Event Causality Extraction (ECE) aims to extract the cause-effect event pairs from the given text, which requires the model to possess a strong reasoning ability to capture event causalities. However, existing works have not adequately exploited the interactions between the cause and effect event that could provide crucial clues for causality reasoning. To this end, we propose an Implicit Cause-Effect interaction (ICE) framework, which formulates ECE as a template-based conditional generation problem. The proposed method captures the implicit intra- and inter-event interactions by incorporating the privileged information (ground truth event types and arguments) for reasoning, and a knowledge distillation mechanism is introduced to alleviate the unavailability of privileged information in the test stage. Furthermore, to facilitate knowledge transfer from teacher to student, we design an event-level alignment strategy named Cause-Effect Optimal Transport (CEOT) to strengthen the semantic interactions of cause-effect event types and arguments. Experimental results indicate that ICE achieves state-of-the-art performance on the ECE-CCKS dataset.

1 Introduction

Event Causality Extraction (ECE) is an emerging yet challenging natural language processing (NLP) task, which requires extracting the whole event structure of cause and effect events, including event type and event arguments. As shown in Figure 1, given the input text, an ECE system is expected to identify the cause event (i.e., the event type is Price Increase, the argument oil plays the role of Product, the argument worldwide as Region), and similarly identify the corresponding effect event. Recognizing event causality can provide support for many downstream NLP tasks, including machine reading comprehension (Berant et al., 2014), intelligent search (Rudnik et al., 2019), future event forecasting (Hashimoto, 2019), and why-question answering (Oh et al., 2017).

Existing studies mainly regard this task as a classification-based problem. Du and Cardie (2020) and Wang et al. (2019) extracted the events and classified the causal relation in a pipelined manner. Wang et al. (2020) leveraged grid tagging to simultaneously extract the events and causality pairs. Cui et al. (2022) also designed a dual grid tagging scheme, which aims at modeling the correlations between different event arguments. Compared to the classification-based methods, generation-based models can take full advantage of Pre-trained Language Models (PLMs) by designing flexible prompt templates (Ma et al., 2022; Liu et al., 2022). As a result, there is a trend to cast the extraction task as a conditional generation problem.

Although generation-based approaches have
achieved remarkable success, the limited interactions between the cause-effect events impede the model’s ability to reason effectively between events. Intuitively, the privileged information (Xu et al., 2020), which stands for the ground truth information of event types or event arguments, could provide valuable knowledge for inferring causal clues. Take Figure 1 as an example, if the model has already known the cause event is Price Increase, and the Product and Region are oil and worldwide in this event, the intra-event interactions could benefit the extraction of effect event. Similarly, if the model has known Price Increase causes Demand Increase during training, the intra-event interactions may help the model to capture event arguments more accurately. By incorporating different kinds of privileged information, the model could make full use of the implicit interactions and be guided to extract causal clues. However, due to the unavailability of such privileged information in practice, incorporating it naively will result in inconsistencies between training and test phases and may affect the model performance. Several methods (Liu et al., 2017; Wei et al., 2021) in other NLP fields have provided solutions to overcome this problem, but they are not applicable when ECE is modeled as a generative problem.

Moreover, generation-based methods are typically trained via maximum likelihood estimation (MLE) (Salakhutdinov, 2015), which maximizes the likelihood of the next word conditioned on its previous ground truth words. Then, it leverages cross-entropy loss to measure the difference at each position of the target sequence. Nevertheless, since MLE only emphasizes strict word-level alignment, it struggles to consider the semantic information from the perspective of event type or event arguments. For instance, as shown in Figure 1, when training the word new from the effect event argument new energy, MLE ignores new energy is a whole unit as the Industry, which results in a partial loss of semantics. The event type and event arguments from the cause-effect event pairs could also be regarded as different wholes, and by interacting with each other, the model could implicitly incorporate such event-level semantic information.

In this paper, we propose an Implicit Cause-Effect interaction (ICE) framework for ECE to address the above issues. Specifically, we formulate the ECE task as a template-based conditional generation problem, which takes the context and prompt template as the input, and decodes event causality and event structure from the generated sequence. To capture the implicit intra- and inter-event interactions, we feed different privileged information to the input template and train two well-informed teacher models. Then a student model is driven by imitating the behaviors of teachers to narrow the input difference of training and test phases through knowledge distillation (Hinton et al., 2015). Furthermore, to facilitate knowledge transfer and strengthen the interactions between cause and effect events, we design a Cause-Effect Optimal Transport (CEOT) mechanism by treating the event type and event arguments as model units, which could implicitly incorporate the event-level semantic information. In summary, the contributions of this paper are as follows:

1) This work proposes an ICE framework, which models event causality extraction in a generative paradigm and incorporates privileged knowledge for reasoning.

