e-CARE: a New Dataset for Exploring Explainable Causal Reasoning

Li Du, Xiao Ding*, Kai Xiong, Ting Liu, and Bing Qin
Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, China
{ldu, xding, kxiong, tliu,qinb}@ir.hit.edu.cn

Abstract

Understanding causality has vital importance for various Natural Language Processing (NLP) applications. Beyond the labeled instances, conceptual explanations of the causality can provide deep understanding of the causal facts to facilitate the causal reasoning process. However, such explanation information still remains absent in existing causal reasoning resources. In this paper, we fill this gap by presenting a human-annotated explainable CAusal REasoning dataset (e-CARE), which contains over 21K causal reasoning questions, together with natural language formed explanations of the causal questions. Experimental results show that generating valid explanations for causal facts still remains especially challenging for the state-of-the-art models, and the explanation information can be helpful for promoting the accuracy and stability of causal reasoning models.

1 Introduction

Causal reasoning is one of the most central cognitive abilities of human beings (Waldmann and Hagmayer, 2013; Jonassen et al., 2008), which enables one to understand the observed facts and predict the future. However, although recent causal reasoning models have achieved impressive performances on certain hand-crafted datasets, there still remains a considerable gap compared to human performances, as they cannot achieve stable performances across different datasets and are susceptible to adversarial attacks (McCoy et al., 2019; Poliak et al., 2018; Gururangan et al., 2018).

One key factor leading to such drastic contrast is that, present causal reasoning models only learn to induce empirical causal patterns that are predictive to the label, while human beings seek for deep and conceptual understanding of the causality to explain the observed causal facts. The conceptual explanations can not only serve as a touchstone to examine whether the underlying causal mechanism has been thoroughly understood, but it can also in turn support the causal reasoning process. As illustrated in Figure 1, observing the causal fact C_1: *adding rock into hydrochloric acid* causes E_1: *rock dissolved*, one may further ask *why such a causal relationship exists* and reach the plausible *conceptual explanation* that *Acid is corrosive*, which goes beyond the isolated facts and reaches the conceptual nature to reveal the principle of the causal mechanism.

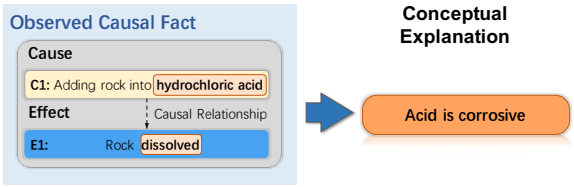

Figure 1: Conceptual explanations of observed causality can be helpful for understanding the unseen causal facts.

However, despite the critical importance of conceptual explanations in causal reasoning, there is still a lack of such an explainable causal reasoning dataset. To fill this gap, we contribute an explainable CAusal REasoning dataset (e-CARE),together with a new causal explanation generation task, and a novel Causal Explanation Quality (CEQ) evaluation metric.

The e-CARE dataset is constructed by crowdsourcing and contains over 21K multiple-choice causal reasoning questions, which makes e-CARE the largest human-annotated commonsense causal reasoning dataset to the best of our knowledge. In addition to the causal reasoning question itself, e-CARE also provides a free-text-formed conceptual explanation for each causal question to explain why the causation exists. On this basis, we propose a new causal explanation generation task that requires models not only to choose the correct causal fact but also to generate the ex-

*Corresponding author

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 432 - 446

planation for the choice. In addition, to directly measure the quality of generated explanations, we propose a novel causal explanation quality evaluation metric (namely, CEQ score). Compared to conventional text generation evaluation metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) which mainly evaluate the textual or semantic similarity between generated explanations with golden annotations, CEQ score focuses on evaluating how much promotion an explanation can bring to understanding the causal mechanism. The dataset is publicly available at https://github.com/Waste-Wood/e-CARE/.

Experimental results demonstrate that the causal questions of e-CARE are still challenging for the state-of-the-art (SOTA) pretrained language models, indicating the effectiveness of the e-CARE dataset in evaluating the causal learning ability of models. In addition, the explanation signal received in the training process can enhance the performance and the stability of the reasoning model, while the SOTA baselines still have trouble in explaining the causal facts at a conceptual level. These analyses highlight the importance of the conceptual explanations in causal reasoning, and suggest an avenue for future researches.

2 Related Work

2.1 Commonsense Causal Reasoning Datasets

Existing commonsense causal reasoning corpora differ in their annotation guidelines and how they are constructed: (1) whether the corpus is automatically constructed or built by human annotation; (2) whether the annotation unit of the corpus is word-level, phrase-level, or sentence-level.

To obtain abundant causal knowledge, a natural way is extracting causal knowledge using heuristic rules from large-scale open-domain web text corpora (Luo et al., 2016; Li et al., 2020; Sap et al., 2019). However, the reporting bias may challenge both the coverage and quality of the extracted causal knowledge.

Different from automatic construction, human annotation can endow datasets with higher precision. A line of work focuses on providing word-level causality knowledge (Girju et al., 2007; Mostafazadeh et al., 2016; Do et al., 2011; Hendrickx et al., 2019). However, a word is not a complete semantic unit, which may limit the integrity of causal expressions and lead to ambiguity. To address this issue, other datasets are constructed to provide phrase-level (Caselli and Vossen, 2017; Bethard and Martin, 2008; Mirza et al., 2014; Dunietz et al., 2017) and sentence-level (Ning et al., 2019; Roemmele et al., 2011) causal knowledge. Among these datasets, COPA (Roemmele et al., 2011) has become a widely adopted benchmark. Nevertheless, the size of COPA is rather limited, which may result in overfitting and arouse concerns about the confidence of the results.

Dataset	Anno. Unit	Size	Expl.
Automatically-Built Dataset			
CausalNet (Luo et al., 2016)	W	11M	N
CausalBank (Li et al., 2020)	P	314M	N
Human-Annotated Dataset			
SemEval-2007 T4 (Girju et al., 2007)	W	220	N
CaTeRS (Mostafazadeh et al., 2016)	W	488	N
EventCausalityData (Do et al., 2011)	W	580	N
SemEval-2010 T8 (Hendrickx et al., 2019)	W	1,003	N
ESC (Caselli and Vossen, 2017)	P	117	N
T-CBank (Bethard and Martin, 2008)	P	271	N
CausalTimeBank (Mirza et al., 2014)	P	318	N
BECauSE 2.0 (Dunietz et al., 2017)	P	1,803	N
TCR (Ning et al., 2019)	S	172	N
COPA (Roemmele et al., 2011)	S	1,000	N
e-CARE	**S**	**21K**	**Y**

Table 1: A list of previous commonsense causal reasoning datasets. In the column "Annotation Unit", "W", "P" and "S" are abbreviation of word, phrase and sentence, respectively. "Expl." is the abbreviation of "Explanation".

