Advancing Event Causality Identification via Heuristic Semantic Consistency Inquiry Network

Anonymous ACL submission

Abstract

Event Causality Identification (ECI) focuses on extracting causal relations between events in texts. Existing methods primarily utilize causal features and external knowledge to identify causality. However, such approaches fall short in two dimensions: (1) the causal features between events in a text often lack explicit clues, and (2) external knowledge may introduce bias, while specific problems require specific analyses. In light of these issues, we introduce a novel Semantic Consistency Inquiry (SemCI) to the ECI task and propose the **H**euristic **Sem**antic **C**onsistency **D**iscriminator (HSemCD), a model that is both straightforward and effective. HSemCD utilizes a *Cloze* Analyzer to facilitate a gap-filling game, aiming to help uncover the semantic dependency in the context. Subsequently, it assesses the semantic consistency between the fill-in token and the given sentence to detect the existence of causality. Through this assessment, HSemCD reveals the causal relations between events indirectly. Comprehensive experiments validate the effectiveness of HSemCD, which surpasses previous state-of-the-art methods on three widely used benchmarks.

1 Introduction

The challenging task of Event Causality Identification (ECI) of natural language understanding (NLU) aims to catch causal relations between event pairs in a text. For instance, "*Strong **winds** knocked down power lines, causing a **blackout**.*", an ECI model should identify the presence of a causal relation between event pair ***(winds, blackout)***. This task is important in language understanding and exhibits a wide range of application values (Oh et al., 2013, 2017; Berant et al., 2014; Mostafazadeh et al., 2016).

The conventional approach for ECI is a binary classification model that takes a triplet (sentence, event-1, event-2) as input and judges the existence or not of a causal relation between the two events, as illustrated at the top of Figure 1. To enhance the performance on the ECI task, various improvements have been implemented. Aside from the early feature-based methods (Hashimoto et al., 2014; Ning et al., 2018; Gao et al., 2019), several recently proposed representation-based methods have shown better ECI capabilities, including external knowledge enhanced methods (Liu et al., 2021; Cao et al., 2021), Pre-trained Language Models (PLMs) based methods (Shen et al., 2022), and data augmentation boosted methods (Zuo et al., 2021b). Recently, leveraging external prior knowledge to augment the model represents commonly. However, it also introduces potential bias. See the instance mentioned above, event pairs ***(winds, blackout)*** seem to be no direct causal relation, while it is reasonable to infer a causality considering the given context. Upon analysis, we can observe a semantic dependency: ***winds*** $\xrightarrow{\text{knocked down}}$ ***power lines*** $\xrightarrow{\text{causing}}$ ***blackout***. This finding means that the

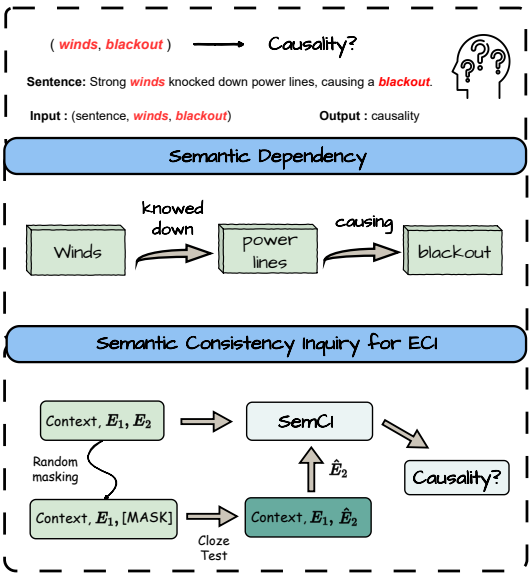


Figure 1: Introduction of the ECI task, along with our motivation.

causal relations between events within a given sentence also manifest as a form of context-dependent semantic reliance. Thus, we claim that the ECI task can be reformulated as a semantic consistency classification task between two events following their respective mappings within the same context.

To this end, we present Semantic Consistency Inquiry (SemCI) as an alternative solution for the ECI task. The goal of SemCI is to explore implicit causal relationships guided by contextual semantic consistency analysis. To meet the SemCI, we propose a model, namely **H**euristic **Sem**antic **C**onsistency **D**iscriminator (HSemCD). HSemCD comprises three primary modules: a *Cloze* Analyzer, a Semantic Consistency Encoder, and a Causality Discriminator. Specifically, HSemCD first utilizes the *Cloze* analyzer to generate a fill-in token within the context. Meanwhile, the semantic consistency encoder captures semantic dependencies by encoding the source sentence. Finally, the causality discriminator assesses whether there is a causal dependency between the given event pairs by evaluating whether the introduction of the fill-in token maintains semantic consistency with the original sentence. The main contributions of this work are summarized as follows:

