Enhancing Event Causality Identification with Counterfactual Reasoning

Feiteng Mu, Wenjie Li
The Department of Computing, The Hong Kong Polytechnic University, Hong Kong
{csfmu,cswjli}@comp.polyu.edu.hk

Abstract

Existing methods for event causality identification (ECI) focus on mining potential causal signals, i.e., causal context keywords and event pairs. However, causal signals are ambiguous, which may lead to the context-keywords bias and the event-pairs bias. To solve this issue, we propose the *counterfactual reasoning* that explicitly estimates the influence of context keywords and event pairs in training, so that we are able to eliminate the biases in inference. Experiments are conducted on two datasets, the result demonstrates the effectiveness of our method.

1 Introduction

Event causality identification (ECI) aims to identify causal relations between event pairs. For example, given the sentence "The *earthquake* generated a *tsunami*.", an ECI system should identify that a causal relation holds between the two mentioned events, i.e., earthquake $\xrightarrow{cause}$ tsunami. A good ECI system is able to discover a large number of causal relations from text and hence supports lots of intelligence applications, such as commonsense causal reasoning (Luo et al., 2016), narrative story generation (Mostafazadeh et al., 2016), and many others.

Existing methods focus on mining potential causal signals, including *causal context keywords* (Liu et al., 2020; Zuo et al., 2021a) and *causal event pairs* (Zuo et al., 2020, 2021b; Cao et al., 2021), to enhance ECI. By mining potential causal signals, these methods improve the coverage of unseen events and causal relations, which is the reason for their success. However, they face the risk of amplifying the role of potential signals, resulting in biased inference.

Due to the polysemy of language, causal signals are ambiguous. The occurrence of those signals does not always indicate that causality is established. That is, ambiguous *context keywords* and *event pairs* may lead to the **context-keywords bias** and the **event-pairs bias** in ECI. Specifically, in most cases, *"(earthquake, tsunami)"* in the training set occurs as a causal event pair, but in the sentence which is from the development set, as shown in Table 1, this event pair is not causal. Similarly, ambiguous keywords, such as "generate", do not always indicate causality (Xie and Mu, 2019a,b). Relying heavily on those ambiguous signals may make an ECI model learn the spurious correlation (Pearl, 2009) between ambiguous signals and labels. In other words, existing methods may overfit those ambiguous causal signals in training, and tends to predict a causal relation once the ambiguous signals appear when inference.

Sentence	Label
A **6.1-magnitude earthquake** which hit the Indonesian province of Aceh on Tuesday killed at least one person, injured dozens and destroyed buildings, sparking panic in a region devastated by the quake-triggered **tsunami** of 2004.	0

Table 1: The example comes from the development set of EventStroyLine (Caselli and Vossen, 2017).

To solve this issue, it is intuitively to explicitly estimate the influence of context keywords and event pairs in training, so that we can mitigate those biases in inference. Motivated by this idea and existing dataset-debiasing works (Niu et al., 2021; Tian et al., 2022; Qian et al., 2021), we introduce *factual* and *counterfactual* reasoning for ECI. The *factual* reasoning takes the entire samples as input, which captures the combined features between context keywords and the event pairs, with the side-effect of learning features of biases. The *counterfactual* reasoning considers the two situations where only context keywords or event pairs are available. Intuitively, in counterfactual reasoning, a model can only make predictions based on context keywords or event pairs, so that the biases can be identified. In inference, we use counterfactual reasoning to estimate context-keywords bias and event-pairs bias, then subtract the biases from the factual predictions. To achieve this goal, we must locate the exact position of context keywords in a sen-

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 967–975

tence[1]. But this is difficult because it requires extensive manual annotation. To avoid this, we adopt a model-based strategy. Considering the powerful feature extraction ability of pre-trained language models (PLMs), if we feed an event-removed sentence into PLMs, PLMs should be able to pay the most attention to the important context keywords. Based on this assumption, we split a sentence into two exclusive parts: an event-masked context and an event pair. They are fed into the counterfactual reasoning module to learn the context-keywords bias and event-pairs bias.

To summarize, we consider the spurious correlation problem in ECI, which may make an ECI model overfit on ambiguous causal signals. To mitigate this problem, we propose a counterfactual reasoning mechanism for ECI. To the best of our knowledge, this is the first work that studies ECI from a counterfactual perspective. We conduct extensive experiments on two benchmark datasets. The result shows that our method is effective.