In this paper, we introduce an *e*xplainable CAusal REasoning dataset (e-CARE). As shown in Table 1, to the best of our knowledge, e-CARE is the largest human-annotated causal reasoning dataset. With more than 21,000 instances, the e-CARE dataset can serve as a more reliable benchmark. Furthermore, compared to previous work, e-CARE can provide additional explanation information, which plays a critical role in learning the underlying mechanism of causal knowledge.

2.2 Explainable Textual Inference

Recently, an increasing amount of datasets have been proposed to address the explainability of textual inference tasks, such as textual entailment inference (Camburu et al., 2018), question-answering (QA) (DeYoung et al., 2019; Perez et al., 2019) and multi-hop QA (Ye et al., 2020). The form and content of the explanations vary with the nature of specific tasks.

The QA task requires a model to answer the question based on evidences within given texts. Therefore, the explanation for this task should de-

Number	Train	Dev	Test	Total
Causal Questions	14,928	2,132	4,264	21,324
Uniq. Explanations	10,491	2,102	3,814	13,048

Table 2: Corpus level statistics of the e-CARE dataset. Uniq. Explanations refer to the explanations that only correspond to a single causal fact.

Premise: Tom holds a copper block by hand and heats it on fire.
Ask-for: Effect
Hypothesis 1: His fingers feel burnt immediately. (✓)
Hypothesis 2: The copper block keeps the same. (×)
Explanation: **Copper is a good thermal conductor.**

Table 3: An instance from the e-CARE dataset.

scribe where and how an answer can be found (Wiegreffe and Marasović, 2021). The explanations can have various forms, including answer-bearing sentences (Perez et al., 2019), structured information connecting the question and answer (Hancock et al., 2018; Ye et al., 2020), or even human-annotated free-formed sentences (Camburu et al., 2018; Rajani et al., 2019). In contrast, the multi-hop QA task requires the model to infer the correct answer through multiple reasoning steps. Hence, the explanation of this task needs to provide the specific reasoning paths (Wiegreffe and Marasović, 2021; Jhamtani and Clark, 2020).

Our work is quite different from previous work. We notice that all of these previous work only offer explanations that explain a specific question. Whereas we aim at providing a conceptual understanding of the causality, which has the potential to explain *a set of related causal observations*, rather than only explain a specific causal fact.

3 e-CARE: an Explainable Causal Reasoning Dataset

e-CARE contains a total of 21,324 instances, corresponding to 13,048 unique explanations. This also makes e-CARE the largest human-annotated commonsense causal reasoning benchmark. The corpus-level statistics of the e-CARE dataset are shown in Table 2.

As shown in Table 3, each instance of the e-CARE dataset is constituted by two components: (1) a multiple-choice causal reasoning question, composed of a premise and two hypotheses, and one of the hypotheses can form a valid causal fact with the premise; (2) a conceptual explanation about the essential condition that enables the existence of the causal fact. For example, as Table 3 shows, the *explanation* points out the nature of copper that *Copper is a good thermal conductor*, so that holding copper on fire will make fingers feel burnt immediately. The appendix provides more discussion about the explanations within e-CARE. On this basis, we introduce two tasks:

Causal Reasoning Task We formulate the causal reasoning task as a multiple-choice task: given a premise event, one needs to choose a more plausible hypothesis from two candidates, so that the premise and the correct hypothesis can form into a valid causal fact.

Explanation Generation Task It requires the model to generate a free-text-formed explanation for a given causal fact (composed of a premise and the corresponding *correct* hypothesis).

3.1 Data Annotation

To construct the e-CARE dataset, we start by collecting statements that describe conceptual understandings of world knowledge. Then given a statement, we ask different annotators to generate causal facts that can be explained by the statement, and build causal questions based on these causal facts. This is because we hope to provide conceptual explanations with more generality, that can explain a set of correlated causal facts, instead of only applicable to a certain isolated causal fact. Moreover, the statements can serve as clues to help the annotators to come up with causal facts.

Collecting Potential Explanations Two key issues remain in collecting statements as potential explanations: (1) what kind of statements can be potential conceptual explanations of the causal facts; (2) where to find the appropriate statements.

For the first question, Jonassen et al. (2008) concluded that, in general, the explanation of causality mainly describes three categories of information: (1) the nature or attributes of the objectives involved in the causal facts; (2) forces or actions that cause changes and drive transient motions; (3) the goals, intentions, motives or purposes of the causal agents. In addition, to be the conceptual explanation of a causal fact, the statement should be able to involve with a category of objects or people, but not only focus on a specific object or person (Sembugamoorthy and Chandrasekaran, 1986).

Following these principles, we notice that there are already several available knowledge bases containing statements about such generic world knowledge, including ConceptNet (Speer

and Havasi, 2013), WordNet (Fellbaum, 2010), Atomic (Sap et al., 2019) and GenericsKB (Bhakthavatsalam et al., 2020). However, ConceptNet and WordNet are structured knowledge graphs, containing only triplet-structured statements with a limited number of predicates. The scope of Atomic is limited in the activities of human beings. Compared to these knowledge bases, GenericsKB is an open-domain, large-scale knowledge base, containing rich generic world knowledge described in free-form text. Therefore, we collect the statements from GenericsKB to ensure the coverage and diversity of the potential explanations.

Specifically, we filter out the statements in GenericsKB with low reliability, and the statements that may disobey the above-mentioned three principles. More details are provided in the Appendix. Thereafter, a total of 19,746 statements are left to form into a potential explanation set, which is further provided to the annotators to generate the causal questions.

Annotating Causal Reasoning Questions Given the potential explanation set, annotators were recruited to generate corresponding causal questions. Specifically, a causal question is generated by two steps:

First, an annotator was presented with a statement as a potential explanation, and was instructed to write a causal fact (composed of a cause and an effect), so that the causal fact can be interpreted by the given statement. In this step, a key issue is controlling the quality of generated causal facts. Thus we demonstrated illustrative examples to guide the annotators to avoid the following mistakes:

(1) The created cause and effect are not in a valid causal relationship;

(2) The created causal fact cannot be explained by the provided statement;

(3) There are factual errors or imaginary contents in the created causal facts.

