ExCAR: Event Graph Knowledge Enhanced 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

Prior work infers the causation between events mainly based on the knowledge induced from the annotated causal event pairs. However, additional evidence information intermediate to the cause and effect remains unexploited. By incorporating such information, the logical law behind the causality can be unveiled, and the interpretability and stability of the causal reasoning system can be improved. To facilitate this, we present an **E**vent graph knowledge enhanced e**x**plainable **CA**usal **R**easoning framework (**ExCAR**). ExCAR first acquires additional evidence information from a large-scale causal event graph as logical rules for causal reasoning. To learn the conditional probabilistic of logical rules, we propose the Conditional Markov Neural Logic Network (CMNLN) that combines the representation learning and structure learning of logical rules in an end-to-end differentiable manner. Experimental results demonstrate that ExCAR outperforms previous state-of-the-art methods. Adversarial evaluation shows the improved stability of ExCAR over baseline systems. Human evaluation shows that ExCAR can achieve a promising explainable performance.

1 Introduction

Causal reasoning aims at understanding the general causal dependency between the cause and effect (Luo et al., 2016). Causality is commonly expressed by humans in the text of natural language, and is of great value for various Artificial Intelligence applications, such as question answering (Oh et al., 2013), event prediction (Li et al., 2018), and decision making (Sun et al., 2018).

Previous work mainly learns causal knowledge from manually annotated causal event pairs, and achieves promising performances (Luo et al., 2016; Xie and Mu, 2019a; Li et al., 2019). However, recent works have questioned the seemingly superb performance for some of these studies (McCoy et al., 2019; Poliak et al., 2018; Gururangan et al., 2018). Specifically, training data may contain exploitable superficial cues that are correlative of the expected output. The main concern is that these works have not learned the underlying mechanism of causation so that their inference models are not stable enough and their results are not explainable.

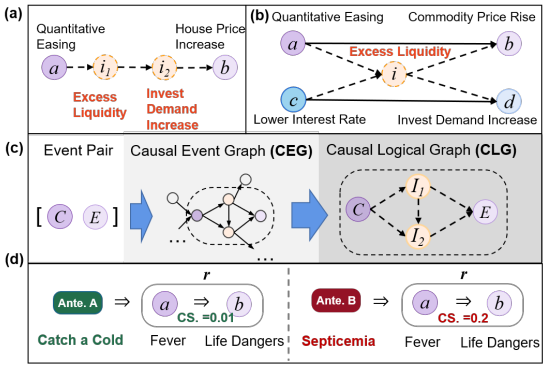

Figure 1: (a) Without the evidence event i, we can hardly reveal the implicit causation between a and b. (b) The absent of evidence events may restrict the performance of event-pair based methods. (c) Given an event pair, the ExCAR framework obtains evidence events from an event graph and conducts causal reasoning using the additional evidence events. (d) The causal strength (cs) of the same rule can vary with different antecedents. We define such phenomenon as *superimposed causal effect*.

While we notice that there is plentiful evidence information outside the given corpus that can provide more clues for understanding the logical law of the causality. Figure 1 (a) exemplifies two clues I_1 : *Excess Liquidity* and I_2: *Invest Demand Increase* for explaining how *a: Quantitative Easing* gradually leads to *b: House Price Increases*.

Without these important evidence information, on the other hand, as illustrated in Figure 1 (b),

*Corresponding author

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 2354–2363
August 1–6, 2021.

the causal relationship between $\langle a, d\rangle$ and between $\langle c, b\rangle$ could not be deducted from the known causation between $\langle a, b\rangle$ and between $\langle c, d\rangle$. In contrast, with intermediate event I in hand, according to the transitivity of causality (Hall, 2000), the logic chain of $\langle a \Rightarrow i \Rightarrow d\rangle$ and $\langle c \Rightarrow i \Rightarrow b\rangle$ could be naturally derived from the observed logic chain $\langle a \Rightarrow i \Rightarrow b\rangle$ and $\langle c \Rightarrow i \Rightarrow d\rangle$.

To fully exploit the potential of the evidence information, we present an **E**vent graph knowledge enhanced e**x**plainable **CA**usal **R**easoning (**ExCAR**) framework. In particular, as illustrated in Figure 1 (c), given an input event pair $\langle C, E\rangle$, ExCAR firstly retrieves external evidence events such as I_1, I_2 from a large-scale causal event graph (CEG, a causal knowledge base constructed by us), and defines the causation between C, I_1, I_2, E as a set of logical rules (e.g., $r_i = (E_i \Rightarrow I_i)$), which rules are useful representations for the causal reasoning task because they are interpretable and can provide insight to inference results.

