Mastering Context-to-Label Representation Transformation for Event Causality Identification with Diffusion Models

Hieu Man[1], **Franck Dernoncourt**[2], **Thien Huu Nguyen**[1,3]
[1] Department of Computer Science, University of Oregon, USA
[2] Adobe Research, USA
[3] VinAI Research, Vietnam
{hieum,thienn}@uoregon.edu, franck.dernoncourt@adobe.com

Abstract

To understand event structures of documents, event causality identification (ECI) emerges as a crucial task, aiming to discern causal relationships among event mentions. The latest approach for ECI has introduced advanced deep learning models where transformer-based encoding models, complemented by enriching components, are typically leveraged to learn effective event context representations for causality prediction. As such, an important step for ECI models is to transform the event context representations into causal label representations to perform logits score computation for training and inference purposes. Within this framework, event context representations might encapsulate numerous complicated and noisy structures due to the potential long context between the input events while causal label representations are intended to capture pure information about the causal relations to facilitate score estimation. Nonetheless, a notable drawback of existing ECI models stems from their reliance on simple feed-forward networks to handle the complex context-to-label representation transformation process, which might require drastic changes in the representations to hinder the learning process. To overcome this issue, our work introduces a novel method for ECI where, instead abrupt transformations, event context representations are gradually updated to achieve effective label representations. This process will be done incrementally to allow filtering of irrelevant structures at varying levels of granularity for causal relations. To realize this, we present a diffusion model to learn gradual representation transition processes between context and causal labels. It operates through a forward pass for causal label representation noising and a reverse pass for reconstructing label representations from random noise. Our experiments on different datasets across multiple languages demonstrate the advantages of the diffusion model with state-of-the-art performance for ECI.

Introduction

Event Causality Identification (ECI) is an active research problem in Information Extraction of Natural Language Processing (NLP) whose goal entails predicting causal relations between event mentions in text. For example, in the sentence "*The hurricane caused significant property damage to his house.*", an ECI system should identify the causal relation between the events "*hurricane*" and "*damage*, i.e., "*hurricane*" $\xrightarrow{\text{cause}}$ "*damage*". Serving as an important step for event structure understanding of text, ECI models can introduce useful information for different NLP applications such as question answering (Oh et al. 2016), machine reading comprehension (Berant et al. 2014), and event forecasting (Hashimoto 2019).

While the early work for ECI has applied feature-based methods (Do, Chan, and Roth 2011; Ning et al. 2018), recent work have employed various deep learning architectures to realize state-of-the-art performance for this problem (Kadowaki et al. 2019; Liu, Chen, and Zhao 2020; Zuo et al. 2021b). As such, the first step in the current deep learning models for ECI often involves a transformer-based encoding network, such as BERT (Devlin et al. 2019) and RoBERTa (Liu et al. 2019), that help induce initial contextualized representations of event mentions in the input texts. Afterward, the ECI models tend to explore different components to further enrich the event context representations from the transformer-based networks, resulting in the final event representations for causality prediction. For instance, (Tran Phu and Nguyen 2021) leverages graph convolutional networks with rich text structures to produce the enriched event representations for ECI. Another example includes (Liu, Chen, and Zhao 2020) that augments input texts with relevant background knowledge retrieved from external sources to enhance event context representations. Consequently, the final step of contemporary ECI models is to transform the event context representations into causal label representations that will be used to compute logits scores for possible causal labels for training and inference. Ideally, the causal label representations are expected to capture pure information for the causal labels to facilitate label score computation and prediction.

As such, the dominant approach in ECI models is to employ a straightforward feed-forward network, characterized by its simplicity with zero, one, or a few layers, to directly convert the event context representations into the causal label representations for causality prediction (Gao, Choubey, and Huang 2019; Tran Phu and Nguyen 2021; Chen et al. 2022). While this approach can be convenient for implementation, a critical issue arises. The computed event context representations of ECI models might still involve complicated and noisy information. Attempting to perform context-to-label representation transformation directly with feed-forward networks can demand a drastic change, posing sig-

nificant challenges for the learning process. Hence, the resulting causal label representations will likely preserve noisy information to hinder causality prediction performance. This problem becomes particularly noticeable for ECI, where the challenge lies in predicting causal relationships for pairs of event mentions that are far part in the input text (i.e., long input context). Such extensive context can encompass a multitude of irrelevant structures that are not helpful for the causal connection between the two event mentions. Coupled with the high capabilities of transformer-based encoding models, these extraneous structures are prone to persist in the event context representations induced by ECI models. Filtering out these irrelevant structures to uncover clear causal label information is not a straightforward task for the feed-forward networks. For instance, in document-level ECI, when the two input event mentions are in separate sentences of a document, numerous irrelevant context words can be presented between two input event mentions to introduce significant noise for event context representations for ECI.

To address this limitation in previous ECI models, we argue that the context-to-label representation transformation for ECI should be carried out more gradually. This involves applying incremental adjustments to the event context representations, systematically eliminating irrelevant components at various levels of detail to unveil more effective causal label representations. This intuition motivates us to develop a diffusion model to generate causal label representations from event context representations that facilitates progressive fine-grained refinements of the representations via a series of denoising steps. In particular, a denoising diffusion probabilistic model (DDPM) (Ho, Jain, and Abbeel 2020) aims to learn a procedure to generate representations or samples from a distribution, involving two major processes. In the forward process, provided samples from the distribution are incrementally added with Gaussian noise to transform them into a Gaussian distribution (i.e., the diffusion process). In contrast, the reverse process attempts to learn a reserve procedure that can generate samples of the distribution from random noise via a series of adjustment steps. At each step in the reverse process, the model maintains the current intermediate data point and predicts an adjustment to be applied to the transition to the next step.

