Hierarchical Selection of Important Context for Generative Event Causality Identification with Optimal Transports

Hieu Man[1], **Chien Van Nguyen**[1], **Nghia Trung Ngo**[1], **Linh Van Ngo**[2], **Franck Dernoncourt**[3], **Thien Huu Nguyen**[1]
[1] Department of Computer Science, University of Oregon, USA
[2] Hanoi University of Science and Technology, Vietnam
[3] Adobe Research, USA
{hieum,chienn,nghian,thienn}@uoregon.edu
linhnv@soict.hust.edu.vn, franck.dernoncourt@adobe.com

Abstract

We study the problem of Event Causality Identification (ECI) that seeks to predict causal relation between event mentions in the text. In contrast to previous classification-based models, a few recent ECI methods have explored generative models to deliver state-of-the-art performance. However, such generative models cannot handle document-level ECI where long context between event mentions must be encoded to secure correct predictions. In addition, previous generative ECI methods tend to rely on external toolkits or human annotation to obtain necessary training signals. To address these limitations, we propose a novel generative framework that leverages Optimal Transport (OT) to automatically select the most important sentences and words from full documents. Specifically, we introduce hierarchical OT alignments between event pairs and the document to extract pertinent contexts. The selected sentences and words are provided as input and output to a T5 encoder-decoder model which is trained to generate both the causal relation label and salient contexts. This allows richer supervision without external tools. We conduct extensive evaluations on different datasets with multiple languages to demonstrate the benefits and state-of-the-art performance of ECI.

Keywords: Event Causality Identification, Hierarchical Optimal Transport, Generative Models

1. Introduction

Aiming to identify causal relations between event mentions in text, Event Causal Identification (ECI), serves as an important task in Information Extraction (IE) to reveal event structures for text understanding. For instance, in the sentence "*The dam collapse caused severe flood in the neighborhood.*", ECI models need to predict the causal relation between the two event mentions/triggers "*collapse*" and "*flood*", i.e., "*collapse*" $\xrightarrow{\text{cause}}$ "*flood*". When deployed for real-world applications, an ECI system can provide useful information for different natural language processing (NLP) tasks such as machine reading comprehension (Berant et al., 2014), question answering (Oh et al., 2016), and event forecasting (Hashimoto, 2019).

A key challenge for ECI models is effectively capturing text context information for two input events to facilitate causal relation prediction. In the literature, sentence-level ECI models only focus on cases where two input events are presented in the same sentences (Do et al., 2011; Hashimoto, 2019; Zuo et al., 2020; Shen et al., 2022; Man et al., 2024). In contrast, document-level ECI models extend the focusing context to allow the event mentions to appear in different sentences of a document, potentially involving long distances with greater challenges for context encoding (Gao et al., 2019; Tran and Nguyen, 2021; Chen et al., 2022). To be clear, this work addresses document-level ECI to achieve the most flexibility for context modeling.

A majority of previous work has formulated ECI as a classification problem by employing a discriminator on top of a pre-trained encoder language models such as BERT (Devlin et al., 2019). However, classification-based methods for ECI cannot leverage the semantics of labels and their dependencies with important context words to boost prediction performance (Man et al., 2022b). To this end, a few recent ECI work has explored a new generation-based approach for ECI to produce state-of-the-art performance (Man et al., 2022b; Shen et al., 2022). In such methods, a model was trained to generate an output containing the label rather than classify inputs into predefined categories. Further, important context words in the input texts for causal prediction can also be included in the output texts for generation, serving as a complementary task to aid ECI (Man et al., 2022b). However, current generative ECI models can only solve sentence-level ECI due to their requirements to consume entire input texts that cannot accept longer context beyond the pre-defined limitations of generative PLMs (Man et al., 2022b; Shen et al., 2022). In addition, the important context words for generation in previous ECI work are obtained via dependency parsing tools (i.e., using words in the dependency paths between two events) (Man

LREC-COLING 2024, pages 8122–8132

et al., 2022b) or human annotation (i.e., human-annotated cue words)(Shen et al., 2022), which might not be ideal for training ECI models in practice. However, dependency parsing tools might not be perfect and the dependency path-based heuristics might not always effectively extract important context words in input texts for causal prediction, especially for document-level ECI where dependency parsing for documents is less well-defined. On the other hand, human annotation for important context words can be expensive and less practical in different domains and languages.

To address the aforementioned limitations, we introduce a novel generative model for ECI based on a hierarchical Optimal Transport alignment to select important contexts in the input document. First, we formulate the identification of important contexts as an OT problem (Peyre and Cuturi, 2019) between event pairs and the document to automatically extract relevant contexts without using third-party tools or manual annotation. Specifically, we solve the OT problem at both sentence and word levels to hierarchically align event pairs to sentences and words in the document. We then identify salient contexts based on the OT alignments.

Second, we propose a new learning paradigm to enhance important context extraction and ECI performance. The generative PLMs will consume the selected important sentences as input, which is often much shorter than the whole document due to the sparsity of OT alignment (Swanson et al., 2020), and then aim to generate both salient context words and causal labels to achieve richer training signals for ECI. To further improve the OT context selection, we optimize these components with reinforcement learning, rewarding selections that improve ECI performance.

Finally, motivated by the benefits of background knowledge for causal prediction between events (Kadowaki et al., 2019; Liu et al., 2020), we propose retrieving sentences expressing relevant event background knowledge from ConceptNet for input documents to enhance our generative models. However, there may be multiple potentially relevant sentences, and it is unclear which are directly useful for the causal prediction task. To address this, we include the knowledge-retrieved sentences in the OT framework for sentence selection to determine their importance for ECI to select those most useful for ECI.