- We propose the Semantic Consistency Inquiry (SemCI) as a potential alternative solution to the ECI task, highlighting the significance of context-dependent semantic dependency analysis in detecting causal relations.
- We introduce a Heuristic Semantic Consistency Discriminator (HSemCD) to implement the SemCI. HSemCD offers simplicity, effectiveness, and represents the first attempt to extend the semantic consistency inquiry to the ECI task.
- The experimental results on three widely used datasets demonstrate that HSemCD achieves 2.1%, 10.8%, and 4.6% improvements in terms of the F1 score compared to the previous SOTA methods, confirming its effectiveness.

2 Related Work

Identifying causal relations between events at the document level and sentence level is challenging and has attracted massive attention in the past few years. In this paper, we focus on the sentence-level ECI. Early methods mainly utilize explicit causal patterns (Hashimoto et al., 2014; Riaz and Girju, 2014a), lexical and syntactic features (Riaz and Girju, 2013, 2014b), and causal indicators or signals (Do et al., 2011; Hidey and McKeown, 2016) to identify causal relations.

Recently, several representation-based methods utilizing Pre-trained Language Models (PLMs) have improved the performance of the ECI task. To address the lack of training data for ECI, (Zuo et al., 2020, 2021b) proposed data augmentation methods, which can generate additional training data to alleviate the problem of overfitting. With the intuition that commonsense causal relations are helpful for ECI, (Liu et al., 2021; Cao et al., 2021) incorporated external knowledge from the knowledge graph ConceptNet (Speer et al., 2017) to enrich the representations derived from PLMs. However, the performance of external knowledge-based methods is closely related to the consistency between the target task domain and the knowledge bases utilized, which can introduce potential bias and result in vulnerabilities in such approaches. Different from previous methods, (Man et al., 2022) introduced a dependency path generation approach for ECI, explicitly enhancing the causal reasoning process. (Hu et al., 2023) exploited two types of semantic structures, namely event-centered structure and event-associated structure, to capture associations between event pairs.

3 Preliminaries

3.1 Problem Statement

Let $\boldsymbol{S} = [S_1, \cdots, S_n] \in \mathbb{R}^{1\times|S|}$ refer to a sentence with $|S|$ tokens, where each token S_i is a word/symbol, including special identifiers to indicate event pair (S_{e_1}, S_{e_2}) in causality. Traditional ECI models determine if there exists a causal relation between two events by focusing on event correlations, which can be written as $\mathcal{F}(\boldsymbol{S}, S_{e_1}, S_{e_2}) = \{0, 1\}$. Actually, correlation does not necessarily imply causation, but it can often be suggestive. Therefore, this study investigates the Semantic Consistency Inquiry (SemCI) as a potential alternative solution to the ECI task. For clarity, we introduce two fundamental problems:

Cloze Test. We denote a mask indicator as $m = [m_1, \cdots, m_{|S|}\} \in \{0, 1\}^{1\times|S|}$, where $m_i = 0$ if S_i is event token, otherwise $m_j = 1, j \in [1, \cdots, |S|], j \neq i$. We use $\hat{\boldsymbol{S}}$ instead of $\boldsymbol{S}$ to explicitly represent the incomplete sentence, i.e, $\hat{\boldsymbol{S}} = m\boldsymbol{S}$. The *Cloze* test in this study is to develop a contextual semantic-based network $\boldsymbol{\Omega}(\cdot)$ to fill

in the masked word, i.e., $\boldsymbol{\Omega}(\hat{\boldsymbol{S}}) \mapsto S_m$, where S_m denotes the generated fill-in token.

Heuristic Consistency Discrimination. Given the input tuple $(\boldsymbol{S}, S_m)$, a discriminator $\boldsymbol{\mathcal{D}}(\cdot)$ aims to examine the presence of semantic consistency in sentence $\boldsymbol{S}$ with regard to S_m, i.e., $\boldsymbol{\mathcal{D}}(\boldsymbol{S}, \boldsymbol{S_m}) \in \{0, 1\}$. Since S_m is the generated fill-in token of $\boldsymbol{\Omega}(\hat{\boldsymbol{S}})$, the alignment of S_m with the source sentence $\boldsymbol{S}$ in logical and semantic context can serve as an indication of a causal relationship, i.e., $\boldsymbol{\mathcal{D}}(\boldsymbol{S}, \boldsymbol{\Omega}(\hat{\boldsymbol{S}})) \Leftrightarrow \mathcal{F}(\boldsymbol{S}, S_{e_1}, S_{e_2})$.

