Semantic Structure Enhanced Event Causality Identification

Zhilei Hu, Zixuan Li[*], Xiaolong Jin[*], Long Bai, Saiping Guan,
Jiafeng Guo and **Xueqi Cheng**
School of Computer Science and Technology, University of Chinese Academy of Sciences;
CAS Key Laboratory of Network Data Science and Technology,
Institute of Computing Technology, Chinese Academy of Sciences.
`{huzhilei19b, lizixuan, jinxiaolong, bailong18b, guansaiping}@ict.ac.cn`
`{guojiafeng, cxq}@ict.ac.cn`

Abstract

Event Causality Identification (ECI) aims to identify causal relations between events in unstructured texts. This is a very challenging task, because causal relations are usually expressed by implicit associations between events. Existing methods usually capture such associations by directly modeling the texts with pre-trained language models, which underestimate two kinds of semantic structures vital to the ECI task, namely, event-centric structure and event-associated structure. The former includes important semantic elements related to the events to describe them more precisely, while the latter contains semantic paths between two events to provide possible supports for ECI. In this paper, we study the implicit associations between events by modeling the above explicit semantic structures, and propose a **Sem**antic **S**tructure **In**tegration model (**SemSIn**). It utilizes a GNN-based event aggregator to integrate the event-centric structure information, and employs an LSTM-based path aggregator to capture the event-associated structure information between two events. Experimental results on three widely used datasets show that SemSIn achieves significant improvements over baseline methods.

1 Introduction

Event Causality Identification (ECI) is an important task in natural language processing that seeks to predict causal relations between events in texts. As shown in the top of Figure 1, given the unstructured text and event pair (***shot***, ***protect***), an ECI model needs to identify that there exists a causal relation between two events, i.e., ***protect*** $\xrightarrow{cause}$ ***shot***. ECI is an important way to construct wide causal connections among events, which supports a variety of practical applications, such as event prediction (Hashimoto, 2019), reading comprehension (Berant et al., 2014), and question answering (Oh et al., 2013, 2017).

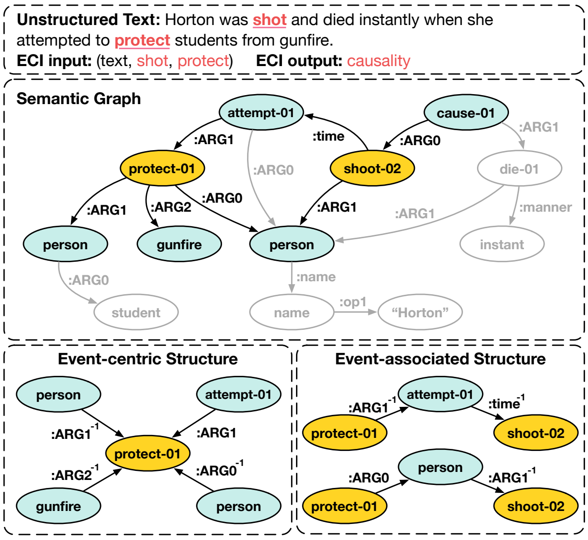


Figure 1: An example of the ECI task, as well as the semantic graph and semantic structures corresponding to the unstructured text. The orange nodes denote events.

ECI is challenging because most causal relations are expressed by texts implicitly, which requires the model to understand the associations between two events adequately. Existing methods directly model the texts with the Pre-trained Language Model (PLM) (Liu et al., 2020; Cao et al., 2021; Zuo et al., 2020, 2021b). They mainly rely on the ability of PLMs, which cannot capture associations between events comprehensively. To enrich the associations between events, some methods (Liu et al., 2020; Cao et al., 2021) introduce external knowledge, such as events in ConceptNet (Speer et al., 2017) that are related to focused events. Other methods (Zuo et al., 2020, 2021b) utilize the data augment framework to generate more training data for the model. However, the above methods are far from fully modeling the associations among events in the texts.

Actually, texts contain rich semantic elements and their associations, which form graph-like semantic structures, i.e., semantic graphs. Figure 1

[*] Corresponding authors.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 10901–10913
July 9-14, 2023 ©2023 Association for Computational Linguistics

shows the semantic graph generated by the Abstract Meaning Representation (AMR) (Banarescu et al., 2013) parser for the corresponding text, where the nodes indicate events, entities, concepts, and other semantic elements, while edges with semantic roles describe the associations among semantic elements. For example, "protect-01" indicates the specific sense of the verb "protect" in the PropBank (Palmer et al., 2005) [1]. "ARG0", "ARG1" and "ARG2" indicate different semantic roles. In this semantic graph, we exploit two kinds of structures vital to the ECI task, namely, event-centric structure and event-associated structure. As shown in the bottom left of Figure 1, the event-centric structure consists of events and their neighbors, which describes events more precisely by considering their arguments and corresponding roles. For example, besides event "protect-01", "person (Horton)" and "person (student)" are also important semantic elements, and their corresponding semantic roles can supply the extra information for the event. As shown in the bottom right of Figure 1, the event-associated structure contains semantic paths between two events, and each path contains the core semantic elements. The composition of these elements indicates the possible semantic relations between events and provides supports for ECI. For example, the path "protect-01 $\xrightarrow{\text{:ARG0}}$ person $\xrightarrow{\text{:ARG1}^{-1}}$ shoot-02" indicates that "person (Horton)" protects somebody first and then was shot. Events "protect-01" and "shoot-02" share the same participant, and there may exist a causal relation between them.

