Improving Event Causality Identification via Self-Supervised Representation Learning on External Causal Statement

Xinyu Zuo[1,2], **Pengfei Cao**[1,2], **Yubo Chen**[1,2], **Kang Liu**[1,2], **Jun Zhao**[1,2],
Weihua Peng[3] and **Yuguang Chen**[3]

[1]National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3]Beijing Baidu Netcom Science Technology Co., Ltd
`{xinyu.zuo,pengfei.cao,yubo.chen,kliu,jzhao}@nlpr.ia.ac.cn`
`{pengweihua,chenyuguang}@baidu.com`

Abstract

Current models for event causality identification (ECI) mainly adopt a supervised framework, which heavily rely on labeled data for training. Unfortunately, the scale of current annotated datasets is relatively limited, which cannot provide sufficient support for models to capture useful indicators from causal statements, especially for handing those new, unseen cases. To alleviate this problem, we propose a novel approach, shortly named CauSeRL, which leverages external causal statements for event causality identification. First of all, we design a self-supervised framework to learn context-specific causal patterns from external causal statements. Then, we adopt a contrastive transfer strategy to incorporate the learned context-specific causal patterns into the target ECI model. Experimental results show that our method significantly outperforms previous methods on EventStoryLine and Causal-TimeBank (+2.0 and +3.4 points on F1 value respectively).

1 Introduction

Event causality identification (ECI) aims to identify causal relations between events in texts, which can provide crucial clues for deep textual understanding (Girju, 2003; Oh et al., 2013, 2017). For example in Figure 1, an ECI system should identify two causal relations in S_1 with mentioned events: **noticed**$_{E1}$ $\xrightarrow{cause}$ **alerted**$_{E3}$ and **alerted**$_{E3}$ $\xrightarrow{cause}$ **ran**$_{E2}$.

To date, most existing methods regard this task as a classification problem and usually train ECI models on annotated data (Hashimoto et al., 2014; Riaz and Girju, 2014b; Mirza and Tonelli, 2016; Hu and Walker, 2017b; Gao et al., 2019). However, the scale of current annotated datasets are relatively limited, where the so far largest dataset EventStoryLine (Caselli and Vossen, 2017) only contains 258 documents, 4316 sentences, and 1770 causal
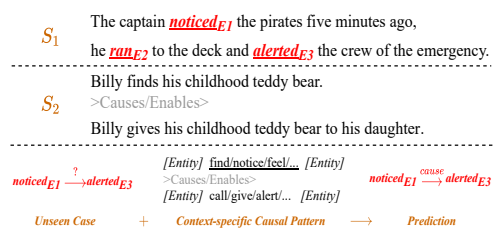

Figure 1: S_1 is a labeled data that contains unseen causal events and their statement when training; S_2 is an external causal statement; The bottom illustrates the context-specific causal pattern in S_2 could help identify the causality of unseen events in S_1.

event pairs. As a result, on the limited annotated examples, existing ECI models could not easily capture useful indicators from causal statements, especially for handing those new, unseen cases.

To address this problem, Liu et al. (2020) employed external event-related knowledge bases (KBs) to enhance the causality inference, where those KBs store inherent causal relations between some given events. For those unseen events and unlabeled causalities in KBs, Liu et al. (2020) proposed a mention-mask based reasoner to enhance the causal statement representation. However, such mention-mask based reasoner is still trained on the human-annotated examples solely. It will still suffer from data limitations and have no capacity to handling unseen contexts. Moreover, Zuo et al. (2020) improved the performance of ECI with the distantly supervised labeled training data. However, their models are still limited to the unsatisfied qualities of the automatically generated data.

To address the insufficient annotated example problem, we employ a large number of external causal statements (Sap et al., 2018; Mostafazadeh et al., 2020) that can support adequate evidence of context-specific causal patterns (Liu et al., 2020)

for understanding event causalities. For example in Figure 1, the context-specific causal pattern support by an external causal statement S_2 is helpful for identifying the causality of event $noticed_{E1}$ and event $alerted_{E3}$ in S_1, which is unseen when only training with labeled data. However, different from annotated examples for the ECI task, there are no event annotations in the external causal statements. As a result, it is difficult for the models to learn context-specific causal patterns from them to identify event causalities. To resolve this issue, inspired by Grill et al. (2020), we design a self-supervised representation learning framework to learn enhanced causal representations from external causal statements. Specifically, we iteratively sample two external causal statements, then take each of them as a target to learn the commonalities among them. Intuitively, we believe that the learned commonalities between different causal statements through self-supervision reflect such context-specific causal patterns which are helpful for identifying event causalities in the unseen cases.

Moreover, to incorporate the learned context-specific causal patterns from external causal statements into the target ECI model, we employ a contrastive transfer strategy. In specific, we regard the self-supervised representation learning module as a teacher model that *masters* abundant external causal statements, and the target ECI model as a student model. Methodologically, we make the representation of the causal events encoded by the student model should be close to the causal representation grasped by the teacher model, and keep the representation of the non-causal events away from it. In this way, the mutual information between the teacher and student models could be maximized (Tian et al., 2020). Then the learned context-specific causal patterns could be naturally transferred into the ECI model and the generalization could be improved.

In experiments, we evaluate our model on two benchmarks. The experimental results show that our model achieves SOTA performance. Then, concrete proofs show that the effectiveness of our self-supervised contrast-based framework for context-specific causal patterns learning and transfer.

