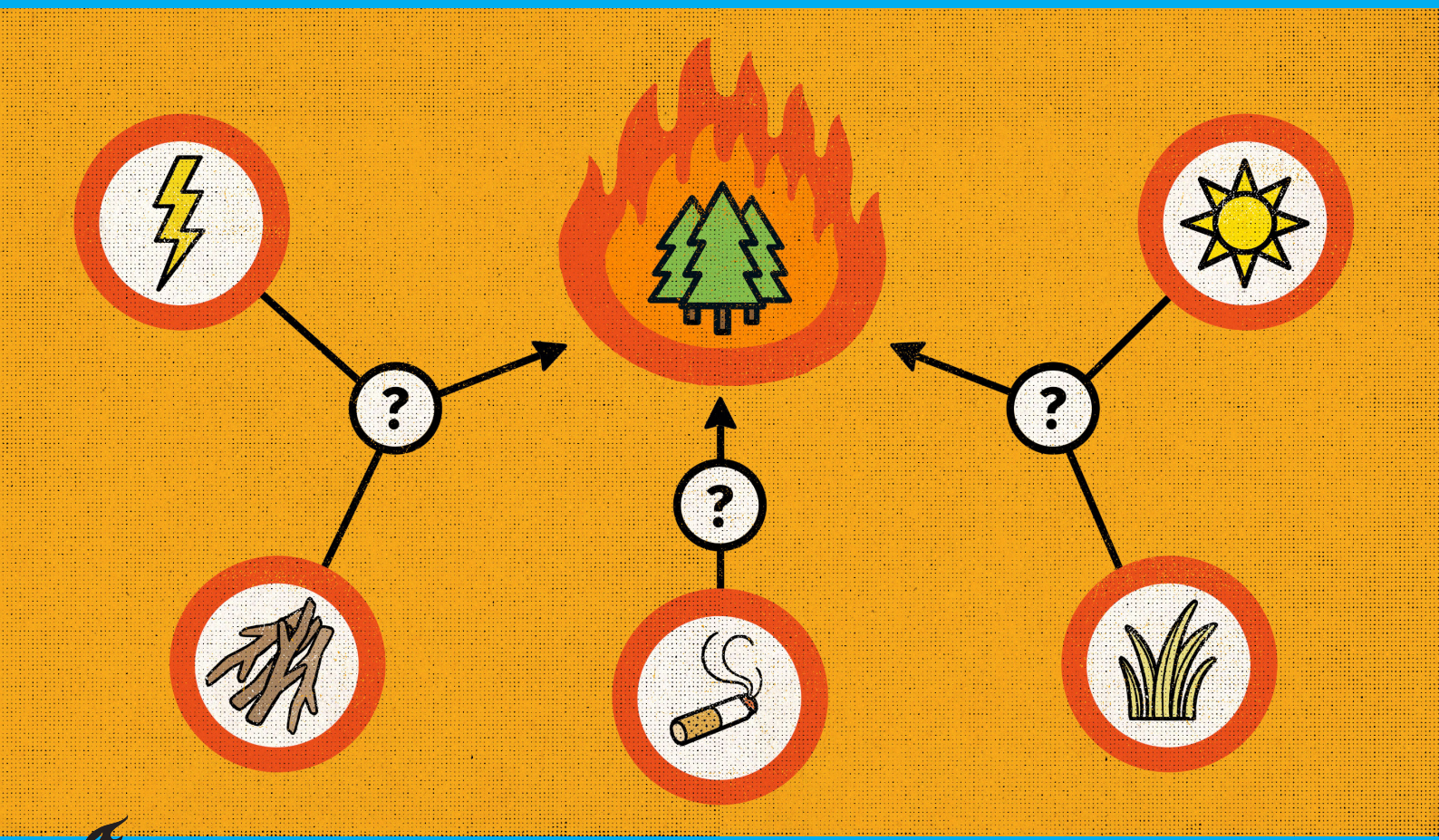
Master Thesis

Toward Fine-grained Causality Reasoning and Question Answering

Zhen Wang



TU Delft

Master Thesis

Toward Fine-grained Causality Reasoning and Question Answering

by

Zhen Wang

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Wednesday June 29, 2022 at 9:45 AM.

Student number:	5220750	
Project duration:	October 1, 2021 – June 1, 2022	
Thesis committee:	Prof. dr. ir. Geert-Jan Houben,	Web Information Systems, chair
	Dr.ir. Jie Yang,	Web Information Systems, daily supervisor
	Dr.ir. Xucong Zhang,	Computer Vision Lab, graduation committee

This thesis is confidential and cannot be made public until June 29, 2022.

An electronic version of this thesis is available at `http://repository.tudelft.nl/`.

TU Delft

Preface

This thesis represents the end of my master's study at the Delft University of Technology. It is not easy to fly thousands of kilometers from China to the Netherlands, especially to live and study in an unfamiliar country and a brand new university. However, with the help of my teachers, classmates, roommates, and all the kind people in my daily life, I enjoy a happy and meaningful two years. Here, I would like to show great thankfulness to my thesis supervisor - Prof. Geert-Jan Houben and daily supervisor - Dr. Jie Yang, for giving me the help and kind advice in my master thesis, without which I cannot make my research idea come true. And I also want to thank Peide Zhu, a Ph.D. candidate in WIS, for giving me lots of advice in writing the thesis. In addition, I want to thank all my friends in Prof. Schermerhornstraa, who helped me a lot when I first arrived in the Netherlands. Lastly, I thank my family for supporting my study out of China. Without their encouragement and financial support, I could not have completed my master's studies.

Zhen Wang
Delft, June 2022

Abstract

This thesis mainly studies the causality in natural language processing. 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. In view of the lack of a dataset that can be used for causality-related research, in this thesis, we first build a first-of-its-kind, fine-grained causal reasoning dataset - FineCR, that contains new causality relations such as enable and prevent, with the help of human annotators. 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. To study current NLP models' ability to deal with the causality-related dataset and to figure out the problems that still exist, we define a series of NLP tasks based on FineCR, including causality detection, causality event extraction and causality question answering. Our experimental results with state-of-the-art deep learning models prove that there is still much room for improvement on those causal reasoning tasks. We found that those models have different shortcomings for different tasks. For the causality detection task, current classification models are easily affected by keywords, while the model cannot accurately extract the events for the causality event extraction task. And for the causality question answering task, it is sometimes difficult for the model to find the corresponding answer due to its inability to understand the semantics well. Those discoveries indicate the need to design better solutions to event causality research. In conclusion, our novel datasets and tasks provide a challenging benchmark for evaluating models' causal ability, and the experimental results shed light on future directions for improving neural language models.

Contents

List of Figures

List of Tables

1 Introduction

Causality [66], which consists of a pair of cause and effect, is an essential way of understanding the world. The event causality represents a causal relation between two events. Both the cause and effect are an event. Causality can be established from objective natural factors, such as the rotation of the earth causing the change of day and night, or from operating rules of human society, such as disobeying traffic rules resulting in being fined by traffic police. The reasonable expectation of the possible causes and consequences of an event is essential for rational decisions, such as when oil is in short supply, the price of oil will go up, thus more and more people turn to electric cars. Sometimes we need to find the cause of when one thing has happened, thus we can take corresponding measures before the next thing happens. For example, a fire broke out in a house, and the investigator found this was because of the lack of a smoke alarm. To prevent the fire next time, we need to install smoke alarms in each room.

Causality in natural language processing (NLP) has received much research attention in recent years [21, 22, 74, 75]. It has been shown that causal reasoning entails a new goal of building more powerful AI systems beyond making predictions using statistical correlations [37, 52, 78]. However, it is still challenging for current deep learning models to conduct causal inference between real-world events due to their complexity. Yet, there is huge potential for machines to be able to understand and reason about causality in text. All the written information recorded by our humans from ancient to the present contains many causal facts. If we can extract and leverage valid causal relationships from these document texts, we can then use them to build a huge causal reasoning graph that can improve performance of various of NLP applications. Therefore, understanding causal relations between events in a document is an essential 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 from text [40, 58, 61]. However, a single "cause" relationship cannot cover a plethora of causal concepts in the real-world scenarios 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 [95], it is important to understand possible fine-grained relationships between two events from three different causal perspectives, including *cause*, *enable*, and *prevent*. As Steven Solman et al.

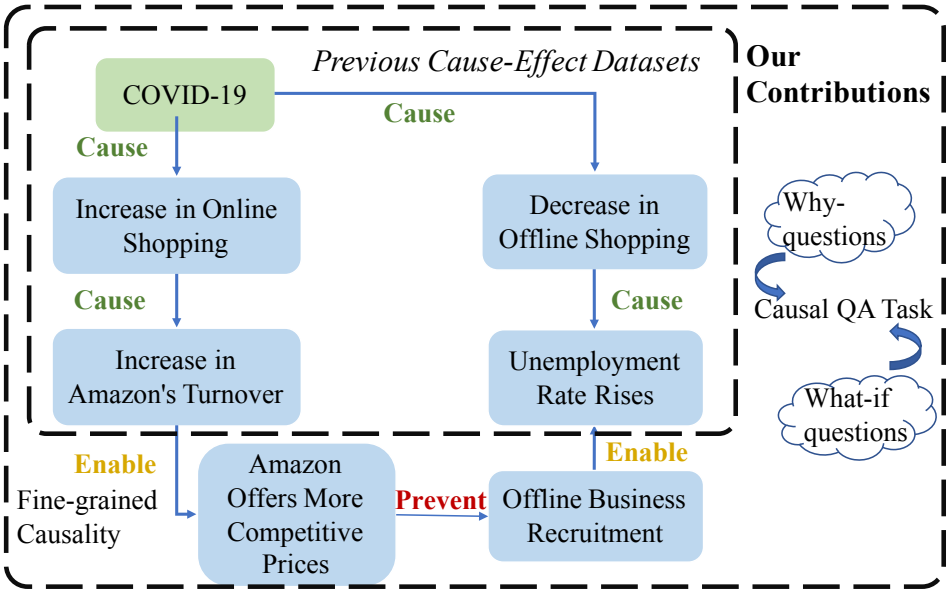


Figure 1.1: Comparison of existing "cause-only" researches (inner box) and more ideal causality graph with three kinds of causal relationships - CAUSE, ENABLE and PREVENT (outer box). More kinds of causal relationships can bring more events into the event graph and make the graph more complete.