To make use of the above semantic structures in texts to carry out the ECI task, we propose a new **Sem**antic **S**tructure **In**tegration model (SemSIn). It first employs an AMR parser to convert each unstructured text into a semantic graph and obtains the above two kinds of semantic structures in that graph. For the event-centric structure, SemSIn adopts an event aggregator based on Graph Neural Networks (GNN). It aggregates the information of the neighboring nodes to the event nodes to obtain more precise representations of the events. For the event-associated structure, SemSIn utilizes a path aggregator based on Long Short-Term Memory (LSTM) network. It encodes the compositional semantic information in the paths and then integrates the information of multiple paths with an attention mechanism. With the above representations of the events and paths as input, SemSIn conducts ECI with a Multi-Layer Perceptron (MLP).

In general, the main contributions of this paper can be summarized as follows:

- We exploit two kinds of critical semantic structures for the ECI task, namely, event-centric structure and event-associated structure. They can explicitly consider the associations between events and their arguments, as well as the associations between event pairs.

- We propose a novel Semantic Structure Integration (SemSIn) model, which utilizes an event aggregator and a path aggregator to integrate the above two kinds of semantic structure information.

- According to experimental results on three widely used datasets, SemSIn achieves 3.5% improvements of the F1 score compared to the state-of-the-art baselines.

2 Related Work

Identifying causal relations between events has attracted extensive attention in the past few years. Early methods mainly rely on the causal association rule (Beamer and Girju, 2009; Do et al., 2011) and causal patterns (Hashimoto et al., 2014; Riaz and Girju, 2010, 2014a; Hidey and McKeown, 2016). Some following methods exploit lexical and syntactic features to improve performance (Riaz and Girju, 2013, 2014b).

Recently, most of works apply PLM to conduct ECI (Liu et al., 2020; Cao et al., 2021; Zuo et al., 2021b). Although PLM has a strong ability for capturing the associations among tokens in the texts, they are not capable of this task because the associations between events are implicit. To enhance PLM, recent works try to introduce external knowledge. Liu et al. (2020) proposed a method to enrich the representations of events using commonsense knowledge related to events from the knowledge graph ConceptNet (Speer et al., 2017). Cao et al. (2021) further proposed a model to exploit knowledge connecting events in the ConceptNet for reasoning. Zuo et al. (2021b) proposed a data augmented method to generate more training samples. Instead of introducing external knowledge to enhance the abilities of the ECI model, we attempt to dive deep into the useful semantic structure information in the texts.

[1] A corpus annotated with verbs and their semantic roles.

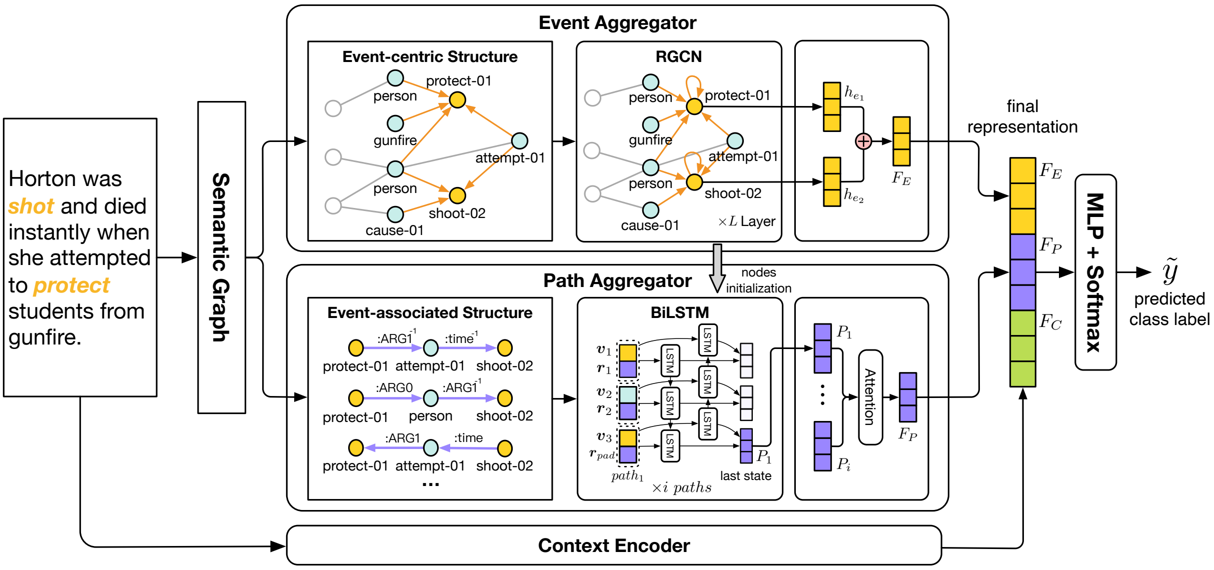


Figure 2: An illustration diagram of the proposed SemSIn model.