In summary, the contributions are as follows:

- We propose a novel approach, shortly named **CauSeRL**, which could leverage external causal statements to identify the causalities between events.

- First of all, we design a self-supervised framework to learn context-specific causal patterns from external causal statements. Then, we adopt a contrastive transfer strategy to incorporate the learned context-specific causal patterns into target ECI model for identification.

- Experimental results on two benchmarks show that our model achieves the best performance.

2 Related Work

Event Causality Identification Up to now, identifying the causality implied in the text has attracted more and more attention (Hu and Walker, 2017a; Riaz and Girju, 2014b; Hashimoto et al., 2014; Riaz and Girju, 2014a, 2010; Do et al., 2011; Hidey and McKeown, 2016; Beamer and Girju, 2009; Hu et al., 2017; Hu and Walker, 2017b). Recently, some benchmarks on the event causality have been released. Mirza et al. (2014), Mirza and Tonelli (2016) extracted causal relation of events with a rule-based multi-sieve approach incorporating with event temporal relation. Mirza and Tonelli (2014) annotated the Causal-TimeBank of event causal relations. Caselli and Vossen (2017) annotated the EventStoryLine Corpus for event causality identification in 320 short stories based on the temporal and causal relations annotated dataset (Mostafazadeh et al., 2016). Dunietz et al. (2017) presented BECauSE 2.0, a new version of the BECauSE (Dunietz et al., 2015) of causal relation and other seven relations.

Based on the above benchmarks, Gao et al. (2019) modeled document-level structures to identify the causalities of events. Liu et al. (2020) identified event causalities with the mention masking generalization and external KBs. Zuo et al. (2020) improved the performance of ECI with the distantly automatically labeled training data. However, these methods only rely on a small scale of labeled data. In this paper, we introduce external causal statements to help identify event causalities.

Self-Supervised Representation Learning Self-supervised representation learning cares about producing good features generally helpful for many tasks (Weng, 2019). Wu et al. (2018) proposed MemoryBank, which stores representations of all the data and samples a random set of keys as negative examples. He et al. (2020) provided a framework, MoCo, of unsupervised learning visual representation as a dynamic dictionary
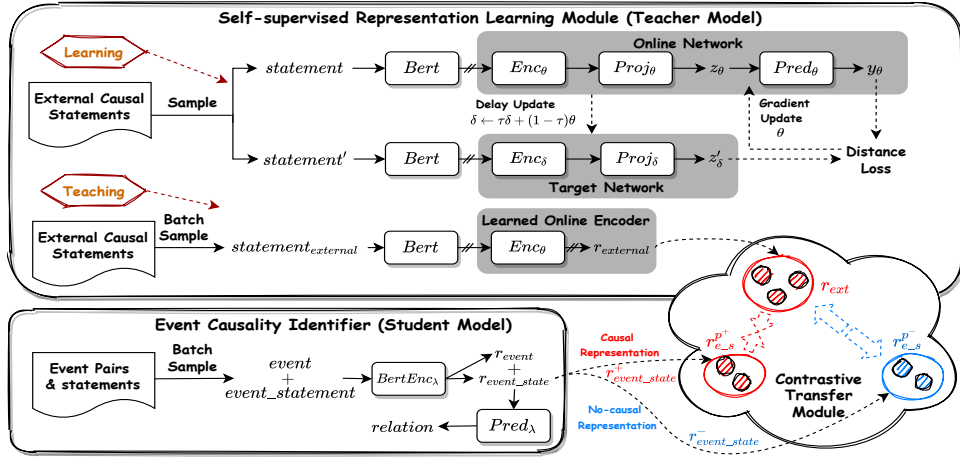
Figure 2: The learning and transfer processes of the proposed CauSeRL for ECI. "$//$" means stop-gradient.

look-up. Chen et al. (2020) proposed the SimCLR which learns representations for visual inputs by maximizing agreement between differently augmented views of the same sample via a contrastive loss. Grill et al. (2020) claimed a novel representation learning framework relies on two neural networks, BYOL, without using negative samples. CURL (Srinivas et al., 2020) applies the above ideas in reinforcement learning. Inspired by them, we design a self-supervised framework to learn context-specific causal patterns from external causal statements and adopt a contrastive transfer strategy to incorporate them into target ECI model.

3 Methodology

As shown in Figure 2, the whole pipeline process of CauSeRL is divided into two major stages.

- **Self-supervised causal representation learning** (**SelfRL**, Sec. 3.1). In this stage, we design a *self-supervised representation learning module* to learn enhanced causal representations by iteratively sampling two external causal statements, taking each of them as a target to learn their commonalities which reflect context-specific causal patterns.

- **Contrastive representation transfer** (**ConRT**, Sec. 3.2). In this stage, we employ a *contrastive transfer module* to transfer the learned context-specific causal patterns into the ECI target model, the *event causality identifier*, via incorporating the enhanced causal representations from SelfRL.

3.1 Self-Supervised Causal Representation Learning (SelfRL)

SelfRL aims to train a module that masters context-specific causal patterns from external causal statements by learning their enhanced causal representation with a self-supervised framework.