To adapt the diffusion model for the ECI problem, our forward process also involves adding Gaussian noise sequentially to the causal label representations to reach Gaussian distribution. As such, the causal label representations will be computed directly from the provided causal labels during the training step to achieve clean information. For the reverse process, to generate a causal label representation for an input event mention pair from a random noise, the prediction of adjustment at each denoising step for transition will be conditioned on not only the current representation, but also the event context representation from the encoder model for the input. Due to the access to event context representation, we can train the diffusion model so the predicted adjustment at each step can realize and filter some irrelevant information from the event context representation for causality prediction. Based on the predicted adjustment for the current step, the representation in the next step can achieve higher quality (i.e., closer to the expected label representation). After a sequence of adjustments in the reverse process, the final representation is expected to possess clean causal label information to effectively predict causal labels. To our knowledge, this is the first diffusion model proposed for ECI in the literature.

Consequently, to accomplish this goal for the diffusion model, we propose to control the adjustment predictions by encouraging the intermediate representations produced at each step in the reverse process to be predictive of the causal label for the input during the course of training. The rationale is to ensure that the important information for causal labels is not eliminated from the representations in the reverse adjustment process, thus guiding the predicted adjustments to only remove irrelevant features for causality prediction. Accordingly, to implement this adjustment controlling idea, our method suggests using the intermediate representations in the reverse process to predict the golden causal label in the training process. Finally, we perform extensive evaluations for our method over several ECI benchmark datasets. The experiments demonstrate the advantages of our diffusion-based model for ECI, achieving state-of-the-art performance over different datasets and languages.

Model

There are two major components in our diffusion-based model for ECI: (i) Event Representation Learning to encode event context in input text, and (ii) Context-to-Label Representation Transformation with a diffusion model (Ho, Jain, and Abbeel 2020; Nichol and Dhariwal 2021). In the following, we will first present necessary background for the diffusion component in our model. Our overall ECI model with the two components will be introduced afterward.

Diffusion Background

A diffusion model, specifically a denoising diffusion probabilistic model (DDPM) (Ho, Jain, and Abbeel 2020), is characterized as a Markov chain that undergoes training through variational inference to generate samples from a data distribution after finite time. There are two processes in a diffusion model: the diffusion/forward process to perturb data samples from the distribution and the reverse process to generate data samples from random noise. Given a sample x_0 from the data distribution $q(x_0)$, the forward pass in a diffusion model involves a noising process to compute T latent variables $x_1, \ldots, x_T$ (of the same dimension as x_0) by adding diagonal Gaussian noise at each time step t ($1 \leq t \leq T$):

$$q(x_1, \ldots, x_T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1})$$

$$x_t \sim q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I})$$

Here, α_t is a hyper-parameter to control the noise added at step t and $\mathbf{I}$ is the identity matrix. As noted in (Ho, Jain, and Abbeel 2020), this setting allows us to sample x_t at an

arbitrary step x_t directly from x_0 using the marginal distribution:

$$x_t \sim q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I})$$

where $\bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s$. Further, by using the reparameterization trick (Ho, Jain, and Abbeel 2020), we can obtain x_t via:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon_t, \text{with } \epsilon_t \sim \mathcal{N}(0, \mathbf{I})$$

Under this framework, it has been shown that x_T is nearly an isotropic Gaussian distribution if T is large enough and α_t is scheduled well over t. As such, if we can compute the reverse distribution $q(x_{t-1}|x_t)$, we can generate a sample x_0 from $q(x_0)$ by first sampling $x_T \sim \mathcal{N}(0, \mathbf{I})$ and then performing a series of denoising steps by incrementally sampling from $x_{t-1} \sim q(x_{t-1}|x_t)$ (i.e., the reverse process). However, as $q(x_{t-1}|x_t)$ relies on the entire data distribution, it is approximated via a Gaussian distribution $p_\theta(x_{t-1}|x_t)$ with parameters θ to estimate the mean μ_θ and variance Σ_θ:

$$x_{t-1} \sim p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

In this work, we utilize a fixed variance $\Sigma_\theta(x_t, t) = (1-\alpha_t)\mathbf{I}$, which has been shown to achieve the best results in (Ho, Jain, and Abbeel 2020). For the mean $\mu_\theta(x_t, t)$, we can also re-parameterize it via the noise prediction network $\epsilon_\theta(x_t)$:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right) \quad (1)$$

To this end, we can sample $x_{t-1} \sim p_\theta(x_{t-1}|x_t)$ for the reverse process using $\epsilon_\theta(x_t, t)$ and $z \sim \mathcal{N}(0, \mathbf{I})$:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right) + z\sqrt{1-\alpha_t}$$

To train the reverse process with the noise prediction network ϵ_θ, we can consider the combination of q and p as a variational auto-encoder (Kingma and Welling 2013), and the variational lower bound can be used for optimization. However, with the re-parameterization trick, (Ho, Jain, and Abbeel 2020) suggests that training can be achieved effectively by minimizing a simple mean squared error between the predicted noise $\epsilon_\theta(x_t, t)$ and the ground-truth sampled Gaussian noise ϵ_t: $\mathcal{L}_{diff}(\theta) = ||\epsilon_\theta(x_t, t) - \epsilon_t||^2$.