2 Counterfactual ECI

Previous ECI methods may overfit the ambiguous context keywords and event pairs, making biased inferences. We use counterfactual reasoning to eliminate this issue. Our method is depicted in Figure 1, which consists of a factual reasoning module and a counterfactual reasoning module.

2.1 Factual Reasoning Module

Factual reasoning learns the influence of entire ECI samples, following the traditional ECI paradigm. Here we present two classical methods.

Fine-tuning PLMs For ECI We first fine-tune PLMs as a basic backbone. Given a sentence with a mentioned event pair (denoted as e_1 and e_2), we use PLMs, e.g., BERT (Devlin et al., 2018), to encode the sentence and the event pair. Then the embeddings of [CLS], e_1 and e_2[2] are concatenated and applied with a non-linear transformation to obtain the hidden representation of the factual reasoning:

$$\mathbf{h}_{\text{ECI}} = tanh(\mathbf{W}_f^\top([\mathbf{h}_{[CLS]}; \mathbf{h}_{e_1}; \mathbf{h}_{e_2}])), \quad (1)$$

where $\mathbf{W}_f^\top \in \mathcal{R}^{3d\times d}$, $\mathbf{h}_{\text{ECI}} \in \mathcal{R}^d$, d is the hidden size of BERT. $\mathbf{h}_{\text{ECI}}$ is then projected with a linear layer $\mathbf{W}_p^\top \in \mathcal{R}^{d\times 2}$ to make a binary classification:

$$P_{\text{ECI}} = softmax(\mathbf{W}_p^\top \mathbf{h}_{\text{ECI}}). \quad (2)$$

[1]The positions of event pairs are already annotated.

[2]An event is annotated as a text span, so the average-pooling operation is applied to obtain the event embedding.

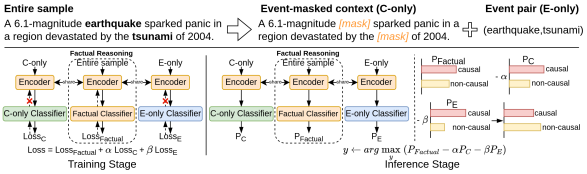

Figure 1: In the upper part, we split a sample into an event pair and an event-masked context. In the bottom part, we show the training and inference process of our method.

Knowledge-Enhanced ECI Existing works prove that knowledge is helpful for ECI. So we develop a knowledge-enhanced backbone. Given (e_1, e_2), we retrieve the related knowledge tuples[3] for e_1 and e_2 respectively, namely $K_{e_i} = \{\tau_{e_i}^1, \tau_{e_i}^2, \cdots, \tau_{e_i}^{N_i}\}$, where $i = 1, 2$ denotes the event index, $\tau = (h, t)$ denotes a knowledge tuple (head, tail), N_1 and N_2 is the number of knowledge tuples. We obtain the knowledge-enhanced features of e_1 and e_2 by average-pooling on the embeddings of corresponding knowledge tuples:

$$\mathbf{h}_{e_i}^K = \frac{1}{N_i}\sum_{j=1}^{N_i} \mathbf{W}_k^\top [\mathbf{h}_{e_i}^j; \mathbf{t}_{e_i}^j], \quad (3)$$

where $i = 1, 2$, $\mathbf{h}$ and $\mathbf{t}$ denote the embeddings of a tuple (h, t), $\mathbf{W}_k \in \mathcal{R}^{2d\times d}$ is trainable. Then the knowledge-enhanced event representations $\mathbf{h}_{e_1}^K$ and $\mathbf{h}_{e_2}^K$ are concatenated with $\mathbf{h}_{\text{ECI}}$ (Equation 1), and input into a MLP to make a binary classification:

$$P_{\text{ECI}}^K = softmax(\text{MLP}([\mathbf{h}_{\text{ECI}}; \mathbf{h}_{e_1}^K; \mathbf{h}_{e_2}^K])). \quad (4)$$

Finally, the cross-entropy loss is applied to P_{ECI} and P_{ECI}^K to train the two backbones. Factual reasoning learns combined features between the context and the event pair, but biases may be entangled into the combined features. Next, we propose counterfactual reasoning to capture the entangled biases.

2.2 Counterfactual Reasoning Module

To estimate the context-keywords bias and the event-pairs bias in training, we split a sentence into two exclusive parts: an event-masked context and an event pair. For each part, we use counterfactual reasoning to estimate the corresponding bias.