Self-Supervised Representation Learning Module We design a self-supervised module to capture the context-specific causal patterns from external causal statements via learning their enhanced causal representation. However, there are no ECI-specific event annotations in the external causal statements, which makes them unable to be directly used as training data to train the ECI model. To handle this problem, inspired by Grill et al. (2020), we iteratively sample two external causal statements, take each of them as a target to learn their commonalities, that is, the causal representations, which reflect context-specific causal patterns.

In specific, as shown in Figure 2, we configure two networks for SelfRL, an online network, and a target network. The target network provides regression targets to train the online network which makes it learn the commonalities among two input causal statements, that is, the causal representations reflecting different context-specific causal patterns. Structurally, the online network is defined as a set of weights θ which is comprised of three submodules: an *encoder* Enc_θ, a *projector* $Proj_\theta$ and a *predictor* $Pred_\theta$. And the target network has the same architecture as the online network, but no predictor and uses a different set of weights δ.

In specific, we iteratively sample two external causal statements, initially encode them by BERT (Devlin et al., 2019), and input them into two net-

works respectively. After encoding and projection, the online network and target network respectively output a projection z_θ and z'_δ. Then the online network outputs a prediction y_θ, and takes the following mean square error between ℓ_2-normalized $\bar{y}_\theta$ and $\bar{z}'_\delta$ as the training objective to learn the commonalities of two causal statements, that are regarded as the context-specific causal patterns.

$$\mathcal{L}_{\theta,\delta} \triangleq \left\| \bar{y}_\theta - \bar{z}'_\delta \right\|_2^2 = 2 - 2 \cdot \frac{\langle y_\theta, z'_\delta \rangle}{\left\| y_\theta \right\|_2 \cdot \left\| z'_\delta \right\|_2}, \tag{1}$$

$$\bar{y}_\theta \triangleq y_\theta / \left\| y_\theta \right\|_2, \; \bar{z}'_\delta \triangleq z'_\delta / \left\| z'_\delta \right\|_2. \tag{2}$$

To reduce the bias, we symmetrize the $\mathcal{L}_{\theta,\delta}$ by swapping the input causal statements of the online and target networks to compute $\widetilde{\mathcal{L}}_{\theta,\delta}$.

Learning of SelfRL For the learning of SelfRL, at each step, as shown in Algorithm 1, we minimize the $\mathcal{L}_{\theta,\delta}^{tea}$ to stochastic gradient update the online network respect to the parameters θ only. For the target network, the parameters δ are an exponential moving average of the parameters θ of the online network (Lillicrap et al., 2016):

$$\mathcal{L}_{\theta,\delta}^{tea} = \mathcal{L}_{\theta,\delta} + \widetilde{\mathcal{L}}_{\theta,\delta}, \tag{3}$$

$$\theta \leftarrow \eta_{tea} \nabla_\theta \mathcal{L}_{\theta,\delta}^{tea}, \tag{4}$$

$$\delta \leftarrow \tau\delta + (1 - \tau)\theta, \tag{5}$$

where, η_{tea} is the learning rate of the online network, and $\tau \in [0, 1]$ is the decay rate that determines the degree of the movement of θ to δ. As shown in Figure 2, when learning, BERT is only used to provide an initial representation for the input statements, and its parameters are not updated.

According to the theoretical analysis by Grill et al. (2020), the addition of a predictor on the online network and the usage of a slow-moving average of the online parameters as the target network encourage SelfRL to encode a more informative causal representation of commonalities within the online projection and avoids collapsed solutions[1].

3.2 Contrastive Representation Transfer (ConRT)

ConRT aims to incorporate the context-specific causal patterns learned in SelfRL from external

---

[1]In this paper, collapse solution means that the model encodes all input statements as the same representation. The slow-moved target network keeps the predictor of the online network always near-optimal, thus avoiding the collapse.

Algorithm 1 Two stages training of CauSeRL.

Require: External causal statements $\mathcal{C}$ for teacher model and event pairs with statements $\mathcal{P}$ for student model.
Training:
1: **Stage:** CAUSAL REPRESENTATION LEARNING
2:    **for** each batch $\mathcal{C}_{bat} \in \mathcal{C}$ **do**   ▷ *Learning of SelfRL*
3:       **for** any two causal statements $\in \mathcal{C}_{bat}$ **do**
4:          One for online another for target;
5:          Get y_θ from $Pred_\theta$ in online network;
6:          Get z'_δ from $Proj_\delta$ in target network;
7:          Swap two statements into two networks;
8:          Get symmetrical y_θ and z'_δ;
9:          Compute $\mathcal{L}_{\theta,\delta}$ and $\widetilde{\mathcal{L}}_{\theta,\delta}$;
10:       **end for**
11:       Compute batch $\mathcal{L}_{\theta,\delta}^{tea}$ in equation (3);
12:       Stochastic gradient update θ in equation (4);
13:       Slow-moving update δ in equation (5);
14:    **end for**
15: **end Stage:**
16:
17: **Stage:** CONTRASTIVE REPRESENTATION TRANSFER
18:    **for** each batch $\mathcal{P}_{bat} \in \mathcal{P}$ **do** ▷ *Learning of identifier*
19:       **for** any event pair with statement $\in \mathcal{P}_{bat}$ **do**
20:          Get r_{event} and r_{event_state} from $BertEnc_\lambda$;
21:          Predict the causality of two events in one pair;
22:       **end for**
23:       Compute batch $\mathcal{L}_\lambda^{stu}$ in equation (6);
24:       Sample $\mathcal{C}_{bat} \in \mathcal{C}$;
25:       Get $r_{external}$ of $c \in \mathcal{C}_{bat}$ from learned Enc_θ;
26:       Get $r_{event\_state}^+$, $r_{event\_state}^-$ from r_{event_state};
27:       Get mapped $r_{e\_s}^{p^+}$, $r_{e\_s}^{p^-}$ and r_{ext};
28:       Compute $\mathcal{L}_\lambda = \mathcal{L}_\lambda^{stu} + \mathcal{L}_\lambda^{con}$ in equation (8);
29:       Stochastic gradient update λ in equation (9);
30:    **end for**
31: **end Stage:**