Event Representation Learning

Following previous work (Liu, Chen, and Zhao 2020; Tran Phu and Nguyen 2021), we formulate ECI as a binary classification task, aiming to predict whether a causal relation exists between two input event mentions. Specifically, let e_1 and e_2 be the two event mentions/trigger words in the input text D where e_1 and e_2 can appear in the same sentence (i.e., sentence-level ECI) or different sentences (i.e., document-level ECI). To facilitate the text encoding with pre-trained language models (PLMs), for each event mention e_i ($i \in \{1, 2\}$), we select a window of five sentences, consisting of the hosting sentence of e_i in D along with two previous and following sentences to create a context C_i for e_i. Here, we do not select a sentence twice in C_1 and C_2. The contexts C_1 and C_2 are then concatenated to form a single context C for the input event mentions. Note that the order of the sentences in the input text D will be preserved in C. In the next step, we utilize the pre-trained RoBERTa model (Liu et al. 2019) to encode the context C. To represent the input event mentions e_1 and e_2, we obtain the hidden vectors r_1 and r_2 (respectively) for their first sub-tokens in the last layer of RoBERTa (called event context representations). Here, for convenience, we use d to denote the dimensionality of the hidden vectors in the last layer of RoBERTa, i.e., $|r_1| = |r_2| = d$.

Context-to-Label Representation Transformation

Given the event context representations r_1 and r_2, a typical approach to perform ECI is to transform these representations into causal label representations using a simple feed-forward network, seeking to capture clean information about the causal labels to compute the logits scores and probability distribution for possible labels. However, a crucial challenge associated with this approach pertains to the potential abundance of irrelevant context words or information within the context C for the input event mentions e_1 and e_2. Since the event context representations r_1 and r_2 are computed using self-attention from the encoder over this context, they might encompass numerous noisy structures that cannot be easily filtered out to produce clean causal label representations for ECI. In light of this, the utilization of feed-forward networks to convert context representations into label representations would entail a drastic change, which may not be effectively learned, thereby retaining noisy information within the causal label representations for ECI. This problem is even more significant for input event mentions that are widely separated in the context as it might introduce more irrelevant words. To address this problem, we introduce a diffusion model to transform event context representations to causal label representations for ECI. The key advantage of the diffusion model is the ability to decompose the context-to-label transformation process into multiple steps to achieve more gradual representation changes for better learning. At each step, the representation transition will be controlled to filter irrelevant features for causality prediction to produce effective label representations for ECI.

Diffusion Model: To train the diffusion model, we first obtain the causal label representation that we expect the model to generate for the input context C and event mentions e_1 and e_2 (i.e., the sample x_0 of the distribution $q(x_0)$ in the diffusion framework). To this end, we create a simple sentence S_{lbl} to encapsulates the causal relation l between the input events e_1 and e_2 in the following format:

$$S_{lbl} = \texttt{[CLS]}\ \textit{There is}\ l\ \textit{between}\ e_1\ \textit{and}\ e_2$$

, where l can be "*a causal relation*" or "*no causal relation*" to indicate the gold relation label between e_1 and e_2 in the input. Next, S_{lbl} is also encoded by the RoBERTa model, and the last-layer hidden vectors $v_1, v_2, \ldots, v_m$ of the first sub-tokens of the $m = 9$ words in S_{lbl} are stacked into a matrix $V = [v_1; v_2; \ldots; v_m]$ of size $m \times d$, serving as our causal

label representation for the input. Note that by using only the representations for the first sub-tokens of the words, we can ensure the same dimensions for V across different input examples. Furthermore, by including the event mentions e_1 and e_2 within our sentence S_{lbl} for label representation, we release the diffusion model from the task of discarding information pertaining to the input event mentions from the context representation. This simplifies the learning process to boost the performance while still preventing the introduction or irrelevant information into S_{lbl} to provide cleaner causal label representation.

Treating V as a sample from the causal label representation distribution (i.e., $x_0 = V_0 = V$), the forward pass of our diffusion model computes T representations V_t (of size $m \times d$) by incrementally adding adjustments ϵ_t in the form of Gaussian noise to V_0:

$$V_t = \sqrt{\bar{\alpha}_t} V_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon_t \text{ with } \epsilon_t \sim \mathcal{N}(0, \mathbf{I})$$

Subsequently, to train the reverse process that generates V from random noise, the original formulation of the diffusion model will require estimating the distribution $p_\theta(V_{t-1}|V_t)$ for reverse sampling. However, as our model's objective is to infer the causal label representation V from the input event context, the reverse process in our diffusion model instead learns the distribution $p_\theta(V_{t-1}|V_t, r_1, r_2)$. This distribution conditions not only on the previous representation V_t but also on the context representations r_1 and r_2 of the event mentions, thus facilitating the removal of the irrelevant features from the context representations for our ECI problem. To this end, following (Ho, Jain, and Abbeel 2020), we compute $p_\theta(V_{t-1}|V_t, r_1, r_2)$ via the neural network $\mu_\theta(V_t, r_1, r_2, t)$ for the mean:

$$p_\theta(V_{t-1}|V_t, r_1, r_2) = \mathcal{N}(V_{t-1}; \mu_\theta(V_t, r_1, r_2, t), (1 - \alpha_t)\mathbf{I})$$

