Knowledge Enhanced Event Causality Identification with Mention Masking Generalizations

Jian Liu[1,2], **Yubo Chen**[1,2] and **Jun Zhao**[1,2]
[1] National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing, 100190, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, 100049, China
{jian.liu, yubo.chen, jzhao}@nlpr.ia.ac.cn

Abstract

Identifying causal relations of events is a crucial language understanding task. Despite many efforts for this task, existing methods lack the ability to adopt background knowledge, and they typically generalize poorly to new, previously unseen data. In this paper, we present a new method for event causality identification, aiming to address limitations of previous methods. On the one hand, our model can leverage external knowledge for reasoning, which can greatly enrich the representation of events; On the other hand, our model can mine event-agnostic, context-specific patterns, via a mechanism called event mention masking generalization, which can greatly enhance the ability of our model to handle new, previously unseen cases. In experiments, we evaluate our model on three benchmark datasets and show our model outperforms previous methods by a significant margin. Moreover, we perform 1) cross-topic adaptation, 2) exploiting unseen predicates, and 3) cross-task adaptation to evaluate the generalization ability of our model. Experimental results show that our model demonstrates a definite advantage over previous methods.

1 Introduction

Event causality identification (ECI) aims to identify *causal relation* of events in texts. For example, in a sentence S1 (shown in Figure 1): "The **earthquake** generates a **tsunami** that rose up to 135 feet", an ECI system should identify that a causal relationship holds between the two mentioned events, i.e., **earthquake** $\xrightarrow{\text{cause}}$ **tsunami**. ECI supports a wide range of intelligent applications including why-question answering [Girju, 2003; Oh *et al.*, 2016], future event/scenario forecasting [Hashimoto *et al.*, 2014], machine reading comprehension [Berant *et al.*, 2014], and others.

To date, various approaches have been proposed for ECI, ranging from the early feature based methods [Do *et al.*, 2011; Hashimoto *et al.*, 2014; Ning *et al.*, 2018; Gao *et al.*, 2019] to the recent representation based methods [Kadowaki *et al.*, 2019]. While, existing methods typically train ECI models on human annotated examples solely, and they generally lack the ability to leverage background knowledge for reasoning. Moreover, owing to the small size of training data (for example, the largest ECI corpus contains less than 300 documents [Caselli and Vossen, 2017]), existing ECI methods suffer from over-fitting issue and have difficulty in handling new, previously unseen cases.

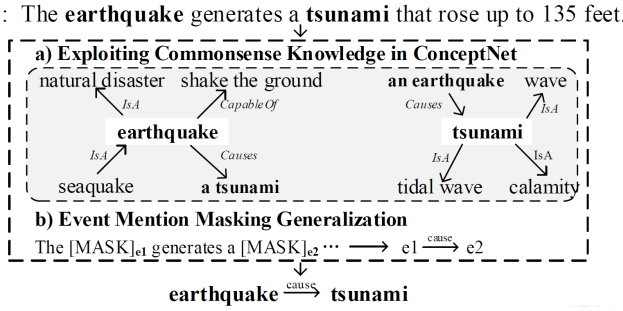

Figure 1: Illustrations of a) exploiting commonsense knowledge and b) mention masking generalization for ECI.

To address the limitations of previous methods, we propose a new approach for ECI, featured by its ability to: 1) explicitly leverage external (commonsense) knowledge for reasoning, which can build more expressive representations for events; and 2) mine event-agnostic, context-specific patterns for reasoning, which results in a decent generalization ability of our model to tackle new, previously unseen examples.

Specifically, one key component of our model is a *knowledge-aware causal reasoner*, which can exploit background knowledge in external knowledge bases (KBs) to enhance the reasoning process. We prefer CONCEPTNET [Speer *et al.*, 2017] as the external KB, which contains abundant semantic knowledge of concepts (represented as words or phrases). For example, in CONCEPTNET, the encoded knowledge associated with "earthquake" includes "earthquake" $\xrightarrow{\text{IsA}}$ "natural disaster", "earthquake" $\xrightarrow{\text{Causes}}$ "a tsunami" and others (as shown in Figure 1 a)). Such knowledge can be used to enrich the representations of events for a more accurate event causality inference. For example, the knowledge-aware causal reasoner may directly predict **earthquake** $\xrightarrow{\text{cause}}$ **tsunami** in S1 based on the semantic knowledge "earthquake" $\xrightarrow{\text{Causes}}$ "a tsunami" encoded in CONCEPTNET.