3 The SemSIn Model

In this section, we introduce the proposed SemSIn model. Figure 2 illustrates the overall architecture of the SemSIn model. Given an input text, SemSIn first uses a pre-trained AMR parser to obtain the corresponding semantic graph of the text. Then, the event-centric structure and the event-associated structure constructed from the semantic graph, as well as the original text, are fed into the following three components respectively: (1) Event aggregator aggregates the event-centric structure information into the representation of the event pair. (2) Path aggregator captures the event-associated structure information between two events into the path representation. (3) Context encoder encodes the text and obtains the contextual representation of the event pair. With the above representations as input, SemSin conducts binary classification to get the final results with an MLP layer. Next, we will first introduce the construction process of the semantic graph and then present these three main components in detail.

3.1 Semantic Graph Construction

The core motivation of SemSIn is to model the implicit associations between events by introducing explicit semantic structures. To get explicit semantic structures from texts, SemSIn employs an AMR parser to convert the original text into an AMR graph, which contains fine-grained node and edge types (Zhang and Ji, 2021).

In the AMR graph, the nodes indicate specific semantic elements and the edges indicate the semantic roles among them. Table 1 lists the used semantic roles in AMR graph. We then add inverse edges to all the edges in the AMR graph to form the final semantic graph, making it reachable between any two nodes. Formally, a semantic graph is defined as $G = (V, E, R)$, where V, E and R are the sets of nodes, edges and role types, respectively.

Semantic Roles	Types
ARG0, ARG1, ARG2, $\cdots$	Core Roles
op1, op2, op3, op4	Operators
manner, instrument, topic, $\cdots$	Means
time, year, weekday, $\cdots$	Temporal
Other semantic roles	Others

Table 1: Semantic roles in the AMR graphs (Zhang and Ji, 2021)

3.2 Event Aggregator

Identifying the causal relation between two events requires the model to comprehensively understand what each event describes. Existing methods use the event mentions in the text to represent the events, which cannot highlight the semantic elements related to the events. Besides the event mentions, events usually have associations with their arguments mentioned in the text. Similarly, event arguments also have associations with some related semantic elements. Therefore, to model this kind of association, SemSIn obtains the event-centric structure from the constructed semantic graph by simply using the L-hop subgraph of the focused event, where L is a hyperparameter.

Node Representation Initialization: To initialize the representations of nodes in the event-centric structure, a rule-based alignment tool [2] is first employed to align AMR nodes to the tokens in the text. For the AMR nodes that have the corresponding tokens in the text, their initialized representations are obtained by averaging the representation vectors of all tokens aligned to the nodes. For example, given a node, the start and end positions of its corresponding tokens are a and b, respectively. Its representation vector is calculated by:

$$\boldsymbol{h} = \frac{1}{|b-a+1|} \sum_{k=a}^{b} \boldsymbol{x}_k, \tag{1}$$

where $\boldsymbol{x}_k$ is the representation of the token k. A PLM, BERT (Devlin et al., 2019), is applied to encode the sequence of tokens. For those nodes without corresponding tokens in the original text (i.e., auxiliary nodes added by the AMR parser, such as "name" and "cause-01" in Figure 1), their representations are randomly initialized.

Semantic Information Aggregation: The graph convolutional network has the property of aggregating the information of neighbor nodes to the specific node, which is suitable to model the event-centric structure. In addition, the types of edges in the semantic graph also contain special information that can be used to distinguish the relations between nodes. Therefore, we apply a Relational Graph Convolutional Network (RGCN) (Schlichtkrull et al., 2018) to aggregate semantic information from L-hop neighbors of the focused events. Specifically, the message passing at layer $l \in [0, L-1]$ is conducted as follows:

$$\boldsymbol{h}_i^{l+1} = \sigma \left(\sum_{r \in R} \sum_{j \in N_i^r} \frac{1}{c_{i,r}} \boldsymbol{W}_r^l \boldsymbol{h}_j^l + \boldsymbol{W}_0^l \boldsymbol{h}_i^l \right), \tag{2}$$

where R denotes the set of the role types; N_i^r denotes the set of the neighbors of node i under relation $r \in R$; $c_{i,r}$ is a normalization constant equal to $|N_i^r|$; $\boldsymbol{h}_i^l$ and $\boldsymbol{h}_j^l$ denote the l^{th} layer representations of the nodes i and j, respectively; $\boldsymbol{W}_r^l$ and $\boldsymbol{W}_0^l$ are the weight matrices for aggregating features from different relations and self-loop in the l^{th} layer; σ is an activation function (e.g., ReLU); $\boldsymbol{h}_i^0$ and $\boldsymbol{h}_j^0$ are the initialized representations of the nodes introduced above. After aggregating the event-centric structure information, the representations of e_1 and e_2 are denoted as $\boldsymbol{h}_{e_1}$ and $\boldsymbol{h}_{e_2}$. In addition, to eliminate the effect of the relative position of the two events, we sum up the representations of the two events to obtain $\boldsymbol{F}_E^{(e_1,e_2)}$, the representation of the event pair,

$$\boldsymbol{F}_E^{(e_1,e_2)} = \boldsymbol{h}_{e_1} + \boldsymbol{h}_{e_2}. \tag{3}$$

3.3 Path Aggregator

Besides the associations between events and their arguments, identifying the causal relation requires the model to discover the association between two events. The paths in the semantic graph between two events can reflect this kind of association. SemSIn thus first finds paths between two events in the semantic graph to form the event-associated structure. Then, SemSIn encodes it via BILSTM and path attention to get the representations of the paths.