---

causal statements into the identifier. As aforementioned, the goal of SelfRL is learning the commonalities among different external causal statements, which does not make the representation learning module have the ability to distinguish the causal and non-causal statements directly. Therefore, we employ a contrastive transfer module to teach the learned context-specific causal patterns to the event causality identifier for training.

Event Causality Identifier Event causality identification is formulated as a sentence level binary classification problem. Specifically, we design a classifier based on BERT (Devlin et al., 2019) to build our identifier. The input is an event pair and its statement. As shown in Figure 2, we take representation of events r_{event} and their contextual statement r_{event_state} encoded by $BertEnc_\lambda$ as the input of top MLP predictor. Finally, the output is a binary vector to indicate the causal relation of the input two events expressed by their statement. The parameters of the identifier are defined as λ and the optimization function is the following

Resource	Original Causal Statement Form	Converted Causal Statement Form
GLU-SPE	Billy finds his childhood teddy bear $> Cause/Enable >$ Billy gives his childhood teddy bear to his daughter	Billy finds his childhood teddy bear, billy gives his childhood teddy bear to his daughter.
GLU-GEN	Someone_A finds Something_A $> Cause/Enable >$ Someone_A gives Something_A to Someone_B	Someone_A finds Something_A, Someone_A gives Something_A to Someone_B.
ATOMIC	PersonX follows PersonY into room $> oWant >$ to know why PersonX is following them	PersonX follows PersonY into room, to know why PersonX is following them.
DISTANT	Fisk was shot to death by his mistress's new lover and Fisk's ex-business partner.	Fisk was shot to death by his mistress's new lover and Fisk's ex-business partner.

Table 1: The original and converted form (the input form of SelfRL) of different causal statements from three resources. GLU-SPE and GLU-GEN denote the specific and general statements from GLUCOSE respectively.

classification cross-entropy function:

$$\mathcal{L}_\lambda^{stu} = \text{CROSSE}(\text{MLP}([\boldsymbol{r}_{event}; \boldsymbol{r}_{event\_state}])). \quad (6)$$

Contrastive Transfer Module As aforementioned, inspired by Tian et al. (2020), we employ a contrastive transfer strategy to transfer the *"knowledge"* mastered by the teacher (self-supervised representation learning module), that is the context-specific causal patterns, to the student (event causality identifier), which helps the latter to identify the event causalities. The key idea of contrastive transfer is intuitional: maximize the mutual information between the teacher and the student (Tian et al., 2020). Methodologically, we make the representation of the statements of causal events encoded by the student model should be close to the causal representation grasped by the teacher model. By contrast, we keep the representation of the statements of non-causal events away from it.

As shown in Figure 2, at each training step of identifier, we sample a batch of external causal statements into the learned Enc_θ of the online network to obtain their causal representation $\boldsymbol{r}_{ext}$ for teaching. At the same time, we also sample a batch of event pairs with their statements into the $BertEnc_\lambda$ of identifier to obtain the statement representation $\boldsymbol{r}_{event\_state}$ of each event pair. Among one batch, $\boldsymbol{r}_{event\_state}$ consists of the $\boldsymbol{r}_{event\_state}^+$ of causal event pairs and the $\boldsymbol{r}_{event\_state}^-$ of non-causal event pairs. After mapping $\boldsymbol{r}_{external}$, $\boldsymbol{r}_{event\_state}^+$ and $\boldsymbol{r}_{event\_state}^-$ into a same space, we obtain $\boldsymbol{r}_{ext}$, $\boldsymbol{r}_{e\_s}^{p+}$ and $\boldsymbol{r}_{e\_s}^{p-}$ respectively. After that, we make $\boldsymbol{r}_{e\_s}^{p+}$ be close to $\boldsymbol{r}_{ext}$ in the contrastive loss function:

$$\mathcal{L}_\lambda^{con} = \frac{1}{|\mathcal{P}^+|} \sum_{p^+ \in \mathcal{P}^+} \log \frac{e^{(\mathcal{D}(\boldsymbol{r}_{e\_s}^{p+}, \boldsymbol{r}_{ext})/T)}}{\sum_{p \in P} e^{(\mathcal{D}(\boldsymbol{r}_{e\_s}^p, \boldsymbol{r}_{ext})/T)}}, \quad (7)$$

where, P^+ and P are the causal event pairs and all event pairs in one batch respectively, T is a temperature that adjusts the concentration level,

and $\mathcal{D}$ is the ℓ_2-distance function to measure the distance of two representation.