In the causal fact generation process, each statement is randomly distributed to 1-3 annotators, so that we can find some statements that could explain multiple causal facts. Note that, in this process, we do not assume all statements are necessary to be a valid explanation. In other words, we do not require that the annotators must generate a causal fact for each given statement. Instead, we leave it to the judgment of annotators. In this way, the unreliable statements can be further excluded to promote the quality of our dataset.

Model	Dev	Test
Random	50.1	50.1
GPT2 (Radford et al., 2018)	57.17	56.30
RoBERTa (Liu et al., 2019)	58.38	56.42
BERT (Devlin et al., 2019)	56.19	54.45

Table 4: Model's accuracy (%) of choosing the correct hypothesis without the premise.

After the generation of causal facts, an ask-for indicator $a \in$ ["cause", "effect"] was randomly generated, where $a =$ "cause" ("effect") means that the cause (effect) event is the hypothesis, and the effect (cause) event is the premise of the causal question, respectively. Then given the ask-for indicator, in order to control the grammar and writing style consistency, the same annotator was prompted to write a distract cause (effect) as the implausible hypothesis according to the ask-for indicator. In this process, the annotators were instructed to create the implausible hypothesis as close as possible to the true hypothesis, meanwhile prevent creating uninformative distractors (such as simply adding a "not" into the true hypothesis).

3.2 Refinement and Analysis of the e-CARE Dataset

A significant challenge in dataset construction is avoiding introducing *superficial cues* into the dataset (Gururangan et al., 2018; Poliak et al., 2018), which refers to the unintentional features that leak the label information. To address this issue, following Bhagavatula et al. (2019) and Sakaguchi et al. (2020), we employ an adversarial filtering algorithm to replace the implausible hypotheses that can easily be distinguished with the correct hypotheses using the superficial clues. More details about the adversarial filtering are provided in the Appendix. As Table 4 shows, after the adversarial filtering, without the existence of the premise, the SOTA pretrained language models can hardly distinguish two candidate hypotheses, which indicates that to predict the correct label, a model must understand the causal relationship between the premise and hypothesis, rather than only depend on the superficial cues within the two hypotheses.

After the refinement, we evaluate the quality of the annotated causal questions and collected explanations through crowdsourcing. We assess the quality of causal questions by testing if there is agreement among human raters on the answer of causal questions. Specifically, we randomly sampled 200 causal questions from e-CARE, and en-

listed 10 annotators to answer the causal questions. In this process, each causal question was evaluated by three annotators. When answering the causal questions, the raters were allowed to choose an additional option "None of the above" if neither hypothesis was deemed plausible. The human annotators achieve a 92% accuracy with a high agreement (Cohen's κ = 0.935) (Cohen, 1960).

To validate the quality of explanations, we enlisted volunteers to determine whether or not the explanations can explain corresponding causal facts. In total 200 causal facts with corresponding explanations were sampled and distributed to 10 volunteers, and each explanation was evaluated by three volunteers. After the evaluation, on average 89.5% of the explanations were deemed as valid (Cohen's κ = 0.832), showcasing the quality of the explanations in e-CARE.

4 Causal Explanation Quality (CEQ) Score

A number of automatic scores have been proposed to evaluate the quality of generated explanations, such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004). However, these metrics evaluate the quality of the generated explanations only through comparing the textual or semantic similarity between the generated explanations and the golden annotation. Alternatively, an ideal causal explanation quality evaluation metric should directly measure if the causal fact is appropriately explained by the explanation.

Hence, we propose a novel causal explanation quality evaluation metric (namely, CEQ score) as a step towards directly measuring the quality of generated explanations. We devise the CEQ score based on the consideration that a better explanation should provide more information for understanding the causality, so that the prediction model can more accurately estimate the reasonableness of the causal fact. Previous literature characterized such reasonableness as the *causal strength* of the given causal fact (Roemmele et al., 2011; Luo et al., 2016), where the *causal strength* is a score in $[0, 1]$. Hence, in theory, for a valid causal fact, its causal strength should be equal to 1. Given a valid causal fact, an explanation should help to increase its estimated causal strength to the ground-truth value 1.

Therefore, we can evaluate the quality of a generated explanation by measuring the increase of causal strength brought by the explanation. Specifically, let C, E, and X denote the cause, the effect and the generated explanation, respectively. Formally, the CEQ score is defined as:

$$\text{CEQ} = \Delta_{\text{cs}} = \text{cs}(C, E|X) - \text{cs}(C, E), \quad (1)$$

where $\text{cs}(C, E)$ is the original causal strength between C and E; $\text{cs}(C, E|X)$ is the causal strength after involvement of the additional explanation information. The explanation enhanced causal strength $\text{cs}(C, E|X)$ is defined as:

$$\text{cs}(C, E|X) = max[\text{cs}(C + X, E), \text{cs}(C, E + X)], \quad (2)$$

where "+" denotes the string concatenate operation. Therefore, the CEQ score is positively related to the *increase of causal strength* between C and E after the involvement of the explanation X.

In this paper, we employ a widely-adopted model-agnostic method proposed by Luo et al. (2016) to calculate the causal strength. The model-agnostic nature enable us to avoid reliance on certain models and keep the fairness of evaluation. Specifically, the phrase-level causal strength is derived through synthesizing the word-level causality.

$$\text{cs}(C_A, E_B) = \frac{1}{N_{C_A} + N_{E_B}} \sum_{w_i \in C_A, w_j \in E_B} \text{cs}(w_i, w_j), \quad (3)$$

where (C_A, E_B) is an arbitrary causal fact; N_{C_A} and N_{E_B} are the number of words within C_A and E_B, respectively; $\text{cs}(w_i, w_j)$ is the causal strength between word w_i and w_j, which is estimated from a large corpus as:

$$\text{cs}(w_i, w_j) = \frac{\text{Count}(w_i, w_j)}{\text{Count}(w_i)\text{Count}(w_j)^{\alpha}}, \quad (4)$$

where α is a penalty coefficient and Luo et al. (2016) empirically set $\alpha = 0.66$.