2) The proposed method implicitly captures the intra- and inter-event interactions through knowledge distillation, and employs a CEOT strategy to strengthen the semantic interactions of cause and effect events.

3) Experimental results show that our model achieves state-of-the-art performance, improving the F1-score by 8.39% on the ECE evaluation benchmark.

2 Related Work

Event Causality Extraction. ECE is derived from the previous event causality identification (ECI), which aims to recognize the causal relations between the given events in text (Zuo et al., 2021; Phu and Nguyen, 2021). Early methods mainly focus on syntactic and lexical features (Gao et al., 2019), causality patterns (Hidey and McKeown, 2016), and statistical causal clues (Hu and Walker, 2017). Recent works seek to employ external knowledge (Liu et al., 2020; Cao et al., 2021) or prompt-based models (Shen et al., 2022; Liu et al., 2023) for this task. But these methods only identify the causality of events expressed by a word or phrase, without considering the event type and event arguments. Cui et al. (2022) first proposed the ECE task and exploited the argument correlations to extract event causality and event structure. Some variant methods from relation extraction have also been applied to this task (Wang et al., 2020; Wei et al., 2020).
However, these works fail to model the implicit cause-effect interactions, making it difficult to extract causal clues.

**Knowledge Distillation.** This technique is first proposed by Hinton et al. (2015), which aims to transfer knowledge from a well-trained teacher to a student model. Jiao et al. (2020) and Sanh et al. (2019) used knowledge distillation for compressing large-scale pre-trained language models. Wu et al. (2021) and Li et al. (2022) adopted multiple-teacher knowledge distillation to improve the effectiveness of distillation. Wei et al. (2023a) proposed to incorporate related arguments knowledge through knowledge distillation for event argument extraction. Nevertheless, they are designed for classification-based methods and struggle to migrate to the ECE task under the generative pattern.

**Optimal Transport.** OT has a wide range of applications in NLP domains (Chen et al., 2019; Xu et al., 2021; Wei et al., 2023b). Li et al. (2020) proposed using optimal transport to tackle the exposure bias issue in training generative models by maximum likelihood estimation. Zhou et al. (2022) modeled events in the sequence as units and adopted optimal transport to explicitly extract the event semantics for generating temporally-ordered event sequences. In this paper, we employ optimal transport to improve the semantic interactions of event type and event argument for cause and effect events.

3 Methodology

Our ICE framework formulates ECE as a template-based generation problem and implicitly incorporates privileged information for reasoning. Under this paradigm, we train two well-informed teacher models by incorporating different privileged information into model inputs. Then we adopt a knowledge distillation mechanism to drive a student model to capture implicit cause-effect interactions, which could alleviate the difference of unavailable privileged information in the test stage. During the training phase, a CEOT strategy is adopted to improve the semantic interactions of cause-effect events and promote the training of the student. The overview of ICE is shown in Figure 2.

### 3.1 Task Formulation

The goal of ECE is to extract event causality and event structure from the text. Formally, given a context, ECE aims to extract a set of cause-effect event pairs \( \{E_{ca_i}, E_{ef_i}\}_{i=1}^{s} \), where \( E_{ca_i} \) and \( E_{ef_i} \) indicate the cause and effect event of the \( i \)-th pair, respectively. The event structure \( E = (t, A) \) contains the event type \( t \) and the argument set \( A \). Each argument in \( A \) corresponds to a role.

### 3.2 Generative Template-based ECE Model

**Template Creation.** At the input stage, we first construct a specific task-related template for ECE. Following Ma et al. (2022) and Liu et al. (2021), we design a soft template that contains learnable pseudo tokens and slots for all components we require extracting. Figure 2(c) shows the ECE template in our model, where \(<\text{Cause}>, <\text{Effect}>, <\text{type}>, </\text{type}>\), etc. are specific learnable pseudo tokens. Then we concatenate the context and template, and feed them into a Transformer-based model to generate the output sequence, where the slots in the template will be filled with concrete event type or event arguments of the cause event and effect event.