Learning of Event Causality Identifier For the training of event causality identifier, we add contrastive loss to the basic classification loss, which could guide the identifier to learn context-specific causal patterns implied in the enhanced causal representation from SelfRL. As shown in Algorithm 1, we minimize the L_λ and stochastic gradient update the λ as following:

$$\mathcal{L}_\lambda = \mathcal{L}_\lambda^{stu} + \mathcal{L}_\lambda^{con}, \quad (8)$$
$$\lambda \leftarrow \eta_{stu} \nabla_\lambda \mathcal{L}_\lambda, \quad (9)$$

where, η_{stu} is the learning rate of the identifier. For evaluation, we predict the causality of input event pair without the contrastive transfer module. Additionally, the T in $\mathcal{L}_\lambda^{con}$ indirectly plays a role in adjusting the influence weight of $\mathcal{L}_\lambda^{stu}$ and $\mathcal{L}_\lambda^{con}$. In specific, for teaching, we take the learned Enc_θ of the online network as the encoder, freeze its parameters, to provide the enhanced causal representation of the external causal statements for contrastive representation transfer.

4 Experiments

4.1 Experimental Setup

Dataset and Evaluation Metrics for ECI Our experiments are conducted on two main benchmarks, including: **EventStoryLine** v0.9 (ESC) (Caselli and Vossen, 2017) described above; and (2) **Causal-TimeBank** (CTB) (Mirza and Tonelli, 2014) which contains 184 documents, 6813 events, and 318 causal event pairs. Same as previous methods, we use the last two topics of ESC as the development set for two datasets. For evaluation, we adopt Precision (P), Recall (R), and F1-score (F1) as evaluation metrics. We conduct 5-fold and 10-fold cross-validation on ESC and CTB respectively, same as previous methods. All the results are the average of three independent experiments.

Data Preparation for Self-Supervised Causal Representation Learning We take four types of external causal statements from three resources. Table 1 illustrates the original form and the converted input form of SelfRL (Sec. 3.1) of the causal statements from three different resources.

- **GLUCOSE** (Mostafazadeh et al., 2020): a large-scale dataset of implicit commonsense knowledge, encoded as causal explanatory mini-theories inspired by cognitive psychology. Each GLUCOSE explanation is stated both as a specific statement (grounded in a given context, **GLU-SPE** in Table 1) and a corresponding general rule (applicable to other contexts, **GLU-GEN** in Table 1).

- **ATOMIC** (Sap et al., 2018): an atlas of machine commonsense, as a step toward addressing the rich spectrum of inferential knowledge that is crucial for commonsense reasoning.

- **DISTANT** (Zuo et al., 2020): the automatically labeled training data for ECI via distant supervision that expresses the causal semantics between events.

Parameters Settings In implementations, all the BERT modules are implemented on BERT-Base architecture[2], which has 12-layers, 768-hiddens, and 12-heads. We employ the one-layer BiLSTM (Hochreiter and Schmidhuber, 1997) as Enc_θ and Enc_δ. For parameters, we set the learning rate of SelfRL (η_{tea}) and identifier (η_{stu}) as 1e-5 and 2e-5 respectively. The size of the space in the contrastive transfer module and the hidden layer of BiLSTM are both set as 50. And we respectively set the decay rate τ of moving average in SelfRL and the temperature of the contrastive loss $\mathcal{L}_\lambda^{con}$ are 0.996 and 0.1 tuned on the development set. Moreover, we also tune the batch size of SelfRL and identifier as 48 and 16 respectively on the development set. And we apply the early stop and AdamW gradient strategy to optimize all models. We also adopt a negative sampling rate of 0.6 for the training of identifier, owing to the sparseness of positive examples in the ECI datasets.

Compared Methods Same as previous methods. For ESC, we prefer 1) **S-Path** (Cheng and Miyao, 2017), a dependency path based sequential method

---

Methods	P	R	F1
EventStoryLine			
S-Path (Cheng and Miyao, 2017)	34.0	41.5	37.4
S-Fea (Choubey and Huang, 2017)	32.7	44.9	37.8
LR+ (Gao et al., 2019)	37.0	45.2	40.7
ILP (Gao et al., 2019)	37.4	55.8	44.7
BERT	36.0	56.8	44.1
KnowDis (Zuo et al., 2020)	39.7	66.5	49.7
MasG (Liu et al., 2020)	41.9	62.5	50.1
KnowDis+CauSeRL (**Ours**)	40.1	68.9	50.7*
MasG+CauSeRL (**Ours**)	40.8	68.0	51.0*
CauSeRL$_{DISTANT}$ (**Ours**)	39.9	67.3	50.1*
CauSeRL$_{ATOMIC}$ (**Ours**)	41.0	68.1	51.2*
CauSeRL$_{GLU-GEN}$ (**Ours**)	41.4	67.8	51.4*
CauSeRL$_{GLU-SPE}$ (**Ours**)	**41.9**	**69.0**	**52.1***
Causal-TimeBank			
Rule-B (Mirza and Tonelli, 2014)	36.8	12.3	18.4
Data-D (Mirza and Tonelli, 2014)	67.3	22.6	33.9
VerR-C (Mirza, 2014)	**69.0**	31.5	43.2
BERT	39.5	44.5	41.9
MasG (Liu et al., 2020)	36.6	55.6	44.1
KnowDis (Zuo et al., 2020)	42.3	60.5	49.8
MasG+CauSeRL (**Ours**)	42.6	62.5	50.7*
KnowDis+CauSeRL (**Ours**)	42.5	66.0	51.7*
CauSeRL$_{DISTANT}$ (**Ours**)	41.6	63.9	50.4*
CauSeRL$_{ATOMIC}$ (**Ours**)	42.8	67.0	52.2*
CauSeRL$_{GLU-GEN}$ (**Ours**)	43.0	66.8	52.3*
CauSeRL$_{GLU-SPE}$ (**Ours**)	43.6	**68.1**	**53.2***