[77] state in the paper "A Causal Model Theory of the Meaning of Cause, Enable, and Prevent", The notion that the causal model involves a relationship from A to B is expressed by the phrase "A causes B." "A enables B" implies that the model has a connection from A to B, that A represents a category of occurrences required for B, and that B has a secondary cause. "A prevents B" means that the model has a relationship from A to B and that A lowers the probability of B. A model, or an event graph as shown in Figure 1.1, can only be virtually complete if it has all three relationships. 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.

Besides the granularity of causal relationships, another problem existing in previous causality-related researches is that not all causal relationships can be automatically extracted. Previous researches usually leverage the conjunction words such as "because" and "since" to find causality event pairs. Although connectives can be used to find causal relationships in a text efficiently, a large number of causal relationships are hidden with no indication by salient conjunction words. Currently, no research can extract that hidden causality. To find these hidden causal relationships, human annotations are essential, since We can use human commonsense to extract potential causality.

To extract fine-grained and implicit causal relationships from the textual data, we construct a large-scale, hand-labeled, fine-grained causal reasoning dataset (**FineCR**) in the financial domain to measure the model's ability to extract these three kinds of event relations. We first collect many English-based financial analysis reports written by professional

financial analysts. We selected the sections in those documents that contain the most causal relationships to construct FineCR. Human annotators are hired to first select sentences containing causal relationships from those documents, then extract all the causal triples from these sentences. Finally, we use template generation plus manual calibration to generate massive question-answering pairs that require causal inference. The resulting dataset, FineCR, consists of 25,193 cause-effect pairs and 24,486 question-answering pairs, in almost all questions involving "why"[64] and "what-if" scenarios belonging to three fine-grained causalities. Our dataset can also potentially benefit downstream applications such as financial analysis [16] and BioNLP [13].

With FineCR, we define a series of causality-related tasks, including causality detection task, causality event extraction task, and causality question answering task. The causality detection task aims at detecting whether a given sentence has any causality relationships, while the causality event extraction task aims to extract all the causality triplets in a sentence, where the causality triplet consists of *a cause event*, *causal category*, and *an effect event*. The goal of causality question answering is to evaluate models' ability to answer questions that need some causal reasoning.

We explore several state-of-the-art neural models to establish the benchmark performance on different tasks with FineCR. 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 series of causality tasks based on that. We then conduct a detailed error analysis on each task and find that, due to current neural-based language models' inability to understand semantics, there is still a long way to solve the causality-related problems. Our contribution can be summarised as follows:

- We leverage human annotators to build the first-of-its-kind dataset that can be used for causality-related research, which makes researchers now possible to study causality in NLP.
- We first define the series of causality tasks, including causality detection task, causality event extraction task, and causality question answering task. These tasks can be used in many places in our life.
- 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 some potential research opportunities, especially in developing "causal-thinking" methods.

2

Background and Related Work

In this chapter, we will first introduce some critical notions about this research and then review some relevant previous research as well as the state-of-the-art NLP models. This thesis brings together two exciting ideas – event causality and causal question answering – and in what follows, we briefly introduce the existing relevant works and datasets to the present work.

2.1. Current Research of Event Causality

There is an extensive literature on causal inference techniques using non-text datasets [21, 38, 62, 66], and a line of work focusing on discovering the causal relationship between events from textual data [17, 25, 60]. Previous efforts lie on the graph-based event causality detection tasks [17, 51, 87] and the event-level causality detection tasks [20, 27, 57]. However, causal reasoning for text data with a particular focus on fine-grained causality between events has been 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.

2.2. Current Research of Question Answering

Question answering (QA) [29] aims to provide correct answers to questions based on context or knowledge. QA is a traditional research direction that was proposed half a century ago. People hope to help with everyday life by teaching the program how to answer questions like real people. Traditional QA systems integrate some information retrieval techniques to find answers. With the development of deep learning, computer programs can tackle more complex problems. At the same time, in the era of deep learning, more and more datasets are being proposed to measure the capabilities of QA models. These datasets, in turn, facilitate the development of deep learning QA models.

A comprehensive understanding of those datasets and benchmarks is essential before further research about QA. Therefore, in this paper, we investigate some of the most commonly used datasets nowadays and categorize them according to the capabilities of the QA models involved. Meanwhile, according to the different modes designed, we divide them into three categories: textual QA, image QA, and video QA. In textual QA, all the corpora involved are presented in textual form. A typical sample in textual QA consists of a question, an answer, and a paragraph that contains the answer. Image QA and video QA are generally

referred to as Visual Question Answering (VQA) [1]. In image QA, the question and answer are usually in textual form, while the context is an image. And in video QA, the question and answer are the same, but the context is a video clip. Since our proposed causal QA is based on textual QA, in Table 2.1, we present the current most used textual question answering datasets.

It can be found from the table that there is a large group of QA datasets that only use the multi-choice or entity as the answer, while our FineCR dataset uses a span of text as the answers. Compared to multi-choice and entity, span-based answers are more challenging for deep learning models, since answer spans are usually much longer and more complex. Compare FineCR to other QA datasets that use span as the answer type, such as SQuAD or DROP, we can find from the table that those previous datasets do not ask and answer questions about causality. The SQuAD is about reading comprehension, and DROP focuses more on multi-hop questions. However, since causality plays a vital role in our daily life, the study of causal QA is essential. The causality question answering dataset proposed in this thesis is a good complement to this research field as a standard to measure the QA model's ability to answer causal questions.

2.3. Transformer-based NLP models

In this thesis, the models we used are mainly developed based on the *Transformer* [90] architecture which proposed by Google in 2017. Nowadays, Transformer-based models are dominate almost all the NLP tasks, such as text classification, sentiment analysis, question answering, etc.

2.3.1. Basic Architecture

The basic architecture of *Transformer* is shown in figure 2.1. On the left is the encoder block and on the right is the decoder block. The encoder is responsible for converting the input alphabetic sentence into matrix vectors, and the decoder is responsible for re-decoding the encoded vectors into an alphabetic sentence.

The key part of *Transformer* model is the self-attention mechanism, which is used in the Multi-Head Attention (MHA) module in figure 2.1. A more detailed MHA is shown in figure 2.2. On the left is the scaled dot-product attention (SDA) module, and on the right is the multi-head attention module composed of multiple SDAs with some linear layers. The calculation of SDA is shown in Equation 2.1.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.1}$$

where *Q, K, V* are all the same word embedding matrix, represent the three inputs of *Transformer*. Through SDA, different words in the same sentence can find which words in the sentence are closest to themselves, so that the model can deduce the correct output more efficiently. This kind of weighted sum between word vectors is called *Attention*. While the role of MHA is to increase the expression space and ability of the model through more layers of SDA to enhance the learning ability of *Transformer*.

Dataset	Answer Type	Size	Domain	Evaluate Ability
ARC[10]	Multi-Choice	7,787	Science	Reasoning
BoolQ [9]	Bool	16K	Wikipedia	Reasoning
BioASQ [89]	Span	282	Biomedical	Articles Indexing
CaseHOLD [105]	Multi-Choice	53,137	Law	Pre-training
bABi [93]	Bool/Entity	40K	Open Domain	Reasoning
CBT [28]	Entity	20K	Children's Book	Model Memory
CliCR [83]	Entity	105K	Medical	Domain Knowledge
CNN and Daily Mail [76]	Entity	311K	News	Text Summarization
CODAH [6]	Multi-choice	4,149	Open Domain	Commonsense
CommonsenseQA [85]	Multi-choice	12,247	ConceptNet	Commonsense
ComplexWebQuestions [84]	Entity	34,689	Freebase	Multi-hop
ConditionalQA [81]	Entity/Span	9983	Public Policy	Multi-hop
COPA [72]	Multi-choice	1000	Commonsense	Reasoning
CoQA [70]	Entity	127K	Open Domain	Conversation
DROP [18]	Span	96K	Wikipedia	Multi-hop
FinQA [7]	Number/Span	8,281	Finance	Multi-hop
HotpotQA [100]	Entity	113K	Wikipedia	Multi-hop
JD Production QA [23]	Generation	469,953	E-commerce	Domain Knowledge
LogiQA [54]	Multi-choice	8,678	Exam	Reasoning
MCTest [71]	Multi-choice	2,000	Fictional Story	RC
Mathematics Dataset [73]	Numeric	2.1×10^6	Mathematics	Calculate
MS MARCO [63]	Generation	1,010,916	Web pages	Search
MultiRC [39]	Multi-choice	6K	Multiple Domain	Multi-hop
NarrativeQA [42]	Span	46,765	Story	Full Document
Natural Questions [44]	Span/Passage	323,045	Wikipedia	Search
NewsQA [88]	Span	100,000	CNN news	RC
OpenBookQA [59]	Multi-choice	6000	Science Facts	Reasoning
PIQA [2]	Multi-choice	21,000	Physical	Physical
PubMedQA [33]	Multi-choice	1K	Medical	Summarization
QASPER [12]	Extractive	5,049	NLP papers	Reasoning
QuAC [8]	Multi-choice Generation	100K	Wikipedia	Dialog
QUASAR [15]	Span	43,000	StackOverflow/Trivia	search
RACE [46]	Multi-choice	100,000	Exam	RC
ReClor [103]	Multi-choice	6138	Exam	Logical
SCDE [43]	Exam	6K	Exam	RC
SimpleQuestions [3]	Entity	100K	Freebase	Knowledge
SQuAD [68, 69]	Span	130,319	Wikipedia	RC
TriviaQA [34]	Span	650K	Open Domain	RC
TweetQA [96]	Generation	13,757	Tweet	RC
WikiHop [92]	Multi-choice	51,318	Wikipedia	Multi-hop
WikiQA [99]	Sentence	3,047	Wikipedia	RC

Table 2.1: Statistics of textual QA datasets. "RC" means "reading comprehension".