**Target Output Sequence.** For the cause-effect event pair in context, we construct the target output sequence \( Y \) for conditional generation by filling the ground truth cause event and effect event into the template. Note that when there is more than one argument corresponding to a role, they will be concatenated by a special token \(<\text{and}>\). If some roles of the event have no arguments, the corresponding positions in the target output sequence will be filled with \(<\text{None}>\).

**Training.** In the training process, we adopt the Transformer-based pre-trained language model BART (Lewis et al., 2020) as our basic model architecture, which consists an encoder and a decoder.

\[
H_e = \text{Encoder}(X) \\
H_d = \text{Decoder}(Y; H_e) 
\]

where \( X \) denotes the concatenation of context and template. The training target is to maximize the likelihood of the next token conditioned on the previous ground truth tokens in the sequence:

\[
\mathcal{L}_{\text{gen}} = - \sum_{l=1}^{L} \log p(Y_l|Y_{<l}, X) 
\]

where \( L \) is the length of the target output sequence.

**Inference.** After obtaining the generated sequence, we decode the event type and event arguments of cause-effect events from corresponding slots with a rule-based matching algorithm. Then we check
Figure 2: (a) The overview of the ICE learning framework. (b) The process of Cause-Effect Optimal Transport (CEOT). (c) The creation of the template. The cause event is marked in orange and the effect event is marked in green. Underlined tokens indicate slots for event types and event arguments of cause-effect events. And tokens with <angle brackets> denote specific pseudo tokens.

whether the event type is in the predefined event type set and whether each event argument is a span of the context.

### 3.3 Teacher-Student Distillation Learning

Based on the generative paradigm, we adopt a teacher-student distillation learning framework to capture the implicit intra- and inter-event interactions, which could implicitly incorporate the privileged information for event causality reasoning and narrow the inconsistencies between training and test phases.

Specifically, we train two well-informed teachers, including an event argument extractor and an effect event extractor. The event argument extractor aims to extract event arguments with the event type of cause event and effect event given, and the effect event extractor seeks to recognize event type and event arguments of effect event with knowing the information of cause event. With template-based generation formulation, we could introduce the privileged information into the model in a flexible way. As shown in Figure 2, for the event argument extractor, we construct a knowledge-enriched template by filling the ground truth event types of the cause and effect event in the raw template. Likewise, for the effect event extractor, we fill in the ground truth event type and arguments of the cause event to form a new knowledge-enriched template. We give an example for the construction of knowledge-enriched templates in Appendix B. Next, we concatenate the context and the knowledge-enriched templates as inputs to train teachers.

In the knowledge distillation stage, we use $H_{S_i}^l$ to denote the hidden states of the $i$-th decoder layer of the student model, and $H_{T_i}^l$ to denote the teacher’s. We adopt the mean squared error (MSE) to encourage the student model to match the hidden states of corresponding layers of the decoder in the teacher:

$$L_{mse} = \sum_{l=1}^{L} \sum_{i=1}^{N} \text{MSE}(H_{S_i}^l, H_{T_i}^l)$$  \hspace{1cm} (3)

where $N$ is the number of decoder layers. We also employ KL-Divergence to encourage the student to match the probability distribution of the teacher over the next possible word at each position:

$$L_{kl} = \sum_{l=1}^{L} \text{KL}(p_{S_i}^l, p_{T_i}^l)$$  \hspace{1cm} (4)

where $p_{S_i}^l$ and $p_{T_i}^l$ are probability distributions of student and teacher over the next possible token at position $l$. Please note that the teachers and student exploit the same model architecture and training objectives, but do not share model parameters. And the parameters of teachers are fixed during the training of the student. The overall loss for training...
student model with a single teacher is:

\[ L_{kd} = L_{gen} + \alpha L_{mse} + \beta L_{kl} \]  

(5)

where \( \alpha \) and \( \beta \) are weight coefficients.

3.4 Cause-Effect Optimal Transport

To improve knowledge transfer, we seek to model semantic information from the perspective of event type or event arguments and introduce a CEOT strategy to promote interactions of cause and effect events, which is achieved by event-level alignment of teacher and student representations.

Optimal transport defines a distance metric between two probability measures on a domain. Given two discrete probability measures \( \mu = \sum_{i=1}^{n} u_i \delta_{x_i} \) and \( \nu = \sum_{j=1}^{m} v_j \delta_{y_j} \), where \( \delta_{x_i} \) is the Dirac function centered on \( x \), the weights \( u = \{ u_i \} \) and \( v = \{ v_j \} \) satisfy the constraints \( \sum_{i=1}^{n} u_i = \sum_{j=1}^{m} v_j = 1 \).