Table 2: Results of event causality identification on two benchmarks. Bold denotes best results; * denotes a significant test at the level of 0.05;

that models the context between events to identify causality; 2) **S-Fea** (Choubey and Huang, 2017), a sequence model explores complex human designed features for ECI; 3) **LR+** and **ILP** (Gao et al., 2019), document-level models adopt document structures for ECI.

For CTB, we prefer 1) **Rule-B**, a rule-based system; 2) **Data-D**, a data driven machine learning based system; 3) **VerR-C**, a verb rule based model with data filtering and causal signals enhancement. These models are designed by Mirza and Tonelli (2014; 2014) for ECI. For both two datasets, 1) we build a baseline **BERT** (our basic proposed event causality identifier); 2) We prefer **MasG** (Liu et al., 2020), a BERT-Large based SOTA model with mention masking generalization; 3) **KnowDis** (Zuo et al., 2020) improved the performance of ECI with the distantly labeled training data.

To make a fair comparison, we employ CauSeRL to retrain MasG and KnowDis to illustrate the effectiveness of our proposed approach for ECI on other methods. In specific, 1) **MasG+CauSeRL**: we retrain MasG with L_λ^{con} based on the CLU-SPE. To be consistent with other BERT-based compared models, we re-construct MasG based on BERT-Base rather than the original BERT-Large of MasG;

Methods	P	R	F	▽
CauSeRL$_{GLU\text{-}SPE}$	**41.9**	**69.0**	**52.1***	-
$Enc_{\theta-init}$ + ConRT	39.1	63.6	48.4*	-3.7
$BertEnc_{\lambda-init}$ + ConRT	38.9	63.1	48.1*	-4.0
BERT	36.0	56.8	44.1	-
BERT+SelfRL$_{finetune}$	38.5	60.9	47.2*	+3.1

Table 3: Ablation results of the self-supervised causal representation learning (SelfRL, Sec. 3.1) of ECI on EventStoryLine. * denotes a significant test at the level of 0.05; ▽ means the points lower than CauSeRL or higher than BERT in the upper and lower parts respectively; $Enc_{\theta-init}$ + ConRT denotes a varietal CauSeRL that removes SelfRL, directly employs an initial Enc_{θ} of the online network to encode external causal statements into ConRT and trains it meanwhile; $BertEnc_{\lambda-init}$ + ConRT denotes a varietal CauSeRL that removes SelfRL, directly employs a same initial $BertEnc_{\lambda}$ of identifier to encode external causal statements into ConRT and trains it meanwhile; BERT+SelfRL$_{finetune}$ denotes a varietal CauSeRL that removes ConRT (Sec. 3.2), and takes the learned Enc_{θ} of the online network as the initial encoder of identifier on the BERT baseline model.

Methods	P	R	F	▽
CauSeRL$_{GLU\text{-}SPE}$	**41.9**	**69.0**	**52.1***	-
$Enc_{\theta-freeze}$ + SelfRL	37.8	59.9	46.4*	-5.7
$Enc_{\theta-finetune}$ + SelfRL	38.5	60.9	47.2*	-4.9
BERT	36.0	56.8	44.1	-
BERT + ConRT$_{Enc_{\theta}}$	39.1	63.6	48.4*	+4.3

Table 4: Ablation results of the contrastive representation transfer (ConRT, Sec. 3.2) of ECI on EventStoryLine. * denotes a significant test at the level of 0.05; ▽ means the points lower than CauSeRL or higher than BERT in the upper and lower parts respectively; $Enc_{\theta-freeze}$ + SelfRL denotes a varietal CauSeRL that removes ConRT, and takes the frozen learned Enc_{θ} of the online network as the encoder of identifier; $Enc_{\theta-finetune}$ denotes a varietal CauSeRL that removes ConRT, and takes the learned Enc_{θ} of the online network as the initial encoder of identifier; BERT + ConRT$_{Enc_{\theta}}$ denotes a varietal CauSeRL that removes SelfRL (Sec. 3.1), directly employs an initial Enc_{θ} of the online network to encode external causal statements into ConRT and trains it meanwhile.

2) **KnowDis+CauSeRL**: we regard the automatically distantly labeled causal sentences generated by KnowDis as causal statements to learn in SelfRL, and transfer to KnowDis.

CauSeRL$_{External\text{-}Statement}$: To further illustrate the ability of CauSeRL to learn the context-specific causal patterns for the ECI task, we make CauSeRL learn from four types of external causal statements shown in Table 1 for identifying the causalities between events. $External\text{-}Statement$ denotes what kind of external causal statements.