Pearl (2001) pointed out that the underlying logic of causality is a probabilistic logic. The advantage of using a probabilistic logic is that by equipping logical rules with probability, one can better model statistically complex and noisy data. However, learning such probabilistic logical rules in the causal reasoning scenario is quite difficult —- it requires modeling the *superimposed causal effect* for each logical rule. Different from first-order logical rules induced from some knowledge graphs, the probability of the logical rule (i.e. the causal strength of the cause-effect pair) in causal reasoning is uncertain, which varies with different antecedents. For example, as shown in Figure 1 (d), with the antecedent A: *Catch a cold*, a *fever* can hardly lead to *life danger*. While if *fever* is caused by the antecedent B: *Septicemia*, it can result in *life danger* with a high probability.

To address this issue, we further propose a **C**onditional **M**arkov **N**eural **L**ogic **N**etwork (CMNLN) for learning the conditional causal dependency of logical rules in an end-to-end fashion. Specifically, CMNLN first decomposes the logical rules set derived from the CEG into several distinct logic chains and learns a distributed representation for each logic chain in an embedding space. Subsequently, CMNLN estimates the conditional probability of each logical rule by an antecedent-aware potential function. Then CMNLN computes the probability of each logic chain by multiplying the probabilities of logical rules in the chain. Finally, CMNLN predicts the causality score of the input event pair based on the disjunction of chain-level causality information.

Experimental results show that our approach can effectively utilize the event graph information to improve the accuracy of causal reasoning by more than 5%. Adversarial evaluation and human evaluation show that ExCAR can achieve stable and explainable performance. The code is released at https://github.com/sjcfr/ExCAR.

2 Background

2.1 Task Formalization

In this paper, both the COPA (Luo et al., 2016) and the C-COPA causal reasoning task are defined as a multiple-choice task. Specifically, as the following example shows, given a premise event, one needs to choose a more plausible cause (effect) from two hypothesis events.

Example:

Premise: The company lost money.

Ask-for: Cause

Hypothesis 1: Its products received favorable comments.

Hypothesis 2: Some of its products were defective.

Therefore, the causal reasoning task could be formalized as a prediction problem: given a cause-effect event pair $\langle C, E\rangle$ composed by the premise event and one of the hypothesis events, the prediction model is required to predict a score measuring the causality of the event pair.

2.2 Causal Event Graph

CEG is a large-scale causal knowledge base constructed by us, from which we can retrieve a set of additional evidences for a given cause-effect event pair $\langle C, E\rangle$. Formally, CEG is a directed acyclic graph and can be denoted as $G = \{V, R\}$, where V is the node set, R is the edge set. Each node $V_i \in V$ corresponds to an event, while each edge $R_{ij} \in R$ denotes that there is a causal relationship between the ith event and jth event.

2.3 Rule-based Reasoning Using Markov Logic Network

In this paper, to enhance the explainability and stability of causal reasoning, we cast the causal reasoning problem as a rule based reasoning task. Specifically, given an input causal event pair $\langle C, E\rangle$, we retrieve a set of evidence events from the CEG. The evidence events together with C and E further form

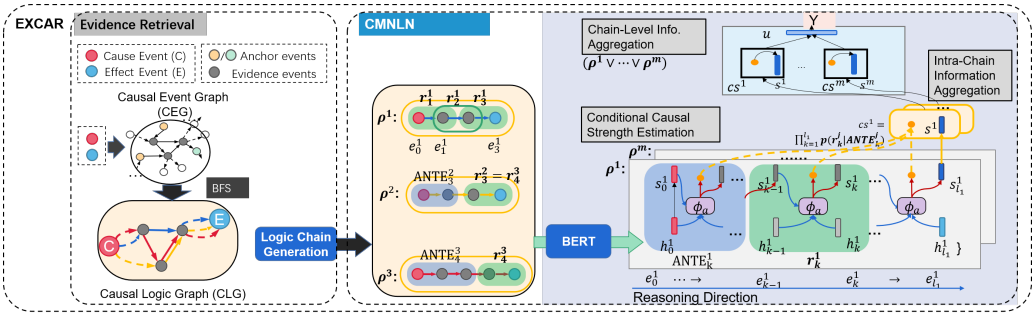

Figure 2: Illustration of the ExCAR framework and the architecture of CMNLN.

into a set of causal logical rules, where a rule describes the causal relationship between two events. Formally, a rule $r_i = (e_{i_1} \Rightarrow e_{i_2})$, where $\Rightarrow$ is a logical connective indicating the causal relationship between two events e_{i_1} and e_{i_2}. With regard to these causal logical rules, the causal mechanism can be revealed and the causal reasoning can be conducted in an explainable way.

However, the underlying logic is a probabilistic logic. Markov Logic Network (MLN) (Pearl, 1988) can model such uncertainty by assigning each causal rule a causal strength, which measures the probability that this rule holds true. Let $P(r_i)$ denote the causal strength of rule r_i. MLN estimates $P(r_i)$ using a potential function $\phi(r_i)$. Thereafter, the causality score Y is predicted by simply multiplying the causal strength of obtained rules:

$$P(Y) = \frac{1}{Z}\prod_i P(r_i) = \frac{1}{Z}\prod_i \phi(r_i), \qquad (1)$$

where $\frac{1}{Z}$ is a normalization constant.