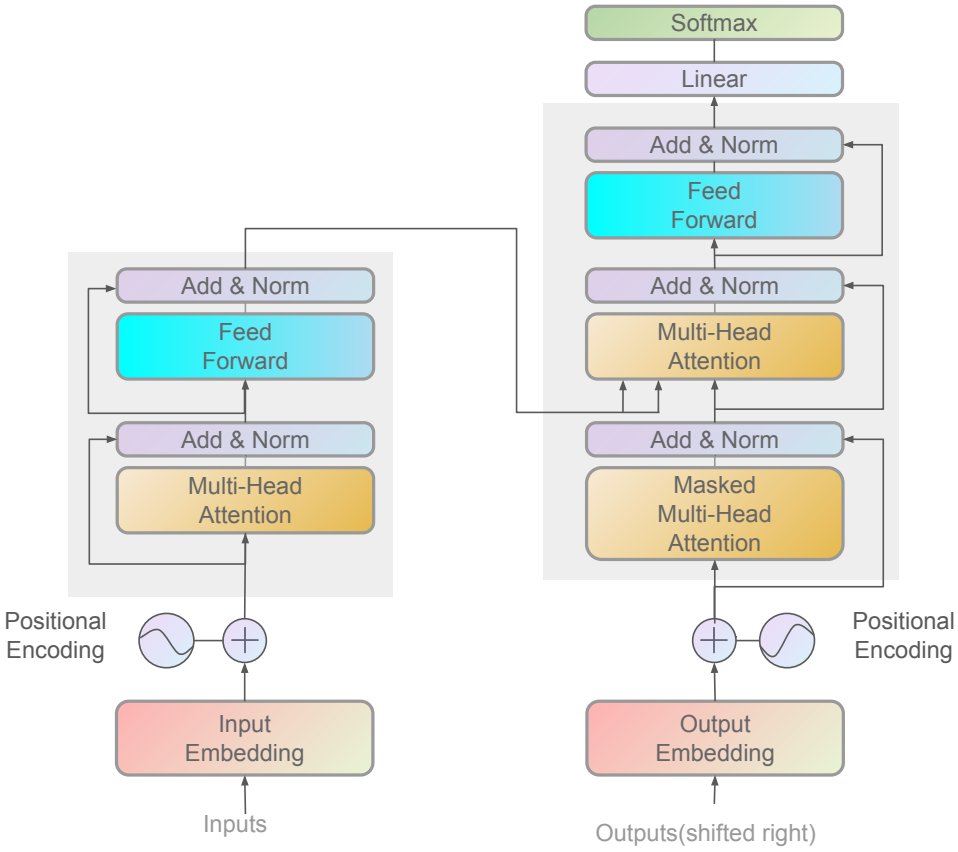


Figure 2.1: Architecture of Transformer.

2.3.2. Pre-training, Fine-tuning and BERT

Although the original *Transformer* model has already surpassed the previous LSTM [30] model, the emergence of *BERT* [14] has made the NLP researchers completely turn to the study of Transformer, and even affected the development of computer vision. *BERT* is made up of multiple encoder layers in *Transformer*. The training of *BERT* mainly consists of two phases: the pre-training phase and fine-tuning phase.

In the pre-training phase, the authors use two kinds of methods: Masked Language Model (MLM) and Next Sentence Prediction (NSP). In MLM, 15% tokens in the input sentence is masked by the special token `[MASK]`, and the training goal of MLM is to make the model correctly recover what the masked tokens are. And in NSP, the input is a pair of sentences concatenated by a special token `[SEP]`, and the training goal is to make the model correctly judge whether the two input sentences are truly adjacent in the original article. Through pre-training with a massive of easily accessible unlabeled texts, *BERT* adjusted its parameters and made it learn some basic text knowledge and semantic knowledge, laying a good foundation for later specific tasks.

In the fine-tuning phrase, the pre-trained *BERT* is trained by a specific NLP dataset with its training set and then tested on the validation set and testing set. Since the training tasks of the pre-training phase (MLM and NSP) are quite different from the specific downstream NLP tasks, such as question answering or text classification, the fine-tuning phase is necessary.

Through detailed experiments, the authors of *BERT* prove that the *BERT* model with

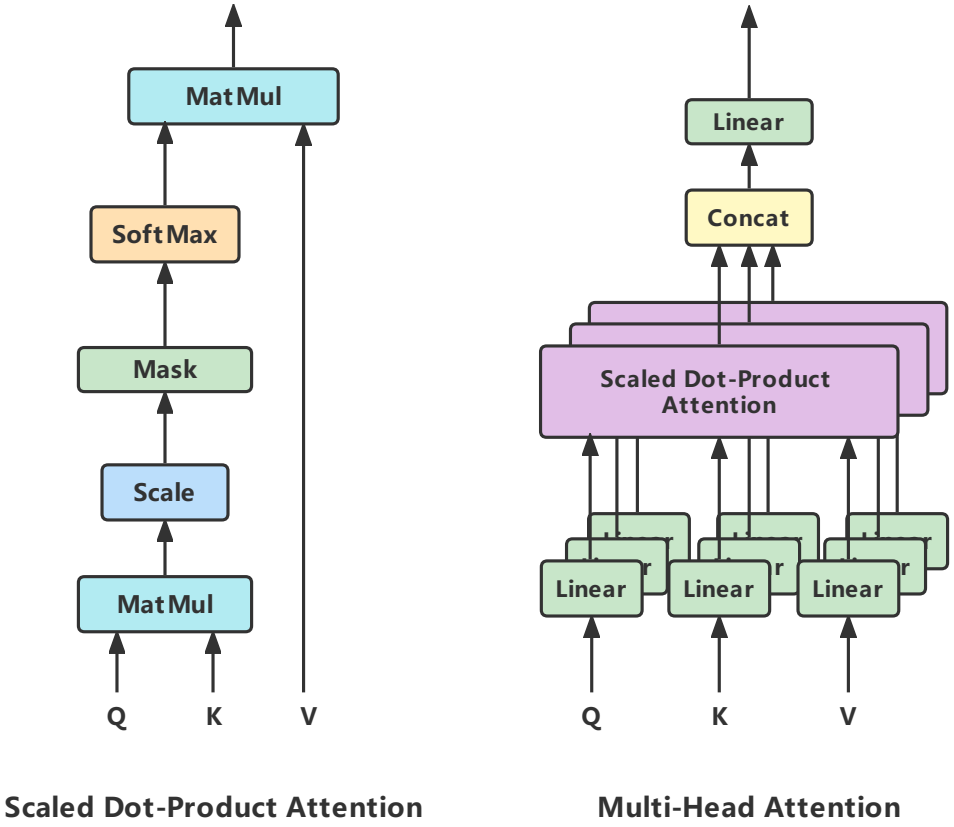


Figure 2.2: Architecture of Multi-Head Attention.

pre-training not only outperforms LSTM, but also outperforms the original *Transformer* model. This demonstrates the importance of pre-training for deep learning models.

2.3.3. Variants of BERT

With the advent of *BERT* leading to research on pre-training tasks, there are more studies based on *BERT* and different pre-training tasks. *RoBERTa* [56] uses larger pre-training corpus and removed the NSP task. The better model performance proves the value of *BERT*, and demonstrates that the pre-training model still has great potential to be exploited.

Different from *BERT* and *RoBERTa*, the pre-training task in *SpanBERT* [35], where the mask operation always masks some neighboring tokens, is specifically designed for span prediction. SpanBERT can be used in span-based NLP tasks such as extractive question answering. *SpanBERT* demonstrates that incorporating the features of downstream tasks into pre-training tasks can significantly improve model performance for those specific tasks.

BART [48] is a model mainly aiming at natural language generation (NLG). It keeps the original *Transformer* architecture. The MLM in *BART* is quite different from the *BERT*. *BART* masked tokens in sentences using spans that matched the Poisson distribution. About 30% of tokens are masked. Another pre-training task is sentence permutation, which is reordering several sentences that had been input out of order. After these two pre-training tasks, *BART* achieved state-of-the-art (SOTA) on many NLG tasks.

The structure of *T5* [67] is also similar to the original *Transformer* structure. The biggest innovation and contribution of *T5* is using generation tasks to unify various NLP tasks, including natural language understanding (NLU) and NLG. For example, usually, the output

of a sentiment analysis task is a “True” or “False” label, which is determined by a binary linear classification layer. But the output is the label itself in *T5*. This makes it possible for different NLP tasks to share the same parameters, allowing *T5* to be trained with plenty of labeled datasets, even if these datasets are originally designed for different tasks.

3

Datasets and Tasks

In this chapter, we will introduce where we collected our dataset and how we leverage human annotators to create this unique causality related dataset. We first introduce the source of the data and how we filtered the raw data to the needed parts. Then we present the data annotation process in detail and give some examples to help understand. Finally, we introduce how we guarantee the quality of the final dataset through some additional quality control methods.

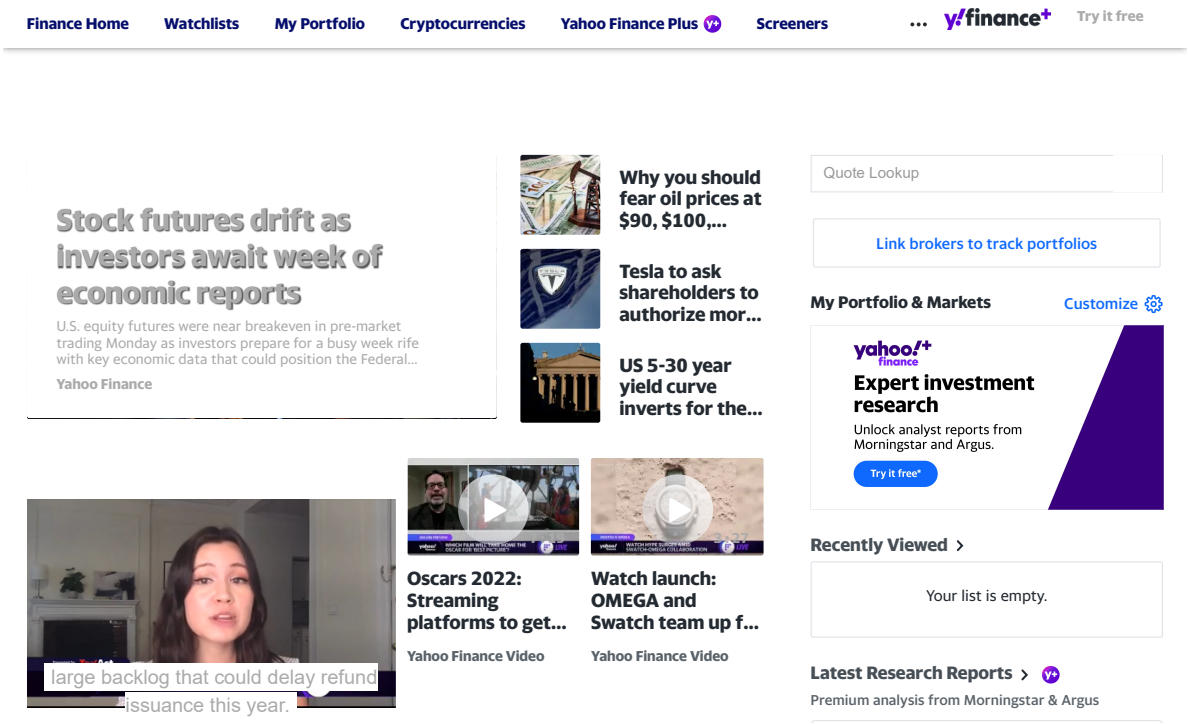


Figure 3.1: Yahoo finance website.