Under this setting, the OT distance is formalized as the following problem (Luise et al., 2018):

\[
\begin{align*}
\mathcal{L}_{ot}(\mu, \nu) &= \min_{M \in \Pi(u, v)} \sum_{i=1}^{n} \sum_{j=1}^{m} M_{ij} \cdot c(x_i, y_j) \\
&= \min_{M \in \Pi(u, v)} \langle M, C \rangle
\end{align*}
\]

(6)

where \( \Pi(u, v) = \{ M \in \mathbb{R}^{n \times m} | M 1_m = u, M^T 1_n = v \} \), \( 1_n \) denotes an \( n \)-dimensional all-one vector, \( C \) is cost matrix defined as \( C_{ij} = c(x_i, y_j) \), \( M \) is the transportation plan, and \( \langle M, C \rangle = \text{Tr}(M^T C) \) denotes the Frobenius inner product.

Xie et al. (2019) proposed an approximate algorithm IPOT to solve Eq. 6, which is illustrated in Appendix A. After solving \( M \), we use OT distance as loss to update model parameters.

Specifically, the representation of the template for student is denoted as \( H^S_e \), which is obtained from the last hidden state of the BART encoder of student corresponding to the template. We first partition \( H^S_e \) into \( n \) groups, where each group corresponds to the representation of the slot together with specific tokens before and behind it (e.g., \(<\text{type}>\text{TypeOfCause}\;<\text{/type}>\) belong to a group). Then we average the representations of tokens in each group to obtain the sequence \( K^S_e = \{ h^S_{e_i} \}_{i=1}^{n} \), where \( h^S_{e_i} \) denotes the representation of the \( i \)-th group, \( n \) is the number of groups. Similarly, we can get the sequence \( K^T_j = \{ h^T_{e_i} \}_{i=1}^{m} \) by group and average the template representations of teacher.

\[
\begin{array}{c|c|c|c}
\text{Split} & \#\text{Sents} & \#\text{Pairs} & \#\text{Events} \\
\hline
\text{Train} & 5600 & 6318 & 12636 \\
\text{Dev} & 700 & 791 & 1582 \\
\text{Test} & 700 & 799 & 1598 \\
\end{array}
\]

Table 1: Statistics of the ECE-CCKS dataset.

We use cosine distance as the cost function and adopt the IPOT algorithm to compute the OT loss:

\[
\mathcal{L}_{dot} = \text{IPOT}(K^S_e, K^T_j)
\]

(7)

Meanwhile, the last hidden state of the BART decoder for teacher and student is denoted as \( H^T_d \) and \( H^S_d \). We first use a linear layer followed by an argmax function to decode the output sequence. Then the representations are divided into several groups based on the event type or event argument together with specific tokens before and behind it.

Note that when some specific tokens are missing in the output, we remove the corresponding groups. Likewise, we can obtain two sequences \( K^S_j \) and \( K^T_j \) via an average operation and compute the OT loss:

\[
\mathcal{L}_{dot} = \text{IPOT}(K^S_d, K^T_d)
\]

(8)

The overall loss of the model is defined as:

\[
\mathcal{L}_o = \mathcal{L}_{kd} + \lambda \mathcal{L}_{dot} + \eta \mathcal{L}_{dot}
\]

(9)

where \( \lambda \) and \( \eta \) are weight coefficients.

3.5 Training Objectives

Since the two teachers exhibit different difficulties in extracting the same sample, motivated by Zhang et al. (2022), we use adaptive weights to control the importance of different teachers for a specific sample:

\[
w_k = 1 - \frac{\exp(L^k_{gen})}{\sum_{j} \exp(L_{gen}^j)}
\]

(10)

where \( L_{gen}^k \) denotes the prediction loss of the \( k \)-th teacher. The total loss under the multi-teacher knowledge distillation framework is calculated as:

\[
\mathcal{L} = \sum_k w_k \mathcal{L}_o^k
\]

(11)

4 Experiments

4.1 Experiment Setup

Dataset. Our experiments are conducted on the ECE-CCKS (Cui et al., 2022) dataset, which is
Table 2: Overall performance compared to the state-of-the-art methods on the test set. P, R, and F1 denote precision (%), recall (%), and F1-score (%). The best results are denoted in bold.