To demonstrate the benefits of the proposed ECI method, we conduct extensive experiments on different benchmark datasets over different languages to produce state-of-the-art performance for ECI. To our knowledge, this is the first work to explore Optimal Transport and background knowledge retrieval for generative ECI.

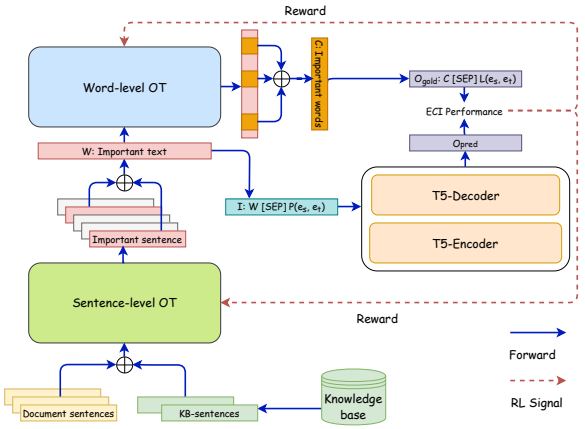

Figure 1: Illustration of our HOTECI framework for ECI. It consists of components for Sentence-level Optimal Transport (left-lower part), Word-level Optimal Transport (left-upper part), and Generative ECI Model (right part).

2. Model

Given an input text S and two event mentions e_s and e_t, ECI models need to predict whether there is a causal relation between the two events. We explore a generative approach for ECI where sequence-to-sequence architectures will be used for the core causal model. The design of a sequence-to-sequence model requires designing a prompt for input and output sequences I and O respectively: the model will consume the input I and attempt to generate the output sequence O in a defined template. In particular, our input sequence I for generative ECI involves a text W to capture necessary context information for the input event mentions e_s and e_t. Following (Man et al., 2022b), we include a prompt $P(e_s, e_t)$ to specify the causal relation prediction task for e_s and e_t in the input sequence:

$$I = W : P(e_s, e_t)$$

with $P(e_s, e_t) =$ *"Is there a causal relation between e_s and e_t?"*. The design of W with important context sentence selection using OT alignments will be described in Section 2.1.

For the output sequence O of generative ECI, our method first obtains a sequence C to contain the most important context words in S for the causal relation prediction of e_s and e_t. Also, let $L(e_s, e_t)$ be either the word *"Yes"* or *"No"* to show the existence of causal relation between e_s and e_t in S. To this end, our output sequence for generative ECI will be the concatenation:

$$O = C : L(e_s, e_t)$$

In this way, our generative ECI model will be trained simultaneously for two highly related tasks, i.e., important context word generation and event causality prediction, thus achieving multi-task training to boost performance for ECI. The word-level OT to select important context words for C will be discussed in Section 2.2. We employ the pre-trained encoder-decoder model T5 (Raffel et al., 2020) to solve the designed sequence-to-sequence problem.

In the end, our model was trained by using a novel training strategy, which will be discussed in Section 2.3, that alternates between updating the generative ECI model and the context selection components. Specifically, we use the standard supervised word prediction objective to train the T5 model while we apply the REINFORCE algorithm (Williams, 1992) to train the OT-based context selection components to select the contexts that enhance the performance of the generative ECI model. For convenience, the proposed model in our work is called **HOTECI** (Hierarchical OT for ECI). Figure 1 provides an overview for our HOTECI model.

2.1. Sentence-level Optimal Transport

As discussed earlier, the input sequence I for our generative ECI model involves a text W to present context information for the two input events e_s and e_t. Previous generative ECI models (Man et al., 2022b) use the whole input text S as W, which can exceed the length limit of PLMs. (Man et al., 2022a) try to address this by iteratively selecting important sentences from S as W. However, their approach can lead to suboptimal solutions due to the number of selected sentences is fixed for all data as a hyperparameter. To address these issues, we propose casting the sentence selection into an optimal transport problem between candidate sentences and host-event sentences. The resulting OT alignment scores each candidate sentence based on its relevance to the input events. This allows adaptive selection of the most informative context sentences for each event pair, without fixing the number of sentences. This way, our model can adaptively select the most informative sentences for each event pair, without fixing the number of sentences. Moreover, our proposed framework can reduce the context length significantly, as the OT alignment is sparse (Swanson et al., 2020).

Let $S = \{s_1, \ldots, s_N\}$ be the input text with N sentences. Let s_s and s_t contain the input events e_s and e_t respectively (possibly the same sentence). Our goal is to select the most relevant sentences in S to facilitate predicting the causal relation between e_s and e_t. We treat s_s and s_t as anchor sentences and define the candidate set $X = S \setminus \{s_s, s_t\}$. We also define the anchor set $Y = \{s_s, s_t, s_{\text{null}}\}$, with s_{null} being a null sentence. Our framework casts context selection as an optimal transport problem between X and Y. By solving this alignment, sentences in X aligned to s_s or s_t are considered highly relevant context and selected. Sentences aligned to s_{null} are discarded as irrelevant.

The application of OT requires definitions of two distributions over X and Y that will be aligned under the transportation cost function T_{sent}:

$$T_{sent} : X \times Y \rightarrow \mathbb{R}_+$$

In this work, we propose to obtain the transportation cost T_{sent} via contextual semantic similarity while distributions for X and Y will be computed via the distance information. Our motivation is to prefer the sentences in X that have similar contextual semantics and are closer physically to the anchor sentences s_s and s_t in S due to their potential to capture relevant context for e_s and e_t.