However, there still remains two challenges for rule-based causal reasoning using MLN: 1) MLN defines potential functions as linear combinations of some hand-crafted features; 2) MLN cannot model the influence of antecedents of rules. Different from MLN, in this paper, we propose a Conditional Markov Neural Logic Network, which works on the embedding space of logic rules to model the conditional causal strength of rules.

3 Method

As shown in Figure 2, ExCAR consists of two components. Given an event pair $\langle C, E\rangle$, ExCAR employs an evidence retrieval module to retrieves evidence events from a prebuilt causal event graph to generate a set of logical rules. Then ExCAR conducts causal reasoning based on the logical rules using a Conditional Markov Neural Logic Network.

3.1 Evidence Events Retrieval

Given an event pair $\langle C, E\rangle$ outside the causal event graph, to obtain the evidences from the CEG, we first locate the cause and effect in the CEG. Intuitively, semantically similar events would have similar causes and effects, and share similar locations in the CEG. To this end, we employ a pretrained language model ELMo (Peters et al., 2018) to derive the semantic representation for events in the CEG, as well as the cause and effect event. Then events in the CEG which are semantically similar to the input cause and effect event can be found using cosine similarity of the semantic representations. These events can serve as anchors for locating the cause and effect event. Then as Figure 2 shows, taking the anchors of the cause event as start points, and taking the anchors of the effect event as end points, the evidence events can be retrieved by a Breadth First Search (BFS) algorithm.

After the retrieving process, the cause, effect and evidence events constitute a causal logical graph (CLG) $G^* = \{V^*, R^*\}$, where V^* and R^* is the node set and edge set, respectively. Each node e_i within V^* is an event, each edge r_j within R^* describes the causal relationship between two events. Taking G^* as the input, the following causal reasoning process is equipped with a set of logical rules for revealing the behind causal mechanism.

3.2 Conditional Markov Neural Logic Network

3.2.1 Overview

Given the CLG, we can derive a set of causal logical rules for supporting the causal reasoning process. However, as Figure 1 (d) shows, the causal strength of a rule may vary with different antecedents, where the antecedent can be an event, a simple rule or a complex of single rules. For clarity, we denote the *antecedent* of a rule r_i as $\mathbf{ANTE}_i$. Influenced by a certain antecedent, the

causal strength of a rule can be described by a conditional probability $P(r_i|\mathbf{ANTE}_i)$.

As shown in Figure 2, a single rule derived from the CLG can have multiple antecedents, and each of these antecedents can have its own influence on the causal strength of the rule. To address this issue by exploiting the effectiveness of neural models in representation learning, we propose the CMNLN that works on the embeddings of logical rules. To model the superimposed causal effect of rules, CMNLN regards the CLG as a composition of distinct causal logic chains $\{\rho^1, \cdots, \rho^m\}$, and predicts causality score through combining information of each causal logic chain. Hence, within each causal logic chain, we can estimate a chain-specific causal strength for each rule $r_k^j \in \rho^j$, using an antecedent-aware potential function. Then CMNLN aggregates the intra-chain causation information and inter-chain causation information to derive the causality score.

3.2.2 Logic Chain Generation

For supporting the following reasoning process, we first explore the CLG to generate all possible causal logic chains $\{\rho^1, \cdots, \rho^m\}$. As shown in Figure 2, $\rho^j = \{r_1^j \wedge, \cdots, \wedge r_{l_j}^j\}$ describes a serial of transitive causal logical rules starting from the cause event C and ending at the effect event E.

Considering that each rule $r_k^j \in \rho^j$ is composed by two events e_{k-1}^j and e_k^j , a causal logic chain ρ^j with l_j rules contains totally $l_j + 1$ events $\{e_0^j, \cdots, e_{l_j}^j\}$, where e_0^j and $e_{l_j}^j$ are the cause event C and the effect event E, respectively. Taking C and E as the start and end point respectively, we can enumerate all distinct causal logic chains in the CLG using a Depth First Searching algorithm.

3.2.3 Event Encoding

A BERT-based encoder (Devlin et al., 2019) is employed to encode all events within each causal logic chain into chain-specific distributed embeddings.

Specifically, for a causal logic chain ρ^j containing l_{j+1} events $\{e_0^j, \cdots, e_{l_j}^j\}$, we first process the event sequence into the form of: $[\mathbf{CLS}]\ \mathbf{e_0^j} \cdots [\mathbf{CLS}]\ \mathbf{e_k^j} \cdots [\mathbf{CLS}]\ \mathbf{e_{l_j}^j}$.

After that, the processed event sequence is fed into BERT. We define the final hidden state of the **[CLS]** token before each event as the representation of the corresponding event. In this way, we obtain an event embedding set $H = \{\mathbf{h}_0^j, \cdots, \mathbf{h}_{l_j}^j\}$, where $\mathbf{h}_k^j \in \mathbb{R}^d$is the embedding of the kth event within the causal logic chain ρ^j. Note that, $\mathbf{h}_0^j$ is the representation of the cause event C, and $\mathbf{h}_{l_j}^j$ is the representation of the effect event E.