3.2 Basic Technique

The multi-head attention mechanism is the core part of Transformer (Vaswani et al., 2017) and has been widely adopted for sequential knowledge modeling. It measures the similarity scores between a given query and a key, whereafter formulating the attentive weight for a value. The canonical formulation can be conducted by the scaled dot-product as follows:

$$\begin{aligned} \text{MHA}(A, B) &= \text{Concat}(H^1, \cdots, H^h), \\ \text{where } H^i &= softmax(\frac{QK^T}{\sqrt{d_h}})V, \\ \text{and } Q &= AW_Q, \{K, V\} = B\{W_K, W_V\}, \end{aligned} \tag{1}$$

herein, $W_{\{Q,K,V\}} \in \mathbb{R}^{d \times d_h}$ are head mapping parameters. Typically, the multi-head attention mechanism can be categorized into two types: (1) when $A = B$, the attention mechanism focuses on the relationship between different elements within the same input; (2) when $A \neq B$, the attention mechanism captures the relationship between elements from different inputs.

4 Methodology

4.1 Overview

In this section, we present our proposed HSemCD model, which reformulates the ECI task as a semantic consistency classification problem. Figure 2 provides an overview of the HSemCD, which comprises three primary components: a *Cloze* Analyzer, a Semantic Consistency Encoder, and a Causality Discriminator. The main distinguishing feature of our approach from other methods lies in the fact that we leverage reading comprehension to its full extent within the generative model, eliminating the need for additional prior knowledge and prioritizing simplicity and efficiency.

4.2 *Cloze* Analyzer

It is reasonable to believe that a well-trained deep generative model is powerful in context awareness (Goswami et al., 2020). In light of this, we adopt a straightforward approach of randomly masking one event from the event pair, and then predicting this event. This approach is inspired by the literary puzzle known as *Cloze*, which plays a crucial role in our framework. The *Cloze* facilitates the prediction of the most appropriate fill-in token for the masked word, revealing the probable semantic relationships within the given context.

Input Embedding Layer aims to encode sentences into a latent space. Given a sentence $\boldsymbol{S} = [S_1, \cdots, S_{e_1}, \cdots, S_{e_2}, \cdots, S_n]$, we correlate a $\hat{\boldsymbol{S}} = \boldsymbol{S} \odot \boldsymbol{M}_{mask}$, where $\odot$ denotes the element-wise product and $\boldsymbol{M}_{mask} = \{m_{1:n}\} \in \{0, 1\}^n$ indicates the randomly masked word. If $m_i = 0$, it means the S_i word is masked, which can be either S_{e_1} or S_{e_2}. In order to adhere to the *Cloze* puzzle setting, we utilize two pairs of specification symbols <e_1>, </e_1> and <e_2>, </e_2> to mark S_{e_1} and S_{e_2} in source sentence $\boldsymbol{S}$. Importantly, the masked word does not have the marker, thus resulting in $|\hat{\boldsymbol{S}}| = |\boldsymbol{S}| - 2$.

The input embedding layer encodes the $\boldsymbol{S}, \hat{\boldsymbol{S}}$ associated with its position. The word embeddings are trained along with the model and initialized from pre-trained BERT word vectors with a dimensionality of $d = 1024$. The specification symbol <e_*> and $[mask]$ are mapped to the appointed tokens, and their embeddings are trainable with random initialization. The position embedding is computed by the $sine$ and $cosine$ functions proposed by Transformer. Finally, the outputs of a given sentence from this layer are the sum of the word embedding and position embedding, namely X and $\hat{X}$ for simplicity, respectively. The latter corresponds to a sentence with the masked word. Notably, $X \in \mathbb{R}^{(n+4) \times d}$, $\hat{X} \in \mathbb{R}^{(n+2) \times d}$.

Semantic Completion Block receives $\hat{X}$ as input, aiming to fill in the blank that is marked by $[mask]$ (i.e., $\hat{x}_m$). Inspired by the Transformer, it is a stack of the basic building block [MHA + norm + feed-forward layer + norm], as illustrated in the upper part of Figure 2. The main idea of this block is to take advantage of the $\hat{x}_m$ as a query, then fill the man-made gap. The process can be formulated as:

$$\tilde{\boldsymbol{c}} = \text{MHA}(\hat{x}_m, \hat{X}), \tag{2}$$

$$\boldsymbol{c} = (ReLU(LN(\tilde{\boldsymbol{c}}) + \hat{x}_m)W_{c_1} + b_{c_1})W_{c_2} + b_{c_2}, \tag{3}$$