3.1. Data Source

We collected a financial analyst report dataset from Yahoo Finance, as shown in figure 3.1 [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. In figure 3.2, we present an example of our used financial analyst report. Each report is composed of multiple sections. We select the most valuable chapters with the most cause-and-effect event pairs, including "Business Strategy & Outlook", "Economic Moat", "Fair Value and Profit Drivers", and "Risk and Uncertainty", where there contains most of causality contents.

3.2. Crowdsourcing

The annotation platform used in this work is introduced in Fig. 3.3. As follows, we provide the detailed annotation instructions used for training the human annotators. Also, we show the annotation of some actual examples stored in our dataset.

3.2.1. General Instruction

This is an 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 the given passages.

3.2.2. 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.

3.2.3. 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.

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

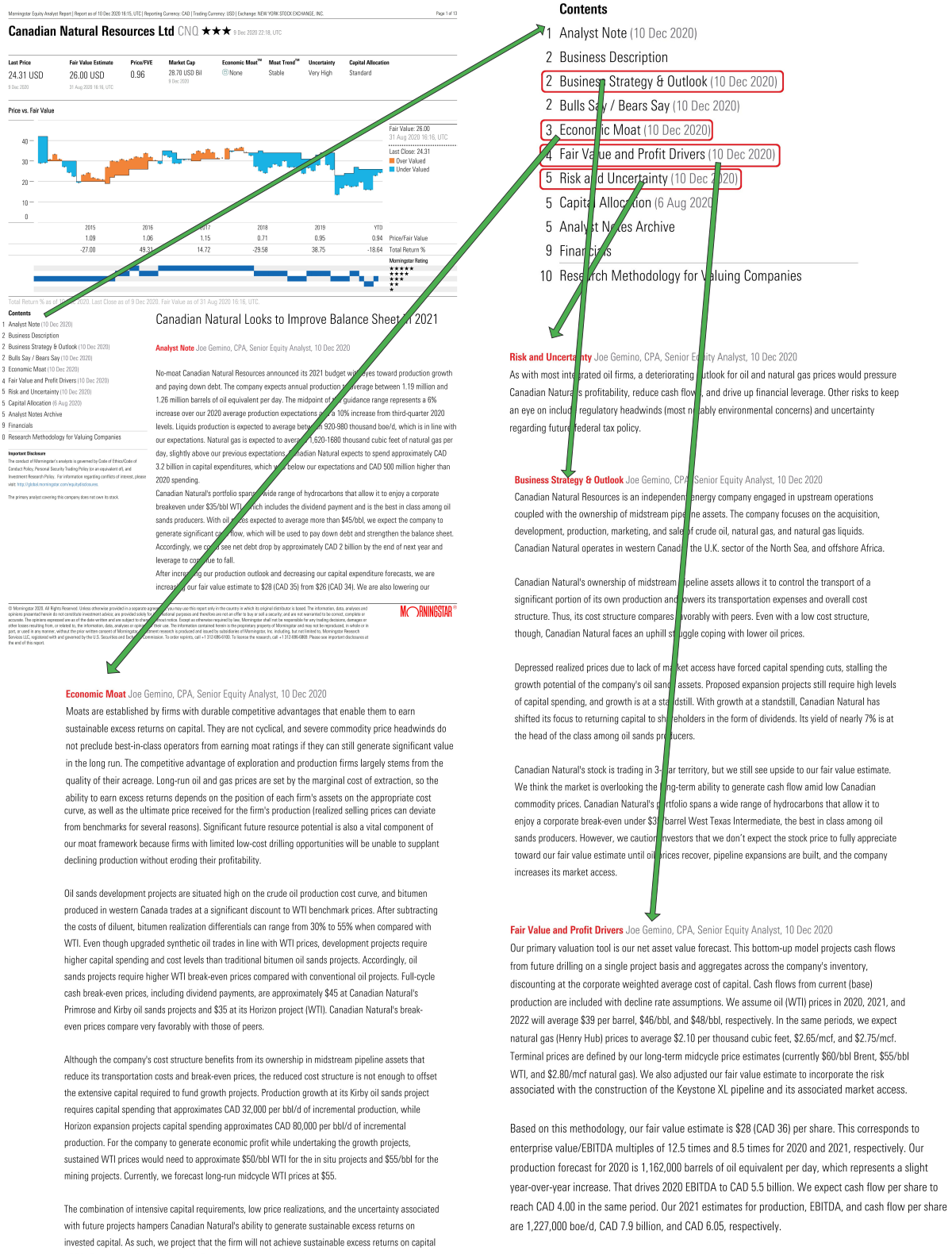


Figure 3.2: An example of financial analyst report.

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.

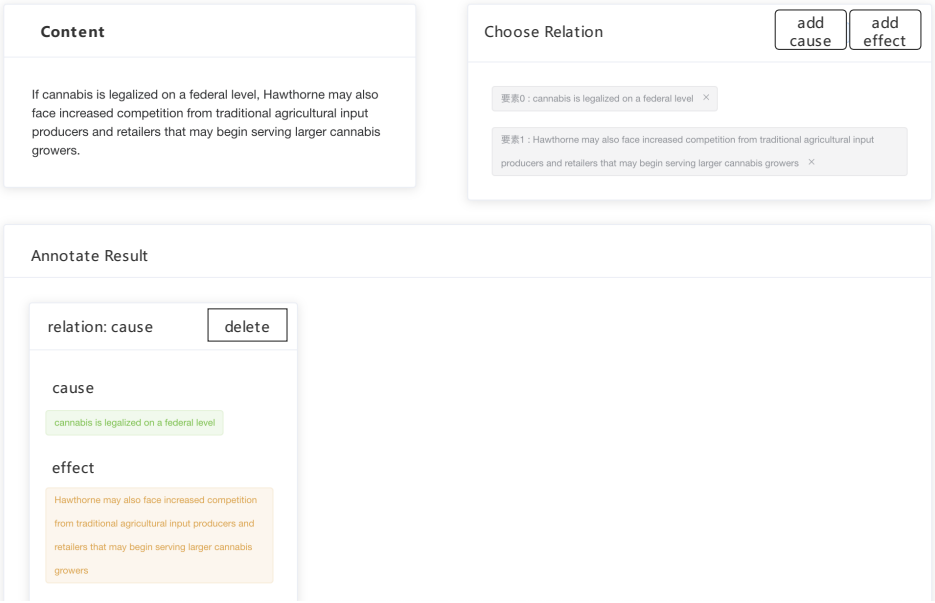


Figure 3.3: The annotation platform provided the crowdsourcing company for collecting annotations for fine-grained causality event extraction and causality question answering.

3.2.4. Quality Control

To ensure high quality, we restricted the participants to experienced human annotators with relevant records. For each task, we conducted pilot tests before the crowd-sourcing work officially began to receive feedback from quality inspectors and revise 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 annotators 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. Additionally, we sort the dataset in chronological order because the future data is not expected to be used for predictions.

3.2.5. An Example

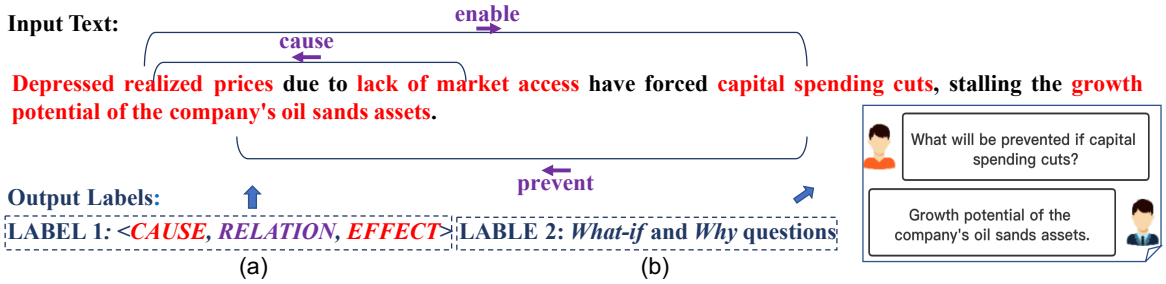


Figure 3.4: Illustration of our crowdsourcing tasks using an example that contains all three types of causal relationship.

In Figure 3.4, we present an example of the annotated data. This is a positive sentence that contains all the three kinds of causality relationships. Each red part represents an event, and the connecting line between the two red and red represents the type of causal relationship between the two events. The starting point of the line represents the cause event, and the focal point of the line represents the effect event. At the same time, we generate corresponding question-answer pairs based on the causality triples (<*cause, relationship, effect*>), which are used to examine the ability of the model to answer causality questions.

3.3. Dataset and Tasks

After the dataset annotation in Section 3.2, we can then build our causality dataset called - **FineCR** (**Fine**-grained **C**ausal **R**easoning Dataset). The whole pipeline of this thesis is shown in Fig. 3.5. We define three tasks on our FineCR dataset and build strong benchmark results for each task. 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 preprocessing steps required crowd-sourcing efforts were carried out to prepare the raw dataset.