<table>
<thead>
<tr>
<th>Method</th>
<th>EAE P</th>
<th>EAE R</th>
<th>EAE F1</th>
<th>CET P</th>
<th>CET R</th>
<th>CET F1</th>
<th>ECE P</th>
<th>ECE R</th>
<th>ECE F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novel-tagging</td>
<td>59.40</td>
<td>28.47</td>
<td>38.49</td>
<td>49.79</td>
<td>61.70</td>
<td>55.11</td>
<td>51.52</td>
<td>26.75</td>
<td>35.22</td>
</tr>
<tr>
<td>CasECE</td>
<td>36.88</td>
<td>36.72</td>
<td>36.80</td>
<td>58.26</td>
<td>59.70</td>
<td>58.97</td>
<td>31.30</td>
<td>41.81</td>
<td>35.80</td>
</tr>
<tr>
<td>Pair-tagging</td>
<td>47.08</td>
<td>46.49</td>
<td>46.79</td>
<td>55.78</td>
<td>62.95</td>
<td>59.14</td>
<td>39.24</td>
<td>47.69</td>
<td>43.05</td>
</tr>
<tr>
<td>DualCor</td>
<td>58.05</td>
<td>47.60</td>
<td>52.31</td>
<td>61.75</td>
<td>58.19</td>
<td>59.92</td>
<td>48.56</td>
<td>44.85</td>
<td>46.63</td>
</tr>
<tr>
<td>BART-ECE</td>
<td>64.62</td>
<td>50.09</td>
<td>56.43</td>
<td>69.43</td>
<td>60.83</td>
<td>64.84</td>
<td>54.74</td>
<td>47.21</td>
<td>50.70</td>
</tr>
<tr>
<td>Student</td>
<td>67.81</td>
<td>54.37</td>
<td>60.35</td>
<td>67.57</td>
<td>59.20</td>
<td>63.11</td>
<td>55.47</td>
<td>49.76</td>
<td>52.46</td>
</tr>
<tr>
<td>MKD</td>
<td>68.83</td>
<td>53.59</td>
<td>60.26</td>
<td>68.29</td>
<td>59.82</td>
<td>63.78</td>
<td>57.02</td>
<td>49.04</td>
<td>52.73</td>
</tr>
<tr>
<td>CE-Pipeline</td>
<td>66.50</td>
<td>52.69</td>
<td>58.80</td>
<td>69.00</td>
<td>60.45</td>
<td>64.44</td>
<td>55.01</td>
<td>48.36</td>
<td>51.47</td>
</tr>
<tr>
<td>TA-Pipeline</td>
<td>67.01</td>
<td>53.55</td>
<td>59.53</td>
<td>68.86</td>
<td>60.33</td>
<td>64.31</td>
<td>54.27</td>
<td>47.98</td>
<td>50.93</td>
</tr>
<tr>
<td>ICE (Ours)</td>
<td><strong>69.69</strong></td>
<td><strong>54.57</strong></td>
<td><strong>61.21</strong></td>
<td><strong>70.00</strong></td>
<td><strong>61.33</strong></td>
<td><strong>65.38</strong></td>
<td><strong>59.13</strong></td>
<td><strong>51.44</strong></td>
<td><strong>55.02</strong></td>
</tr>
</tbody>
</table>

4.2 Baseline Methods

Classification-based Method. (1) Novel-tagging is introduced to ECE by combining causalities, event types, event roles, and argument span into a unified label space (Zheng et al., 2017). (2) CasECE is a pipelined method inspired by Wei et al. (2020), which first extracts the cause event and then recognizes the effect event conditioned on the former prediction. (3) Pair-linking is a grid tagging method based on Wang et al. (2020), which uses event-type-level pair linking as conditional information for token-pair linking to extract event arguments. (4) DualCor (Cui et al., 2022) designed a dual grid tagging scheme to capture the argument correlations for ECE.

Generation-based Method. Since there is no generative method for ECE in existing studies, we implement the following baselines: (1) BART-ECE achieves event causality extraction with a template-based conditional generation method in natural language form. (2) Student is the generative model introduced in Section 3.2. (3) MKD employs a multi-teacher knowledge distillation framework where each teacher is trained without privileged information. (4) CE-Pipeline is a pipelined generation model, which first extracts the cause event and then predicts the effect event based on the cause. (5) TA-Pipeline is also a pipelined generation model that first identifies the event type of the cause and effect event, conditioned on which to detect the event arguments.