With the intuition that the fewer hops in the path, the stronger information it contains to reflect the association between two events, we choose the shortest path between two event nodes in the semantic graph to form the event-associated structure. This operation can avoid introducing redundant information and improve efficiency. Besides, we add the reverse path for each semantic path. Formally, if there is a path denoted as $(v_1, r_1, v_2, \cdots, v_{n-1}, r_{n-1}, v_n)$, the corresponding reverse path is $(v_n, r_{n-1}, v_{n-1}, \cdots, v_2, r_1, v_1)$.

Path Encoding: The compositional semantic information of the semantic elements and roles in paths can provide possible supports to the causal relation. Recently, recurrent neural networks have been widely used in processing sequence data such as path information (Wang et al., 2019; Huang et al., 2021). Therefore, we apply a BiLSTM to better encode each path in the event-associated structure and output its representations. Here, the initialized representations of all nodes are obtained by applying the RGCN to the whole semantic graph, while the representations of relations are randomly initialized and updated during the training process. To convert multi-hop paths into a sequence of vectors, we concatenate node and relation representation vectors as the input at each state. For example, the sequence is organized as $[(\boldsymbol{v}_1, \boldsymbol{r}_1); (\boldsymbol{v}_2, \boldsymbol{r}_2); \cdots; (\boldsymbol{v}_n, \boldsymbol{r}_{pad})]$, where $\boldsymbol{v}_i$ denotes the representation of the node i; $\boldsymbol{r}_i$ denotes the representation of the relation i; $\boldsymbol{r}_{pad}$ denotes the representation of the special PAD relation added to

[2] RBW Aligner in https://github.com/bjascob/amrlib

the last state. Then, the representation $\boldsymbol{P}_i$ of this path can be obtained:

$$\boldsymbol{P}_i = \text{BiLSTM}\left[(\boldsymbol{v}_1, \boldsymbol{r}_1); \cdots; (\boldsymbol{v}_n, \boldsymbol{r}_{pad})\right]. \quad (4)$$

Path Attention: There may exist multiple paths with the same shortest length. Different paths reflect different semantic information. Thus, to distinguish the importances of different paths, SemSIn adopts an attention mechanism to integrate the information of multiple paths. The query for attention is the representation of the event pair $\boldsymbol{F}_E^{(e_1,e_2)}$, which is obtained from the event aggregator. Both key and value are the representation $\boldsymbol{P}_i$ of the path:

$$\alpha_i = \frac{(\boldsymbol{F}_E^{(e_1,e_2)}\boldsymbol{W}_Q)(\boldsymbol{P}_i\boldsymbol{W}_K)^T}{\sqrt{d_k}}, \quad (5)$$

$$\boldsymbol{F}_P^{(e_1,e_2)} = \sum_i \text{Softmax}(\alpha_i)(\boldsymbol{P}_i\boldsymbol{W}_V), \quad (6)$$

where $\boldsymbol{W}_Q$, $\boldsymbol{W}_K$ and $\boldsymbol{W}_V$ are parameter weights; α_i denotes the salient score for path i to event pair $\boldsymbol{F}_E^{(e_1,e_2)}$; $\boldsymbol{F}_P^{(e_1,e_2)}$ is the integrated representation of multiple paths.

3.4 Context Encoder

Besides the above semantic structure information, the contextual semantic information is proved to be useful for ECI (Cao et al., 2021). Thus, we adopt an extra context encoder to encode the tokens of the text and obtain the contextual semantic representation of the event pair. Specifically, we first add two pairs of special markers <e1></e1> and <e2></e2> to indicate the boundaries of the two event mentions. Two special tokens [CLS] and [SEP] are also added to indicate the beginning and end of the whole text, respectively. To model the representations of the tokens in the context encoder and event aggregator separately, here we adopt another BERT model to encode the context. Following Liu et al. (2020), we use the representations of the tokens <e1> and <e2> as the representations of the two events, i.e., e_1 and e_2. And the representation of the token [CLS] is adopted as that of the whole text. In order to achieve sufficient interaction between the events and their corresponding contexts, we apply a linear layer and an activation function to obtain more accurate representations of the events:

$$\tilde{\boldsymbol{u}}_i = \tanh(\boldsymbol{W}_u[\boldsymbol{u}_{[CLS]}||\boldsymbol{u}_i] + \boldsymbol{b}_u), \quad (7)$$

where $||$ represents the concatenation operation. $\boldsymbol{u}_{[CLS]}$ and $\boldsymbol{u}_i$ are the representations of the whole text and e_i, $i \in (1, 2)$, respectively. $\boldsymbol{W}_u$ and $\boldsymbol{b}_u$ are the weight matrix and the bias, respectively.

We again sum up the representations of the two events as the representation of the event pair:

$$\boldsymbol{F}_C^{(e_1,e_2)} = \tilde{\boldsymbol{u}}_1 + \tilde{\boldsymbol{u}}_2, \quad (8)$$

where $\boldsymbol{F}_C^{(e_1,e_2)}$ is the contextual representation of the event pair and will be used for further computation.