As mentioned in Section 3.2.2, we first ask the annotator to find which sentences contain causality and which sentences not from a document. Sentences with causality are marked as “Positive”, while sentences without causality are marked as “Negative”. After the annotation we obtain a binary classification dataset - **FineCR-D** and define the first task - **CausalDet** (**Causal**ity **Det**ection Task). In this task, the model should predict whether a sentence contains any causality relationships.

With those positive sentences, we then construct the second dataset - **FineCR-E** using the annotated causality triples and define the second task - **CausalExt** (Fine-grained **Causal**ity **Ext**raction Task). In this task, the model should correct extraction all the causality triples contained in a sentence.

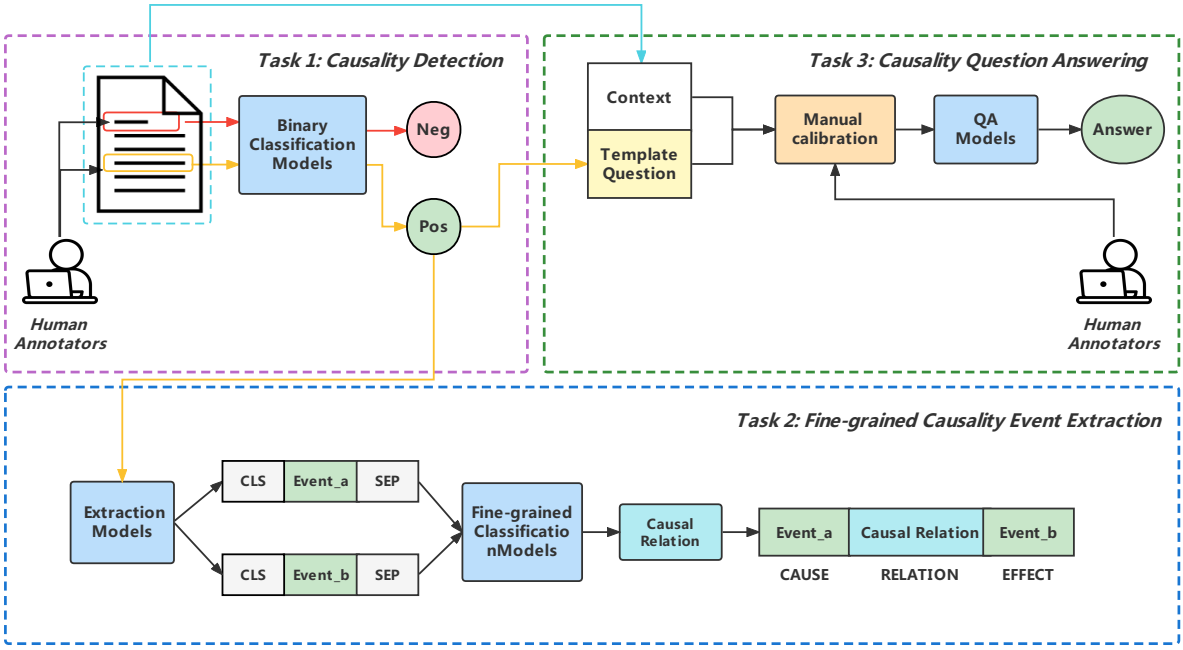


Figure 3.5: The pipeline of experiments based on the FineCR dataset.

Finally, we use the template as well as human manual calibration to generate a causality question answering dataset - **FineCR-Q** based on the causality triples, and design the third task - **CausalQA** (**Causal**ity **Q**uestion **A**nswering Task) based on this dataset. In this task, the model needs to find the correct answer given the question and related context.

For hold-out evaluation, we split each of the 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. In the following chapters, we will describe each dataset and task with the corresponding experiments in detail.

4

Causality Detection

This chapter mainly describes the task of **causality detection**. We first introduce the definition of this task and give some basic information about the dataset for this task, including some statistics and several concrete examples. Then we present the results of experiments using some classification models. In response to these results, we carried out both quantitative and qualitative analysis and hoped our experiments could have a specific enlightening effect on some follow-up research.

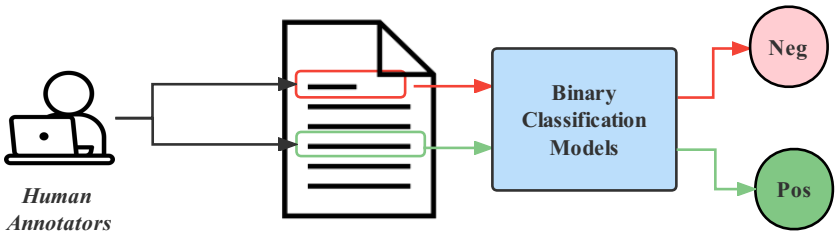


Figure 4.1: The architecture of causality detection model.

4.1. CausalDet Introduction

The goal of the causality detection task is that given a sentence, the model should figure out whether the sentence contains at least one causality events pair. CausalDet is a binary classification task similar to sentiment analysis.

4.2. FineCR-D Introduction

FineCR-D is a binary classification dataset containing the positive and negative samples to do the causality detection task. Each positive sample is a sentence that contains at least one causal relationship (one of CAUSE, ENABLE, and PREVENT), while a negative sample is one sentence that does not contain any causal relationship. In Table 5.1 we list some positive samples that contain causality as well as some negative samples that do not contain any causality.

In Table 4.2 we list the detail statistics about the causality detection dataset and also its comparison to FinCausal [58]. Our dataset significantly outperforms FinCausal in the number of samples, which was previously the largest causality detection related dataset. More importantly, FinCausal only includes the *cause-effect* relationship, while in our dataset, the

Sentence	Label
MercadoLibre faces limited online retail competition in the region, with Amazon, eBay, and local players like B2W generating less unique visitors in its core markets, though there are several solid competitors in the payments segment.	Positive
Offering secure and convenient payment settlement between sellers and buyers on and off the MercadoLibre platform, MercadoPago attracts more sellers if more buyers sign up to use the services; the more sellers offer MercadoPago's payment solutions, the more popular the service will become among buyers.	Positive
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.	Negative
Despite first-quarter weather headwinds, we expect solid demand and pricing recovery to support 13% revenue growth in 2021, including industrial end-market improvement, a strong intermodal rebound, and healthy cross-border refined product shipments.	Negative

Table 4.1: Examples in FineCR-D.

relationship includes *cause*, *enable*, and *prevent*, which allows our dataset to be used to study a wider range of causal relationships.

We observe no significant difference between positive and negative examples in the average token numbers, which shows that predictive models are difficult to learn from shortcut features [47, 79, 80] (e.g., the instance length) during the training process. Furthermore, our dataset contains 846 multi-sentence samples, and 3,017 text chunks have more than one causal relation in one instance, which requires a complex reasoning process to get the correct answer, even for a human.

Metric	Counts
FineCR-D	
#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
#Total Instances	51,025
FinCausal [58]	
#Total Instances	3031

Table 4.2: Statistics for FineCR-D and the comparison to FinCausal[58].

4.3. Model

We consider using both classical deep learning models – CNN-Test [41] and HAN [102] – and Transformer-based models downloaded from Huggingface[1] – BERT [14], RoBERTa [56], and SpanBERT [35] – as predictive models.

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. Our methods are built on the recently advanced Transformer architectures [94] with the framework provided by Huggingface[2]. As follows, we introduce the detailed implementation of deep neural methods on three tasks completed on the FineCR dataset.

4.4. Metrics

The F1-score and accuracy are used for evaluating the causality detection task. The Macro F1-score is defined as the mean of label-wise F1-scores:

$$\text{Macro F1-score} = \frac{1}{N}\sum_{i=0}^{N} \text{F1-score}_i \tag{4.1}$$

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

4.5. Experimental Results

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

The causality detection result is shown in Table 4.3. 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 [58] dataset can reach the human-level result (F1 = 97.75), providing

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

indirect evidence that our dataset is more challenging caused of more complex causality instances.

4.6. Error Analysis

In Table 4.4, we present some correctly classified sentences and also some falsely classified sentences. From the table, we can find that causality keywords play an important role in the classification results. While keywords (such as since, because, or limit) can help the model do classification faster, it can lead to serious prediction errors in some cases. For example, in the first two samples (sentences 1 and 2), based on the keywords “because” and “since”, the model predicts those two sentences are positive, which means they contain causality relationships. However, in the examples of wrong judgment, the last two sentences (sentences 7 and 8) are judged as positive because they contain relevant keywords. But the correct label should be negative because in these sentences, “as” and “since” do not indicate a relationship, but are part of a phrase that belongs to “view as” and “since the beginning of” respectively. And when the keyword does not appear, such as in sentences 5 and 6, although there are causal relationships in the sentences, because the current NLP models still cannot understand the language semantics well, these causal relationships cannot be found and thus lead to error classification.

Those experimental results prove that when the deep learning models are used for NLP classification tasks, they can easily fall into the situation of only learning salient keywords. This phenomenon has also been found in previous studies related to sentiment analysis [98]. When there is no keyword in the sentence, the classification fails because the model cannot really understand the semantic knowledge of causality.

Index	Sentence	Ground Truth	Model Prediction
	Correct Prediction		
1	**Because** it offers a full suite of sterilization solutions, from the sale of individual sterilizers to full outsourcing, Steris benefits regardless of how hospitals choose to manage sterilization.	Positive	Positive
2	Even though the number of firms with an NRSRO designation has increased (to about 10) **since** the introduction of the Credit Rating Agency Reform Act of 2006, we believe the network effect has been strong enough to result in limited traction of other rating agencies.	Positive	Positive
3	Over the past decade, Beyond's research and development team has delivered several plant-based meat breakthroughs as well as continuous improvements to existing products.	Negative	Negative
4	We increased our five-year capital expenditure forecast by $700 million, to $8.7 billion, when we advanced our estimates by one year to 2021-25.	Negative	Negative
	Wrong Prediction		
5	The tendency for manufacturers to reduce the cost of aircraft design by using derivative (such as Airbus' A320neo and Boeing's 737 MAX), rather than clean-sheet new aircraft, reduces the probability that a particular component will be recompeted in the redesign.	Positive	Negative
6	Previously we rated the Dutch banking system as only Fair, with the poor performance of the Dutch banks during the 2008 financial crisis a major consideration. positive	Positive	Negative
7	ING DiBa, ING's digital German bank generates return on equity in what we view **as** the eurozone's least attractive banking market.	Negative	Positive
8	The strong share price performance from Italian banks **since** the beginning of 2017 would suggest that the market is currently less concerned about the risks they face, but we are not convinced.	Negative	Positive

Table 4.4: Some classification results of causality detection experiment.