Using the reparameterization trick for μ_θ as in Equation 1, we can instead leverage the adjustment prediction network $\epsilon_\theta(V_t, r_1, r_2, t)$ to obtain $\mu_\theta(V_t, r_1, r_2, t)$ for reserve sampling. Consequently, we can employ the mean square error between ϵ_θ and ϵ_t to train our diffusion model to generate the causal label representation V for ECI:

$$\mathcal{L}_{diff} = ||\epsilon_\theta(V_t, r_1, r_2, t) - \epsilon_t||^2$$

In the test time, by starting from $V_T \sim \mathcal{N}(0, \mathbf{I})$ and then following the reverse process to sample from $p_\theta(V_{t-1}|V_t, r_1, r_2)$ L times, we can generate the representation $V = V_{T-L}$ and use it as the causal label representation for our ECI model.

Adjustment Prediction Network: We employ a transformer encoder network with K layers for the adjustment prediction model $\epsilon_\theta(V_t, r_1, r_2, t)$, aiming to compute a more label-oriented representation with less irrelevant context features for V_t (K is a hyper-parameter). To form the input for ϵ_θ, we consider V_t as a sequence of m vectors of d dimensions. The time step t is also transformed into an embedding vector $\bar{t}$ using the sinusoidal embedding in (Ho, Jain, and Abbeel 2020). Here, we ensure that $\bar{t}$ also has d dimensions by feeding the sinusoidal embedding into a learnable two-layer feed-forward network. Afterward, the representation vectors r_1, r_2 and $\bar{t}$ will be prepended to V_t to create the input vector sequence for the transformer network for ϵ_θ. Finally, based on the resulting sequence of hidden vectors in the last layer of the transformer network, we retain the last m vectors as the output for $\epsilon_\theta(V_t, r_1, r_2, t)$. These m vectors correspond to the vectors from V_t in the input and can be used as the output V_{t-1} for the next sampling step. Due to the self-attention in the transformer network, the intermediate representation V_t can interact with the event context representations r_1 and r_2 to discard irrelevant features.

Training and Inference

We jointly train our ECI and diffusion models in this work. For the ECI model, we concatenate the event context representations r_1 and r_2 and the first vector $V[0]$ of the causal label representation V to form the feature vector to predict the causal relation between e_1 and e_2. Here, $V[0]$ corresponds to the representation of the `[CLS]` token in S_{lbl}. In particular, the concatenation is sent into a two-layer feed-forward network with softmax in the end FF_{ECI} to compute a distribution over two possible outcomes for the binary causality prediction: $P_{ECI}(\cdot|e_1, e_2, C) = FF_{ECI}([r_1, r_2, V[0]])$. The negative log-likelihood function is employed to train the model:

$$\mathcal{L}_{ECI} = -\log P_{ECI}(y|e_1, e_2, C)$$

where y is the golden label for the causality of e_1 and e_2.

In addition, to control the reverse process for gradual context-to-label representation transformation, we aim to encourage each adjustment step to incrementally remove irrelevant context features from the intermediate representations V_t to achieve more effective label representations. As such, we implicitly realize this goal by ensuring the maintenance of event causality information between e_1 and e_2 in the intermediate representations V_t along the way, thus forcing the adjustments to focus on irrelevant information for causal prediction. To accomplish this maintenance, we propose to further utilize the intermediate representations V_t to predict golden label y in the training process.

In particular, given the input event mentions e_1 and e_2, we first sample a time step t from the uniform distribution $\mathcal{U}(1, T)$. Afterward, we obtain the representation V_{t-1} from our reverse distribution $p_\theta(V_{t-1}|V_t, r_1, r_2)$. Consequently, we feed the concatenation of r_1, r_2, and $V_t[0]$ into FF_{ECI} to obtain a distribution of two possible causality labels: $P^t_{inter}(\cdot|e_1, e_2, D) = FF_{ECI}([r_1, r_2, V_t[0]])$. As such, we will optimize the negative log-likelihood function

$$\mathcal{L}^t_{inter} = -\log P^t_{inter}(y|e_1, e_2, D)$$

to preserve causal label information between e_1 and e_2 for the intermediate representations V_t.

Finally, the training objective for our ECI model is:

$$\mathcal{L} = \mathcal{L}_{ECI} + \frac{1}{t}\mathcal{L}^t_{inter} + \mathcal{L}_{diff}$$

Note that when t is large, the intermediate representation V_t might still involve numerous irrelevant features, and optimizing $\mathcal{L}^t_{inter}$ might cause the model to rely on such noisy information for prediction. We thus scale $\mathcal{L}^t_{inter}$ by $1/t$ in the overall loss to limit the influence for our loss in such cases.

Experiments

Evaluation Datasets

We assess our diffusion model for ECI, named **DiffusECI**, on two English benchmark datasets: EventStoryLine (ESL) (Caselli and Vossen 2017) and Causal-TimeBank (CTB) (Mirza 2014). These datasets have been widely used in previous ECI research (Gao, Choubey, and Huang 2019; Liu, Chen, and Zhao 2020; Tran Phu and Nguyen 2021). Specifically, ESL (version 0.9) comprises 258 annotated documents spanning 22 topics, containing 4316 sentences and 5334 event mentions. There are 7805 intra-sentence mention pairs and 46521 inter-sentence mention pairs, out of which 1770 and 3855 pairs (respectively) are positive examples with causal relations. We follow the same data split and setting in previous work for ESL (Liu, Chen, and Zhao 2020; Tran Phu and Nguyen 2021), which reserve the last two topics as development data and perform 5-fold cross-validation evaluation on the remaining 20 topics. On the other hand, the CTB dataset consists of 184 annotated documents, encompassing 6813 events. Within CTB, there are 7608 event mention pairs, out of which 318 are positive examples. Following previous work (Liu, Chen, and Zhao 2020; Zuo et al. 2021b), we use the same data split with 10-fold cross-validation for the evaluation on CTB.