3.2.4 Chain-specific Conditional Causal Strength Estimation

Given one of the causal logic chains $\rho^j = (r_1^j \wedge, \cdots, \wedge r_{l_j}^j)$ and corresponding event representations $H = \{\mathbf{h}_0^j, \cdots, \mathbf{h}_{l_j}^j\}$, CMNLN estimates the chain-specific causal strength for each rule using an antecedent-aware potential function.

For a rule $r_k^j \in \rho^j$, we define the chain-wise antecedent of r_k^j as $(r_1^j \wedge r_2^j \wedge, \cdots, \wedge r_{k-1}^j)$, and denote it as $\mathbf{ANTE}_k^j$. Therefore, with regard to $\mathbf{ANTE}_k^j$, we can derive the chain-specific causal strength using an antecedent-aware potential function as:

$$P(r_k^j|\mathbf{ANTE}_k^j) = \phi_a(r_k^j, \mathbf{ANTE}_k^j). \quad (2)$$

Considering that each logical rule $r_k^j$is composed of two events $e_{k-1}^j$and e_k^j, the input of $\phi_a(\cdot)$ is the distributed representation of $\mathbf{ANTE}_k^j$, and the embedding of $e_{k-1}^j$and e_k^j. We denote the representation of $\mathbf{ANTE}_k^j$as $\mathbf{s}_k^j$, and describe the specific process for deriving $\mathbf{s}_k^j$ in the following section.

Given $\mathbf{s}_k^j$, $\mathbf{h}_{k-1}^j$ and $\mathbf{h}_k^j$, to model the influence of $\mathbf{ANTE}_k^j$, we first derive antecedent-aware representations of e_{k-1}^j and e_k^j using an MLP:

$$\mathbf{h'}_{k-1}^j = \tanh(\mathbf{W}_c[\mathbf{s}_k^j||\mathbf{h}_{k-1}^j] + \mathbf{b}_c), \quad (3)$$

$$\mathbf{h'}_k^j = \tanh(\mathbf{W}_e[\mathbf{s}_k^j||\mathbf{h}_k^j] + \mathbf{b}_c), \quad (4)$$

where $\cdot||\cdot$ is the concatenate operation, and $\mathbf{W}_c$, $\mathbf{W}_e \in \mathbb{R}^{d\times 2d}$ are two different weight matrix modeling the influence of $\mathbf{s}_k^j$ on e_{k-1}^j and e_k^j, respectively.

Then based on the antecedent-aware event representations $\mathbf{h'}_{k-1}^j$ and $\mathbf{h'}_k^j$, we calculate the conditional causal strength of r_k^j as:

$$\phi_a(r_k^j, \mathbf{ANTE}_k^j) = \sigma(\mathbf{h'}_{k-1}^j \mathbf{W}_{cs} \mathbf{h'}_k^j), \quad (5)$$

where $\mathbf{W}_{cs} \in \mathbb{R}^{d\times d}$ are trainable parameters, and σ is a sigmoid function.

Antecedent Representation Along with the estimation of conditional causal strength, the representation of antecedents are also recursively updated. Specifically, at the first reasoning step, we initialize $\mathbf{s}_0^j$ with $\mathbf{h}_0^j$. At the kth reasoning step, $\mathbf{s}_k^j$ is obtained based on $\mathbf{s}_{k-1}^j$, the conditional causal strength $P(r_k^j|\mathbf{ANTE}_k^j)$, and the embedding of events within r_k^j :

$$\mathbf{s}_k^j = \tanh(P(r_k^j|\mathbf{ANTE}_k^j)\mathbf{W}_u[\mathbf{h}_{k-1}^j||\mathbf{h}_k^j]) + \mathbf{s}_{k-1}^j, \quad (6)$$

where $\mathbf{W}_u \in \mathbb{R}^{d\times 2d}$ is a parameter matrix.

3.2.5 Intra-Chain Information Aggregation

We aggregate the intra-chain causality information to derive a distributed representation and a chain-level causal strength for each causal logic chain.

We notice that, in the conditional causal strength estimation process, at the l_jth reasoning step, $\mathbf{ANTE}^j_{l_j+1}$ actually includes all the rules within ρ^j. Hence, we utilize the representation of $\mathbf{ANTE}^j_{l_j+1}$ as the representation of ρ^j, which we denote as $\mathbf{s}^j$.