3.5 Model Prediction

We concatenate the representations obtained from the above three components as the final representation of each event pair:

$$\boldsymbol{F}_{(e_1,e_2)} = \boldsymbol{F}_E^{(e_1,e_2)}||\boldsymbol{F}_P^{(e_1,e_2)}||\boldsymbol{F}_C^{(e_1,e_2)}. \quad (9)$$

Then, $\boldsymbol{F}_{(e_1,e_2)}$ is fed into the softmax layer for classification,

$$\boldsymbol{p} = \text{softmax}(\boldsymbol{W}_f\boldsymbol{F}_{(e_1,e_2)} + \boldsymbol{b}_f), \quad (10)$$

where $\boldsymbol{p}$ is the probability indicating whether there is a causal relation between two events; $\boldsymbol{W}_f$ and $\boldsymbol{b}_f$ are trainable parameters.

3.6 Parameter Learning

For the classification task, the model generally adopts the cross-entropy loss function and treats all samples equally. However, most of the samples without causality are easily predicted and these samples will dominate the total loss. In order to pay more attention to samples that are difficult to predict, we adopt focal loss (Lin et al., 2017) as the loss function of our model:

$$J(\Theta) = -\sum_{(e_i,e_j)\in E_s} \beta(1-\boldsymbol{p})^{\gamma} log(\boldsymbol{p}), \quad (11)$$

where Θ denotes the model parameters; (e_i, e_j) denotes the sample in the training set E_s. Besides, to balance the importance of positive and negative samples, we add the loss weighting factor $\beta \in [0, 1]$ for the class "positive" and $1 - \beta$ for the class "negative".

4 Experiments

4.1 Datasets and Metrics

We evaluate the proposed SemSIn on two datasets from EventStoryLine Corpus v0.9 (ESC) (Caselli and Vossen, 2017) and one dataset from Causal-TimeBank (Causal-TB) (Mirza et al., 2014), namely, ESC, ESC* and Causal-TB.

ESC [3] contains 22 topics, 258 documents, and

[3] https://github.com/tommasoc80/EventStoryLine

5334 event mentions. The dataset is processed following Gao et al. (2019), excluding aspectual, causative, perception, and reporting event mentions, most of which are not annotated with any causality. After processing, there are 7805 intra-sentence event mention pairs in the corpus, of which 1770 (22.7%) are annotated with a causal relation. The same as previous methods (Gao et al., 2019; Zuo et al., 2021b), we use the documents in the last two topics as the development set, and report the experimental results by conducting 5-fold cross-validation on the remaining 20 topics. The dataset used in the cross-validation evaluation is partitioned as follows: documents are sorted according to their topic IDs, which means that the training and test sets are cross-topic. Under this setting, the data distributions of the training and test sets are inconsistent, and the generalization ability of the model is mainly evaluated.

ESC* is another data partition setting for the ESC dataset, which is used in Man et al. (2022). In this dataset, documents are randomly shuffled based on their document names without sorting according to their topic IDs. Thus, the training and test sets have data on all topics. Under this setting, the data distributions of the training and test sets are more consistent, and it can better reflect the performance of the model under the same distribution of data. In real data, some causal event pairs are mostly appeared in topic-specific documents, because the event type is related to the topic of the document. This phenomenon inspires us to split the dataset in two different ways, i.e., cross-topic partition (ESC) and random partition (ESC*).

Causal-TB [4] contains 183 documents and 6811 event mentions. There are 9721 intra-sentence event mention pairs in the corpus, of which 298 (3.1%) are annotated with a causal relation. Similar to Liu et al. (2020), we conduct 10-fold cross-validation for Causal-TB.

Evaluation Metrics. For evaluation, we adopt widely used Precision (P), Recall (R), and F1-score (F1) as evaluation metrics.

4.2 Expeimental Setup

Implementation Details. In the experiments, we use the pre-trained AMR parser parse_xfm_bart_large v0.1.0 [5]. The PLM used in this paper is BERT-base (Devlin et al., 2019) and it is fine-tuned during the training process. The representation dimension of nodes and relations is set to 768, the same as the representation dimension of tokens. The NetwokX toolkit [6] is adopted to obtain the shortest path between two events. The learning rate of the model is set to 1e-5 and the dropout rate is set to 0.5. We perform grid search on the number of the RGCN layers, and it is experimentally set to 3. γ in focal loss is set to 2. β is set to 0.5 and 0.75 for ESC and Causal-TB, respectively. The batch size is set to 20 for all the three datasets. The AdamW gradient strategy is used to optimize all parameters. Due to the sparsity of causality in the Causal-TB dataset, we use both positive and negative sampling strategies for training. The positive sampling rate and negative sampling rate are set to 5 and 0.3, respectively.

Method	P	R	F1
LSTM (Cheng and Miyao, 2017)	34.0	41.5	37.4
Seq (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
KnowDis (Zuo et al., 2020)	39.7	66.5	49.7
MM (Liu et al., 2020)	41.9	62.5	50.1
CauSeRL (Zuo et al., 2021a)	41.9	69.0	52.1
LSIN (Cao et al., 2021)	47.9	58.1	52.5
LearnDA (Zuo et al., 2021b)	42.2	**69.8**	52.6
SemSIn	**50.5**	63.0	**56.1**
T5 Classify* (Man et al., 2022)	39.1	**69.5**	47.7
GenECI* (Man et al., 2022)	59.5	57.1	58.8
SemSIn*	**64.2**	65.7	**64.9**

Table 2: Experimental results on ESC and ESC*. * denotes experimental results on ESC*.