Furthermore, we evaluate our model's performance on MECI (Lai et al. 2022), a recent dataset designed for multilingual ECI. MECI provides annotations for ECI in text across five different languages, namely English, Danish, Spanish, Turkish, and Urdu. The documents in MECI are sourced from Wikipedia, and the annotation schema follows that used in ESL. This dataset provides both intra-sentence and inter-sentence examples. To ensure a fair comparison, we adopt the same data split for training/dev/test data portions for each language, as established in (Lai et al. 2022), for the evaluation process.

Hyperparameters

Our model utilizes the base version of RoBERTa for the encoder, involving 12 transformer layers, 12 heads, and $d = 728$ for the hidden vector size. We use the development data of ESL to tune the hyperparameters for our DiffusECI model. The tuning process returns the following values: $K = 8$ transformer layers with 8 heads and 768 hidden dimensions for the adjustment prediction network ϵ_θ, 32 for the minibatch size, 768 dimensions for hidden vectors in the feed-forward networks, and $5e$-5 for the learning rate with the AdamW (Loshchilov and Hutter 2019) optimizer. For the diffusion model, we follow the same hyper-parameters in (Ho, Jain, and Abbeel 2020). In particular, the number of diffusion steps T is set to 1000 while $L = 100$ is used for the number of sampling steps in the reverse process. For the nosing hyper-parameters $\alpha_t = 1 - \beta_t$, we employ constants that are increased linearly from $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$. Finally, a reproducibility checklist is provided the Appendix.

Baselines

We compare our model DiffusECI with the state-of-the-art models for ECI in two groups of methods based on whether they utilize the transformer architecture or not. In particular, for the ESL and CTB datasets, we consider the following non-transformer baselines: (1) **LSTM** (Gao, Choubey, and Huang 2019); (2) **Seq** (Gao, Choubey, and Huang 2019) adopted from (Choubey and Huang 2017) for ECI; (3) **LR+** and **LIP** (Gao, Choubey, and Huang 2019): document structure models; and (4) **ML**: a feature-based model in (Mirza 2014).

For the transformer-based models, the following baselines are explored in our comparison: (1) **KnowBERT** (Liu, Chen, and Zhao 2020): a model integrating external commonsense knowledge and mention masking technique; (2) **KnowDis** (Zuo et al. 2020): a distant supervision-based model; (3) **CauSeRL** (Zuo et al. 2021a): a self-supervised method with external causal statements; (4) **LearnDA** (Zuo et al. 2021b): a learnable knowledge-guided data augmentation method with dual learning to generate task-related training data; (5) **RichGCN** (Tran Phu and Nguyen 2021): a graph convolutional network incorporating rich document-level structures; (6) **ERGO** (Chen et al. 2022): a relational graph transformer framework formulating ECI as a node classification problem; (7) **CF-ECI** (Mu and Li 2023): a counterfactual reasoning mdoel to explicitly estimate the influence of context keywords and event pairs for debiasing; (8) **CHEER** (Chen et al. 2023): a graph framework considering the centrality of events and their interactions in a document-level graph; (9) **SemSIn** (Hu et al. 2023): a graph model integrating event-centric and event-associated semantic structures; (10) **SENDIR** (Yuan et al. 2023): a document-level ECI framework using sparse attention and discriminative reasoning; (11) **KADE** (Wu et al. 2023): a model utilizing external knowledge and internal event analogy; (12) **GenECI** (Man, Nguyen, and Nguyen 2022): a generative model jointly generating causal relation and dependency path between input event mentions with reinforcement learning; and (13) **DPJL** (Shen et al. 2022): a derivative prompt-based method.

Comparison

Table 1 reports the performance of DiffusECI and the baseline models over the ESL and CTB datasets. For ESL, similar to previous work (Tran Phu and Nguyen 2021; Wu et al. 2023), we provide the results for different settings for ECI, i.e., intra-sentence, inter-sentence, and both intra- and inter-sentence causality predictions. The most important observation from the table is that the proposed model DiffusECI achieves significantly better performance than the state-of-the-art model DPJL for sentence-level ECI on ESL. DiffusECI also surpasses the best baseline model KADE for inter- and intra+inter-sentence ECI on ESL and intra-sentence ECI on CTB. For example, DiffusECI substantially outperforms KADE by 3% for intra+inter-sentence ECI on ESL and 8% for intra-sentence ECI on CTB. The performance improvements are significant with $p < 0.01$, which clearly shows the advantages of the proposed diffusion model for ECI. Importantly, DiffusECI can accomplish state-of-the-art performance for ECI over different settings without any additional annotation or third-party tools. This is different from recent work on ECI that requires additional