2.2.1 Estimating Context-Keywords Bias

We consider the counterfactual situation where only the event-masked context is available. We input the context into PLMs, and let PLMs automatically attend to the important context keywords. The [CLS] token embedding $\overline{\mathbf{h}_{[CLS]}}$ is used as the representation of the event-masked context. Note that $\overline{\mathbf{h}_{[CLS]}}$

[3]Details can be seen in Appendix A.

is different from $\mathbf{h}_{[CLS]}$ (Equation 1) because the event pair is removed in the current situation. We obtain the hidden state of the current situation by:

$$\overline{\mathbf{h}_C} = tanh(\mathbf{W}_f^\top([\overline{\mathbf{h}_{[CLS]}}; \mathbf{\Phi}_E; \mathbf{\Phi}_E])), \quad (5)$$

where $\mathbf{W}_f$ is the shared parameter (Equation 1), $\mathbf{\Phi}_E \in \mathcal{R}^d$ is a learnable constant, and represents the void input events. The insight of this setting is that if we have no information about the event pair, we would like to make inference by random guess. Then $\overline{\mathbf{h}_C}$ is projected to make binary classification:

$$P_C = softmax(\mathbf{W}_C^\top \overline{\mathbf{h}_C}), \quad (6)$$

where $\mathbf{W}_C$ is trainable, P_C estimates the influence of the context-keywords bias.

2.2.2 Estimating Event-Pairs Bias

Next, we consider the counterfactual situation where only the event pair (e_1, e_2) is available. Through PLMs, we get the event embeddings of $\overline{\mathbf{h}_{e_1}}$ and $\overline{\mathbf{h}_{e_2}}$. Note that $\overline{\mathbf{h}_{e_1}}$ and $\overline{\mathbf{h}_{e_2}}$ is different from $\mathbf{h}_{e_1}$ and $\mathbf{h}_{e_2}$ (Equation 1) because the context is invisible in the current situation. We obtain the hidden state of the current situation by:

$$\overline{\mathbf{h}_E} = tanh(\mathbf{W}_f^\top([\mathbf{\Phi}_C; \overline{\mathbf{h}_{e_1}}; \overline{\mathbf{h}_{e_2}}])), \quad (7)$$

where $\mathbf{\Phi}_C$ is a learnable constant, and represents the void input context. Then $\overline{\mathbf{h}_E}$ is projected with a linear layer to make binary classification:

$$P_E = softmax(\mathbf{W}_E^\top \overline{\mathbf{h}_E}), \quad (8)$$

where $\mathbf{W}_E$ is trainable, P_E estimates the influence of the event-pairs bias.

2.3 Training and De-biased Inference

We jointly train the factual and counterfactual reasoning modules, the final loss is:

$$Loss = Loss_{Factual} + \alpha Loss_C + \beta Loss_E. \quad (9)$$

$Loss_{Factual}$ is over P_{ECI} or P_{ECI}^K. $Loss_C$ is over P_C and $Loss_E$ is over P_E. α and β are two trade-off coefficients that balance the two types of biases. Note that we share the encoding process (Equation 1) between factual and counterfactual modules, but we do not backpropagate $Loss_C$ and $Loss_E$ to the encoder, as shown in Figure 1. This is because we require the counterfactual reasoning module to make predictions only based on the event-masked context or the event pair, and has no information about the missing part.

After training, the counterfactual reasoning module will learn the bias-estimation mechanism. Therefore, we can make de-biased inference by:

$$y \leftarrow \text{argmax}_y(P_{Factual} - \alpha P_C - \beta P_E), \quad (10)$$

where $P_{Factual}$ can be P_{ECI} or P_{ECI}^K.

3 Experiment

3.1 Experimental Settings

Datasets include EventStoryLine (ESL) (Caselli and Vossen, 2017) and Causal-TimeBank (CTB) (Mirza et al., 2014). ESL contains 22 topics, and 1770 of 7805 event pairs are causally related. CTB contains 184 documents, and 318 of 7608 event pairs are causally related. We conduct the 5-fold and 10-fold cross-validation on ESL and CTB respectively. The last two topics of ESL are used as the development set for two tasks. All of this is the same as previous works for fairness. Evaluation metrics are Precision (P), Recall (R) and F1-score (F1). All parameters are searched according to the F1 on the Dev set. The compared baselines include KMMG (Liu et al., 2020), KnowDis (Zuo et al., 2020), LearnDA (Zuo et al., 2021b), LSIN (Cao et al., 2021) and CauSeRL (Zuo et al., 2021a). When implementing our factual reasoning models, we adopt BERT(base), which is same as previous methods. We denote our two factual backbones as BERT and BERT$_K$. Details about experimental settings can be seen in Appendix B.