In particular, to capture contextual semantics for sentences, we first send each sentence in X and Y into the encoder of the T5 model where the representation vector of the <CLS> token in the last layer is used as the representation for the sentence. For convenience, we rename the elements in X and Y by $X = \{x_1, x_2, \ldots, x_{|X|}\}$ and $Y = \{y_1, y_2, y_3\}$ ($y_3 = s_{null}$). Also, let $\bar{x}_i$ and $\bar{y}_j$ be the representations from T5 for the sentence $x_i \in X$ and $y_j \in Y$. Here, the representation for y_3 (or s_{null}) is obtained via the average of the representations in X: $\bar{y}_3 = \text{average}(\bar{x}_i | x_i \in X)$. The transportation cost $T_{sent}(x_i, y_j)$ for $x_i \in X$ and $y_j \in Y$ is then computed via:

$$T_{sent}(x_i, y_j) = 1 - \text{cosine}(FF_1(\bar{x}_i), FF_1(\bar{y}_j))$$

where FF_1 is a learnable two-layer feed-forward network. For the distribution over X, let ds_i^s and ds_i^t be the numbers of sentences from $x_i \in X$ to s_s and s_t respectively. The distance ds_i from x_i to the anchor sentences s_s and s_t is then defined by: $ds_i = \min(ds_i^s, ds_i^t)$. As such, the distribution $P^X(x_i)$ over X is obtained via softmax:

$$P^X(x_i) = \text{softmax}(ds_i | x_i \in X)$$

For Y, its distribution $P^Y(y_j)$ will be uniform due to the equal importance of s_s, s_t, and s_{null} in our model. Given $P^X(x_i)$, $P^Y(y_j)$, and $T_{sent}(x_i, y_j)$, the optimal joint alignment/distribution $\pi^*_{sent}(x_i, y_j)$ over X and Y with the marginals $P^X(x_i)$ and $P^Y(y_j)$ can be achieved via solving the optimization of OT problem[1]. As such, the distribution $\pi^*_{sent}(x_i, y_j)$ is a matrix; its element (x_i, y_j) represents the probability of transforming $x_i \in X$ to $y_j \in Y$ in the optimal alignment.

[1]The Sinkhorn algorithm (Peyre and Cuturi, 2019) is employed to solve the entropy-based approximation of OT.

Using the optimal alignment $\pi^*_{sent}(x_i, y_j)$, the probability for a sentence $x_i \in X$ to be aligned with an anchor sentence (i.e., s_s or s_t) $\pi^*_{sent}(x_i)$ is computed:

$$\pi^*_{sent}(x_i) = \pi^*_{sent}(x_i, y_1) + \pi^*_{sent}(x_i, y_2)$$

We also consider $\pi^*_{sent}(x_i)$ as a probability for $x_i \in X$ to be selected for important context in our model. Eventually, we select the sentences in X with probability $\pi^*_{sent}(x_i)$ greater than 0.5 and their concatenation will be used as the context text W in our input sequence I. For convenient computation, we also obtain a distribution Q_{sent} over X to represent the probability that each sentence x_i is selected for important context via the softmax operation: $Q_{sent}(x_i) = \text{softmax}(\pi^*_{sent}(x_i)|x_i \in X)$.

2.2. Word-level Optimal Transport

Given the selected sentences for the context text W, our next step involves selecting important context words in W to form the sequence C for the output sequence O (i.e., $O = C : L(e_s, e_t)$) that will be generated by the generative model to improve performance for ECI. Similar to sentence selection, we consider the input event triggers e_s and e_t as the most important context words for ECI (i.e., the anchor words). We thus propose to solve the word selection problem via an OT framework between anchor words and the other words in W. In particular, let $V = \{e_s, e_t, w_{null}\} = \{v_1, v_2, v_3\}$ be the set of input event triggers along with a special token w_{null} ($v_3 = w_{null}$). Also, let $U = W \setminus \{e_s, e_t\} = \{u_1, u_2, \ldots, u_{|U|}\}$ be the set of non-anchor words in W. To capture relevant context words for e_s and e_t, similar to sentence-level, our method also emphasizes words in U with similar contextual semantics and close distances. To this end, OT methods are also employed to solve the alignment problem between U and V to effectively combine the two information preferences.

In particular, the text W is first sent into the encoder of T5 to obtain representations for every word (using hidden vectors of the first sub-tokens in the last layer). The representations from T5 for the words $u_i \in U$ and $v_j \in V$ are then denoted by $\bar{u}_i$ and $\bar{v}_j$. Here, the representation for v_3 (or w_{null}) is computed via the average over the representations for the words in U, i.e., $\bar{v}_3 = \text{average}(\bar{u}_i|u_i \in U)$. The transportation cost function $T_{word}(u_i, v_j)$ for the OT between U and V is then defined via:

$$T_{word}(u_i, v_j) = 1 - \text{cosine}(FF_2(u_i), FF_2(v_j))$$

with FF_2 as a two-layer feed-forward network. To compute the distributions over U and V for OT, we also employ the uniform distribution $P^V(v_j)$ for V. For U, we compute the numbers of words da^s_i and da^t_i from the word $u_i \in U$ to the event triggers e_s and e_t in W respectively. Afterward, the distance to anchor words $da_i = \min(da^s_i, da^t_i)$ is computed for each $u_i \in U$ and the distribution $P^U(u_i)$ for U is returned from the softmax operation:

$$P^U(u_i) = \text{softmax}(da_i|u_i \in U)$$

To this end, by solving this word-level OT problem, we obtain the OT optimization problem for our alignment problem between U and V for word selection. The solution for this problem from OT methods is called $\pi^*_{word}(u_i, v_j)$, an alignment matrix/joint distribution over U and V.