5

Causality Event Extraction

In this chapter, we mainly describe the task of **fine-grained causality event extraction**. We first introduce the definition of this task and give some basic information about the dataset for this task, including some statistics and several concrete examples. Then we introduce the two-step model we designed to solve this task and present the results of experiments using our model. In response to these results, we carried out both quantitative and qualitative analysis and hoped our experiments could have a specific enlightening on some follow-up research.

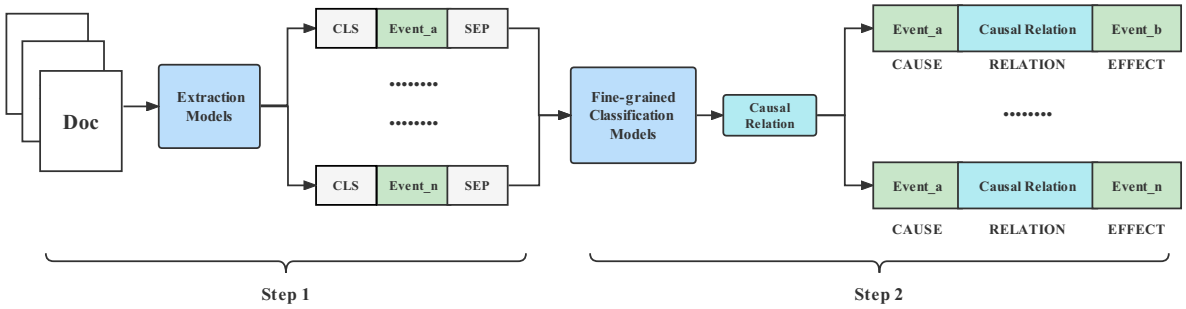


Figure 5.1: The architecture of two-step causality extraction model.

5.1. CausalExt Introduction

In fine-grained causality event extraction task, the model should first correctly extract all the events in a given sentence and then pair them to find all the plausible causalities.

5.2. FineCR-E Introduction

We build the first fine-grained causality event extraction dataset using the annotated causality triplet in Chapter 3. In Table 5.1, we list some samples from this dataset, and the dataset statistics is shown in Table 5.2. For each corpus in FineCR-E, there is at least one corresponding causality relationship, and some samples contain multiple causal relationships. At the same time, in order to make the data set more difficult, some corpus are composed of multiple sentences, one of which contains the cause and the other one contains the effect.

Sentence	Relation
We expect **Neogen will continue to face headwinds through 2021** because of **reduced volumes and price volatility during the ongoing pandemic**.	Cause
We forecast **8% per year volume growth for recreational and food-products** based on **expanded distribution, consumers converting from the black market, and non-consumers becoming consumers**.	Enable
Even so, **annual research and development expense as well as number of patents are not disclosed by Magna**, leaving no way for the **market to judge the extent of technological innovation**.	Prevent
On the production animal side, **rising standards of living in emerging markets** should lead to **wider adoption of meat-heavy diets, driving greater demand for livestock products**.	Enable
Fastly has a more modern approach to running a content delivery network (CDN) that enables it to **offer equivalent or superior service, depending on the task, with fewer points of presence (PoPs) and less commitment of capital**.	Enable
Additionally, the pandemic has caused **many millennials to consider moves to the suburbs, either into suburban apartments or their own single-family homes**, which limits the **potential short-term demand for new urban apartments**.	Prevent

Table 5.1: Examples in the causality extraction dataset. **Red** is the cause and the **green** is the effect.

Metric	Counts
Cause-Effect Event Pairs	
#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

Table 5.2: Statistics for FineCR-E.

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 5.3: Error analysis for fine-grained classifications.

5.3. Model

To do this task, we design a two-step model shown in Figure 6.1. The first step is the event extraction step, in which we use models to extract all feasible events within the given text. After getting those events, we add *[CLS]* and *[SEP]* tokens to the beginning and end of each event separately. Then we input all the extracted events into a classification model. The second step is to predict whether there is a causality relationship between the two events and, if so, to which causality relationship belongs.

Because two events appear in a sentence in a sequence, to distinguish this order, we further divide the relationship into "relation" and "relation_by", as shown in Table 5.3. For relation prediction, we always use the event that occurs first in the sentence as the first input to the model, and the following event as the second event input. If an event occurs after, but is a cause event, then the predicted relationship will end with a "_by". After the above two steps, we can get the causality triplet, consisting of two events and the causality relationship between them.

The classification model we used is the same in Section 4.3. And the extraction model we used including BERT [14], RoBERTa [56], and SpanBERT [35]. 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.

5.4. Metrics

In event extraction step, we use F1 and exact match (EM) to measure the model performance. Given the extracted event sentence and the ground-truth sentence, EM is **true** if and only if the extracted sentence is exactly the same as the ground-truth sentence, including the types of words that are formed and the order between words, otherwise it is **false**. And different to the F1 in Section 4.4, the F1 score used to measure the similarity between two sentences is defined as:

$$
\begin{aligned}
\text{Overlap} &= BAG(P\_S) \bigcap BAG(G\_S) \\
\text{Precise} &= \frac{len(Overlap)}{len(BAG(P\_S))} \\
\text{Recall} &= \frac{len(Overlap)}{len(BAG(G\_S))} \\
\text{F1-score} &= \frac{2 * Precise * Recall}{Precise + recall}
\end{aligned}
$$

where BAG means bag of word set, P_S is the predicted sentence, G_S is the ground-truth sentence, len means the number of words in a set. The F1 in step 2 is the same as in Section 4.4.

5.5. Experimental Results

The results of the fine-grained event causality extraction task are shown in Table 5.4. 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.

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
-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
-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
-large	85.50	**57.40**	86.33	**60.26**	73.65	68.15	64.43	72.93	**23.01**	**21.78**
Human	-	-	94.32	81.34	-	-	88.61	90.53	-	-

Table 5.4: 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.

We find that the large Transformer-based models [90] 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.

5.6. Error Analysis

5.6.1. Event Extraction

We first present the error analysis of the first event extraction step in our model in Table 5.5. The first three are examples of correct extraction, and the last three are examples of some extraction errors. In the first error example, it can be seen that the cause predicted by the model has a missing premise - “In Spain, its second-largest market, ”. Although the answer predicted by the model contains the most important part, “Orange does not benefit from cost advantages”, it lacks an essential premise, which makes the extracted answer not very accurate. Because the conclusion can only be established under this specific condition (in Spain), it is not necessarily the case if you change the country to the Netherlands.

While in the second error example, the model mistakenly included connectives “Because”, which should not be part of the event. A similar error can also be observed in the last example. The effect mispredicted by the model includes “we believe”, and the clause that explicitly explains the “sticky area” is not included, which is “like automotive safety systems, industrial controls, and server platforms”. Without this explanation, we may not know what the “sticky area” represents.

The above error examples illustrate why it is easy for the model to find where an event is in a sentence roughly, but it can be challenging to extract the event very precisely. Merely using the existing data to train the event extraction model may not achieve very high accuracy in extraction.

5.6.2. Relationship Classification

For fine-grained classification in the second step, a more detailed error analysis by using the best-performed RoBERTa-Large model is given in Table 5.3. 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.

Sentence	P_Cause	P_Effect
Correct Extraction		
The **continued reinvestment in R&D supports Hardie's strong brand equity** and thus **perpetuates the price premium that Hardie's range attracts**.	**continued reinvestment in R&D supports Hardie's strong brand equity**	**perpetuates the price premium that Hardie's range attracts**
As the number of contacts increases, **more developers will choose to develop mini programs for Weixin or mobile QQ**, and **existing users can benefit from having access to additional mini programs**.	**more developers will choose to develop mini programs for Weixin or mobile QQ**	**existing users can benefit from having access to additional mini programs**
Robinson is a highly attractive source of freight opportunities, given **its ability to aggregate fragmented demand across a broad customer base of shippers**.	**its ability to aggregate fragmented demand across a broad customer base of shippers**	**Robinson is a highly attractive source of freight opportunities**
Wrong Extraction		
In Spain, its second-largest market, Orange does not benefit from cost advantages, since **its customer base is smaller than the incumbent Telefonica**.	**its customer base is smaller than the incumbent Telefonica**	**Orange does not benefit from cost advantages**
Because **the liquid could be anything from water to a dangerous chemical**, **choosing the different metals for the seal faces and understanding how they will interact with fluid running through the pump is critical to proper functioning of the seal**.	~~**Because**~~ **the liquid could be anything from water to a dangerous chemical**	**choosing the different metals for the seal faces and understanding how they will interact with fluid running through the pump is critical to proper functioning of the seal**
Furthermore, we believe **the design wins and resulting share gains in sticky areas like automotive safety systems, industrial controls, and server platforms** arise from **customer switching costs**.	**customer switching costs**	~~**we believe**~~ **the design wins and resulting share gains in sticky areas**

Table 5.5: Examples in the causality extraction experiment. **Red** is the cause and the **green** is the effect. **P_Cause** mean the predicted cause and **P_Effect** means the predicted effect.

6

Causality Question Answering

In this chapter, we mainly introduce the task of causality question answering. We first give the definition of this task and how we constructed the corresponding dataset, and at the same time, we give some basic statistics and some examples of the dataset. Then we use current state-of-the-art QA models to experiment on the dataset and give the experimental results. For these results, we carried out both quantitative analysis and qualitative analysis. Finally, we compare our dataset with previous QA datasets, highlight our dataset's uniqueness and research value, and hope to inspire some research on causality problems in the QA field.