Model	ESL (Intra-sentence)			ESL (Inter-sentence)			ESL (Intra + Inter)			CTB (Intra)		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
LSTM (2019)	34.0	41.5	37.4	13.5	30.3	18.7	17.6	33.9	23.2	-	-	-
Seq (2019)	32.7	44.9	37.8	11.3	29.5	16.4	15.5	34.3	21.4	-	-	-
LR+ (2019)	37.0	45.2	40.7	25.2	48.1	33.1	27.9	47.2	35.1	-	-	-
LIP (2019)	38.8	52.4	44.6	35.1	48.2	40.6	36.2	49.5	41.9	-	-	-
ML (2014)	-	-	-	-	-	-	-	-	-	67.3	22.6	33.9
KnowBERT (2020)	41.9	62.5	50.1	-	-	-	-	-	-	36.6	55.6	44.1
KnowDis (2020)	39.7	66.5	49.7	-	-	-	-	-	-	42.3	60.5	49.8
CauSeRL (2021a)	41.9	69.0	52.1	-	-	-	-	-	-	43.6	68.1	53.2
LearnDA (2021b)	42.2	69.8	52.6	-	-	-	-	-	-	41.9	68.0	51.9
RichGCN (2021)	49.2	63.0	55.2	39.2	45.7	42.2	42.6	51.3	46.6	39.7	56.5	46.7
ERGO (2022)	57.5	72.0	63.9	-	-	-	-	-	-	62.1	61.3	61.7
CF-ECI (2023)	47.1	66.4	55.1	-	-	-	-	-	-	50.5	59.9	54.8
CHEER (2023)	59.9	69.9	62.6	45.2	52.1	48.4	49.7	53.3	51.4	56.4	69.5	62.3
SemSIn (2023)	64.2	65.7	64.9	-	-	-	-	-	-	52.3	65.8	58.3
SENDIR (2023)	65.8	66.7	66.2	33	90	48.3	37.8	82.8	51.9	65.2	57.7	61.2
KADE (2023)	61.5	73.2	66.8	52.1	74.2	60.5	51.9	70.6	59.8	56.8	70.6	66.7
GenECI (2022)	59.5	57.1	58.8	-	-	-	-	-	-	60.1	53.3	58.3
DPJL (2022)	65.3	70.8	67.9	-	-	-	-	-	-	63.6	66.7	64.6
DiffusECI (ours)	65.3	78.3	**71.4**	61.9	59.9	**60.9**	63	64.1	**63.5**	87.7	66.1	**75.4**

Table 1: Model performance on ESL and CTB. The performance improvements of DiffusECI over the baselines are significant with $p < 0.01$.

Model	English			Danish			Spanish			Turkish			Urdu		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
XLM-RoBERTa	48.7	59.9	53.7	35.9	36.2	36.0	50.6	49.1	49.9	44.0	59.4	50.5	40.4	43.2	41.8
KnowBERT	39.3	42.6	40.9	31.4	11.4	16.7	39.9	28.4	33.2	36.5	46.7	41.0	41.1	22.2	28.9
RichGCN	50.6	68.0	58.1	31.9	50.0	38.9	50.7	55.0	52.8	50.5	64.6	56.7	37.7	56.0	45.1
DiffusECI (ours)	70.1	68.3	**69.2**	42.7	53.3	**47.4**	62.9	50.2	**55.8**	52.6	66.5	**58.7**	58.1	52.5	**55.2**

Table 2: Model performance on the test sets of MECI for different languages.

resources to secure good performance, such as human annotation for causal signals in DPJL or dependency parsing in GenECI and RichGCN.

Moreover, Table 2 provides a comparison between DiffusECI and the existing models (Lai et al. 2022) on the test sets of the multilingual MECI dataset. To make DiffusECI suitable for the multilingual context, we translate the English template S_{lbl} for label representation computation into the target languages. For the encoder model, we utilize the multilingual version of RoBERTa, specifically the base version of XLM-RoBERTa (Conneau et al. 2020). The table illustrates that DiffusECI also substantially outperforms the baseline models across all the languages in MECI. The observed performance improvements are significant for all languages ($p < 0.01$), thereby further showcasing the effectiveness of our model for multilingual ECI.

Ablation Study

To shed light on the impact of the designed diffusion model for DiffusECI, we evaluate the performance of several ablated baselines and variants:

• **RoBERTa**: This model concerns the complete removal of the diffusion model that directly sends the concatenation of the context representations r_1 and r_2 into a two-layer feed-forward network with softmax in the end to compute the label distribution for prediction.

Model	P	R	F1
RoBERTa	39.3	45.5	42.2
FF	54.2	59.1	56.5
TRANS	55.5	60.1	57.7
No-Event	55.6	62.1	58.7
DiffusECI (full)	63.0	64.1	63.5

Table 3: Model's performance for ablation study, computed for the intra+inter-sentence examples in ESL.

• **FF**: This baseline replaces the diffusion model in DiffusECI with a 8-layer feed-forward network (i.e., similar to the number of layers in the adjustment prediction network ϵ_θ). The feed-forward network will also aim to transform the context representations r_1 and r_2 into the causal label representation $V[0]$ for the label sentence S_{lbl}. The mean squared error between $V[0]$ and the predicted label represen-

Model	English		Danish		Spanish		Turkish		Urdu	
	# ≤ 15	# > 15	# ≤ 15	# > 15	# ≤ 15	# > 15	# ≤ 15	# > 15	# ≤ 15	# > 15
XLM-RoBERTa	64.0	11.5	44.9	20.8	54.5	20.9	61.7	19.1	46.1	14.8
DiffusECI	71.6	46.8	55.0	41.4	62.9	40.0	65.8	43.9	56.9	38.3
Δ	7.6	**35.3**	10.1	**20.6**	8.4	**19.1**	4.1	**24.8**	10.8	**23.5**

Table 4: Model performance on the test sets of MECI for different languages. # represents the number of words between e_1 and e_2 in the input text. Δ indicates the performance diference between DiffusECI and XLM-RoBERTa.

tation is used to train this network in the training step while the predicted label representation from the network will be combined with r_1 and r_2 to predict causal relation at the test time.