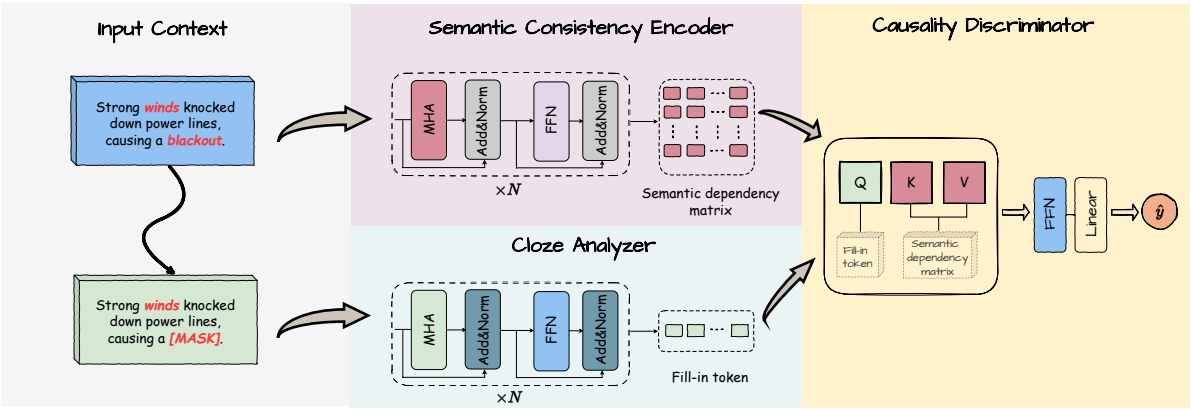


Figure 2: Overview of the proposed HSemCD.

where $\{W_*, b_*\}$ are learnable parameters, $\boldsymbol{c} \in \mathbb{R}^{1\times d}$ is the output of this block, i.e., the generated filled-in word.

4.3 Semantic Consistency Encoder

The Semantic Consistency Encoder enables the model to leverage the entire input field that facilitates comprehensive information reception and fosters an understanding of the logical relationships between words. It receives X as input to establish the semantic dependencies present in the entire sentence. It is also a stack of the basic building block: [MHA+norm+feed-forward layer+norm], which is formalized as:

$$\tilde{\boldsymbol{H}} = \text{MHA}(X, X), \tag{4}$$

$$\boldsymbol{H} = (ReLU(LN(\tilde{\boldsymbol{H}}) + X)W_{H_1} + b_{H_1})W_{H_2} + b_{H_2} \tag{5}$$

where $\{\boldsymbol{H}, \tilde{\boldsymbol{H}}\} \in \mathbb{R}^{(n+4)\times d}$ are sentence representations that assimilate intricate semantic connections among words.

4.4 Causality Discriminator

According to our motivation, we conduct a causality inquiry between the fill-in token and the semantic dependency matrix by utilizing cross attentive network, namely:

$$\boldsymbol{z} = \text{MHA}(\boldsymbol{c}, \boldsymbol{H}). \tag{6}$$

After that, we obtain the $\boldsymbol{z} \in \mathbb{R}^{1\times d}$ as the result of the inquiry. A two-layer feed-forward network transforms it to the causality classifier as:

$$\boldsymbol{y}_z = (ReLU(\boldsymbol{z}W_{in} + b_{in})W_{out} + b_{out}), \tag{7}$$

where $\{W_*, b_*\}$ are learnable parameters.

4.5 Training Criterion

We adopt the cross-entropy loss function to train the HSemCD:

$$J(\Theta) = -\sum_{(s_{e_1}, s_{e_2})\in \boldsymbol{S}} \boldsymbol{y}_{(s_{e_1}, s_{e_2})} \log\left(softmax(\boldsymbol{y}_z W_y + b_y)\right), \tag{8}$$

where Θ denotes the model parameters, $\boldsymbol{S}$ refers to the full sentence in the training set, (s_{e_1}, s_{e_2}) are the events pairs and the $\boldsymbol{y}_{(s_{e_1}, s_{e_2})}$ is a one-hot vector that indicates the gold relationship between s_{e_1} and s_{e_2}. That is, we utilize the $\boldsymbol{y}_{s_{(*)}}$ to guide whether the generated fill-in token is causally related to the source sentence $\boldsymbol{S}$, i.e., $\boldsymbol{y}_{S_{(*)}} \rightarrow \boldsymbol{y}_z \rightarrow \mathcal{D}(\boldsymbol{S}, \Omega(\hat{\boldsymbol{S}}))$.