5 Experiments and Results

We examine the performance of state-of-the-art pretrained language models on the causal reasoning task and the explanation generation task. Furthermore, we investigate the specific role of explanations in causal reasoning by: (1) a predict-and-generate experiment, which requires models to conduct the causal reasoning task and generate corresponding explanations simultaneously; (2) a stability analysis using adversarial attacks.

Model	AVG-BLEU	ROUGE-l	PPL	CEQ	Human Evaluation (%)
GRU-Seq2Seq	18.66	21.32	33.71	0.024	0
GPT2 (Radford et al., 2019)	32.04	31.47	7.14	0.105	20.0
Human Generation	**35.51**	**33.46**	-	**0.144**	**89.5**

Table 6: Model performance on the explanation generation task.

Model	Accuracy (%)
GPT2 (Radford et al., 2019)	69.51
RoBERTa (Liu et al., 2019)	70.73
BART (Lewis et al., 2020)	71.65
XLNET (Yang et al., 2019)	74.58
BERT (Devlin et al., 2019)	75.38
ALBERT (Lan et al., 2019)	74.60
Human Performance	**92.00**

Table 5: Performance of pretrained language models on the test set of the causal reasoning task.

	Corr. Coef with Human Eval.	P-value
AVG-BLEU	0.032	0.749
ROUGE-l	0.021	0.836
CEQ	**0.247**	**0.013***

Table 7: Pearson Correlation coefficients between human evaluation and automatic scores. "*" denotes P-value < 0.05.

5.1 Causal Reasoning

Settings We cast the causal reasoning task as a prediction problem: The input of the model is a candidate causal fact composed of a premise and one of the corresponding candidate hypotheses. The output is a score measuring the reasonableness of the candidate causal fact. We evaluate the causal reasoning ability of several SOTA pretrained language models, including discriminative pretrained language models BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019), and ALBERT (Lan et al., 2019); as well as autoregressive generative pretrained language models GPT2 (Radford et al., 2019) and BART (Lewis et al., 2020), which can also be adapted to the predictive causal reasoning task. In this section and the following parts, all experiments are conducted using the base-sized version of the pretrained language models. Additional details about experimental settings are provided in the Appendix.

Results As shown in Table 5, ALBERT achieves the highest accuracy of 73.86% on the causal reasoning task of e-CARE. However, ALBERT can achieve an accuracy of 86.0% on the widely adopted causal reasoning benchmark COPA by our implementation. This is mainly because, on one hand, previous causal reasoning datasets are too small to evaluate the genuine reasoning ability of the model. On the other hand, previous datasets may provide some superficial cues for the reasoning models to achieve superb performances. In contrast, e-CARE is the largest causal reasoning dataset that can provide enough test instances to evaluate the actual ability of the model. Moreover, in the annotating process of e-CARE, we introduced an adversarial filtering process to avoid the influence of superficial cues on the performances of reasoning models. Hence, we believe that e-CARE dataset can serve as a new benchmark for effectively evaluating models' causal reasoning ability. We also notice that human beings can achieve an accuracy of 92.00% on the e-CARE dataset. The large gap between the human performance and the pretrained language models suggests that the causal reasoning questions provided in our dataset still remain challenging, and calls for more powerful causal reasoning models.

5.2 Explanation Generation

We investigate whether the model can generate correct explanations for given valid causal facts by training a GRU-based Seq2Seq model (Chung et al., 2014), and finetuning a generative pretrained language model GPT2 (Radford et al., 2019) on the e-CARE dataset. Both models take the concatenation of the cause and effect as input. Please refer to the Appendix for more details.

Evaluation Metrics We automatically evaluate the quality of generated explanations using average-BLEU (n=4) (Papineni et al., 2002), ROUGE-l (Lin, 2004), Perplexity (Horgan, 1995), together with our proposed CEQ score.

Human Evaluation We also assess the quality of model-generated explanations through human evaluation. Specifically, we sampled 200 explanations generated by each method. Then three workers were shown with the generated explanations, together with corresponding causal facts, and were asked to label whether the generated explanation can explain the corresponding causal fact.

Quantitative Results As shown in Table 6, 89.5% of human-written explanations are found to be valid, while the generative pretrained language

Causal Facts	(Generated) Explanation	Human Annotation	CEQ
Cause: He was infected with gram-positive bacteria. *Effect:* The doctor raised the lysozyme in his body.	Lysozyme is a chemical produced by the bacterium Bacillus thuringiensis to kill bacteria.	Lysozyme destroys cell wall of bacteria.	0.193
Cause: Pneumonia pervaded in this area. *Effect:* Many people died of Pneumonia.	Pneumonia is a leading infectious disease caused by a bacterium called Pseudomonas aeruginosa.	Pneumonia is the most common infectious disease leading to death.	0.069
Cause: Tom wanted to prevent cancer. *Effect:* The doctor told him to eat more foods containing Vitamin C.	Vitamin C is a naturally occurring amino acid found in many foods.	Vitamin C slow the growth of small cell lung cancer cells.	0.012

Table 8: Examples of explanations generated by GPT2. We highlighted the factual mistakes within the generated explanations and the totally irrelevant explanation in yellow and pink, respectively.

Model	Accu (%)	AVG-BLEU	ROUGE-l	CEQ	Human Eval. (%)	△Accu. (%) after Adv. Attack
$GPT2_{CR}$	69.51	-	-	-	-	-6.40
$GPT2_{EG}$	-	32.04	31.47	0.035	20.0	-
$GPT2_{CR\text{-}EG}$	**71.06**	**34.83**	**34.22**	**0.042**	**26.5**	**-5.49**

Table 9: Model performance on the test set of Joint Causal Reasoning and Explanation Generation task.

model GPT2 only achieves a correctness of 20.0%. The last row of Table 6 reports the score of held-out human-written explanations, which serves as a ceiling for model performance. The significant gap indicates that, although GPT2 can achieve impressive performance on various natural language generation tasks, it still remains especially challenging for GPT2 to deeply understand the causal facts and then generate explanations like human beings. This may be one of the main obstacles hindering the further improvement of present causal reasoning models.

Moreover, we measure the similarity between the automatic scores with the results of human evaluation using the Spearman correlation coefficient. As Table 7 shows, ROUGH-l and average-BLEU barely have a correlation with the results of human evaluation. This is because average-BLEU and ROUGH-l only implicitly evaluate the quality of generated explanations by measuring the textual similarity with the golden annotations. Compared to average-BLEU and ROUGH-l, the CEQ score has a significant positive relationship with the human evaluation results. This indicates the efficiency of the CEQ score in evaluating the quality of generated explanations.