Given the chain-specific conditional causal strength for each rule within ρ^j, we can calculate a chain-level causal strength cs^j for ρ^j by multiplying the conditional causal strength of the rules:

$$cs^j = \prod_{k=1}^{l_j} P(r^j_k|\mathbf{ANTE}^j_k) = \prod_{k=1}^{l_j} \phi_a(r^j_k, \mathbf{ANTE}^j_k). \quad (7)$$

Then we normalize the chain-level causal strengths as:

$$\hat{cs}^j = \mathrm{softmax}_j(cs^j). \quad (8)$$

3.2.6 Aggregating Chain-level Information for Predicting Causality Score

Finally, we obtain the disjunction of chain-level causality information to predict the causality score Y. Intuitively, a causal logic chain with higher causal strength should have a stronger influence on Y. Therefore, we aggregate the chain-level information through calculating a linear combination of logic chain representations $\{s^1 \cdots, s^m\}$ using the normalized causal strengths $\{\hat{cs}^1, \cdots, \hat{cs}^m\}$:

$$\mathbf{u} = \Sigma_j \hat{cs}^j \cdot s^j \quad (9)$$

where $\mathbf{u} \in \mathbb{R}^{1\times d}$ is a final state carrying information from the disjunction of $\{\rho^1, \cdots, \rho^m\}$.

The causality score Y is predicted based on $\mathbf{u}$:

$$Y = \mathrm{softmax}(\mathbf{W}_y\mathbf{u} + \mathbf{b}_y), \quad (10)$$

where $\mathbf{W}_y$ and $\mathbf{b}_y$ are trainable parameters.

3.3 Training

In the training process, we introduce a causal logic driven negative sampling to improve the reliability of conditional causal strength estimation. In particular, if there exists a rule $r_i = (e_{i_1} \Rightarrow e_{i_2})$ within the CLG, due to the unidirectionality of causality, we can derive a corresponding false rule $r_F = (e_{i_2} \Rightarrow e_{i_1})$. From the CLG, we can also generate a wrong antecedent for the false rule through random sampling. Hence, ideally, the conditional causal strength of these false rules should equal 0. In addition, we also combine the unidirectionality of causality with the transitivity of causality to generate false rules with more complex patterns (e.g.: if $e_1 \Rightarrow e_2 \Rightarrow e_3$, then we can induce a $r_F = (e_3 \Rightarrow e_1)$). By sampling false rules and training the potential functions of these false rules $\phi_a(r_F, \mathbf{ANTE}_F)$ to be zero, the reliability of conditional causal strength estimation can be enhanced.

With regard to the causal logic driven negative sampling process, the loss function of CMNLN is defined as:

$$L = L_{\text{Causality\_Score}} + \lambda L_{\text{Conditional\_CS}}, \quad (11)$$

where both $L_{\text{Causality\_Score}}$ and $L_{\text{Conditional\_CS}}$ are cross entropy loss, measuring the difference between the predicted and ground truth causality score, and between the predicted and the ideal conditional causal strength, respectively; λ is a balance coefficient.

4 Experiments

4.1 Construction of C-COPA Dataset

To evaluate the robustness of the ExCAR framework, we build an additional Chinese commonsense causal reasoning dataset C-COPA.

The C-COPA dataset is built upon a large-scale web news corpus SogouCS (Wang et al., 2008) by human annotation. We start the annotation process from manually extracting causal event pairs from raw texts within the corpus. Given a causal event pair, we first randomly generate an ask-for indicator, where ask-for ∈ [“effect”, “cause”]. Then the ask-for indicator are used to decide whether the cause or effect event to be the premise or plausible hypothesis. Given the premise, an implausible effect (cause) events is generated by a human annotator. As a result, the same as the COPA dataset, each instance within the C-COPA consists a premise event p, a plausible and an implausible hypothesis event h^+ and h^-, and an ask-for indicator a.

Three Chinese volunteers are enlisted for validating the dataset. Agreement between volunteers is high (Cohen’s K = 0.923). Instances with diverged results between volunteers are removed from the dataset. After the annotation process, a total of 3,258 instances are left and we randomly split these instances into two equal-sized parts as the development set and the test set, respectively.

4.2 Construction of Causal Event Graph

Before constructing the CEG, we have to collect a sufficient number of causal event pairs. To this

end, we harvest English causal event pairs from the CausalBank Corpus (Li et al., 2020), which contains 314 million commonsense causal event pairs in total. While the Chinese causal event pairs are collected from a raw web text corpus crawled from multiple websites date from 2018 to 2019, and filtered with keywords. More details could be found in the Appendix.

Then an English and a Chinese CEG are build based on the corresponding causal event pair corpus. To balance the computation burden and coverage of the event graph, we build the English and the Chinese CEG based on 1,500,000 Chinese and 1,5000,000 English causal event pairs randomly sampled from the whole corpus, respectively.

4.3 Experimental Settings

Given a cause or effect event, we find three most textually similar events from the causal event graph, and employ them as the anchors. In the evidence retrieving process, we limit the maximum searching depth of BFS to 3, and restrict the size of evidence event set to be no more than 8. We employ the pre-trained BERT-base model as the event encoder, which encodes each input event to a 768-dimension vector. On both datasets, for each instance, 5 negative rules are sampled to facilitate the estimation of conditional causal strength. Model is trained with the balance coefficient λ of 0.1.