Similar to the sentence selection component, we use the score $\pi^*_{word}(u_i)$:

$$\pi^*_{word}(u_i) = \pi^*_{word}(u_i, v_1) + \pi^*_{word}(u_i, v_2)$$

to represent the probability for the word $u_i \in U$ to be selected for important context words. As such, the words u_i in U with probability score $\pi^*_{word}(u_i)$ greater than 0.5 will be selected for the context C in our output sequence. Finally, we also obtain a distribution Q_{word} over U to capture selection probability for each word u_i via: $Q_{word}(u_i) = \text{softmax}(\pi^*_{word}(u_i)|u_i \in U)$.

2.3. Training

Using the input and output sequences I and O, one way to train the encoder-decoder model T5 is to only optimize the negative log-likelihood loss:

$$\mathcal{L}_{likelihood} = -\log P(O|I)$$

where the probability $P(O|I)$ is computed via the returned distributions from the decoder. However, this approach cannot update the parameters for the OT components for context selection (i.e., the learnable networks FF_1 and FF_2 for T_{sent} and T_{word}) due to the discreteness of the selected sentences and words in I and O. To this end, we further employ the REINFORCE algorithm (Williams, 1992) to enable training of OT-based context selection components.

In particular, let $W = \{w_1, \ldots, w_{N_s}\}$ be the selected sentence set for the context text W and $C = \{c_1, \ldots, c_{N_w}\}$ be the selected word set. Also, let $\hat{L}(e_s, e_t)$ be the word generated by the decoder of T5 after its encoder has consumed the input I and the decoder has produced "C :" for the output. The reward R for context selection in W and C for the input and output sequences is set to 1 if the generated label $\hat{L}(e_s, e_t)$ is the same as the golden one $L(e_s, e_t)$, and 0 otherwise. The REINFORCE loss to train our model is thus:

$$\mathcal{L}_{RL} = -R \log P(W, C|S)$$

. Here, $P(W, C|S)$ is computed via the selection distributions Q_{sent} and Q_{word}:

$$P(W, C|S) = \prod_{i=1}^{N_s} P(w_i|S) \prod_{j=1}^{N_w} P(c_j|W)$$
$$= \prod_{i=1}^{N_s} Q_{sent}(w_i) \prod_{j=1}^{N_w} Q_{word}(c_j)$$

As the parameters for the OT-based components are involved in $P(W, C|S)$, $\mathcal{L}_{RL}$ will allow us to update the parameters for learning.

To this end, we design an alternating training procedure for our model. In each iteration with a batch of data, our model first uses the current parameters to obtain the selected context W and C to form the input and output sequences I and O. The T5 model for generative ECI is then updated using the loss $\mathcal{L}_{likelihood}$. Afterward, the parameters for the OT components for sentence and word selections will be updated using the loss $\mathcal{L}_{RL}$.

2.4. Background Knowledge Retrieval

Background knowledge has been shown to be helpful for ECI in previous work, especially for implicit relations (Kadowaki et al., 2019; Liu et al., 2020). Accordingly, Therefore, we propose to use ConceptNet (Speer et al., 2017) as a source of background knowledge for our generative ECI model. ConceptNet is a large-scale knowledge graph that contains concepts (including events) and their relations. Some of the relations in ConceptNet are relevant for ECI, such as *"Causes"*, *"Causes Desire"*, *"Created By"*, and *"DerivedFrom"*. Each relation in ConceptNet can be represented as a triple of the form $< Concept_1, Relation, Concept_2 >$. To use ConceptNet for ECI, we first identify the concepts in ConceptNet that match the event triggers in the input document S. Then, we retrieve the triples from ConceptNet that involve at least one of these concepts. These triples represent the background knowledge about the events in the document. Next, we convert each triple $< Concept_1, Relation, Concept_2 >$ into a natural language sentence of the form "$Concept_1 Relation Concept_2$". We denote the set of these sentences as $B = \{b_1, \ldots, b_{|B|}\}$. Finally, we append the sentences in B to the original document to form an enriched input S that contains both the document and the background knowledge.

3. Experiments

Datasets and Hyperparameters: Following previous work (Gao et al., 2019; Liu et al., 2020; Man et al., 2022b), we employ two widely-used English datasets for ECI to evaluate our model HOTECI, i.e., EventStoryLine (ESL) (Caselli and Vossen, 2017) and Causal-TimeBank (CTB) (Mirza, 2014). ESL (version 0.9) includes 258 documents in 22 topics with 4316 sentences and 5334 event mentions. There are 7805 intra-sentence and 46521 inter-sentence mention pairs in ESL; 1770 and 3855 of them have causal relation respectively. Following the same data split in prior work (Liu et al., 2020; Tran and Nguyen, 2021), the last two topics of ESL is used as development data; the other 20 topics are leveraged for 5-fold cross-validation evaluation. Also, CTB presents 184 documents with 6813 events; there are 7608 event mention pairs with 318 positive examples for causal relation. We use the same data split as previous work (Liu et al., 2020; Zuo et al., 2021b) with 10-fold cross-validation for the evaluation on CTB.