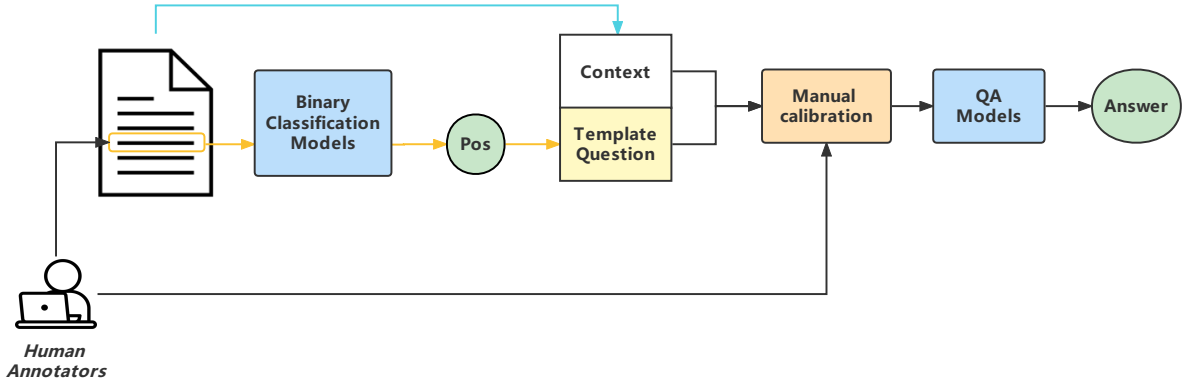


Figure 6.1: The architecture of question answering model.

6.1. CausalQA Introduction

In causality question answering task, for a given question and its corresponding question, model should correctly give the answer. This task can help us research how to deal with the question answering related to causality.

6.2. FineCR-Q Introduction

6.2.1. Example

Figure 6.2 shows a sample of the FineCR-Q. Given the causality triplet, We first use "**relationship**" to fix the type of problem, for example here because the relationship is "**prevent**", the type

Causality Triplet: <Cause, Relationship, Effect>

Context: However, Dell Technologies derives most of its revenue from challenging commoditized markets that affect its ability to generate excess economic return on invested capital. The amalgamation of various technology brands into an end-to-end product portfolio for the IT environment provides Dell Technologies with substantial cross-selling and upselling opportunities as it broadens its product portfolio to encompass the overarching market shift to hybrid cloud environments. Altogether, we expect difficult pricing environments for servers and PCs to inhibit consolidated operating results as Dell Technologies attempts to move forward as a unified company. Dell Technologies' three reportable segments are its infrastructure solutions group, or ISG; client solutions group, or CSG; and VMware. In fiscal 2020, net revenue of $92 billion was split as 37% from ISG, 50% from CSG, and 14% from VMware.

Question: What will **prevent** Dell's ability to generate excess economic return on invested capital?

Answer: Dell Technologies derives most of its revenue from challenging commoditized markets

Figure 6.2: A sample from FineCR-Q.

of problem is "**What will prevent**". Then we randomly use one of the **cause** and **effect** as the question and the other as the answer. Here we use "**Dell's ability ...**" as the question and use "**Dell Technologies drives ...**" as the answer. After the questions are generated with the template, the annotators are asked to check and manually calibrate the ungrammatical questions.

6.2.2. Statistics

Metric	**Counts**
CausalQA Pairs	
#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 6.1: Statistics for FineCR-Q.

In Table 6.1, we present the statistics of FineCR-Q. unlike other QA datasets [79, 80] that can easily benefit from the test-train overlap as revealed by [49, 55, 91], our dataset is sorted in chronological order so that the future test data could be theoretically challenging to coincide with the training set. This allows us to obtain greater insight into what extent models can generalize.

6.2.3. Datasets Comparison

Table 6.2 compares our dataset with datasets in the domain of both event causality and question answering (QA). FinCausal [57] dataset is the most relevant to ours, which developed a relatively small dataset from the Edgar Database[1] focusing on the simple "cause"

[1] `https://www.sec.gov/edgar/searchedgar/`

Datasets	Event Extraction	Causal Reasoning	Fine-grained Causality	Span-based QA	Sources
FinCausal [58]	✔	✔	✗	✗	Finance
COPA [72]	✗	✔	✗	✗	Open
SQuAD [69]	✗	✗	✗	✔	Wikipedia
LogiQA [53]	✗	✔	✗	✗	Examination
HotpotQA [101]	✗	✗	✗	✔	Wikipedia
DROP [19]	✗	✗	✗	✔	Wikipedia
DREAM [82]	✗	✔	✗	✗	Examination
RACE [45]	✗	✔	✗	✗	Examination
FineCR-Q (Ours)	✔	✔	✔	✔	Finance

Table 6.2: Comparisons of our FineCR-Q and related public QA datasets.

relation only and do not contain QA tasks. In addition, existing popular question answering datasets [11, 53, 82] mainly focus on *what*, *who*, *where* and *when* questions, making their usage scenarios somewhat limited. SQuAD [68, 69] consists of factual questions concerning Wikipedia articles, and some unanswerable questions are involved in SQuAD2.0. Although some datasets contain the causal reasoning tasks [11, 45, 82], none of them consider answering questions by text span. Span-based question answering problems have gained wide interest in recent years [31, 50, 101]. HotpotQA [101] focuses on multi-hop QA where the question can only be answered through analyzing multiple documents. The answers in the [19] may come from different spans of a passage and require some combination technologies to get the correct answer. Compared with these datasets, 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 of "Why" and "What-if" questions. Our task is similar to the machine reading comprehension setting [31] 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.

6.3. Model

We perform a causality question answering task by leveraging six Transformer-based pre-trained models provided by Huggingface [94] on our dataset, including BERT-base, BERT-large, RoBERTa-base, RoBERTa-large, RoBERTa-base-with-squad, and RoBERTa-large-with-squad[2]. Furthermore, pre-trained seq2seq models such as T5 [67], or BART [48] 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.

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

6.4. Metrics

F1 and EM are used as the metrics in CausalQA task and they are the same as CausalExt step 1 as described in Section 5.4.

6.5. Experimental Results

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 6.3: 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 [3] only.

The results of CausalQA are given in Table 6.3, 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 mainly 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 in 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-labeled 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. However, further research is required, as to what extent models can actually benefit from the additional data for the generalization is hard to be evaluated.

6.6. Error Analysis

Table 6.4 presents a qualitative analysis for causality question answering, where we highlight the question and answer parts extracted from the raw context. Human annotators label the gold answers while the BERT-based model generates the output answers. The model answers the first three questions correctly, while the last two instances show two typical patterns prone to errors. In the first incorrect example, the model outputs "*targeted*

Context	Question	Gold	Output
(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
Incorrect Predictions Examples			
(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 6.4: 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.

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

6.7. Challenging by CausalQA

Dataset	Method	F1	ACC	EM
SQuAD1.1 [68]	LUKE [97]	95.7	-	90.6
SQuAD2.0 [69]	IE-Net [24]	93.2	-	90.9
DROP [19]	QDGAT [5]	88.4	-	-
HotpotQA [101]	BigBird-etc [104]	95.7	-	90.6
Reasoning Based Datasets				
LogiQA[53]	DAGAN [32]	-	39.3	-
FineCR-Q (Ours)	RoBERTa-SQuAD	84.7	85.6	61.6

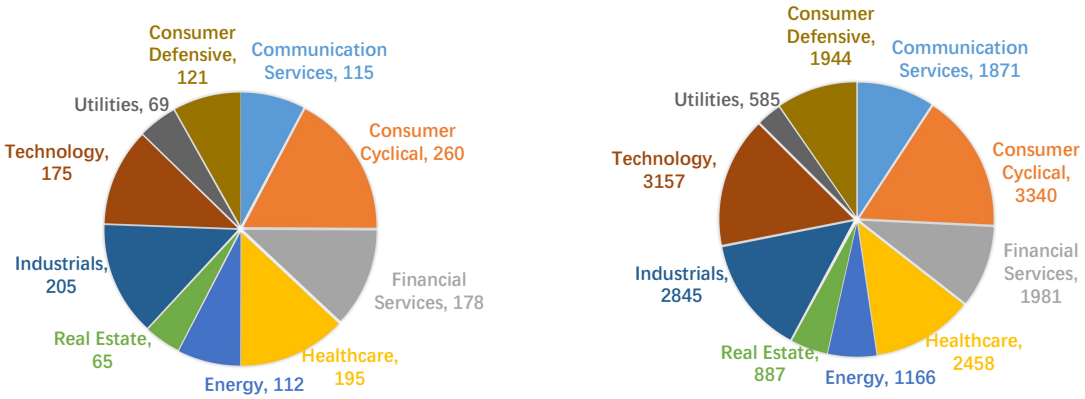
Table 6.5: The comparison of best performance between our dataset and other popular QA datasets.

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 6.5. 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.

7

Out-of-Domain in Causality

In this chapter, we mainly examine the impact of the Out-of-domain (OOD) problem in causality-related research. We first describe how we constructed a causality dataset containing different economic domains from the meta-information we collected earlier. We used this dataset to conduct some OOD experiments and then carried out some quantitative analysis to reveal the existence of OOD in causality research.



(a) Companies Count

(b) Reports Count

Figure 7.1: Sector distributions on companies and reports.

7.1. 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. 7.1, 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 for out-of-domain evaluations (see 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 [4, 86].

7.2. Experimental Results

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 [65]. 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 the out-of-distribution data [26, 36, 78]. 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 7.1 and Table 7.2.

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 7.1: 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 7.2: Out-of-domain test results of the Span-Large model for the cause-effect extraction task.

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 a 59.69 F1 Score when testing on the Consumer Cyclical data while decreasing to 47.95 on technology companies. Moreover, in the cause-effect extraction task, the model trained with the data from the Consumer Cyclical domain achieves an 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.

8

Conclusion and Discussion

In this chapter, we summarize the previous chapters, mainly including the findings brought by the dataset and experimental results. At the same time, we analyze the limitations of this thesis and reveal some possible future research directions.

8.1. Findings

In this thesis, we construct the first-of-its-kind human-annotated high-quality causality dataset - FineCR, in response to the lack of research on causality in NLP. Instead of using salient conjunction words, we leverage human annotators to find all the causality relationships contained in the text, which makes our dataset more accurate and comprehensive. The dataset we constructed makes it now possible to study causality in NLP. For the application of causality-related research in our real life, we design a total of three tasks, including causality detection, causality event extraction, and causality question answering.