4.3 Overall Performance

The experimental results are reported in Table 2. We can draw the following conclusions: (1) The proposed method achieves the best performance, outperforming the previous state-of-the-art model, DualCor, by 8.90%, 5.46%, and 8.39% on EAE, CET, and ECE in terms of F1-score. The significant improvements demonstrate the effectiveness of our
(2) The generation-based methods generally produce better performance than classification-based methods. This suggests that the generation-based model could make full use of the knowledge in PLMs and exhibit strong advantages on ECE. (3) Compared with Student, our method performs better on the three tasks. We credit the reason to that the student model has difficulty in extracting implicit causal clues, while ICE can equip the model with more implicit event causality reasoning knowledge via a multi-teacher knowledge distillation framework, thus boosting the model performance. (4) With the same model architecture, the performance of ICE exceeds MKD. The improvements indicate that the privileged information could help to train well-informed teachers, which guide the student to capture the intra- and inter-event interactions. (5) CE-Pipeline and TA-Pipeline perform poorly among the generation-based methods. The results illustrate that the pipelined methods suffer from error accumulation problems, while ICE is an end-to-end method with privileged information as the supervision, which could avoid introducing noise or irrelevant information.

### 4.4 Further Discussion

#### Ablation Study

To evaluate the contribution of each component, we conduct ablation studies by removing event argument extractor (w/o EAE), effect event extractor (w/o EEE), Cause-Effect Optimal Transport (w/o CEOT), and adaptive weights (w/o WA). The experimental results are shown in Table 3. We observe that: (1) Removing the event argument extractor or effect event extractor will result in performance decay, demonstrating that both extractors are beneficial for event causality extraction. This is because they could capture the implicit intra- and inter-event interactions to perform event causality reasoning, and the knowledge is transferred to the student via a teacher-student distillation framework. (2) The model performance decreases after removing CEOT, which illustrates this strategy could promote the training process by enhancing the event-level semantic information interactions of cause event and effect event. (3) The adaptive weights mechanism can assign different reliability for each teacher according to the characteristics of the samples, further improving the performance of the student.

### Effect of the Number of Cause-Effect Pairs

To evaluate the effect of different numbers of event causality pairs in a sample, we conduct experiments by dividing the test set into two subsets: **Single**, which indicates the subset with one pair in a sample; **Multiple**, which denotes the subset with more than one pairs. The results are reported in Table 4. It can be observed that: (1) ICE achieves the best performance among all the baselines on Single and Multiple subsets, which shows the generalization ability and robustness of ICE. (2) The methods on Multiple generally suffer from weak performance. The reason may be that multiple event causality pairs impose difficulties in capturing implicit causal clues, leading to extremely low recall. (3) The performance gaps between Student and ICE on the two subsets further demonstrate that the knowledge distillation framework could utilize the privileged information to improve the
The growth of \{global\} RegionOfCause \{gasoline\} ProductOfCause demand is sluggish, and the combination of excess supply and unsatisfactory demand has led to a continuous decline in \{gasoline\} ProductOfEffect cracking spreads this year.

**Example1**

BART-ECE: \{TypeOfCause: Supply Increase, TypeOfEffect: Demand Drop\}
ICE: \{TypeOfCause: Demand Drop, TypeOfEffect: Price Drop\}

The amount of \{silicon\} ProductOfCause used in June increased compared to May, and the increase came from polysilicon, while the supply of \{monosilicon\} ProductOfCause decreased.

**Example2**

Therefore, it is expected that the price of \{monosilicon\} ProductOfEffect will rise after July.

BART-ECE: \{TypeOfCause: Supply Decrease, TypeOfEffect: Price Increase\}
ICE: \{TypeOfCause: Supply Decrease, TypeOfEffect: Price Increase\}

**Table 6:** Case study on the test set. The correctly predicted event type or event arguments are marked in teal, while wrong predictions by BART-ECE are marked in red.

**Figure 3:** Experimental results (F1-score) under different ratios of training data on three tasks.

**Performance in Low-resource Scenarios.** To investigate the model performance in low-resource scenarios, we adopt different low proportions of training data to conduct experiments on the three tasks. As shown in Figure 3, it can be observed that: (1) ICE achieves the best performance under different ratios of training data on three tasks. This suggests that our model can extract event causalities and event structure effectively with a small scale of annotated data, which is more practical to use. (2) ICE using a small amount of data can even exceed DualCor using full data, and the performance gap becomes larger with the decrease of training data. These observations indicate that ICE can elicit knowledge in PLMs and has a strong ability to capture causal clues, which is beneficial for the model to perform the ECE task.