In addition, we evaluate our model on MECI (Lai et al., 2022b), the recent dataset for multilingual ECI that annotates causal event relation for text over five different languages, i.e., English, Danish, Spanish, Turkish, and Urdu. The documents in MECI are based on Wikipedia and the annotation schema follows those for ESL. As such, MECI contains both intra-sentence and inter-sentence examples. To facilitate comparison, for each language, we utilize the same data split for training/dev/test data portions as in (Lai et al., 2022b) in the evaluation.

We tune the hyperparameters for HOTECI on the development data of ESL and leverage the chosen parameters for CTB and MECI datasets for consistency. Our tuning process returns the following hyper-parameters: $2e$-5 for the learning rate with the Adam optimizer; 16 for the mini-batch size; and 512 dimensions for hidden vectors in the feed-forward networks FF_1 and FF_2. Finally, we use the base versions of T5 (Raffel et al., 2020) for the evaluation on ESL and CTB, and multilingual T5, i.e., mT5 (Xue et al., 2021), for MECI.

Baselines: This section compares our model HOTECI with state-of-the-art (SOTA) models for ECI. For ESL, we consider the following baseline methods: (1) **LSTM** (Gao et al., 2019); (2) **Seq** (Gao et al., 2019) adopted from (Choubey and Huang, 2017) for ECI; and (3) **LR+** and **LIP** (Gao et al., 2019): document structure models. For CTB, we evaluate **ML**: a feature-based model in (Mirza, 2014). For both ESL and CTB, we also compare with the following transformer-based models for ECI: (1) **BERT**: a BERT-based baseline in (Zuo et al., 2021b); (2) **KnowDis** (Zuo et al., 2020): a distant supervision-based model; (3) **Know** (Liu et al., 2020): a ConceptNet-based model; (4) **RichGCN** (Tran and Nguyen, 2021): a rich graph convolutional model, (5) **LearnDA** (Zuo et al., 2021b): a data augmentation method, (6) **CauSeRL** (Zuo et al., 2021a): a self-supervised

Model	ESL (Intra-sentence)			ESL (Inter-sentence)			ESL (Intra + Inter)			CTB (Intra-Sentence)		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
LSTM (Gao et al., 2019)	34.0	41.5	37.4	13.5	30.3	18.7	17.6	33.9	23.2	-	-	-
Seq (Gao et al., 2019)	32.7	44.9	37.8	11.3	29.5	16.4	15.5	34.3	21.4	-	-	-
LR+ (Gao et al., 2019)	37.0	45.2	40.7	25.2	48.1	33.1	27.9	47.2	35.1	-	-	-
LIP (Gao et al., 2019)	38.8	52.4	44.6	35.1	48.2	40.6	36.2	49.5	41.9	-	-	-
ML (Mirza, 2014)	-	-	-	-	-	-	-	-	-	67.3	22.6	33.9
BERT (Tran and Nguyen, 2021)	39.2	49.3	43.7	22.3	29.2	25.3	27.3	35.3	30.8	38.5	43.9	41.0
KnowDis (Zuo et al., 2020)	39.7	66.5	49.7	-	-	-	-	-	-	42.3	60.5	49.8
Know (Liu et al., 2020)	41.9	62.5	50.1	-	-	-	-	-	-	36.6	55.6	44.1
RichGCN (Tran and Nguyen, 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
LearnDA (Zuo et al., 2021b)	42.2	69.8	52.6	-	-	-	-	-	-	41.9	68.0	51.9
CauSeRL (Zuo et al., 2021a)	41.9	69.0	52.1	-	-	-	-	-	-	43.6	68.1	53.2
ERGO-BERT (Chen et al., 2022)	49.7	72.6	59.0	-	-	-	-	-	-	58.4	60.5	59.4
ERGO-Longformer (Chen et al., 2022)	57.5	72.0	63.9	-	-	-	-	-	-	62.1	61.3	61.7
CF-ECI (Mu and Li, 2023)	47.1	66.4	55.1	-	-	-	-	-	-	50.5	59.9	54.8
CHEER (Chen et al., 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 (Hu et al., 2023)	64.2	65.7	64.9	-	-	-	-	-	-	52.3	65.8	58.3
SENDIR (Yuan et al., 2023)	65.8	66.7	66.2	33	90	48.3	37.8	82.8	51.9	65.2	57.7	61.2
GenECI* (Man et al., 2022b)	58.7	65.7	61.9	-	-	-	-	-	-	58.6	59.3	58.6
DPJL (Shen et al., 2022)	65.3	70.8	67.9	-	-	-	-	-	-	63.6	66.7	64.6
HOTECI (ours)*	66.1	72.3	**69.1**	81.4	40.6	**55.1**	63.1	51.2	**56.5**	71.1	65.9	**68.4**

Table 1: Model's performance on ESL and CTB. The performance improvement of HOTECI over the baselines is significant with $p < 0.01$. * designates models that use T5.