It is essential to highlight that we do not establish a loss to directly guide the fill-in token. This is because we do not necessitate alignment between the fill-in tokens and the original words. Instead, our objective is to generate tokens through the *Cloze* analyzer based on semantic coherence within the context and then use it to inquire about the presence of a causal relationship. This aligns with our main argument: *the existence of a causal relationship between two events is context-dependent*.

5 Experiments

5.1 Datasets and Evaluation Metrics

We evaluate the HSemCD on widely-used ECI benchmark datasets, including two from EventStoryLine v0.9 (ESC) (Caselli and Vossen, 2017) and one from Causal-TimeBank (CTB) (Mirza et al., 2014), namely ESC, ESC*, and CTB.

ESC[1] contains 22 topics, 258 documents, and 5334 event mentions. The same as (Gao et al., 2019), we exclude aspectual, causative, perception,

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

and reporting event mentions, since most of which were not annotated with any causal relation. After the data processing, there are 7805 intra-sentence event mention pairs in the corpus, 1770 (22.67%) of which are annotated with a causal relation. Identical to the data split in previous methods (Hu et al., 2023; Zuo et al., 2021b), we select the last two topics in ESC as development set and use the remaining 20 topics for a 5-fold cross-validation. Note that the documents are sorted according to their topic IDs under this data partition setting, which means that the training and test sets are cross-topic. Due to the distribution gap between the training and test sets, the generalization ability of the model can be better evaluated.

ESC* is another data partition setting for the ESC dataset, which is adopted in (Man et al., 2022; Hu et al., 2023). Instead of sorting the documents according to their topic IDs, documents are randomly shuffled under this setting. Thus, the distributions of the training and test sets are more consistent, because both two sets contain data on all topics. The experimental results under this setting can better demonstrate the model's ability to identify causal relations in topic-centered documents, which are common in real-world scenarios.

CTB [2] contains 183 documents and 6811 event mentions. Out of 9721 intra-sentence event pairs, 298 (3.1%) pairs are annotated with causal relations. The same as previous methods (Man et al., 2022; Hu et al., 2023), we perform a 10-fold cross-validation on CTB. Given the sparsity of causality in the CTB dataset, we follow existing works to conduct a negative sampling technique for training with the sampling rate of 0.7.

Evaluation Metrics. We utilize the commonly used **P**recision, **R**ecall, and **F1**-score as evaluation metrics.

5.2 Experimental Setup

Implementation Details. Our model is based on the uncased BERT model released by HuggingFace Transformer library [3] and is fine-tuned during the training process. All experiments are accelerated by one piece of Nvidia GeForce RTX 3090. The dimension of hidden units is set to 1024, the batch size is set to 20, and the dropout rate is set to 0.5. The *Cloze* Analyzer and semantic consistency encoder are stacked 24 blocks, with each block having 16 heads in MHA. The gradient strategy used for optimization is AdamW (Loshchilov and Hutter, 2017) with an initial learning rate of $1e-5$, and a total of 100 epochs are utilized to train the model. For reproducibility, the source codes are anonymously available at https://anonymous.4open.science/r/ECI-2658.

Baslines. We compare our proposed HSemCD model with existing state-of-the-art (SOTA) ECI methods, including feature-based methods and methods based on Pre-trained Language Models (PLMs). For the ESC dataset, we adopted the following baselines: **LSTM** (Cheng and Miyao, 2017), a dependency path boosted sequential model; **Seq** (Choubey and Huang, 2017), a sequence model explores manually designed features for ECI. **LR+** and **ILP** (Gao et al., 2019), models considering document-level structure. For the CTB dataset, we select **RB** (Mirza and Tonelli, 2014), a rule-based ECI system; **DD** (Mirza and Tonelli, 2016), a data-driven machine learning-based method; **VR-C** (Mirza, 2014), a verb rule-based model boosted by filtered data and causal signals.

Furthermore, we also contrast HSemCD with the following PLMs-based methods: **MM** (Liu et al., 2021), a commonsense knowledge enhanced method with mention masking generalization; **KnowDis** (Zuo et al., 2020), a knowledge-enhanced distant data augmentation approach; **LearnDA** (Zuo et al., 2021b), a learnable augmentation framework alleviating lack of training data. **LSIN** (Cao et al., 2021), an approach which constructs a descriptive graph to exploit external knowledge; **CauSeRL** (Zuo et al., 2021a), a self-supervised method utilizing external causal statements. **GenECI** and **T5 Classify** (Man et al., 2022), methods that formulates ECI as a generation problem. **SemSIn** (Hu et al., 2023) is the previous SOTA method that leverages event-centric structure and event-associated structure for causal reasoning. Similar to our approach, it also does not utilize external knowledge. **KEPT** (Liu et al., 2023), a study that leverages BERT to integrate external knowledge bases for ECI, sharing the same fundamental structure as ours.