3.2 Overall Result and Ablation Study

The overall result is shown in Table 2. We have the following observations. **(1)** BERT$_K$ has a similar result with compared baselines, and performs better than BERT. This coincides with previous works that knowledge is helpful for ECI. **(2)** Our CF-ECI method achieves consistent improvement when deployed on BERT or BERT$_K$. This shows the effectiveness of our method. **(3)** Compared with the previous methods, our method has a higher precision score. This is because we make a de-biased inference, which is able to reduce the false-positive predictions, hence improve the precision. **(4)** Utilizing knowledge may reduce the precision score, because irrelevant knowledge may be introduced. This coincides with LSIN (Zuo et al., 2021a).

Ablation Study We conduct ablation study to investigate the influence of context-keywords de-biasing (§ 2.2.1) and event-pairs de-biasing (§ 2.2.2). The result is shown in Table 2. No matter what backbone (BERT or BERT$_K$) is used, after ablating "EPB" or "CKB", the ablated variant has a performance drop. This indicates that ambiguous context-keywords and event-pairs have adversely

influence of ECI. By making de-biased inference, our CF-ECI achieves the best performance. In addition, we observe that the context-keywords bias is more severe than the event-pairs bias, which indicates that the trained models tend to use superficially keywords for inference. The possible reason is that this strategy inevitably leverages ambiguous keywords that are potential biases, though it can capture some causal keywords as good evidence.

Models	ESL			CTB		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
KMMG	41.9	62.5	50.1	36.6	55.6	44.1
KnowDis	39.7	66.5	49.7	42.3	60.5	49.8
LearnDA	42.2	**69.8**	52.6	41.9	68.0	51.9
CauSeRL	41.9	69.0	52.1	43.6	**68.1**	53.2
LSIN	47.9	58.1	52.5	51.5	56.2	52.9
This Paper						
BERT	45.8	57.4	50.9	49.8	50.3	50.1
BERT_K	43.2	65.8	52.2	48.3	54.5	51.2
$\text{CF-ECI}_{\text{BERT}}$	**48.7**	59.0	53.4*	**54.1**	53.0	53.5*
$\text{CF-ECI}_{\text{BERT}_K}$	47.1	66.4	**55.1***	50.5	59.9	**54.8**
Ablation Experiment						
$\text{CF-ECI}_{\text{BERT}}$						
: w/o EPB	47.7	57.6	52.2	51.7	53.6	52.6
: w/o CKB	48.0	56.7	52.0	51.1	52.5	51.8
$\text{CF-ECI}_{\text{BERT}_K}$						
: w/o EPB	46.8	63.8	54.0	50.8	56.4	53.4
: w/o CKB	47.0	62.6	53.7	50.2	56.3	53.1

Table 2: The overall and ablation-study result. Scores with **bold** denotes the best results. *: the significant test is conducted using paired t-test between our method and the used backbones, with the level of $p = 0.05$. "CKB" denotes the context-keywords de-biasing. "EPB" denotes the event-pairs de-biasing.

3.3 Further Discussion

Methods	ESL		CTB	
	Dev	Test	Dev	Test
BERT	17.75	16.71	20.47	21.02
$\text{CF-ECI}_{\text{BERT}}$	02.40	02.09	02.71	02.64
BERT_K	17.08	15.70	20.46	21.04
$\text{CF-ECI}_{\text{BERT}_K}$	02.44	02.25	02.81	02.77

Table 3: The model unfairness result (lower is better) on the dev-set and test-set of ESL and CTB.