Model	English			Danish			Spanish			Turkish			Urdu		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
PLM	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
Know	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
HOTECI (ours)	66.6	67.1	**66.8**	50.5	63.7	**56.3**	60.7	60.7	**60.7**	72.5	76.6	**74.5**	59.1	71.0	**64.5**

Table 2: Model's performance on MECI for different languages. The baselines use the base version of the multilingual RoBERTa model, i.e., XLMR (Conneau et al., 2020). PLM is similar to the BERT baseline in Table 1, but replaces BERT with XLMR.

method with external causal statements, (7) **ERGO-BERT** and **ERGO-Longformer** (Chen et al., 2022): relational graph transformer frameworks; (8) **CF-ECI** (Mu and Li, 2023): a counterfactual reasoning mdoel to explicitly estimate the influence of context keywords and event pairs for debiasing; (9) **CHEER** (Chen et al., 2023): a graph framework considering the centrality of events and their interactions in a document-level graph; (10) **SemSIn** (Hu et al., 2023): a graph model integrating event-centric and event-associated semantic structures; (11) **SENDIR** (Yuan et al., 2023): a document-level ECI framework using sparse attention and discriminative reasoning; (12) **GenECI** (Man et al., 2022b): a T5-based generative ECI model, and (13) **DPJL** (Shen et al., 2022): a derivative prompt-based model. Among the baselines, DPJL has the best-reported performance for sentence-level ECI on ESL while SENDIR currently observes the state-of-the-art performance on CTB and inter-sentence ECI over ESL.

Comparison: Table 1 shows the performance of the models on test data of the ESL and CTB datasets. For ESL, similar to previous work (Tran and Nguyen, 2021; Yuan et al., 2023), we report the performance in different scenarios for intra-sentence, inter-sentence, and both intra- and inter-sentence examples. As can be seen, the proposed model HOTECI significantly outperforms the state-of-the-art model DPJL for sentence-level ECI over ESL. HOTECI is also substantially better than the best baseline model SENDIR for inter- and intra+inter-sentence ECI on ESL and CTB (i.e., more than 7% for inter-sentence ECI on ESL and CTB). The improvement is significant with $p < 0.01$, thus clearly demonstrating the benefits of the proposed generative framework for ECI in this work. Notably, HOTECI can achieve state-of-the-art performance for ECI without relying on additional annotation or third-party tools. This is in contrast to recent work on ECI that requires those resources to ensure good performance, e.g., human annotation for causal signals in DPJL or dependency parsing in GenECI and RichGCN.

In addition, Table 2 compares HOTECI with the best-reported models (Lai et al., 2022b) on the multilingual dataset MECI over different languages. Here, to adapt HOTECI to multilingual setting, we translate the simple English prompt $P(e_s, e_t)$ to the target languages to form the input sequences I. For ConceptNet, we utilize its multilingual version (Speer et al., 2017) in RichGCN and HOTECI. It is clear from the table that HOTECI performs significantly better than the baselines over different languages. The improvement gaps are large for all the languages, thus clearly testifying to the effectiveness of hierarchical context selection and generative models for multilingual ECI.

Ablation Study: To evaluate the design of the components in HOTECI, we perform an ablation study over the model architecture. Table 3 reports the performance of the ablated models for HOTECI.

First, we analyze the context selection modules. Removing the reinforcement learning (RL) loss for training the context selection (line 2) decreases performance, demonstrating the importance of RL for optimizing these components. Line 3 shows a poorer performance of HOTECI when background knowledge retrieval from ConceptNet is not employed, thus suggesting its importance for ECI.

Eliminating word selection (WS) with optimal transport (OT) (line 4) substantially reduces accuracy, highlighting the benefits of extracting salient words with OT alignments. Replacing WS with baseline strategies like selecting words similar to the event triggers (lines 5-6) also harms performance. We observe similar trends when ablating the sentence selection (SS) module. Simply using the hosting sentences of the event mentions (line 7) or heuristics like surrounding sentences (lines 8-10) are inferior to SS with OT. This validates OT's ability to extract the most relevant sentences.

Further ablations analyze other modeling choices. Feeding the selected words to the encoder rather than decoder (line 12) is ineffective, confirming the advantages of modeling salient words in the output. Removing distance-based distributions in OT (lines 13-14) also degrades performance, showing the utility of distance-aware alignments.

#	Model	P	R	F1
1	HOTECI (full)	63.1	51.2	56.5
2	- RL loss	61.2	49.8	54.9
3	- Background Knowledge	62.2	50.5	55.7
4	- WS	61.1	47.5	53.4
5	- WS (five most similar words)	61.1	43.9	51.1
6	- WS (ten most similar words)	60.3	43.1	50.3
7	- SS (hosting sentences)	60.8	48.7	54.1
8	- SS (max surrounding sentences)	55.8	48.9	52.1
9	- SS (max most similar sentences)	61.2	50.1	55.1
10	- SS (five surrounding sentences)	58.9	47.6	52.7
11	- SS (five most similar sentences)	60.7	48.4	53.9
12	Selected Words to T5 Encoder	55.8	45.9	50.4
13	Uniform Dist for WS	63.2	49.8	55.7
14	Uniform Dist for SS	66.4	48.6	56.1

Table 3: Ablation study over test data of ESL using intra+inter sentence performance. WS and SS stand for word selection and sentence selection (respectively) with OT.