**Effect of the Prompt Template.** We study how different types of prompt templates affect the model performance by conducting experiments with the following templates: Concatenation Template, Manual Template, and Soft Template. The creation of the three templates is shown in Appendix C and the results are listed in Table 5. We find that the soft template is superior to the manual template and concatenation template, which illustrates the effectiveness of our template. Although the manual template can elicit pre-training knowledge in a cloze formulation, it is labor-intensive and hard to achieve optimal. However, the soft template can avoid this laborious process and take the best advantage of the PLMs.

**Case Study.** We present case studies to further illustrate the performance of the proposed method. As shown in Table 6, for Example1, BART-GEN gives wrong predictions about the event type of the cause event and effect event. The reason may be that BART-ECE has difficulty capturing implicit cause-effect interactions, so it fails to recognize the causality between Demand Drop and Price Drop. For Example2, both methods produce the correct cause-effect event type, while BART-GEN fails to predict the event argument monosilicon of the cause event. We credit the reason to that our model could leverage the event argument extractor and effect event extractor trained with privileged information to guide the training process of the student, thus obtaining better performance in extracting event causality and event structure.
5 Conclusions

In this paper, we propose an Implicit Cause-Effect interaction (ICE) framework to improve the reasoning ability of the model, which tackles ECE in a generative manner. The proposed method incorporates privileged information for reasoning to capture implicit intra- and inter-event interactions, and utilizes a teacher-student learning framework to bridge the gap between training and test stages. Besides, we introduce a Cause-Effect Optimal Transport (CEOT) strategy to improve the event-level semantic interactions of cause and effect events. Experimental results indicate that ICE outperforms all the baselines on the ECE-CCKS dataset, demonstrating the effectiveness of this work.

Limitations

The multi-teacher knowledge distillation mechanism utilized in the ICE framework may increase the computational time during the training process. However, only the student model is leveraged during the test process, and the test time is identical to regular generation-based models. The problem of relatively long training time can be mitigated by strategies such as GPU parallelization. Considering the significant improvement brought by the ICE framework, we believe the cost is acceptable.

Acknowledgements

The research is supported by the National Natural Science Foundation of China under Grant 62206267.

References


Chen Xu, Quan Li, Junfeng Ge, Jinyang Gao, Xiaoyong Yang, Changhua Pei, Fei Sun, Jian Wu, Hanxiao Sun, and Wenwu Ou. 2020. Privileged features distillation at taobao recommendations. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 2590–2598.


A The details of the IPOT algorithm

The details of the IPOT algorithm is shown in Algorithm 1.

Algorithm 1 IPOT algorithm

Input: Feature vectors $S = \{h_i\}^n_{i=1}, S' = \{h'_j\}^m_{j=1}$, Generalized stepsize $\beta$

Output: $\langle M, C \rangle$

1: $\sigma = \frac{1}{m^1}, M^{(1)} = 1_{n^1}^\top$
2: $C_{ij} = c(h_i, h'_j), A_{ij} = e^{-\frac{C_{ij}}{\beta}}$
3: for $t = 1, 2, 3, \cdots$ do
4: $Q = A \odot M^{(t)}$ // $\odot$ is Hadamart product
5: for $k = 1, 2, \cdots, K$ do // $K=1$ in practice
6: $\delta = \frac{1}{m^Q\sigma}, \sigma = \frac{1}{m^Q\delta}$
7: end for
8: $M^{(t+1)} = \text{diag}(\delta) \ Q \ \text{diag}(\sigma)$
9: end for
10: return $\langle M, C \rangle$

B An example of the construction of knowledge-enriched templates

We show an example of the construction of knowledge-enriched templates for EAE and EEE in Table 7. The knowledge-enriched template for EAE is constructed by filling the ground truth event types of the cause and effect event in the raw template; The knowledge-enriched template for EEE is constructed by filling the ground truth event type and arguments of the cause event in the raw template.

C The creation of different types of templates

We construct three types of templates for comparison: (1) Concatenation Template, where all slots for event type and event arguments are concatenated; (2) Manual Template, where event types and event arguments are integrated into templates in natural language form; (3) Soft Template, which is used in our method. And detailed information of the templates is shown in Table 8.