Qualitative Analysis In Table 8, we provide examples of explanations generated by GPT2. We observe that GPT2 can generate a reasonable explanation for some causal facts, while the generated explanations may still contain factual mistakes, or be totally irrelevant to the given causal fact (highlighted in yellow and pink, respectively). This indicate that the explanation generation still remains challenging for the GPT2 model.

5.3 Joint Causal Reasoning and Explanation Generation

To investigate the role of causal explanations in the causal reasoning process, we trained models to jointly conduct these two tasks.

Settings Since this task requires a model to predict a label meanwhile generate an explanation, we conduct the experiments using the GPT2 model, which can be adapted to conduct the predictive causal reasoning task and explanation generation simultaneously. We denote this multi-task fine-tuned GPT2 model as $GPT2_{CR\text{-}GE}$. Details for training $GPT2_{CR\text{-}GE}$ is provided in the Appendix.

To make the performance comparable, when evaluating the performance of $GPT2_{CR\text{-}GE}$ on the causal expatiations generation task, the same as the settings in the explanation generation task, the premise and the *correct* hypothesis are taken as the input of $GPT2_{CR\text{-}GE}$ for generating explanations.

Results We measure the quality of generated explanations using the same automatic scores and human evaluation settings as the Explanation Generation experiment. The performance of causal reasoning is also measured using accuracy. The results are shown in Table 9, where $GPT2_{CR}$ denotes the GPT2 model finetuned for the causal reasoning task, and $GPT2_{EG}$ refers to the GPT2 model fine-tuned for the explanation generation task. We observe that compared with $GPT2_{CR}$, the improved performance of $GPT2_{CR\text{-}EG}$ on causal reasoning indicates that the additional explanation can be helpful for the causal reasoning task, as it prompts model to have a deep understanding of the causal mechanisms. Interestingly, by comparing with $GPT2_{EG}$ and $GPT2_{CR\text{-}EG}$, we find that learning to predict the label can also be helpful for the explanation generation process. This indicates the

synergistic effect of the causal reasoning and the explanation generation on promoting models' understanding of causal mechanism.

5.4 Stability Analysis

Previous studies indicate that models may utilize some superficial cues within the dataset to predict the label. This leads to the vulnerability of models when facing adversarial attacks (Poliak et al., 2018; McCoy et al., 2019). Learning to generate the additional conceptual explanation may promote the understanding of causality to increase the stability of the reasoning model. Hence, we conduct a stability analysis to examine the specific effect of additional explanations.

Following Bekoulis et al. (2018) and Yasunaga et al. (2018), we attack the causal reasoning system by adding a perturbation term on the word embeddings of inputs. The perturbation term is derived using the gradient-based FGM method (Miyato et al., 2016). Table 9 shows the change of causal reasoning accuracy (ΔAccu.) brought by the adversarial attack. For example, $\Delta = -6.40$ means a 6.40% decrease of prediction accuracy after the adversarial attack. We find that, compared to the vanilla GPT2_{CR} model, the explanation enhanced GPT2 model $\text{GPT2}_{\text{CR-EG}}$ demonstrates stronger stability. This suggests that, by training reasoning models to generate correct explanations of the causal facts, the understanding of the causality can be promoted, and then the stability of model performance can be increased.

5.5 Enhancing Pretrained Language Model with e-CARE

Causal knowledge is critical for various NLP applications. In this section, we investigate if the causality knowledge provided by e-CARE can be used as a resource to boost model performance on other causal-related tasks. To this end, we apply transfer learning by first finetuning a BERT model on e-CARE, then adapting the e-CARE-enhanced model (denoted as BERT_{E}) on a causal extraction task EventStoryLine 0.9 (Caselli and Vossen, 2017), two causal reasoning tasks BECauSE 2.0 (Dunietz et al., 2017) and COPA (Roemmele et al., 2011), as well as a commonsense reasoning dataset CommonsenseQA (Talmor et al., 2019). On the EventStoryLine 0.9 dataset, we conduct experiment only on the instances about within-sentence causal relationship. The results are shown in Table 10. We observe

Dataset	Metric	BERT	BERT_{E}
EventStoryLine 0.9	F1 (%)	66.5	**68.1**
BECauSE 2.1	Accu. (%)	76.8	**81.0**
COPA	Accu. (%)	70.4	**75.4**
CommonsenseQA	Accu. (%)	52.6	**56.4**

Table 10: Performance of e-CARE-enhanced BERT.

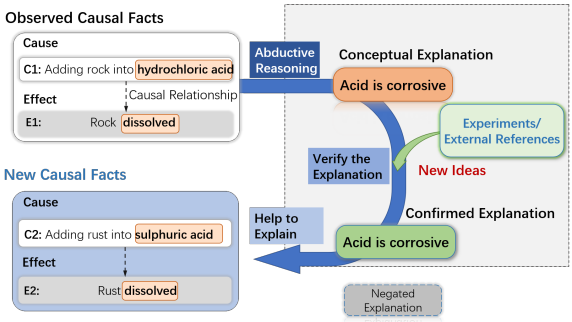

Figure 2: Conceptual explanations of observed causality can be helpful for understanding the unseen causal facts.

that the additional training process on e-CARE can consistently increase the model performance on all four tasks. This indicates the potential of e-CARE in providing necessary causality information for promoting causal-related tasks in multiple domains.

6 Discussion

In this paper, we introduce additional explanation information for the causal reasoning process, and propose a corresponding explanation generation task. Previous literature concluded the explanation generation process as an *abductive reasoning* process (Hanson, 1958; Peirce, 1974) and highlighted the importance of the abdutive explanation generation, as it may interact with the causal reasoning process to promote the understanding of causal mechanism, and increase the efficiency and reliability of causal reasoning.

For example, as Figure 2 shows, one may have an observation that C_1: *adding rock into hydrochloric acid* caused E_1: *rock dissolved*. Through abductive reasoning, one may come up with a conceptual explanation for the observation that *acid is corrosive*. After that, one can confirm or rectify the explanation by experiments, or resorting to external references. In this way, new ideas about causality can be involved for understanding the observed causal fact. Then if the explanation is confirmed, it can be further utilized to support the causal reasoning process by helping to explain and validate other related causal facts,

such as C_2: *adding rust into sulphuric acid* may lead to E_2: rust dissolved. This analysis highlights the pivotal role of conceptual explanation in learning and inferring causality. In this paper, we introduce the e-CARE dataset to provide causal explanations and support future research towards stronger human-like causal reasoning systems.