Analysis: We analyze the contribution of the background knowledge for our generative HOTECI model. In particular, we find that HOTECI selects at least one sentence in the background knowledge sentence pools B for 50.5% of test data examples in ESL while this percentage for selected words in the background knowledge sentences is 20%. These results highlight the important contributions of background knowledge to aid ECI for generative models. In addition, we examine the examples in the test data of ECI that are correctly predicted by HOTECI due to the introduction of background knowledge (i.e., the model without background knowledge fails to predict these examples). Our analysis shows that a majority of such examples require a complicated reasoning processing, involving implicit/common sense knowledge to successfully realize causal relations between events. For example, consider the following input document:

Powerful Quake in Iran Kills 10; 80 Hurt and 7 Villages Damaged. Published: November 28, 2005. A powerful earthquake hit southern Iran on Sunday , causing major destruction in seven villages and killing 10 people, and injuring 80. The tremor shook Oman and the United Arab Emirates as well, forcing many office workers to evacuate their buildings. The official IRNA news agency and the United States Geological Survey said it had a magnitude of 5.9. Iran's seismologic center said the epicenter of the earthquake was in the waters of the Persian Gulf, 35 miles southwest of the port of Bandar Abbas. Iran is on seismic fault lines. A major earthquake killed more than 31, 000 in the city of Bam in central Iran in 2003, and 600 were killed in the city of Zarand in February in an earthquake with a magnitude of 6.4. The tourist Island of Qeshm on the Persian Gulf and seven of its villages were most strongly affected by the quake. [It] hit at 1:53 p.m. local time and was followed by at least four strong aftershocks, IRNA reported. The news agency also reported that one of the major hospitals on the island, in the village of Jeyhian, was destroyed and the village's power lines were cut. The island's airport was also [damage]. Abdolreza Sheikholeslami, the governor of Hormozgan Province, the center of the damaged area, said 40 percent to 70 percent of the buildings in seven villages were destroyed, IRNA reported. Two helicopters began moving the injured to the hospital in Bandar Abbas, and aid workers began distributing food, blankets, and tents in the region, the governor said. Qeshm is Iran's largest island in the Persian Gulf, with a population of 120, 000. The quake jolted several cities in the United Arab Emirates, across the Persian Gulf from Iran. Office workers in Dubai, United Arab Emirates, evacuated several buildings in the city, pouring onto the streets and snarling traffic. There were no reported injuries in Dubai. Mehdi Zareh, director of the seismological center in Tehran, dismissed concerns that the earthquake would cause tsunamis, IRNA reported. "The Persian Gulf is not deep enough so that we can expect tsunamis," he was quoted as saying. Map of Iran highlighting epicenter of earthquake: A tremor shook southern Iran yesterday, causing major damage.

In this document, input event triggers are shown in brackets (i.e., "*It*" and "*damaged*"). The selected sentences from this document of our HOTECI

model are underlined while the selected words are written in red. The selected context W from our HOTECI model for this document is thus as follows:

A powerful earthquake hit southern Iran on Sunday, causing major destruction in seven villages and killing 10 people, and injuring 80. [It] hit at 1:53 p.m. local time and was followed by at least four strong aftershocks, IRNA reported. The island's airport was also [damaged]. A tremor shook southern Iran yesterday, causing major damage. Earthquake causes ruined streets, pipelines , and houses. Shock is a type of earthquake.

Here, the last two sentences (underlined) are the background knowledge sentences retrieved from ConceptNet while the other sentences are selected from the input document S. The selected context words in C are written in red. As such, HOTECI is able to select the first sentence in the example to provide necessary context for the coreference of the pronoun event trigger "*It*" to an earthquake. In addition, HOTECI can include the background knowledge sentence "*Earthquake causes ruined ruined streets, pipelines, and houses.*", which is related to the trigger "*damaged*", to reveal important information for causal relation prediction. In all, it demonstrates the operation of HOTECI and the benefits of background knowledge retrieval in our model.

4. Related Work

The early approaches have explored rule-based (Riaz and Girju, 2014) and feature-based (Beamer and Girju, 2009; Do et al., 2011; Hidey and McKeown, 2016; Ning et al., 2018; Hashimoto, 2019; Gao et al., 2019) models to solve ECI. Recently, the advent of deep learning models has introduced significant advances for ECI. In addition to PLMs, these methods have leveraged different resources to boost ECI performance (Chen et al., 2022), including distant supervision (Zuo et al., 2020), background knowledge (Liu et al., 2020), dependency parsing (Tran and Nguyen, 2021), data augmentation (Zuo et al., 2021b), and external causal statements (Zuo et al., 2021a). However, such previous work has only employed the classification setting for ECI. Due to its relation prediction nature, ECI can also be viewed as a form of the general problem of Relation Extraction in Information Extraction (Pouran Ben Veyseh et al., 2020; Veyseh et al., 2020; Nguyen et al., 2022).

Motivated by the recent success of the generative reformulation for different NLP tasks (Athiwaratkun et al., 2020; Yan et al., 2021; Zhang et al., 2021), there have been two recent works to explore generative models for ECI. In particular, (Shen et al., 2022) introduces a declarative prompt joint learning method using RoBERTa to generate labels while (Man et al., 2022b) studies a generative method based on T5 for ECI. Recently, (Man et al., 2024) explores a diffusion model to generate effective representations for ECI. However, none of these works has considered important context selection to improve generative models for ECI. In addition, we also note some related work that leverage optimal transport to solve NLP problems, e.g., Relation Extraction (Pouran Ben Veyseh et al., 2022; Lai et al., 2022a) and Event Detection (Guzman-Nateras et al., 2022).

The closest work to ours is SCS-EERE (Man et al., 2022a), which models document-level context by iteratively choosing relevant context sentences to address temporal event relation and subevent relation tasks. However, their method has some limitations, such as the fixed number of sentences to select for the whole data. Moreover, SCS-EERE employs a classification setting thus it prevents SCS-EERE from leveraging the semantics of the labels for learning. Additionally, SCS-EERE only explores sentence selection that might still contain irrelevant context words, then can lead to limit in the performance.