4.4 Baselines

Statistical-based Methods

These methods estimate words or phrase level causality from large-scale corpora. Then the causality of an input event pair could be obtained through synthesizing the word or phrase level causality.

• PMI (Jabeen et al., 2014) measures the word-level causality using Point Mutual Information.

• PMI_EX (Gordon et al., 2011) is an asymmetric word-level PMI which takes the directionality of causal inference into consideration.

• CS (Luo et al., 2016) measures word-level causality through integrating both the necessity causality and sufficiency causality.

• CS_MWP (Sasaki et al., 2017) measures the causality between words and prepositional phrases using the CS score.

Pre-trained-model-based Methods

• BERT Wang et al. [2019a] and Li et al. [2019] finetune $\text{BERT}_{\text{base}}$ with different hyper parameters to predict the causality of each $\langle C, E \rangle$ pair.

ExCAR-based Methods

Methods	COPA	C-COPA
PMI (Jabeen et al., 2014)	58.8	56.2
PMI_EX (Gordon et al., 2011)	65.4	62.3
CS (Luo et al., 2016)	70.2	68.9
CS_MWP (Sasaki et al., 2017)	71.2	-
BERT (Wang et al., 2019a)	70.4	72.8
BERT (Li et al., 2019)	73.4	74.5
ExCAR (with CMNLN)	**78.8**	**81.5**
-w/ MLN	76.3	78.0
-w/ fixed-cs	75.0	76.9
-concat	75.4	77.1

Table 1: Accuracy (%) of causal reasoning on the test set of COPA and C-COPA.

We replace the CMNLN layer of ExCAR framework with different reasoning modules and get:

• ExCAR-w/ MLN refers to substitute the CMNLN layer by a classical Markov Logic Network layer.

• ExCAR-w/ fix-cs arbitrarily assign a fixed causal strength 0.5 for each logical rule.

• ExCAR-concat flattens the causal logical graph into a single event sequence and takes the event sequence as input.

4.5 Quantitative Analysis

We list the results on both the COPA dataset and C-COPA dataset in Table 1. We find that:

(1) Statistical-based methods, such as CS (Luo et al., 2016) and CS_MWP (Sasaki et al., 2017) achieve comparable performances with BERT-based methods, this is mainly because they harvest causal knowledge with elaborate patterns from large-scale corpus sized up to 10TB. Training BERT with such causal knowledge may provide potential space for improvement, which is left for future work.

(2) Compared to causal pair based BERT, ExCAR related methods show improved performance. This indicates that incorporating additional evidences from the event graph can be helpful for revealing the causal decision mechanism and then improve the accuracy of causal reasoning.

(3) ExCAR-w/ MLN and ExCAR -w/ CMNLN outperforms ExCAR-concat, which flats the CLG into an event sequence. This shows that exploiting the complex causal correlation patterns between logical rules can be helpful for the causal reasoning task.

(4) ExCAR-w/ MLN and ExCAR -w/ CMNLN shows improved performance compared to ExCAR -w/ fixed-cs. This confirms that neuralizing rules to account for the uncertainty of the logical rules is helpful for the causal reasoning task.

Methods	COPA	C-COPA
BERT (Li et al., 2019)	61.5 ($\Delta = -9.9$)	62.7 ($\Delta = -10.1$)
ExCAR		
-w/ CMNLN	78.2 ($\Delta = -\mathbf{0.6}$)	80.7 ($\Delta = -\mathbf{0.8}$)
-w/ MLN	76.1 ($\Delta = -1.8$)	76.4 ($\Delta = -1.6$)
-w/ fixed-cs	74.3 ($\Delta = -0.7$)	75.9 ($\Delta = -1.0$)
-concat	73.9 ($\Delta = -1.5$)	76.0 ($\Delta = -1.1$)

Table 2: Prediction accuracy (%) after adversary attack.

	Fixed-cs	MLN	CMNLN
Avg. Explainability Score	0.95	1.25	1.43

Table 3: Average explainability score of CMNLN, MLN and unified causal strength on C-COPA.

(5) ExCAR-w/ CMNLN further improves the prediction accuracy compared to ExCAR-w/ MLN, suggesting that by incorporating the antecedent-aware potential function CMNLN can model the conditional causal strength of logical rules for causal reasoning.

4.6 Stability Analysis

In this paper, we propose to enhance the stability of our approach through introducing additional evidence information. We investigate the specific influence of these evidences on the stability of our approach through an adversarial evaluation. Following Bekoulis et al. [2018] and Yasunaga et al. [2018], we attack the reasoning systems by adding a perturbation term on the word embedding of inputs. The perturbation term is derived using a gradient-based method FGM (Miyato et al., 2016).