5.3 Main Results

Table 1 and Table 2 report the performance of different approaches on three datasets, respectively. The best scores are highlighted in **bold**, and the second-best scores are underlined. Specifically, we

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

[3] https://huggingface.co/bert-large-uncased

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., 2021)	41.9	62.5	50.1
CauSeRL (Zuo et al., 2021a)	41.9	69.0	52.1
LSIN (Cao et al., 2021)	49.7	58.1	52.5
LearnDA (Zuo et al., 2021b)	42.2	**69.8**	52.6
SemSIn (Hu et al., 2023)	50.5	63.0	56.1
KEPT (Liu et al., 2023)	50.0	68.8	57.9
HSemCD	**52.2**	68.5	**59.1**
T5 Classify* (Man et al., 2022)	39.1	69.5	47.7
GenECI* (Man et al., 2022)	59.5	57.1	58.8
SemSIn* (Hu et al., 2023)	64.2	65.7	64.9
HSemCD*	**70.9**	**73.0**	**71.9**

Table 1: 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, 2016)	67.3	22.6	33.9
VR-C(Mirza, 2014)	**69.0**	31.5	43.2
MM (Liu et al., 2021)	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
KEPT (Liu et al., 2023)	48.2	60.0	53.5
GenECI (Man et al., 2022)	60.1	53.3	56.5
SemSIn (Hu et al., 2023)	52.3	65.8	58.3
HSemCD	59.1	66.4	**61.0**

Table 2: Experimental results on CTB.

have the following observations.

Overall, HSemCD demonstrates superior performance compared to all baselines in terms of the F1-score, including feature-based and PLMs-based methods, confirming its effectiveness. More specifically, HSemCD surpasses the previous state-of-the-art (SOTA) by significant margins of 1.2, 7.0, and 2.7 in terms of **F1**-score on ESC, ESC*, and CTB, respectively. It is important to point out that HSemCD achieves this new SOTA performance without relying on any prior knowledge or extra information. This further supports our claim that the ECI task should prioritize attention to the semantic consistency problem within the given sentence. It is demonstrated by the improvements in model precision, as evidenced by HSemCD achieving the highest precision score on ESC, ESC*, and a considerably high precision score on CTB.

Compared to the method LearnDA, which achieves the highest **R**ecall score in the ESC dataset (in the top of Table 1), HSemCD shows a significant improvement of 23.7% in **P**recision. This suggests that HSemCD has better reliability in making decisions. It is understandable that the LearnDA performs better recall, as it can generate additional training event pairs out of the training set. We observe similar phenomenon in CTB compared to model CauSeRL. Given that the HSemCD achieves the highest **P**recision and **F1** score, we claim that the HSemCD has the ability to strike a better balance between precision and recall.

Sharing the same fundamental structure of KEPT, HSemCD surpasses it by a large margin on both two datasets. The superior performance of HSemCD could be attributed to the concentration on semantic understanding rather than the inclusion of prior knowledge from external knowledge bases.

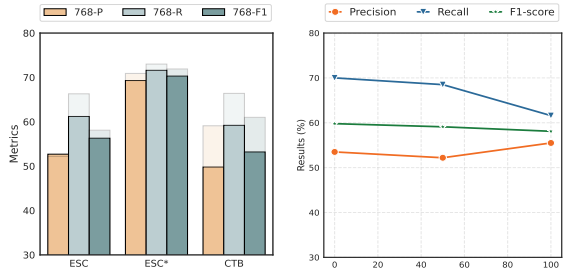
We note that on the ESC* dataset, HSemCD has achieved improvements of 10.4% in precision, 5.0% in recall, and 10.8% in F1-score. These results demonstrate the powerful ability of HSemCD to detect causal relations in topic-centered documents, which is beneficial for real-world applications. Additionally, the results of HSemCD on ESC* are significantly higher than on ESC. This discrepancy arises due to the cross-topic nature of the ESC training and test data, which presents a greater challenge for the model to generalize to unseen topics. Conversely, the ESC* data partition exhibits more consistent distributions between the training and test data, leading to improved performance.

5.4 Sensitivity Analysis

We now discuss the impacts of key hyper-parameters that affect the model performance.

Impact of hidden size. We analyze the impact of model size on two classic dimensions, 768 and 1024, as depicted in Figure 3(a), where the shaded portion corresponds to 1024. It can be observed that the overall performance of HSemCD improves notably when the model size increases, especially for the CTB dataset. This phenomenon can be attributed to the enhanced representation capability brought by higher model dimensions, facilitating a better reading comprehension ability, which is the core part of HSemCD. We can also observe that HSemCD is sensitive to hidden size under low-resource scenarios (CTB) while maintaining good

performance with sufficient annotated data for training (ESC and ESC*).