5. Conclusion

Formulating ECI as a generation-based problem, we present a novel method for ECI that hierarchically selects important context sentences and words in input documents via Optimal Transports. Our method can effectively handle document-level ECI with long context and inter-sentence event mention pairs to achieve state-of-the-art performance on different benchmark datasets. In the future, we will extend our context selection method to improve generative models for other NLP tasks.

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. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

Bibliographical References

Ben Athiwaratkun, Cicero Nogueira dos Santos, Jason Krone, and Bing Xiang. 2020. Augmented natural language for generative sequence labeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Brandon Beamer and Roxana Girju. 2009. Using a bigram event model to predict causal potential. In *CICLing*.

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.

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.

Meiqi Chen, Yixin Cao, Kunquan Deng, Mukai Li, Kun Wang, Jing Shao, and Yan Zhang. 2022. ERGO: Event relational graph transformer for document-level event causality identification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2118–2128, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Meiqi Chen, Yixin Cao, Yan Zhang, and Zhiwei Liu. 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)*, pages 10804–10816, Toronto, 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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Quang Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 294–303, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Lei Gao, Prafulla Kumar Choubey, and Ruihong Huang. 2019. Modeling document-level causal structures for event causal relation identification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1808–1817, Minneapolis, Minnesota. Association for Computational Linguistics.

Luis Guzman-Nateras, Minh Van Nguyen, and Thien Nguyen. 2022. Cross-lingual event detection via optimized adversarial training. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5588–5599, Seattle, United States. Association for Computational Linguistics.

Chikara Hashimoto. 2019. Weakly supervised multilingual causality extraction from Wikipedia. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2988–2999, Hong Kong, China. Association for Computational Linguistics.

Christopher Hidey and Kathy McKeown. 2016. Identifying causal relations using parallel Wikipedia articles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1424–1433, Berlin, Germany. Association for Computational Linguistics.

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.

Kazuma Kadowaki, Ryu Iida, Kentaro Torisawa, Jong-Hoon Oh, and Julien Kloetzer. 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)*, pages 5816–5822, Hong Kong, China. Association for Computational Linguistics.

Viet Lai, Hieu Man, Linh Ngo, Franck Dernoncourt, and Thien Nguyen. 2022a. Multilingual SubEvent relation extraction: A novel dataset and structure induction method. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5559–5570, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Viet Dac Lai, Amir Pouran Ben Veyseh, Minh Van Nguyen, Franck Dernoncourt, and Thien Huu Nguyen. 2022b. MECI: A multilingual dataset for event causality identification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2346–2356, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Jian Liu, Yubo Chen, and Jun Zhao. 2020. Knowledge enhanced event causality identification with mention masking generalizations. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3608–3614. International Joint Conferences on Artificial Intelligence Organization.

Hieu Man, Franck Dernoncourt, and Thien Huu Nguyen. 2024. Mastering context-to-label representation transformation for event causality identification with diffusion models. In *The AAAI Conference on Artificial Intelligence (AAAI)*. Association for the Advancement of Artificial Intelligence.

Hieu Man, Nghia Trung Ngo, Linh Ngo Van, and Thien Huu Nguyen. 2022a. Selecting optimal context sentences for event-event relation extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11058–11066.

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

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

Feiteng Mu and Wenjie Li. 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)*, pages 967–975, Toronto, Canada. Association for Computational Linguistics.

Minh Van Nguyen, Bonan Min, Franck Dernoncourt, and Thien Nguyen. 2022. Joint extraction of entities, relations, and events via modeling inter-instance and inter-label dependencies. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4363–4374, Seattle, United States. 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, K. Torisawa, C. Hashimoto, R. Iida, M. Tanaka, and Julien Kloetzer. 2016. A semi-supervised learning approach to why-question answering. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*.

Gabriel Peyre and Marco Cuturi. 2019. Computational optimal transport: With applications to data science. In *Foundations and Trends in Machine Learning*.

Amir Pouran Ben Veyseh, Franck Dernoncourt, Dejing Dou, and Thien Huu Nguyen. 2020. Exploiting the syntax-model consistency for neural relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8021–8032, Online. Association for Computational Linguistics.

Amir Pouran Ben Veyseh, Minh Van Nguyen, Franck Dernoncourt, Bonan Min, and Thien Nguyen. 2022. Document-level event argument extraction via optimal transport. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1648–1658, Dublin, Ireland. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. In *Journal of Machine Learning Research*.

Mehwish Riaz and Roxana Girju. 2014. In-depth exploitation of noun and verb semantics to identify causation in verb-noun pairs. In *SIGDIAL*.

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 Association for the Advancement of Artificial Intelligence (AAAI)*.

Kyle Swanson, Lili Yu, and Tao Lei. 2020. Rationalizing text matching: Learning sparse alignments via optimal transport. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5609–5626, Online. Association for Computational Linguistics.

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

Amir Pouran Ben Veyseh, Franck Dernoncourt, Dejing Dou, and Thien Huu Nguyen. 2020. A joint model for definition extraction with syntactic connection and semantic consistency. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*.

Ronald J. Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. In *Kluwer Academic*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021. A unified generative framework for various NER subtasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL)*.

Changsen Yuan, Heyan Huang, Yixin Cao, and Yonggang Wen. 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)*, pages 16222–16234, Toronto, Canada. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021. Towards generative aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL)*.

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.