• **TRANS**: This baseline substitutes the diffusion model in DiffusECI with a transformer network of 8 layers as in the adjustment prediction network ϵ_θ to predict the causal label representation V from r_1 and r_2. In particular, TRANS also initializes V_0 of size $m \times d$ randomly with Gaussian noise. Afterward, r_1 and r_2 will be appended to the vector sequence in V_0 to serve as the input for the transformer network. The last m hidden vectors in the last layer V_{pred} will serve as the prediction for V. V_0 and the transformer network can then be trained using the mean square error between V and V_{pred}. Eventually, $V_{pred}[0]$ can be used to predict causal relation as done with $\hat{V}[0]$ in DiffusECI.

• **No-Event**: To demonstrate the benefits of the event mentions e_1 and e_2 in the sentence S_{lbl} for label representation computation, this baseline removes e_1 and e_2 from S_{lbl} in the diffusion model of DiffusECI (preserving the other components).

Table 3 presents model performance for the ablation study over the ESL dataset. Comparing RoBERTa to FF and TRANS, we observe notable performance improvements when explicitly transforming event context representations into causal label representations, as seen in FF and TRANS, as opposed to RoBERTa's implicit transformation. Moreover, DiffusECI enhances FF and TRANS performance substantially, highlighting the diffusion model's role in achieving gradual context-to-label representation transformation and irrelevant feature elimination for ECI. Additionally, excluding event mentions e_1 and e_2 from S_{lbl} in No-Event leads to significant performance reduction in DiffusECI, highlighting their necessity to simplify the context-to-label transformation and boost our model's performance.

Analysis

To obtain further insights for the advantages of DiffusECI, Table 4 compares the performance of Diffusion and XLM-RoBERTa over the test sets of MECI in different languages. Here, XLM-RoBERTa is similar to the RoBERTa model in the ablation study, which predicts the causal relation from the combined context representations r_1 and r_2 (encoded by XLM-RoBERTa) using a feed-forward network. For each language, the table considers performance for examples in two scenarios, depending on whether the number of words between the event mentions e_1 and e_2 in the input text are greater than 15 or not. Apart from DiffusECI outperforming XLM-RoBERTa across different languages and scenarios, a crucial insight from the table is the notably larger performance enhancement of DiffusECI for event mention pairs that are positioned further apart in the input (i.e., exceeding 15 words). This observation implies that the major reason for the better performance of DiffusECI lies in its enhanced ability to effectively eliminate irrelevant features and learn better representations for longer input for ECI. Overall, it further demonstrates the benefits of the diffusion model in our method for the ECI problem.

Related Work

Rule-based and feature-based methods represent the major approach in early research for ECI (Riaz and Girju 2014; Beamer and Girju 2009; Do, Chan, and Roth 2011; Hidey and McKeown 2016; Ning et al. 2018; Hashimoto 2019; Gao, Choubey, and Huang 2019). With the introduction of deep learning methods, the performance of ECI models have been advanced to a higher level (Zuo et al. 2021b; Chen et al. 2022). In addition to the use of pre-trained language models, at the core of such deep learning models characterizes different additional resources to advance the performance for ECI, such as distant supervision data (Zuo et al. 2020), background knowledge (Liu, Chen, and Zhao 2020), dependency parsing (Tran Phu and Nguyen 2021), and external causal statements (Zuo et al. 2021a). Recently, some work has also explored a new formulation for ECI using generative models to demonstrate promising performance (Man, Nguyen, and Nguyen 2022; Shen et al. 2022). However, the common issue of previous ECI models concerns the drastic changes in the processes to transform event context representations to causal label representations, which cannot secure optimal performance. Our work thus introduces the first diffusion model to enhance the context-to-label representation transformation processes, thus boosting the causal prediction performance for ECI.

Conclusion

We introduce the first diffusion model for event causality identification. Our model improves the context-to-label representation transformation for ECI models by decomposing the process into multiple steps to facilitate the representation learning performance. We present a controlling mechanism to encourage the representation transition in each step to focus on incremental irrelevant feature elimination, thus leading to cleaner causal label representations for ECI. We extensively evaluate our model across diverse datasets encompassing different languages and learning scenarios, showcasing its state-of-the-art performance.

Acknowledgements

This research has been supported by the Army Research Office (ARO) grant W911NF-21-1-0112, the NSF grant CNS-1747798 to the IUCRC Center for Big Learning, and the NSF grant # 2239570. This research is also supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract 2022-22072200003. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government.

References

Beamer, B.; and Girju, R. 2009. Using a Bigram Event Model to Predict Causal Potential. In *CICLing*.

Berant, J.; Srikumar, V.; Chen, P.-C.; Vander Linden, A.; Harding, B.; Huang, B.; Clark, P.; and Manning, C. D. 2014. Modeling Biological Processes for Reading Comprehension. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1499–1510. Doha, Qatar: Association for Computational Linguistics.