Table 2 shows the prediction accuracy after adversary attack, and Δ denotes the change of performance brought by adversary attack. For example, Δ = -9.9 means a 9.9% decrease of prediction accuracy after the adversary attack. We find that, compared with event pair based BERT, ExCAR can significantly improve the stability of the prediction accuracy. These results show that by incorporating additional evidence events, ExCAR could reveal the behind causal mechanism to increase the stability of prediction results.

4.7 Human Evaluation for Explainability

We analyze the explainability of our approach quantitatively through human evaluations. In particular, we randomly sample 200 instances from the test set of C-COPA and make prediction using ExCAR. Then we employ three experts to give an explainability score belonging to $\{0, 1, 2\}$ to evaluate whether the causality strengths derived by our approach are reasonable, where 0 stands for unexplainable, 1 stands for moderately explainable and 2 stands for explainable. For comparison, we further introduce two baselines: (1) Markov Logic Network (MLN); (2) Fixed-cs.

The average explainability scores are shown in Table 3, from which we can observe that: (1) The average explainability scores of CMNLN and MLN are higher than that of fixed-cs. This is because, through neuralizing the logical rules and equipping the logical rules with probability, CMNLN and MLN can better model the potential noise in the retrieved evidences, as well as the uncertainty of rules. (2) The explainability score of CMNLN is further higher than that of MLN. This indicates that, CMNLN can model the conditional causal strength of logical rules using the antecedent-aware potential function, and then increase the reasonability of causal strength estimation.

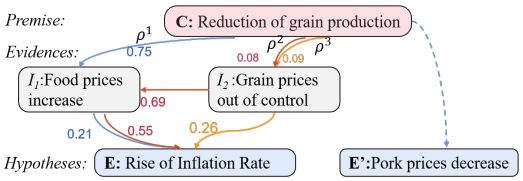

Figure 3: Example of causal reasoning result made by ExCAR.

4.8 Case Study

Figure 3 provides an example of causal reasoning made by ExCAR on C-COPA. Given a cause event *Reduction of grain production*, E: *Rise of Inflation Rate* is more likely to be the effect of the cause. However, it is difficult to directly infer the effect E: *Rise of Inflation Rate* directly from the cause event C:*Reduction of grain production*. Correspondingly, given C and E, ExCAR can obtain evidence events such as I_1: *Food prices increase* and I_2: *Grain prices out of control* from the causal event graph. These results show that ExCAR can obtain relevant evidences and hence choose the correct effect event in an explainable manner.

We also examined the estimated causal strengths. As shown in Figure 3, the causal strength between I_1 and E is higher in the logic chain ρ^2 compared to ρ^1. Intuitively, with the additional antecedent I_2: *Grain prices out of control*, I_1: *Food prices increase* could be more likely to lead to E: *Rise of Inflation Rate*. These results indicate that CMNLN can model the conditional causal strength of rules.

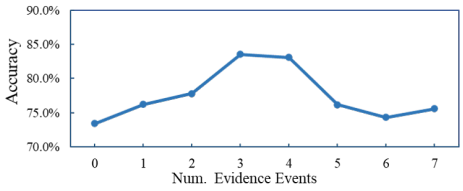

Figure 4: Reasoning accuracy of ExCAR with different number of evidence events on the test set of C-COPA.

4.9 Effect of the Number of Evidence Events

We compare the reasoning accuracy of ExCAR on samples with different numbers of evidence events. Experiments are conducted on the test set of C-COPA. Results are shown in Figure 4. We can find that, when the evidence events number increases from 0 to 3, the reasoning accuracy increases in general, since sufficient evidences are helpful for the reasoning task. However, the accuracy starts to decrease when evidence number exceeds 4. This indicates that noisy evidence events may be obtained. The inclusion of noisy evidence events emphasis the necessity of neutralizing the logical rules, as the symbolic logic based systems cannot accommodate for the noise in the rules.

5 Related Work

5.1 Causal Reasoning

Causal reasoning remains a challenging problem for today's AI systems. Statistical-based methods can provide strong baselines, as they can find some useful cues from large-scale causal corpus. For example, Gordon et al. (2011) measured the causality between words using PMI, and estimated the PMI based on a personal story corpus. While Luo et al. (2016) and Sasaki et al. (2017) further introduced direction information into a causal strength index. Then through synthesizing the word-level causality, the causality between events could be inferred.

Compared to statistical-based methods, deep neural networks enable models to learn the causality between events considering the semantics of events. To this end, Xie and Mu (2019b) devised attention-based models to capture the word-level causal relationships. While Wang et al. and Li et al. (2019) finetuned the pretrained language model BERT on causal event pairs corpus to learn the pairwise causality knowledge between events.

In this paper, we argue that in addition to the event pair itself, causal reasoning also needs to involve more evidence information. To address this issue, we propose a novel inference framework ExCAR, which is able to incorporate the additional evidence events from an event graph for supporting the causal reasoning task.