7 Conclusion

In this paper, we present an *e*xplainable CAusal REeasoning dataset *e*-CARE, which contains over 21K causal questions, together with over 13K unique conceptual explanations about the deep understanding of the causal facts, which also makes e-CARE the largest causal reasoning benchmark. Experimental results show that both the causal reasoning task and especially the explanation generation task remain challenging for the SOTA pretrained language models. Moreover, the additional explanation signal can promote both the prediction accuracy and stability of models, highlighting the vital importance of the conceptual explanations in causal reasoning.

8 Acknowledgments

We thank the anonymous reviewers for their constructive comments, and gratefully acknowledge the support of the New Generation Artificial Intelligence of China (2020AAA0106501), and the National Natural Science Foundation of China (62176079, 61976073).

References

Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. 2018. Adversarial training for multi-context joint entity and relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2830–2836.

Steven Bethard and James H Martin. 2008. Learning semantic links from a corpus of parallel temporal and causal relations. In *Proceedings of ACL-08: HLT, Short Papers*, pages 177–180.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning. In *International Conference on Learning Representations*.

Sumithra Bhakthavatsalam, Chloe Anastasiades, and Peter Clark. 2020. Genericskb: A knowledge base of generic statements. *arXiv preprint arXiv:2005.00660*.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *NeurIPS*.

Tommaso Caselli and Piek Vossen. 2017. The event storyline corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

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

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2019. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*.

Quang Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 294–303.

Jesse Dunietz, Lori Levin, and Jaime G Carbonell. 2017. The because corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 95–104.

Christiane Fellbaum. 2010. Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. Semeval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 13–18.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 107–112.

Braden Hancock, Martin Bringmann, Paroma Varma, Percy Liang, Stephanie Wang, and Christopher Ré. 2018. Training classifiers with natural language explanations. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, page 1884. NIH Public Access.

Norwood Russell Hanson. 1958. *Patterns of discovery: An inquiry into the conceptual foundations of science*, volume 251. CUP Archive.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid O Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2019. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. *arXiv preprint arXiv:1911.10422*.

John Horgan. 1995. From complexity to perplexity. *Scientific American*, 272(6):104–109.

Harsh Jhamtani and Peter Clark. 2020. Learning to explain: Datasets and models for identifying valid reasoning chains in multihop question-answering. *arXiv preprint arXiv:2010.03274*.

David H Jonassen, Ionas, and Gelu Ioan. 2008. Designing effective supports for causal reasoning. *Educational Technology Research and Development*, 56(3):287–308.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

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

Zhongyang Li, Xiao Ding, Ting Liu, J Edward Hu, and Benjamin Van Durme. 2020. Guided generation of cause and effect. IJCAI.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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

Zhiyi Luo, Yuchen Sha, Kenny Q Zhu, Seung-won Hwang, and Zhongyuan Wang. 2016. Commonsense causal reasoning between short texts. In *KR*, pages 421–431.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.

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

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016. Caters: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61.

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2019. Joint reasoning for temporal and causal relations. *arXiv preprint arXiv:1906.04941*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Charles Sanders Peirce. 1974. *Collected papers of charles sanders peirce*, volume 2. Harvard University Press.

Ethan Perez, Siddharth Karamcheti, Rob Fergus, Jason Weston, Douwe Kiela, and Kyunghyun Cho. 2019. Finding generalizable evidence by learning to convince q&a models. *arXiv preprint arXiv:1909.05863*.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 67–81.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, pages 90–95.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: An adversarial winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8732–8740.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

V Sembugamoorthy and B Chandrasekaran. 1986. Functional representation of devices and compilation of diagnostic problem-solving systems. *Experience, memory and Reasoning*, pages 47–73.

Robert Speer and Catherine Havasi. 2013. Conceptnet 5: A large semantic network for relational knowledge. In *The Peoples Web Meets NLP*, pages 161–176. Springer.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. 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 4149–4158.

Michael R Waldmann and York Hagmayer. 2013. Causal reasoning.

Sarah Wiegreffe and Ana Marasović. 2021. Teach me to explain: A review of datasets for explainable nlp. *arXiv preprint arXiv:2102.12060*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Michihiro Yasunaga, Jungo Kasai, and Dragomir Radev. 2018. Robust multilingual part-of-speech tagging via adversarial training. 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 976–986.

Qinyuan Ye, Xiao Huang, and Xiang Ren. 2020. Teaching machine comprehension with compositional explanations. *arXiv preprint arXiv:2005.00806*.

9 More Discussions about the e-CARE Dataset

9.1 The Generality of the Conceptual Explanation

In this paper, we construct the dataset by first obtaining the conceptual explanations, then obtaining the causal questions. This is because, we also hope to find the conceptual explanations with more generality, that that can explain more than one causal fact, but can explain a set of correlated causal facts. Table 11 demonstrate an example of such conceptual explanation. The *explanation* points out the nature of Copper that *Copper is a good thermal conductor*, so that holding copper on fire will make fingers feel burnt immediately. Additionally, the same explanation can also provide insights about another causal fact seemingly totally different from the case in Table 3 (a), that putting copper tubes into computer can promote thermal dispersion. This is because, the conceptual explanation points out the nature of copper, which drives a set of causal facts into existence.

This example demonstrate the usefulness of the conceptual explanations in providing the deep understanding of causality to support the causal reasoning. However, note that in this paper, we do not assume all the statements we collected can explain multiple causal facts. Instead, we resort to the empirical knowledge of human annotators to find such explanations. Specifically, we distribute statements to several annotators, and require each annotator to generate a causal fact that can be explained by the statement. For a certain statement, if it is distributed to multiple annotators and more than one annotator can generate a corresponding causal fact, then we assume that this statement can be a conceptual statement.

9.2 The Exhaustiveness of the Explanations

Another point we wish to elucidate is about the exhaustiveness of the explanations. In this paper, we only aim at providing *plausible* explanations that can explain the causal fact, but do not assume the provided explanations to be exhaustive or self-sufficient.