4.2 Our Method vs. State-of-the-art Methods

Table 2 shows the results of ECI on EventStoryLine and Causal-TimeBank. From the results:

1) Our CauSeRL outperforms all baseline methods and achieves the best performance on F1 value, 52.1% on ESC and 53.2% on CTB respectively. Specifically, CauSeRL outperforms the no-bert (ILP/VerR-C) and bert (MasG/KnowDis) baseline methods by a margin of 7.4%/10.0% and 2.0%/3.4% on two benchmarks respectively. It illustrates the context-specific causal patterns from external causal statements are effective for ECI.

2) Comparing MasG+CauSeRL with MasG, we note that even with BERT-Base, the performance of MasG+CauSeRL is significantly higher than that of MasG based on BERT-Large. This shows that the context-specific causal patterns learned by CauSeRL from external causal statements can ef-

fectively alleviate the limitation of mask generalization only relying on limited labeled causal context.

3) Comparing KnowDis+CauSeRL with KnowDis, we find that CauSeRL could more efficiently make use of the automatically labeled causal statements, which learns their context-specific causal patterns to further enhance the ability of models to identify the causalities between events.

4) Comparing different external causal statements. a) GLU-SPE brings the most significant improvement because the specific causal statements from GLU-SPE have complete text structures that are more similar to ECI labeled data and make models easier to learn. There, all the ablation experiments are conducted on GLU-SPE. b) The effects of GLU-GEN and ATOMIC are similar because these two types of statements are abstract causal structures. Although they are similar to the context-specific causal patterns, it is relatively difficult to understand directly. c) The improvement brought by DISTANT is relatively small because of the effects of the noise from distantly labeled data.

5) Comparing CauSeRL with MasG+CauSeRL, we notice that after removing the ConceptNet knowledge enhancement employed by MasG, the external causal statements could be better learned and transferred. This is because MasG directly flattens the event concept knowledge into the statement sequence, which disrupts the statement structure and affects the understanding of the statement.

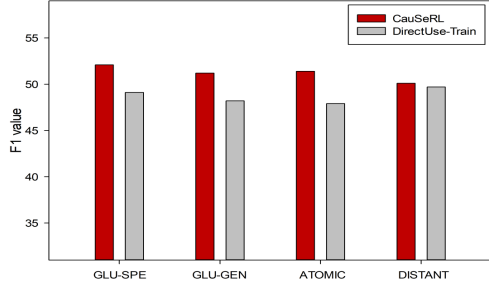6) It is worth noting that the improvement on

Figure 3: Results of event causality identification on EventStoryLine that directly using external causal statements as the training data of ECI task.

Figure 4: Case study of the probability changes with external causal statements enhancement.

the CTB is higher than that of the ESC, because the amount of labeled data of the former is relatively small, and more need for the help of external causal statements. Moreover, compared with the traditional methods based on features or rules, all BERT-based methods demonstrate high recall value, which is benefited from more training data, knowledge and causal statements.

4.3 Effect of Self-Supervised Causal Representation Learning

We analyze the effect of the self-supervised causal representation learning (SelfRL, Sec. 3.1). As shown in Table 3, from the results, 1) after removing SelfRL, the performance of ECI significantly decreases. This illustrates that the context-specific causal patterns learned by SelfRL are important for the ECI model to understand the causality. 2) Comparing BERT+SelfRL$_{finetune}$ with BERT, the Enc_θ that has learned from external causal statements could improve the performance of ECI to a certain extent. This illustrates that SelfRL could effectively capture the context-specific causal patterns in the statements for identification. 3) Comparing $Enc_{\theta-init}$ + ConRT and $BertEnc_{\lambda-init}$ + ConRT, after representation learning, the fine-tuned Enc_θ could further improve the performance of ECI. This indirectly shows that the context-specific causal patterns learned in the SelfRL is generalized.

4.4 Effect of Contrastive Representation Transfer

We analyze the effect of the contrastive representation transfer (ConRT, Sec. 3.2). As shown in Table 4, from the results, 1) after removing ConRT, the performance of ECI also significantly decreases. This illustrates that the learned causal representations from external statements are not suitable for direct application to ECI, and needs to be ef-
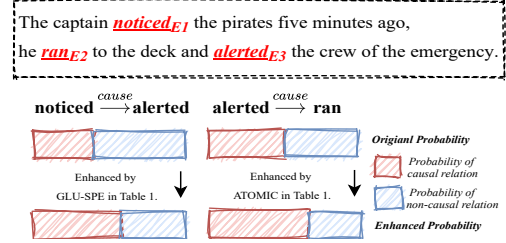
fectively transferred that the ConRT focuses on. 2) Comparing BERT + ConRT$_{Enc_\theta}$ with BERT, even if causal representation learning is not carried out in advance, adopting contrast strategy to directly transfer the context-specific causal patterns could also help the inference of event causality to a certain extent. 3) Comparing $Enc_{\theta-freeze}$ + SelfRL with $Enc_{\theta-finetune}$ + SelfRL, we find that the causal representations encoded by pre-trained BERT and BiLSTM have similar effects. Aforementioned, to avoid collapse solutions (Sec. 3.1), we choose the BiLSTM as an encoder in SelfRL that could be initialized completely independently.

4.5 Effect of the Utilization of External Causal Statement

As shown in Figure 3, we regard external causal statements as positive training data for ECI and directly use them to train the BERT baseline model. In specific, we treat two words that play a predicate role in the syntactic structure of each statement as events. From the results, CauSeRL could more effectively make use of causal statements to help understand the causalities of events. In contrast, directly serving as training data is not effective.