Method	P	R	F1
RB (Mirza and Tonelli, 2014)	36.8	12.3	18.4
DD (Mirza and Tonelli, 2014)	67.3	22.6	33.9
VR-C (Mirza, 2014)	**69.0**	31.5	43.2
MM (Liu et al., 2020)	36.6	55.6	44.1
KnowDis (Zuo et al., 2020)	42.3	60.5	49.8
LearnDA (Zuo et al., 2021b)	41.9	68.0	51.9
LSIN (Cao et al., 2021)	51.5	56.2	52.9
CauSeRL (Zuo et al., 2021a)	43.6	**68.1**	53.2
GenECI (Man et al., 2022)	60.1	53.3	56.5
SemSIn	52.3	65.8	**58.3**

Table 3: Experimental results on Causal-TB.

Baseline Methods. We compare the proposed SemSIn method with two types of existing state-of-the-art (SOTA) methods, namely, feature-based ones and PLM-based ones. For the ESC dataset, the

[4] https://github.com/paramitamirza/Causal-TimeBank

[5] https://github.com/bjascob/amrlib

[6] https://networkx.org/

following baselines are adopted: **LSTM** (Cheng and Miyao, 2017) is a sequential model based on dependency paths; **Seq** (Choubey and Huang, 2017) is a sequential model that explores context word sequences; **LR+** and **ILP** (Gao et al., 2019), they consider the document-level causal structure. For Causal-TB, the following baselines are selected: **RB** (Mirza and Tonelli, 2014) is a rule-based method; **DD** (Mirza and Tonelli, 2014) is a data-driven machine learning based method; **VR-C** (Mirza, 2014) is a verb rule-based model with lexical information and causal signals.

In addition, we also compare SemSIn with typical methods based on PLMs. **KnowDis** (Zuo et al., 2020) is a knowledge enhanced distant data augmentation framework; **MM** (Liu et al., 2020) is a knowledge enhanced method with mention masking generalization; **CauSeRL** (Zuo et al., 2021a) is a self-supervised method; **LSIN** (Cao et al., 2021) is a method that constructs a descriptive graph to explore external knowledge; **LearnDA** (Zuo et al., 2021b) is a learnable knowledge-guided data augmentation framework; **T5 Classify** and **GenECI** (Man et al., 2022) are the methods that formulate ECI as a generation problem.

4.3 Experimental Results

Tables 2 and 3 present the experimental results on the ESC and Causal-TB datasets, respectively. Overall, our method outperforms all baselines in terms of the F1-score on both datasets. Compared with the SOTA methods, SemSIn achieves more than 3.5% and 1.8% improvement on the ESC and Causal-TB datasets, respectively. Note that, although our method does not utilize external knowledge, it still achieves better results than the SOTA methods. The reason is that our method makes better use of the semantic structure information in the texts. The results indicate that the texts still contain a considerable amount of useful information for the ECI task that can be mined and exploited.

Compared with the SOTA method LearnDA in Table 2, SemSIn achieves a significant improvement of 8.3% in precision on the ESC dataset. This suggests that SemSIn can better model the implicit associations between two events. It can be observed that LearnDA has a higher recall score than SemSIn. The possible reason is that LearnDA can generate event pairs out of the training set. Extra training samples make the model recall more samples and get a higher recall score.

Method	**P**	**R**	**F1**	Δ
SemSIn$_{w/o.stru}$	49.8	49.0	49.4	-
SemSIn$_{w/o.path}$	49.3	52.6	50.9	+1.5
SemSIn$_{w/o.cent}$	44.5	63.6	52.4	+3.0
SemSIn	50.5	63.0	**56.1**	**+6.7**

Table 4: Ablation results on ESC. Δ means the improvement of the F1 score relative to SemSIn$_{w/o.stru}$.

To verify the effectiveness of the model on the ESC* dataset, we compare the proposed method with the SOTA T5 Classify and GenECI methods. The results are in the bottom of Table 2. SemSIn achieves 4.7%, 8.6%, and 6.1% improvements of the precision, recall and F1-score, respectively. This again justifies that using semantic structures is beneficial for ECI.

Comparing the results of SemSIn and SemSIn* in Table 2, the experimental results under different settings have a large gap. The results on ESC are significantly higher than those on ESC*. This is because the training and test data for ESC are cross-topic, and data on different topics usually involve diverse events. Dealing with unseen event pairs is difficult, thus it is more challenging to conduct the ECI task on ESC than ESC*.

4.4 Ablation Studies

To illustrate the effect of two kinds of semantic structures, we conduct ablation experiments on the ESC dataset. The results are presented in Table 4. $w/o.stru$ indicates the model predicts event causality without two kinds of semantic structures. $w/o.path$ and $w/o.cent$ indicate without the event-associated structure and without the event-centric structure, respectively.

Impact of the Event-centric Structure. Compared with SemSIn, SemSIn$_{w/o.cent}$ has a 6.0% decrease of the precision score. By considering the event-centric structure information, the model can describe events more accurately. Thus SemSIn$_{w/o.cent}$ is worse than SemSIn. Comparing SemSIn$_{w/o.path}$ with SemSIn$_{w/o.stru}$, SemSIn$_{w/o.path}$ achieves 1.5% improvements of the F1 score. It proves that the associations between events and their arguments are vital for the ECI task. The event-centric Structure information can enhance BERT with the ability to capture these associations.