(a) Hidden size (b) Mask scheme

Figure 3: Sensitivity Analysis.

Impact of mask scheme. HSemCD effectively learns semantic dependency by the guidance of context-dependent *Cloze* test. Meanwhile, the *Cloze* test allows for generating multiple variations of a sentence by masking one of any events. To further figure out the impact of different mask schemes, we evaluate three versions: (1) "100% mask e_1" (0 in x-axis); (2) "100% mask e_2" (100 in x-axis); (3) randomly mask e_1 or e_2 with a 50/50 chance (50 in x-axis). As shown in Figure 3(b), HSemCD manifests stable performance across all schemes in terms of **F1**-score. Furthermore, an interesting phenomenon can be seen: scheme (1) exhibits higher **R**ecall and lower **P**recision compared to scheme (2). In other words, the model under scheme (1) seems to be over-confident, tending to give positive answers, while showing conservative in scheme (2). We hypothesize that this discrepancy originates from the presence of word order in the ESC dataset where the effect is presented before the cause (i.e., effect-before-cause). Specifically, the ratio of *cause-before-effect* sentences to *effect-before-cause* sentences in ESC is 1 : 1.3. Thus, when HSemCD utilizes scheme (1), there is a better chance that the masked event is "effect", inferencing causality with only the "cause" is inherently more challenging. Considering this sample, "Police said 12 people were injured in the riots" with event pairs (injured, riots), we can observe the *effect-before-cause*: riots $\xrightarrow{causing}$ injured. When "injured" (the effect $\&e_1$) is masked, the model only sees the precondition "riots". In this scenario, the consequence of the riots leading to "injured" is merely one of many potentials. For instance, "rescued" for "Police said 12 people were rescued in the riots" also represents a plausible. Conversely, when "riots" (the cause $\&e_2$) is masked, inferring causality with "injured" (the effect $\&e_1$) is relatively easy, resulting in higher precision.

Method	P	R	F1	∇
HSemCD	52.2	68.5	59.1	-
w/o. CA	53.4	64.0	58.0	−1.1
w/o. SemCE	49.5	49.9	49.4	−9.7

Table 3: Ablation results on ESC.

5.5 Ablation Study

We have designed two variants of HSemCD in ESC to investigate the contribution of each component, including: *HSemCD w/o CA*, removes *Cloze* analyzer and utilizes the original event embedding for causality inquiry; *HSemCD w/o SemCE*, removes semantic consistency encoder and directly feeds the generated fill-in token to the classifier. Table 3 illustrates the results. It can be observed first that both CA and SemCE contribute positively to performance improvement. Furthermore, when CA is removed, the performance of HSemCD decreases by 4.5 in terms of **R**ecall but increases by 1.2 in **P**recison. This phenomenon is similar to the presenting mask scheme (2) discussed in Sec. 5.4, when using original words as causal query terms makes the model more conservative. We argue that this approach, which directly adopts original words, overlooks the dedicated semantic dependency analysis of the words in the contextual setting, thereby limiting the judgment of causal dependencies to the computation of semantic similarity. Additionally, we found that after removing the SemCE, the performance of HSemCD drops by 9.7 in **F1**-score. This result emphasizes the significant role of SemCE and also supports our claim that the analysis of semantic consistency aids in the discovery of causal dependencies.

5.6 Interpretability Analysis

We now investigate the interpretability of interactive semantics obtained by HSemCD. To this end, we randomly select two examples from the ESC dataset to visualize the attention heatmap on the causality inquiry process, which is depicted in Figure 4. We first observed that the words generated by the CA exhibit homogeneity with the original terms in the two samples, such as the term pairs (hail, winds) in case 1 are natural disasters, (explosion, riots) in case 2 are events that cause harm. This further suggests that upon replacing the corresponding words with fill-in tokens, the semantics of the two sentences remain remarkably context-

Sentence	Masked	Fill-in	Golden	HSemCD
A goth was being **questioned** on suspicion of **murder** yesterday after his mother and sister were found dead at home.	questioned	investigated	✓	✓
A Kraft Foods plant worker who had been **suspended** for feuding with colleagues, then **escorted** from the building, returned minutes later with a handgun, found her foes in a break room and executed two of them with a single bullet each and critically wounded a third, police said Friday.	escorted	retired	✗	✓

Table 4: Case studies of HSemCD.

similar. This phenomenon indicates that HSemCD possesses the capability to distill semantics centered around event pairs, and that these semantics are congruent with the contextual environment. In addition, we can observe that HSemCD is enabled to focus on the two primary events within a sentence, which suggests that the semantic completion and causality inquiry designed in the HSemCD is able to refine the interactions between model-made event pairs (fill-in token, another event) and their associated semantic regions.