D Compare with ChatGPT

In this section, we conduct experiments to evaluate the performance of ChatGPT on the ECE task. A well-designed prompt template is as follows:

Suppose you are now an event causality extraction model. Given a sentence, please give the cause event and result event respectively, where the event contains the event type and the arguments corresponding to each role. The list of event types is: ['Typhoon', 'Demand Increase', 'Price Decrease', 'Cold Wave', 'Price Increase', 'Other Natural Disasters', 'Supply Decrease', 'Supply Increase', 'Sales Decrease', 'Demand Drop', 'Import Decrease', 'Flood', 'Other Trade Frictions', 'Negative Impact', 'Swine Fever', 'Sales Increase', 'Limited Production', 'Operating Costs Increased', 'Other Livestock Epidemics', 'Positive Impact', 'Drought', 'Operating Cost Decrease', 'Export Decrease', 'Frost', 'Other or Unclear', 'Import Increase', 'Bird Flu', 'Earthquake', 'Anti-dumping Against China', 'Exports Increase', 'Add Tariffs to China', 'Decrease in Product Profits', 'Increase in Product Profits', 'Foot-and-mouth Disease of Pigs', 'Anti-dumping Against Other Countries', 'Unsalable', 'Cattle Foot and Mouth Disease', 'Flash Flood', 'Hail']. The list of event argument roles is: ['Region', 'Industry', 'Product']. Given a sentence: "The worldwide rise of oil prices stimulates the demand for new energy such as Ammonia fuel.", please extract the event type and arguments corresponding to cause and effect event. If no argument corresponds to a role, the argument content returns "None". If multiple arguments corresponds to a role, the arguments are connected with "and".

As shown in Table 9, we can observe that ICE outperforms ChatGPT by a large margin on EAE, CET, and ECE tasks. The results indicate that ChatGPT has difficulty solving such complex event causality extraction tasks without any fine-tuning or training to update parameters.
<table>
<thead>
<tr>
<th>Temp.</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw Temp.</td>
<td>&lt;Cause&gt; &lt;type&gt; TypeOfCause &lt;/type&gt; &lt;region&gt; RegionOfCause &lt;/region&gt; &lt;industry&gt; IndustryOfCause &lt;/industry&gt; &lt;product&gt; ProductOfCause &lt;/product&gt; &lt;Effect&gt; &lt;type&gt; TypeOfEffect &lt;/type&gt; &lt;region&gt; RegionOfEffect &lt;/region&gt; &lt;industry&gt; IndustryOfEffect &lt;/industry&gt; &lt;product&gt; ProductOfEffect &lt;/product&gt;.</td>
</tr>
</tbody>
</table>

Table 7: An example of the construction of knowledge-enriched templates for EAE and EEE.

<table>
<thead>
<tr>
<th>Temp.</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>CA</td>
<td>TypeOfCause RegionOfCause IndustryOfCause ProductOfCause TypeOfEffect RegionOfEffect IndustryOfEffect ProductOfEffect.</td>
</tr>
<tr>
<td>MA</td>
<td>The cause TypeOfCause, the region RegionOfCause, the industry IndustryOfCause, the product ProductOfCause, leads to the effect TypeOfEffect, the region RegionOfEffect, the industry IndustryOfEffect, the product ProductOfEffect.</td>
</tr>
<tr>
<td>SF</td>
<td>&lt;Cause&gt; &lt;type&gt; TypeOfCause &lt;/type&gt; &lt;region&gt; RegionOfCause &lt;/region&gt; &lt;industry&gt; IndustryOfCause &lt;/industry&gt; &lt;product&gt; ProductOfCause &lt;/product&gt; &lt;Effect&gt; &lt;type&gt; TypeOfEffect &lt;/type&gt; &lt;region&gt; RegionOfEffect &lt;/region&gt; &lt;industry&gt; IndustryOfEffect &lt;/industry&gt; &lt;product&gt; ProductOfEffect &lt;/product&gt;.</td>
</tr>
</tbody>
</table>

Table 8: The creation of different types of templates. CA: Concatenation Template, MA: Manual Template, SF: Soft Template.

<table>
<thead>
<tr>
<th>Method</th>
<th>EAE</th>
<th>CET</th>
<th>ECE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
</tr>
<tr>
<td>ChatGPT</td>
<td>13.16</td>
<td>16.26</td>
<td>14.54</td>
</tr>
<tr>
<td>ICE</td>
<td><strong>69.69</strong></td>
<td><strong>54.57</strong></td>
<td><strong>61.21</strong></td>
</tr>
</tbody>
</table>

Table 9: Overall performance compared to ChatGPT on the test set. P, R, and F1 denote precision (%), recall (%), and F1-score (%). The best results are denoted in bold.