Method	P	R	F1
CompGCN	46.8	62.0	53.3
GCN	50.3	56.8	53.4
RGCN	**50.5**	**63.0**	**56.1**

Table 5: Experimental results using different graph encoders on ESC.

Impact of the Event-associated Structure. Compared with SemSIn, SemSIn$_{w/o.path}$ has a 10.4% decrease of the recall score. This indicates that the event-associated structure information can help the model discover more causal clues between two events. Comparing SemSIn$_{w/o.cent}$ with SemSIn$_{w/o.stru}$, SemSIn$_{w/o.cent}$ achieves 3.0% improvements of the F1 score. It proves that the associations between events are vital for this task.

4.5 Sub-module Analysis

Impact of Relations in the Path. In the phase of acquiring semantic paths between two events, we keep only the nodes in the paths and neglect the edges. This method achieves an F1 score of 53.3% on ESC, which is a 2.8% reduction compared to the model that considers both nodes and edges. It suggests that the relations between elements are also useful for identifying causality.

Impact of the Path Attention. In the multi-path information integration phase, we replace the attention mechanism with a method that averages the representations of multiple paths. This approach obtains an F1 score of 54.0% on ESC, which is a 2.1% reduction compared to the model utilizing the attention mechanism. This shows that the "Path Attention" sub-module can effectively aggregate information from multiple paths.

4.6 Graph Encoder Analysis

Impact of the Graph Encoder. To analyze the effect of graph encoders on experimental results, we utilized three different graph encoders, namely, GCN (Kipf and Welling, 2017), CompGCN (Vashishth et al., 2020), and RGCN (Schlichtkrull et al., 2018). The results are shown in Table 5. From the results, we can observe that the best result is achieved with the model using the RGCN graph encoder. This suggests that RGCN has the capability to utilize the edge-type information in the semantic graph more effectively, enabling more accurate aggregation of information from surrounding nodes.

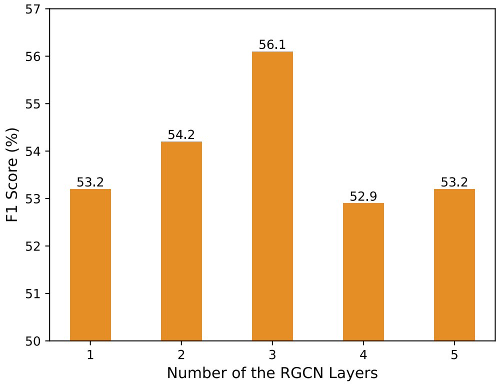
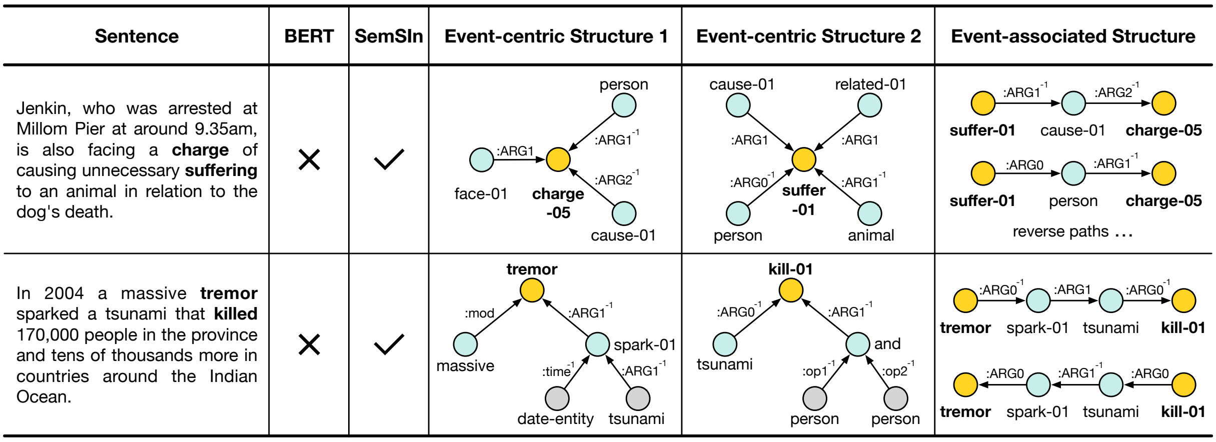


Figure 3: Impact of the number of the RGCN layers on ESC.

Impact of the Number of the RGCN Layers. The number of the RGCN layers L is an important parameter of the model, which means that nodes can aggregate information from their L-hop neighbors through message passing. We evaluate performance of the model with different numbers of the RGCN layers on ESC. The results are shown in Figure 3. A larger L can get better results when $L <= 3$. This is because that events usually have associations with their arguments mentioned in the text and event arguments also have associations with some related semantic elements. Thus introducing a relative large L can describe the events more precisely. It can be observed that the model performance decreases significantly when $L > 3$. The reason may be that the larger L may introduce some noisy nodes or the RGCN encounters the over-smoothing problem (Kipf and Welling, 2016).

