen at the same time.
6. Remember to annotate all event causality pairs. If there is no more pairs, process to the next passage.

B.3 Examples

Here are some annotation examples, please read it before starting your annotation.

Example 1: Moreover, we do not think that DBK's investment banking operation has the necessary scale and set-up to outcompete peers globally or within Europe.

Answer: # Negative

Explanation: This is a sentence that contains no causal relationship between events.

Example 2: In our view, **customers are likely to stay with VMware** because of **knowledge of its product ecosystem as well as the risks and complexities associated with changing virtual machine providers**.

Answer: # Positive

Explanation: This is a causal sentence. There exist two events marked by **yellow** color. You should first annotate the two events and then give them the label according to their relationship, using one of **Cause**, **Enable** and **Prevent**. Here the relationship is **Cause**.

Example 3: **Depressed realized prices** due to **lack of market access** have forced **capital spending cuts**, stalling the **growth potential of the company's oil sands assets**.

Answer: # Positive

Explanation: This is a causal sentence and there exist four events. You need to mark out all four of these events and then pair them up to see if they're related. If so, determine what kind of relationship they belong to.