We experiment with state-of-the-art deep learning models for these tasks and obtain some interesting insights. Experimental results using the state-of-the-art neural language models provide evidence that there is still much room for improvement on causal reasoning tasks. And there is a need to design better solutions to correlation discovery related to event causality analysis and causality QA tasks. After a detailed analysis of the experimental results, we find that those models have different shortcomings for different tasks.

First, for the causal detection task, we find that the results of current classification models are easily affected by keywords. Therefore, when the model has an error understanding of keywords or lacks the keywords, there is a high probability that the model will be classified incorrectly. This problem is an important reason that we propose FineCR. If models could only find causal relationships based on keywords, they would ignore massive hidden causal relationships. Our dataset contains many causal relationships annotated by human semantics and commonsense, making our task more challenging and can significantly prompt the discovery of implicit causality.

While for the causality event extraction task, although the model can roughly find out which events exist in the text with a high probability, the model cannot accurately extract the events (exact match). This is mainly because the model cannot understand the semantics of the text. Thus they may include some irrelevant phrases into events or sometimes miss some important premise phrases. In FineCR, different from previous word-level research, the annotation of events is phrase-level or even sentence-level. This places higher

requirements on the model's ability to extract events precisely.

Finally, we found that it is sometimes difficult for the model to find the corresponding answer for the causality question answering task due to the model's inability to understand the semantics well. Especially when multiple events appear in a sentence simultaneously, the model may not be able to pick the correct one. FineCR is an excellent dataset for measuring models' ability to answer causality questions.

Additionally, we use the collected meta information in our dataset to do an out-of-domain (OOD) experiment across different company sectors. Experimental results prove that OOD issues also exist in causality-related tasks. A model trained in one sector cannot be perfectly transferred to another. This is mainly because there are different causal relationships in different sectors. For example, for online retail companies, the COVID-19 epidemic will lead to a rise in stock prices, while for offline retail companies, the COVID-19 may even lead to the company's closure. The same event can lead to vastly different outcomes.

8.2. Limitations

Although we conduct a comprehensive analysis of the experiments on the dataset, our research in this thesis still suffers from certain limitations. Firstly, since our dataset is merely built on the financial data, it is unclear if the research based on it can generate open domain causality detection in the real world. And the causality events in our dataset are annotated by human annotators, which is small and expensive. If unsupervised methods exist that can extract large amounts of data from textual data, then it is possible to construct a huge causal event graph from this data. This graph will surely help us understand the world better.

For the model part, as mentioned in Section 8.1, our experiments reveal some problems that currently exist in deep learning models. Statistical-based models cannot understand human semantics well, making it difficult for the model's prediction results to achieve human-level accuracy. Actually, this is a problem that has plagued many NLP researchers. Given the length and depth of this thesis, we did not provide a better solution to this problem.

8.3. Future work

There are a lot of subsequence works that can be done in the future. First, we should try to expand our causality research from finance to other domains and finally achieve causality research in the open domain, which will make our research have more significant social value. And we can also try to extract causality triplets from more extensive texts through the automatical pipeline we constructed in this thesis. Thus build a huge causality event graph, which can help us better use causal knowledge to enable humans to make decisions. Finally, given some of the shortcomings of the current deep learning model in causality research, we can study corresponding methods, such as introducing additional knowledge, to assist the model in causal discovery and extraction to achieve better performance.

Bibliography

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[2] Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7432–7439, 2020.

[3] Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. Large-scale simple question answering with memory networks. *arXiv preprint arXiv:1506.02075*, 2015.

[4] Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. *From Opinion Mining to Financial Argument Mining*. Springer Briefs in Computer Science. Springer, 2021. ISBN 978-981-16-2880-1. doi: 10.1007/978-981-16-2881-8. URL https://doi.org/10.1007/978-981-16-2881-8.

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

[6] Michael Chen, Mike D'Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. Codah: An adversarially-authored question answering dataset for common sense. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 63–69, 2019.

[7] Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, et al. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*, 2021.

[8] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*, 2018.

[9] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*, 2019.

[10] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

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

[12] Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. A dataset of information-seeking questions and answers anchored in research papers. *arXiv preprint arXiv:2105.03011*, 2021.

[13] Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii, editors. *Proceedings of the 20th Workshop on Biomedical Language Processing*, Online, June 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.bionlp-1.0.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[15] Bhuwan Dhingra, Kathryn Mazaitis, and William W Cohen. Quasar: Datasets for question answering by search and reading. *arXiv preprint arXiv:1707.03904*, 2017.

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

[17] Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 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, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.183. URL https://aclanthology.org/2021.acl-long.183.

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

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

[20] Mahmoud El-Haj, Paul Rayson, and Nadhem Zmandar, editors. *Proceedings of the 3rd Financial Narrative Processing Workshop*, Lancaster, United Kingdom, 15-16 September 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.fnp-1.0.

[21] 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. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *arXiv preprint arXiv:2109.00725*, 2021.

[22] Lei Gao, Prafulla Kumar Choubey, and Ruihong Huang. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1179. URL https://aclanthology.org/N19-1179.

[23] Shen Gao, Zhaochun Ren, Yihong Zhao, Dongyan Zhao, Dawei Yin, and Rui Yan. Product-aware answer generation in e-commerce question-answering. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 429–437, 2019.

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

[25] Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 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, 2012.

[26] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 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, 2018.

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

[28] Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*, 2015.

[29] Lynette Hirschman and Robert Gaizauskas. Natural language question answering: the view from here. *natural language engineering*, 7(4):275–300, 2001.

[30] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

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

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

[33] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019.

[34] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

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

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

[37] Divyansh Kaushik, Amrith Setlur, Eduard Hovy, and Zachary C Lipton. Explaining the efficacy of counterfactually augmented data. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=HHiiQKWsOcV`.

[38] Katherine Keith, David Jensen, and Brendan O'Connor. 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, 2020.

[39] Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, 2018.

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

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

[42] Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328, 2018.

[43] Xiang Kong, Varun Gangal, and Eduard Hovy. Scde: sentence cloze dataset with high quality distractors from examinations. *arXiv preprint arXiv:2004.12934*, 2020.

[44] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

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

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

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

[48] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 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, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703.

[49] Patrick Lewis, Pontus Stenetorp, and Sebastian Riedel. 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, 2021.

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

[51] Zhongyang Li, Xiao Ding, Ting Liu, J. Edward Hu, and Benjamin Van Durme. 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, 7 2020. doi: 10.24963/ijcai.2020/502. URL https://doi.org/10.24963/ijcai.2020/502. Main track.

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

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

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

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

[56] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[57] Dominique Mariko, Hanna Abi-Akl, Estelle Labidurie, Stephane Durfort, Hugues De Mazancourt, and Mahmoud El-Haj. 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), December 2020. COLING. URL https://aclanthology.org/2020.fnp-1.3.

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

[59] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*, 2018.

[60] Paramita Mirza and Sara Tonelli. 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, 2016.

[61] Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 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, 2014.

[62] Stephen L Morgan and Christopher Winship. *Counterfactuals and causal inference*. Cambridge University Press, 2015.

[63] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. Ms marco: A human generated machine reading comprehension dataset. In *CoCo@ NIPS*, 2016.

[64] Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Ryu Iida, Masahiro Tanaka, and Julien Kloetzer. A Semi-Supervised Learning Approach to Why-Question Answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), 2016. URL https://ojs.aaai.org/index.php/AAAI/article/view/10388.

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

[66] Judea Pearl. *Causality*. Cambridge university press, 2009.

[67] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21 (140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

[68] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[69] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018.

[70] Siva Reddy, Danqi Chen, and Christopher D Manning. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019.

[71] Matthew Richardson, Christopher JC Burges, and Erin Renshaw. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 193–203, 2013.

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

[73] David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. Analysing mathematical reasoning abilities of neural models. *arXiv preprint arXiv:1904.01557*, 2019.

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

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

[76] Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*, 2017.

[77] Steven Sloman, Aron K Barbey, and Jared M Hotaling. A causal model theory of the meaning of cause, enable, and prevent. *Cognitive Science*, 33(1):21–50, 2009.

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

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

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

[81] Haitian Sun, William W Cohen, and Ruslan Salakhutdinov. Conditionalqa: A complex reading comprehension dataset with conditional answers. *arXiv preprint arXiv:2110.06884*, 2021.

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

[83] Simon Šuster and Walter Daelemans. Clicr: A dataset of clinical case reports for machine reading comprehension. *arXiv preprint arXiv:1803.09720*, 2018.

[84] Alon Talmor and Jonathan Berant. The web as a knowledge-base for answering complex questions. *arXiv preprint arXiv:1803.06643*, 2018.

[85] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*, 2018.

[86] Tsun-Hsien Tang, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 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, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3462999. URL https://doi.org/10.1145/3404835.3462999.

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

[88] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*, 2016.

[89] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):1–28, 2015.

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

[91] Cunxiang Wang, Pai Liu, and Yue Zhang. 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, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.251. URL https://aclanthology.org/2021.acl-long.251.

[92] Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302, 2018.

[93] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*, 2015.

[94] 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. 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, October 2020. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/2020.emnlp-demos.6`.

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

[96] Wenhan Xiong, Jiawei Wu, Hong Wang, Vivek Kulkarni, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. Tweetqa: A social media focused question answering dataset. *arXiv preprint arXiv:1907.06292*, 2019.

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

[98] Linyi Yang, Jiazheng Li, Pádraig Cunningham, Yue Zhang, Barry Smyth, and Ruihai Dong. Exploring the efficacy of automatically generated counterfactuals for sentiment analysis. *arXiv preprint arXiv:2106.15231*, 2021.

[99] Yi Yang, Wen-tau Yih, and Christopher Meek. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2013–2018, 2015.

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

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

[102] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 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, 2016.

[103] Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. Reclor: A reading comprehension dataset requiring logical reasoning. *arXiv preprint arXiv:2002.04326*, 2020.

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

[105] Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. When does pretraining help? assessing self-supervised learning for law and the casehold dataset. *arXiv preprint arXiv:2104.08671*, 2021.