(a) *Premise*: Tom held a copper block by hand and heated it on fire. *Ask-for*: Effect *Hypothesis 1*: His fingers felt burnt for a short time. (✓) *Hypothesis 2*: The copper block kept the same. (×) *Explanation*: **Copper is a good thermal conductor.**	**(b)** *Premise*:This computer's heat dispersion performance is bad. *Ask-for*: Effect *Hypothesis 1*: Designers add copper tubes into the computer. (✓) *Hypothesis 2*: Designers put the computer into the ice water. (×) *Explanation*: **Copper is a good thermal conductor.**

Table 11: Two instances from the e-CARE dataset.

9.3 The Relationship between the Unique Explanations and Causal Questions

Due to the practical limits, to ensure the coverage of dataset, only a part of statements are distributed to multiple annotators, as described in Section 3.1.

10 Data Collection Details

10.1 Collection of Explanations

We collect the potential explanations from a commonsense knowledge base GenericsKB (Bhakthavatsalam et al., 2020), which contains naturally occurring generic statements, such as "Trees remove carbon dioxide from the atmosphere", collected from multiple corpora. We first filtered the statements according to their quality score s, which is a human-annotation based metric, provided in the GenericsKB and evaluating the correctness of each statement. To ensure the factual correctness of the potential explanations, we only kept the statements whose quality score are among the highest 1%. In addition, we also excluded the statements including: (1) Overly complex statements. The statements with connective, and statements with more than 20 words are excluded. This is because, by observation, we found that the annotators always struggle with understand and generate plausible causal facts for the over complex explanations. The number 20 is an empirical setting. (2) Statements describing named entities. (3) Statements describing the hypernymy or hyperonymy relationship between the subject and object. For example, the statement *Monkey is a kind of mammal.* describes the hypernymy relationship between the subject monkey and object mammal. This kind of statement does not belong to the three kinds of information that a valid explanation contains, as mentioned in Section 3.1.

After the filtering process, totally 19K statements are remained to be the potential explanations. Note that we do not assume that the statements after the filtering process are necessarily to be valid potential explanation and force the annotators to generate corresponding causal fact(s). Instead, we left the judgment to the annotators. If a statement has already been distributed to three annotators and no annotator can generate a corresponding causal question for this statement, then it is discarded.

10.2 Collection of Causal Questions

We guided the annotators using illustrative examples to avoid the following mistakes:
(1) The generated cause and effect cannot be explained by the statement.

- Wrong Case

  Explanation: Copper is a good The copper block was oxidized and the surface became dark..
  Cause: Tom held a copper block and heated it on fire.
  Effect: The copper block was oxidized and the surface became dark.

- Correct Case

  Explanation: Copper is a good thermal conductor.
  Cause: Tom held a copper block by hand and heated it on fire.
  Effect: His fingers felt burnt for a short time.

(2) The generated "cause" and "effect" do not form a valid causal relationship.

- Wrong Case

  Explanation: Oncologists specialize in the treatment of cancer.
  Cause: Jerry suffered from cancer.
  Effect: Jerry consulted many artists.

- Correct Case

  Explanation: Oncologists specialize in the treatment of cancer.
  Cause: Jerry suffered from cancer.
  Effect: Jerry consulted many oncologists.

(3) The distractor can also form a causal relationship with the premise.

- Wrong Case

  Explanation: Oncologists specialize in the treatment of cancer.
  Cause: Jerry suffered from cancer.
  Effect: Jerry consulted many oncologists.
  Disctractor Cause: Jerry consulted many traditional herbalists.

(4) The generated distractor is uninformative.

- Wrong Case

  Explanation: Copper is a good thermal conductor.
  Cause: Tom held a copper block by hand and heated it on fire.
  Effect: His fingers felt burnt for a short time.
  Disctractor Effect: His fingers did not feel burnt for a short time.

11 Adversarial Filtering

During the annotation process, some superficial clues may be incurred into the dataset, which makes the correct and implausible hypothesis can be distinguished merely using these annotation artifacts. To decrease the influence of potential annotation artifacts, we introduce an Adversarial Filtering algorithm (Bhagavatula et al., 2019) to refine our dataset.

In specific, for an arbitrary causal question $\langle p, a, h^+, h^- \rangle$, where p is the premise, $a \in [\text{“}cause\text{”}, \text{“}effect\text{”}]$ is an ask-for annotator, h^+ and h^- is the correct and wrong hypothesis, respectively, if $\langle p, h^+ \rangle$ and $\langle p, h^- \rangle$ can be easily distinguished by a predictive model, then we replace h^- with another implausible hypothesis $h^{-'}$ sampled from an implausible hypothesis set $\mathcal{H}$, so that $\langle p, h^{-'} \rangle$ is harder to be distinguished from $\langle p, h^+ \rangle$. Where the implausible hypothesis set $\mathcal{H}$ is the collection of all wrong hypotheses within the dataset.

Algorithm 1 provides a formal description of our adversarial filtering algorithm. Specifically, in each iteration i, we randomly split the dataset into a training set $\mathcal{T}_i$ and a validation set $\mathcal{V}_i$. Then a model $\mathcal{M}_i$ is trained on $\mathcal{T}_i$ to update $\mathcal{V}_i$ to make it more challenging for $\mathcal{M}_i$. To this end, given an instance $\langle p_j, a_j, h_j^+, h_{j0}^- \rangle \in \mathcal{V}_i$, we randomly sample K more implausible hypotheses $h_j^- 1', \cdots, h_j^- K'$. Let $\delta_k^{\mathcal{M}_i}$ denotes the difference of model evaluation between $\langle p_j, a_j, h_j^+, h_j^- \rangle$ and $\langle p_j, a_j, h_k^- \rangle$, where $\delta_k^{\mathcal{M}_i} < 0$ means model $\mathcal{M}_i$ favors h_j^+ to be the plausible hypothesis than the implausible hypothesis h_{jk}^-. With probability t_i, we replace h_j^- with the implausible that is hardest to distinguish with h_j^+, i.e., $h_j^- = h_{jl}^-$, $l = \arg\min_l \delta_k^{\mathcal{M}_i}$. In this way, in each iteration, the proportion of easy implausible hypotheses decreases, and then the adversary model is forced to capture more causality knowledge.