Caselli, T.; and Vossen, P. 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*, 77–86. Vancouver, Canada: Association for Computational Linguistics.

Chen, M.; Cao, Y.; Deng, K.; Li, M.; Wang, K.; Shao, J.; and Zhang, Y. 2022. ERGO: Event Relational Graph Transformer for Document-level Event Causality Identification. In *Proceedings of the 29th International Conference on Computational Linguistics*, 2118–2128. Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

Chen, M.; Cao, Y.; Zhang, Y.; and Liu, Z. 2023. CHEER: Centrality-aware High-order Event Reasoning Network for Document-level Event Causality Identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10804–10816. Toronto, Canada: Association for Computational Linguistics.

Choubey, P. K.; and Huang, R. 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*, 1796–1802. Copenhagen, Denmark: Association for Computational Linguistics.

Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; and Stoyanov, V. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451. Online: Association for Computational Linguistics.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Do, Q.; Chan, Y. S.; and Roth, D. 2011. Minimally Supervised Event Causality Identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 294–303. Edinburgh, Scotland, UK.: Association for Computational Linguistics.

Gao, L.; Choubey, P. K.; and Huang, R. 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)*, 1808–1817. Minneapolis, Minnesota: Association for Computational Linguistics.

Hashimoto, C. 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)*, 2988–2999. Hong Kong, China: Association for Computational Linguistics.

Hidey, C.; and McKeown, K. 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)*, 1424–1433. Berlin, Germany: Association for Computational Linguistics.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, Z.; Li, Z.; Jin, X.; Bai, L.; Guan, S.; Guo, J.; and Cheng, X. 2023. Semantic Structure Enhanced Event Causality Identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10901–10913. Toronto, Canada: Association for Computational Linguistics.

Kadowaki, K.; Iida, R.; Torisawa, K.; Oh, J.-H.; and Kloetzer, J. 2019. Event Causality Recognition Exploiting Multiple Annotators' Judgments and Background Knowledge. 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)*, 5816–5822. Hong Kong, China: Association for Computational Linguistics.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Lai, V. D.; Veyseh, A. P. B.; Nguyen, M. V.; Dernoncourt, F.; and Nguyen, T. H. 2022. MECI: A Multilingual Dataset for Event Causality Identification. In *Proceedings of the 29th International Conference on Computational Linguistics*, 2346–2356. Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

Liu, J.; Chen, Y.; and Zhao, J. 2020. Knowledge Enhanced Event Causality Identification with Mention Masking Generalizations. In Bessiere, C., ed., *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, 3608–3614. International Joint Conferences on Artificial Intelligence Organization. Main track.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Man, H.; Nguyen, M.; and Nguyen, T. 2022. Event Causality Identification via Generation of Important Context Words. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, 323–330. Seattle, Washington: Association for Computational Linguistics.

Mirza, P. 2014. Extracting Temporal and Causal Relations between Events. In *Proceedings of the ACL 2014 Student Research Workshop*, 10–17. Baltimore, Maryland, USA: Association for Computational Linguistics.

Mu, F.; and Li, W. 2023. Enhancing Event Causality Identification with Counterfactual Reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 967–975. Toronto, Canada: Association for Computational Linguistics.

Nichol, A. Q.; and Dhariwal, P. 2021. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 8162–8171. PMLR.

Ning, Q.; Feng, Z.; Wu, H.; and Roth, D. 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)*, 2278–2288. Melbourne, Australia: Association for Computational Linguistics.

Oh, J.-H.; Torisawa, K.; Hashimoto, C.; Iida, R.; Tanaka, M.; and Kloetzer, J. 2016. A Semi-Supervised Learning Approach to Why-Question Answering. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*.

Riaz, M.; and Girju, R. 2014. In-depth Exploitation of Noun and Verb Semantics to Identify Causation in Verb-Noun Pairs. In *SIGDIAL*.

Shen, S.; Zhou, H.; Wu, T.; and Qi, G. 2022. Event Causality Identification via Derivative Prompt Joint Learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, 2288–2299. Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

Tran Phu, M.; and Nguyen, T. H. 2021. Graph Convolutional Networks for Event Causality Identification with Rich Document-level Structures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3480–3490. Online: Association for Computational Linguistics.

Wu, S.; Zhao, R.; Zheng, Y.; Pei, J.; and Liu, B. 2023. Identify Event Causality with Knowledge and Analogy. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11): 13745–13753.

Yuan, C.; Huang, H.; Cao, Y.; and Wen, Y. 2023. Discriminative Reasoning with Sparse Event Representation for Document-level Event-Event Relation Extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 16222–16234. Toronto, Canada: Association for Computational Linguistics.

Zuo, X.; Cao, P.; Chen, Y.; Liu, K.; Zhao, J.; Peng, W.; and Chen, Y. 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*, 2162–2172. Online: Association for Computational Linguistics.

Zuo, X.; Cao, P.; Chen, Y.; Liu, K.; Zhao, J.; Peng, W.; and Chen, Y. 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)*, 3558–3571. Online: Association for Computational Linguistics.

Zuo, X.; Chen, Y.; Liu, K.; and Zhao, J. 2020. KnowDis: Knowledge Enhanced Data Augmentation for Event Causality Detection via Distant Supervision. In *Proceedings of the 28th International Conference on Computational Linguistics*, 1544–1550. Barcelona, Spain (Online): International Committee on Computational Linguistics.