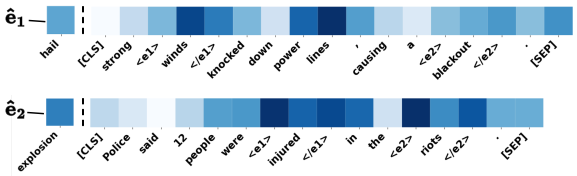


Figure 4: Visualization of the attention heatmap on the causality inquiry process. The token "$\hat{e}_*$" denotes the fill-in token associated with its event position.

5.7 Case Study

We present case studies to further explore the performance of HSemCD in Table 4. We randomly selected two examples from the experiment results. Overall, the *Cloze* Analyzer is designed to predict fill-in tokens that are consistent with the context of a given sentence, i.e., the predicted words "investigated" and "retired", which is aligned with our findings in 5.6. As seen in case 1, a clear semantic dependency can be observed: ***goth*** $\xrightarrow{\text{suspicion}}$ ***murder*** $\xrightarrow{\text{causing}}$ ***questioned***. Since the fill-in tokens are highly related to the source sentence, HSemCD provides a positive final decision. Case 2 servers as an example of a faulty decision. The fill-in token "retired" sharply contrasts with the original word "escorted". This directly leads to the model making erroneous decisions when conducting causal inquiry between the fill-in token and semantic dependency matrix.

6 Conclusions

In this paper, we investigate the Semantic Consistency Inquiry (SemCI) as a potential alternative solution for ECI and propose the **H**euristic **Sem**antic **C**onsistency **D**iscriminator (HSemCD). Our approach leverages a *Cloze* Analyzer to generate a fill-in token that is aware of context correlations, which is then employed as a query term to facilitate heuristic guidance for the model in executing causal inference, grounded in an understanding of semantic dependencies. Experimental evaluations conducted on three widely recognized datasets exhibit the superior performance of HSemCD, while also highlighting the contribution of SemCI in enhancing the ECI task.

Limitations

The limitations of this work can be concluded as follows:

1. HSemCD exhibits sensitivity to the quantity of annotated event pairs available for training. Consequently, it demonstrates reduced accuracy in capturing causal relations within the CTB, as illustrated in Table. 2. As a result, it still needs further improvement when facing low-resource scenarios.

2. While acknowledging the potential for bias introduced by external knowledge, we argue that incorporating commonsense is crucial for ECI, particularly in scenarios with different mask schemes between word order *cause-before-effect* and *effect-before-cause* (see Sec. 5.4). HSemCD concentrates on investigating the effectiveness of semantic consistency for ECI, missing the opportunity to take advantage of commonsense reasoning. Investigating how to integrate commonsense reasoning within the semantic-guided framework presents a promising avenue for future research.

References

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.

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. Association for Computational Linguistics.

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.

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.

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.

Ankur Goswami, Akshata Bhat, Hadar Ohana, and Theodoros Rekatsinas. 2020. Unsupervised relation extraction from language models using constrained cloze completion. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1263–1276, Online. 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, Baltimore, Maryland. Association for Computational Linguistics.

Christopher Hidey and Kathleen 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.

Zhilei Hu, Zixuan Li, Xiaolong Jin, Long Bai, Saiping Guan, Jiafeng Guo, and Xueqi Cheng. 2023. Semantic structure enhanced event causality identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10901–10913, Toronto, Canada. Association for Computational Linguistics.

Jian Liu, Yubo Chen, and Jun Zhao. 2021. 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.

Jintao Liu, Zequn Zhang, Zhi Guo, Li Jin, Xiaoyu Li, Kaiwen Wei, and Xian Sun. 2023. Kept: Knowledge enhanced prompt tuning for event causality identification. *Knowledge-Based Systems*, 259:110064.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

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.

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.

Paramita Mirza and Sara Tonelli. 2016. Catena: Causal and temporal relation extraction from natural language texts. In *The 26th international conference on computational linguistics*, pages 64–75. ACL.

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.

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018. Joint reasoning for temporal and causal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2278–2288, Melbourne, Australia. 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.

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, page 415–424, New York, NY, USA. Association for Computing Machinery.

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.

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.

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.

Shirong Shen, Heng Zhou, Tongtong Wu, and Guilin Qi. 2022. Event causality identification via derivative prompt joint learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2288–2299, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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.