Bias Analysis (Sweeney and Najafian, 2019; Qian et al., 2021) point out that the unfairness of a trained model can be measured by the imbalance of the predictions produced by the model. Following (Qian et al., 2021), we use the metric *imbalance divergence* (ID) to evaluate whether a predicted distribution P is unfair: $ID(P, U) = JS(P||U)$, where $JS(\cdot)$ denotes the JS divergence of P and the uniform distribution U. To evaluate the unfairness of a trained model M, we calculate its ID over all dev or test samples: $ID(M) = \frac{1}{|\mathcal{D}|}\sum_{x\in\mathcal{D}} JS(P(x), U)$, where $P(x)$ can be the output distribution of a factual (§ 2.1) or counterfactual (§ 2.2) model. As shown in Table 3, when deployed on different backbones, our method can obviously and consistently reduce the ID metric. This indicates that our method is helpful to eliminate two kinds of biases.

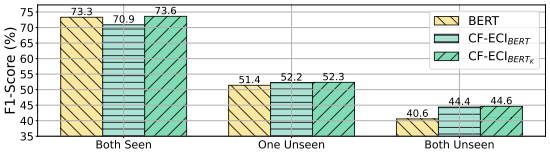

Figure 2: F1 scores (%) of identifying unseen events.

Figure 3: The heatmaps of the predictions by BERT and CF-ECI$_{BERT}$ respectively. Text with the dotted line denotes the annotated events.

Identifying Unseen Events We explore the ability of our method to identify unseen events. We first randomly select 1/3 of ESL documents as the training set, then divide the remaining documents into (1) "Both Seen", where two events of a sample appear in training data; (2) "One Unseen", where only one event of a sample exists in training data; (3) "Both Unseen", where both events are unobserved during training. From Figure 2, we have following observations. (1) CF-ECI has a significant improvement on the "Both Unseen" set, compared with BERT. (2) CF-ECI$_{BERT_K}$ performs better than CF-ECI$_{BERT}$ on the "Both Seen" set.

Visualization We depict the heatmaps of predictions by BERT and CF-ECI$_{BERT}$ respectively, in Figure 3. BERT pays the most attention to the words: "*eqrthquake, spark, quake, tsunami*", and gives a causal prediction with the 97.9% probability. In contrast, CF-ECI$_{BERT}$ dispersedly attends to words and does not find enough causal evidence, hence it gives a non-causal prediction.

4 Related Work

Event Causality Identification There are mainly two types of ECI works: document-level ECI (Gao et al., 2019; Phu and Nguyen, 2021) and sentence-level ECI. In this work, we pay attention to the sentence-level ECI. (Liu et al., 2020) propose to mask event mentions to mine event-agnostic causal patterns. (Zuo et al., 2021a) devises self-supervised methods to learn context-specific causal patterns from external causal statements. (Zuo et al., 2020, 2021b) utilize causal event pairs to find useful data from external resources. Nevertheless, these methods rely on ambiguous causal signals, and may learn the spurious correlations between ambiguous causal signals and labels. Different from these works, we regard ECI from a counterfactual perspective, and devise a counterfactual inference module to the spurious correlations in ECI.

Counterfactual Reasoning Counterfactual data augmentation is a data-level manipulation, which is effective to mitigate biases in datasets (Wei and Zou, 2019; Kaushik et al., 2019). However, it needs extra manual cost of data annotation. A recent trend is counterfactual reasoning, which imagines the situation that what will the prediction be if seeing only the biased part in the input. In this way, the biases can be distilled and eliminated in the inference. This strategy avoids data annotation, and is adopted by many works (Niu et al., 2021; Tian et al., 2022; Qian et al., 2021). Motivated by these works, we devise the counterfactual reasoning module to make a de-baised ECI inference.

5 Conclusion

We discuss the issue of context-keywords and event-pairs biases in ECI. To mitigate this problem, we propose the counterfactual reasoning which explicitly estimates the influence of the biases, so that we can make a de-biased inference. Experimental results demonstrate the significant superiority of our method. The robustness and explainability of our method are also verified by further studies.

6 Limitations

First, we only access limited computation resources and perform continual pre-training from BERT (Devlin et al., 2018), which is not general enough for every event-related reasoning task. Second, counterfactual reasoning makes our approach conservative in identifying causal relationships, so our method has a higher precision. However, some potential causal relationships will be discarded. How to achieve a good trade-off between precision and coverage is a problem. In addition, the way we utilize knowledge is relatively simple, and it is very likely that we have not made full use of knowledge. Designing more complex knowledge-enhanced methods may lead to better results.

7 Ethical Considerations

This work does not involve any sensitive data, but only crowd-sourced datasets released in previous works, including Event-StoryLine (Caselli and Vossen, 2017) and Causal-TimeBank (Mirza et al., 2014). We believe that our research work meets the ethics of ACL.