5.2 Explainable Textual Inference

Explainability has been a long-pursued goal for textual inference systems, as it can help to unveil the decision making mechanism of black-box models and enhance the stability of reasoning, which can be crucial for applications in various domains, such as medical and financial domains. To introduce interpretability in textural inference process, previous studies can be mainly divided into two categories: generating explainable information and devising self-explaining mechanism.

Beyond the task related information, automated generated textual explanations are helpful for justifying the reliability of models. For example, Camburu et al. (2018) and Nie et al. (2019) train multitask learning models to learn to generate explanations for textual entailment inference. On the other hand, the incorporation of relevant external knowledge can not only increase the model performance compared to purely data-driven approaches, but also can be helpful for understanding the model behavior (Niu et al., 2019; Wang et al., 2019b).

Another line of work designs self-explaining models to reveal the reasoning process of models. Attention mechanism was devised to explicitly measure the relative importance of input textual features. Hence, it has been widely employed to enhance the interpretability of deep neural models.

In this paper, to conduct causal reasoning in an explainable manner, we propose to induce a set of logic rules from a pre-built causal event graph, and explicitly model the conditional causal strength of each logical rule. The probabilistic logical rules can provide clues to explain the prediction results.

6 Conclusion

We devise a novel explainable causal reasoning framework ExCAR. Given an event pair, ExCAR is able to obtain logical rules from a large-scale causal event graph to provide insight to inference results. To learn the conditional probabilistic of logical rules, we propose a conditional Markov neural logic network that combines the strengths of rule-based and neural models. Empirically, our method outperforms prior work on two causal reasoning datasets, including COPA and C-COPA. Furthermore, ExCAR is interpretable by providing explanations in terms of probabilistic logical rules.

7 Acknowledgments

We thank the anonymous reviewers for their constructive comments, and gratefully acknowledge the support of the Technological Innovation "2030 Megaproject" - New Generation Artificial Intelligence of China (2018AAA0101901), and the National Natural Science Foundation of China (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.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, pages 9539–9549.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NACCL 2019*, pages 4171–4186.

Andrew S Gordon, Bejan, Cosmin A , and Kenji Sagae. 2011. Commonsense causal reasoning using millions of personal stories. In *IJCAI*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.

Ned Hall. 2000. Causation and the price of transitivity. *The Journal of Philosophy*, 97(4):198–222.

Shahida Jabeen, Xiaoying Gao, and Peter Andreae. 2014. Using asymmetric associations for commonsense causality detection. In *PRICAI*.

Zhongyang Li, , Xiao , Ding, and Ting Liu. 2018. Constructing narrative event evolutionary graph for script event prediction. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4201–4207.

Zhongyang Li, Tongfei Chen, , and Benjamin , Van Durme. 2019. Learning to rank for plausible plausibility. *arXiv preprint arXiv:1906.02079*.

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

Zhiyi Luo, Yuchen Sha, Kenny Q Zhu, Seung-won Hwang, and Zhongyuan Wang. 2016. Commonsense causal reasoning between short texts. In *Fifteenth International Conference on the Principles of Knowledge Representation and Reasoning*.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

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

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

Zheng-Yu Niu, Hua Wu, Haifeng Wang, et al. 2019. Knowledge aware conversation generation with explainable reasoning over augmented graphs. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1782–1792.

Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Motoki Sano, Stijn De Saeger, and Kiyonori Ohtake. 2013. Why-question answering using intra-and inter-sentential causal relations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1733–1743.

Judea Pearl. 1988. Probabilistic reasoning in intelligent systems; networks of plausible inference. Technical report.

Judea Pearl. 2001. Direct and indirect effects. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 411–420.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

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. *arXiv preprint arXiv:1804.08207*.

Shota Sasaki, Sho Takase, Naoya Inoue, Naoaki Okazaki, and Kentaro Inui. 2017. Handling multiword expressions in causality estimation. In *IWCS 2017*.

Yawei Sun, Cheng, Gong , and Yuzhong Qu. 2018. Reading comprehension with graph-based temporal-casual reasoning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 806–817.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019a. Superglue: A stickier

benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, pages 3266–3280.

Canhui Wang, Min Zhang, Shaoping Ma, and Liyun Ru. 2008. Automatic online news issue construction in web environment. In *Proceedings of the 17th international conference on World Wide Web*, pages 457–466.

Xiang Wang, Dingxian Wang, Canran Xu, Xiangnan He, Yixin Cao, and Tat-Seng Chua. 2019b. Explainable reasoning over knowledge graphs for recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5329–5336.

Zhipeng Xie and Feiteng Mu. 2019a. Boosting causal embeddings via potential verb-mediated causal patterns. In *IJCAI*, pages 1921–1927. I Press.

Zhipeng Xie and Feiteng Mu. 2019b. Distributed representation of words in cause and effect spaces. In *IJCAI*, volume 33, pages 7330–7337.

Michihiro Yasunaga, Jungo Kasai, and Dragomir Radev. 2018. Robust multilingual part-of-speech tagging via adversarial training. In *NAACL*.