Algorithm 1 Adversarial Filtering

Input: number of iteration n, dataset $\mathcal{D}_0$, implausible hypothesis set $\mathcal{H}^-$, initial and final temperature parameter t_s and t_e.
Output: dataset $\mathcal{D}_n$

for iteration $i = 1 \rightarrow (n-1)$ **do**
  $t_i = t + e + \frac{t_s - t_e}{1 + e^{0.3(i - 3n/4)}}$
  Random split M_i into training set $\mathcal{T}_i$ and validation set $\mathcal{V}_i$
  Train Model M_i on $\mathcal{T}_i$
  for instance $j \in \mathcal{S}_i$ **do**
    for $h_{jk}^- \in \mathcal{H}_j^-$ **do**
      Calculate $\delta_k^{\mathcal{M}_i}(\langle p_j, a_j, h_j^+ \rangle, \langle p_j, a_j, h_{jk}^- \rangle)$
      $l = \arg\min_l \delta_k^{\mathcal{M}_i}$
      Sample r from a Uniform distribution $U(0, 1)$
      If $r < t_i$ or $\delta_l^{\mathcal{M}_i} < 0$ then $h_j^- = h_{jl}^-$
      Add instance j into $\mathcal{S}_i$
    end for
  end for
end for
$\mathcal{D}_n = \mathcal{S}_n$

We implemented the adversary model using pretrained language model RoBERTa-base (Liu et al., 2019). The AF algorithm is run for 25 iterations and the temperature t_i follows a sigmoid function, parameterized by the iteration number, between $t_s = 1.0$ and $t_e = 0.2$. For each instance, we sampled $K = 20$ more implausible hypotheses from the implausible hypothesis set $\mathcal{H}$.

12 Details of Experiments

12.1 Details of the Causal Reasoning Experiment

Settings In this paper, the causal reasoning task is defined as a multiple-choice problem, which requires the model to choose a more plausible hypothesis from two candidates, so that the premise and hypothesis can form a valid causal fact. Therefore, the causal reasoning task could be formalized as a prediction problem: given a candidate cause fact $\langle cause, effect \rangle$ composed of the premise event and one of the hypothesis events, the prediction model is required to predict a score mea-

Model	Input Format
GPT2	< |startoftext| > C [SEP] E < |endoftext| >
RoBERTa	<s> C <s> E <s>
BART	<s> C <s> E <s>
XLNET	<cls> C <sep> E <sep>
BERT	[CLS] C [SEP] E [SEP]
ALBERT	[CLS] C [SEP] E [SEP]

Table 12: Input format of models in the causal reasoning task.

suring the causality of the event pair. Note that the ask-for indicator decides whether the premise or candidate hypothesis to be the cause or effect, respectively.

To this end, we concatenate the premise with each one of the candidate hypothesis to form two candidate causal facts. Then each of the candidate causal fact is fed into the models, to obtain a probability measuring the plausibility of the candidate causal fact. To satisfy the input format of the pretrained language models, the input candidate causal fact is preprocessed by adding special tokens. Additionally, we adapt GPT2 and BART to predictive causal reasoning task by adding an EOS token to the end of input text, and making predictions based on the representation of the EOS token. The specific input format of the models is listed in Table 12, where C, E denotes the cause and effect of the candidate causal fact, respectively.

Training Details In the causal reasoning task, we optimize all the models with a batch size of 64, learning rate of 1e-5, and the model is finetuned for 3 epochs.

12.2 Details of the Explanation Generation Experiment

Settings In the explanation generation experiment, models are trained to generate an explanation for a given valid causal fact $\langle C, E \rangle$. Hence, the input of GPT2 is formated as:

$$< |startoftext| > C\ [SEP]\ E < |endoftext| >, \quad (5)$$

where < |startoftext| > and < |endoftext| > are two special tokens. The input of the GRU-Seq2Seq model is formated as:

$$< SOS > C\ ,\ E < EOS > . \quad (6)$$

Training Details In the explanation generation task, the GPT2 model is trained with a batch size of 32, learning rate of 1e-5, and the model is finetuned for 10 epochs. For the GRU-Seq2seq model, both the encoder and the decoder contains 2 GRU layers with a dimension of 300×300. The word embedding is initialized using 300-dimension GloVe. During optimazation, the GRU-Seq2seq model is trained for 10 epochs as well.

12.3 Details of Explanation AND Generation Experiment

Settings Given a causal question, we first concatenate the premise with each one of the candidate hypothesis to form two candidate causal facts. Then each of the candidate causal fact is fed into the GPT2 model, to get a distributed representation of the candidate causal fact. Then probability measuring the plausibility of the candidate causal fact is predicted using an MLP based on the distributed representation. After predicting plausibility score of two candidate causal facts, the model is trained to generate an explanation based on only the representation of the candidate causal fact that model thinks is more likely to be valid.

Training Details During the training process, to balance the generation loss and prediction loss, we introduce an balance coefficient λ. Hence, the loss function is formulated as $L = (1-\lambda)L_{\text{Prediction}} + \lambda L_{\text{Generation}}$. We empirically set $\lambda = 0.1$. The batch size and learning rate are also set as 32 and 1e-5, respectively. While different to the explanation generation process, in the Generate And Prediction experiment, the GPT2 model is trained for 5 epochs, as it receives two kinds of supervision signals.

12.4 Details of Transfer Analysis

Settings

All four tasks in the transfer analysis can be formalized as multiple-choice problem. Specifically, the causal event extraction task EventStoryLine requires model to predict whether two phrase-level events within a sentence can form a causal relationship. While in two causal reasoning tasks BECauSE 2.0 (Dunietz et al., 2017) and COPA (Roemmele et al., 2011), models are required to choose a plausible hypothesis, so that the premise and the hypothesis can form a valid causal fact.

Dataset	Input Format
EventStoryLine	[CLS] Statement
BECauSE 2.0	[CLS] C [SEP] E [SEP]
COPA	[CLS] C [SEP] E [SEP]
CommonsenseQA 2.0	[CLS] Q [SEP] A [SEP]

Table 13: Input format of models in the transfer analysis.

The CommonsenseQA (Talmor et al., 2019) task requires model to choose a correct answer for a given question. We list the specific format of the input on these four tasks in Table 13, where C and E denotes the cause and effect, respectively, Q and A denotes the question and answer, respectively.

Training Details To equip model with the causality knowledge within e-CARE, we train a BERT model for 3 epochs, with a batch size of 32 and a learning rate of 1e-5. Then in the following fine-tuning stage, on all four datasets, both BERT and e-CARE enhanced model BERT_{E} are fine-tuned using a grid search with the following set of hyper-parameters:

- batch size: {16, 32}
- number of epochs: {3,5,10}
- learning rate: {1e-6, 1e-5}