4.6 Case Study

As shown in Figure 4, with limited labeled data, the model could not understand the causal relation between event *noticed* and event *alerted*. Fortunately, with the support of the context-specific causal pattern from GLU-SPE in Table 1, the prediction is modified correctly. Moreover, the original model that only trained with limited labeled data is ambiguous about the causal relation between event *alerted* and event *ran*. Influenced by the similar causal statements with the example in Table 1 from ATOMIC, the prediction confidence is improved.

5 Conclusion

We propose a novel approach, CauSeRL, which could leverage external causal statements to identify the causalities of events. First of all, we design a self-supervised framework to learn context-specific causal patterns from external causal statements. Then, we adopt a contrastive transfer strategy to incorporate the learned context-specific causal patterns into the target ECI model for identification. Experimental results on two benchmarks show that our model achieves the best performance.

Acknowledgments

References

Brandon Beamer and Roxana Girju. 2009. Using a bigram event model to predict causal potential. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 430–441. Springer.

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, Vancouver, Canada. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations.

Fei Cheng and Yusuke Miyao. 2017. Classifying temporal relations by bidirectional LSTM over dependency paths. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–6, Vancouver, Canada. Association for Computational Linguistics.

Prafulla Kumar Choubey and Ruihong Huang. 2017. A sequential model for classifying temporal relations between intra-sentence events. pages 1796–1802.

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

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, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2015. Annotating causal language using corpus lexicography of constructions. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 188–196, Denver, Colorado, USA. Association for Computational Linguistics.

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

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

Roxana Girju. 2003. Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering*, pages 76–83, Sapporo, Japan. Association for Computational Linguistics.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. 2020. Bootstrap your own latent - A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, Motoki Sano, István Varga, Jong-Hoon Oh, and Yutaka Kidawara. 2014. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers), pages 987–997, Baltimore, Maryland. Association for Computational Linguistics.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning.

Christopher Hidey and Kathy McKeown. 2016. Identifying causal relations using parallel Wikipedia articles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1424–1433, Berlin, Germany. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9:1735–80.

Zhichao Hu, Elahe Rahimtoroghi, and Marilyn Walker. 2017. Inference of fine-grained event causality from blogs and films. pages 52–58.

Zhichao Hu and Marilyn Walker. 2017a. Inferring narrative causality between event pairs in films. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 342–351, Saarbrücken, Germany. Association for Computational Linguistics.

Zhichao Hu and Marilyn A Walker. 2017b. Inferring narrative causality between event pairs in films. *arXiv preprint arXiv:1708.09496*.

T. Lillicrap, J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and Daan Wierstra. 2016. Continuous control with deep reinforcement learning. *CoRR*, abs/1509.02971.

Jian Liu, Yubo Chen, and Jun Zhao. 2020. Knowledge enhanced event causality identification with mention masking generalizations. In *IJCAI-20*, pages 3608–3614. International Joint Conferences on Artificial Intelligence Organization. Main track.

Paramita Mirza. 2014. Extracting temporal and causal relations between events. pages 10–17.

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

Paramita Mirza and Sara Tonelli. 2014. An analysis of causality between events and its relation to temporal information. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2097–2106, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Paramita Mirza and Sara Tonelli. 2016. CATENA: CAusal and TEmporal relation extraction from NAtural language texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 64–75, Osaka, Japan. The COLING 2016 Organizing Committee.

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, San Diego, California. Association for Computational Linguistics.

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. GLUCOSE: GeneraLized and COntextualized story explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4569–4586, Online. Association for Computational Linguistics.

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, Sofia, Bulgaria. Association for Computational Linguistics.

Jong-Hoon Oh, Kentaro Torisawa, Canasai Kruengkrai, Ryu Iida, and Julien Kloetzer. 2017. Multi-column convolutional neural networks with causality-attention for why-question answering. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 415–424. ACM.

Mehwish Riaz and Roxana Girju. 2010. Another look at causality: Discovering scenario-specific contingency relationships with no supervision. In *2010 IEEE Fourth International Conference on Semantic Computing*, pages 361–368. IEEE.

Mehwish Riaz and Roxana Girju. 2014a. In-depth exploitation of noun and verb semantics to identify causation in verb-noun pairs. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 161–170, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

Mehwish Riaz and Roxana Girju. 2014b. Recognizing causality in verb-noun pairs via noun and verb semantics. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 48–57, Gothenburg, Sweden. Association for Computational Linguistics.

Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2018. ATOMIC: an atlas of machine commonsense for if-then reasoning. *CoRR*, abs/1811.00146.

Aravind Srinivas, Michael Laskin, and Pieter Abbeel. 2020. Curl: Contrastive unsupervised representations for reinforcement learning.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive representation distillation. In *International Conference on Learning Representations*.

Lilian Weng. 2019. Self-supervised representation learning. *lilianweng.github.io/lil-log*.

Zhirong Wu, Yuanjun Xiong, Stella Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance-level discrimination.

Xinyu Zuo, Yubo Chen, Kang Liu, and Jun Zhao. 2020. KnowDis: Knowledge enhanced data augmentation for event causality detection via distant supervision. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1544–1550, Barcelona, Spain (Online). International Committee on Computational Linguistics.