4.7 Case Studies

To well demonstrate how semantic structures can help improve performance, a few cases are studied. Figure 4 shows two cases where causal relations between events are implicit. Here, BERT predicts the wrong answers and SemSIn predicts the correct ones, which demonstrates that leveraging the semantic structure information can effectively enhance ECI. In Case 1, the meaning of the "charge-05" is that "make an allegation or criminal charge against someone". Its event-centric structure information includes "someone is facing a charge", "a person is charged" and "the charge is for causing something to happen", which are the elements directly related to the event. By aggregating the information of these elements, the event is semantically represented more precisely. In Case 2, the causal relation between the two events



Sentence	BERT	SemSIn	Event-centric Structure 1	Event-centric Structure 2	Event-associated Structure
Jenkin, who was arrested at Millom Pier at around 9.35am, is also facing a **charge** of causing unnecessary **suffering** to an animal in relation to the dog's death.	×	✓			
In 2004 a massive **tremor** sparked a tsunami that **killed** 170,000 people in the province and tens of thousands more in countries around the Indian Ocean.	×	✓			

Figure 4: Results of the case study where bold indicates events. × and ✓ indicate wrong and correct predictions, respectively.

"tremor" and "kill" is expressed indirectly through the "tsunami" event. Specifically, it can be deduced using "tremor sparked a tsunami" and "the tsunami killed tens of thousands of people". The model effectively utilizes the event-associated structure information to capture the associations between events.

5 Conclusions

In this paper, we proposed a new semantic structure integration model (SemSIn) for ECI, which leveraged two kinds of semantic structures, i.e., event-centric structure and event-associated structure. An event aggregator was utilized to aggregate event-centric structure information and a path aggregator was proposed to capture event-associated structure information between two events. Experimental results on three widely used datasets demonstrate that introducing semantic structure information helps improve the performance of the ECI task.

Limitations

The limitations of this work can be concluded into two points: (1) To obtain the associations between semantic elements, SemSIn needs to transform the texts into the corresponding semantic graphs. Existing methods can only transform single sentences into semantic graphs, and cannot parse texts containing multiple sentences. Therefore, this method is not suitable for identifying causal relations between events in different sentences. (2) SemSIn only exploits the semantic structures of the texts and does not utilize external knowledge. External knowledge is also important for the ECI task, and simultaneously exploiting semantic structures and external knowledge is a good direction for future studies.

Acknowledgements

The work is supported by the National Natural Science Foundation of China under grant U1911401, the National Key Research and Development Project of China, the JCJQ Project of China, Beijing Academy of Artificial Intelligence under grant BAAI2019ZD0306, and the Lenovo-CAS Joint Lab Youth Scientist Project. We thank anonymous reviewers for their insightful comments and suggestions.

References

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Brandon Beamer and Roxana Girju. 2009. Using a bigram event model to predict causal potential. In *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, pages 430–441, Berlin, Heidelberg. Springer.

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D. Manning. 2014. Modeling biological processes for reading comprehension. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1499–1510, Doha, Qatar. Association for Computational Linguistics.

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. 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, Vancouver, Canada. Association for Computational Linguistics.

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. ACL Press.

Prafulla Kumar Choubey and Ruihong Huang. 2017. A sequential model for classifying temporal relations between intra-sentence events. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1796–1802, Copenhagen, Denmark. Association for Computational Linguistics.

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.

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.

Chikara Hashimoto. 2019. Weakly supervised multilingual causality extraction from wikipedia. 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 2988–2999, Hong Kong, China. Association for Computational Linguistics.

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.

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.

Yafan Huang, Feng Zhao, Xiangyu Gui, and Hai Jin. 2021. Path-enhanced explainable recommendation with knowledge graphs. *World Wide Web*, 24(5):1769–1789.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988.

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

Hieu Man, Minh Nguyen, and Thien Nguyen. 2022. Event causality identification via generation of important context words. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 323–330, Seattle, Washington. Association for Computational Linguistics.

Paramita Mirza. 2014. Extracting temporal and causal relations between events. In *Proceedings of the ACL 2014 Student Research Workshop*, pages 10–17, Baltimore, Maryland, USA. Association for Computational Linguistics.

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.

Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Motoki Sano, Stijn De Saeger, and Kiyonori Ohtake. 2013. Why-question answering using intra- and intersentential 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*, WSDM '17, pages 415–424, New York, NY, USA. Association for Computing Machinery.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational linguistics*, 31(1):71–106.

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.

Mehwish Riaz and Roxana Girju. 2013. Toward a better understanding of causality between verbal events: Extraction and analysis of the causal power of verb-verb associations. In *Proceedings of the SIGDIAL 2013 Conference*, pages 21–30, Metz, France. Association for Computational Linguistics.

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.

Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, pages 593–607. Springer.

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*.

Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha P. Talukdar. 2020. Composition-based multi-relational graph convolutional networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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

Zixuan Zhang and Heng Ji. 2021. Abstract meaning representation guided graph encoding and decoding for joint information extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 39–49, Online. Association for Computational Linguistics.

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.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
In the limitations section.

☒ A2. Did you discuss any potential risks of your work?
Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and introduction section.

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

B ☒ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
No response.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

□ 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)?
No response.

□ 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?
No response.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

□ 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.
No response.

C ☑ Did you run computational experiments?

Section 4.

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

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?
Section 4.2.

☑ 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?
Section 4.1.

☑ 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.)?
Section 4.2.

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.