8 Acknowledgements

We thank the anonymous reviewers for their encouraging feedback. This work is supported by Research Grants Council of Hong Kong(PolyU/15207920, PolyU/15207821) and National Natural Science Foundation of China (62076212).

References

Pengfei Cao, Xinyu Zuo, Yubo Chen, Kang Liu, Jun Zhao, Yuguang Chen, and Weihua Peng. 2021. Knowledge-enriched event causality identification via latent structure induction networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4862–4872, Online. Association for Computational Linguistics.

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.

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

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.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*.

Jian Liu, Yubo Chen, and Jun Zhao. 2020. Knowledge enhanced event causality identification with mention masking generalizations. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3608–3614.

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*.

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.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12700–12710.

Judea Pearl. 2009. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146.

Minh Tran Phu and Thien Huu Nguyen. 2021. Graph convolutional networks for event causality identification with rich document-level structures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3480–3490.

Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. 2021. Counterfactual inference for text classification debiasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5434–5445.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Chris Sweeney and Maryam Najafian. 2019. A transparent framework for evaluating unintended demographic bias in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1662–1667, Florence, Italy. Association for Computational Linguistics.

Bing Tian, Yixin Cao, Yong Zhang, and Chunxiao Xing. 2022. Debiasing nlu models via causal intervention and counterfactual reasoning.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

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

Zhipeng Xie and Feiteng Mu. 2019b. Distributed representation of words in cause and effect spaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7330–7337.

Xinyu Zuo, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Weihua Peng, and Yuguang Chen. 2021a. Improving event causality identification via self-supervised representation learning on external causal statement. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2162–2172, Online. Association for Computational Linguistics.

Xinyu Zuo, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Weihua Peng, and Yuguang Chen. 2021b. LearnDA: Learnable knowledge-guided data augmentation for event causality identification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3558–3571, Online. Association for Computational Linguistics.

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.

A Details about Knowledge Retrieving

Following (Liu et al., 2020), we leverage external knowledge to further improve ECI. We use ConceptNet (Speer et al., 2017) as knowledge base. In ConceptNet, knowledge is structured as graph, where each node corresponds a concept, and each edge corresponds to a semantic relation. For e_1 and e_2, we search their related knowledge, i.e., matching an event with the tokens of concepts in ConceptNet. Events and concepts are Lemmatized with the Spacy toolkit to improve the rate of matching. We only consider 12 semantic relations that are potentially useful for ECI: CapableOf, Causes, CausesDesire, UsedFor, HasSubevent, HasPrerequisite, Entails, ReceivesAction, UsedFor, CreatedBy, MadeOf, and Desires. For each relation, we retrieve at most two knowledge relations according to the weights of relations.

B Details about Experimental Settings

B.1 Compared Baselines

- KMMG (Liu et al., 2020), which proposes a mention masking generalization method and also utilizes the external knowledge.
- KnowDis (Zuo et al., 2020), a data-augmentation method that utilizes the distantly labeled training data.
- LearnDA (Zuo et al., 2021b), a data-augmentation method with iteratively generating new examples and classifying event causality in a dual learning framework.
- LSIN (Cao et al., 2021), a latent-structure induction network to leverage the external knowledge;.
- CauSeRL (Zuo et al., 2021a), a self-supervised framework to learn context-specific causal patterns from external causal corpora.

B.2 Implementation Details

Due to the data imbalance problem, we adopt a over-sampling strategy for training. The early-stop is used due to the small scale of datasets. We use the Adam optimizer and linearly decrease learning rate to zero with no warmup. We use PyTorch toolkit to conduct all experiments on the Arch Linux with RTX3090 GPU. All the hyperparameter for two tasks are searched according to the F1 score on the development set. For reproduction, we set the random seed to 42 for all experiments. The searched parameters for two datasets are shown in Table 4.

Parameters	ESL	CTB
Batch Size	32	32
Learning Rate	5e-5	5e-5
Drop-rate	0.3	0.2
α	0.15	0.25
β	0.35	0.25

Table 4: The used hyperparameters for two datasets.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 5

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3

☑ B1. Did you cite the creators of artifacts you used?
Section 3

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 3.1

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Statistics of Datasets are reported in Appendix B

C ☑ Did you run computational experiments?

Section 3

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix B

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix B.3

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Appendix B.3

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix B.3

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.