This indicates that explicitly introducing external knowledge may benefit the ECI task.

Nevertheless, a potential issue of the above method is that a KB is never complete [Min *et al.*, 2013]; Especially, a KB may lack definitions of newly emerging events. To mitigate this problem, we propose a complementary *mention masking reasoner*, aiming to exploit the event-agnostic clues for reasoning. We motivate our approach by noting that causal statements usually contain event-independent patterns, which are helpful for identifying causality of unseen events. For example, we can distill a causality pattern: "*The* [SLOT] *generates* [SLOT] ... " from S1, which can be used to identify **traffic congestion** $\xrightarrow{\text{cause}}$ **environmental pollution** in a new sentence "The **traffic congestion** generates **environmental pollution** and economic loss". To learn such context-specific patterns, we propose a learning mechanism called event mention masking generalization, which explicitly excludes event information for learning. Methodologically, it replaces event mentions with a placeholder symbol [MASK], and force our model to make predictions based on such mask-containing texts (as shown in Figure 1 b)). This can be seen as adding a hard attention on context information, and thus enhance the ability of our model in handling unseen cases.

Lastly, we build an *attentive sentinel* to allow a trade-off between the aforementioned two components. This trade-off is crucial because in some cases text context should override the background knowledge for the task, and in other cases the opposite is true (For example, despite the sentence "Both of **earthquake** and **tsunami** are natural disasters" contains an event pair of **earthquake** and **tsunami**, it does not express a causal relation according the context).

In experiments, we evaluate our model on three benchmark datasets. We first concern the standard evaluation and show that our model attains state-of-the-art performance. We then estimate the generalization ability of our model by performing i) cross-topic adaptation, ii) exploring unseen predicates, and iii) cross-task adaptation. Our model demonstrates definite advantages over previous methods.

To summarize, we make the following contributions:

- We propose a new approach for ECI, which can leverage external knowledge to enrich representations of events for accurate reasoning. To our best knowledge, this is the first work explicitly introducing external knowledge for this task.
- Moreover, we propose a mention masking generalization mechanism to learn event-agnostic, context-specific patterns. This grants our model a decent generalization ability to handle new, previously unseen data.
- We conduct extensive experiments and show that our model sets up a new state-of-the-art for ECI. Moreover, our approach shows definite advantages over previous ECI methods regarding generalization evaluation.

2 Related Work

2.1 Event Causality Identification

The task of ECI aims to identify causal relations of events in texts, which has attracted a lot of interests among researchers. Earlier methods for ECI are predominantly feature-based, which adopt lexical and syntactic features [Hashimoto *et al.*, 2014; Gao *et al.*, 2019], causality cues (such as "because", "for") [Riaz and Girju, 2014], event co-occurance patterns [Beamer and Girju, 2009; Hu *et al.*, 2017], temporal patterns [Mirza, 2014a; Ning *et al.*, 2018], and others for the task. The very recent work [Kadowaki *et al.*, 2019] employs BERT architecture [Devlin *et al.*, 2019], which can learn context-dependent representations for the task and achieves superior performance. Regarding datasets construction, Do *et al.* [2011] annotated a corpus consisting of 25 documents for evaluation; Mirza [2014a] annotated event causal relations in the TempEval-3 corpus and release a corpus called CausalTimeBank; Caselli and Vossen [2017] had built a corpus called EventStoryLine, which contains 258 documents in total. Hashimoto [2019] exploited weakly supervised method to construct ECI datasets. However, as introduced in Introduction, previous methods typically train a model on the annotated examples only and disregard a lot of background knowledge. Moreover, they generally have difficulty in handling new, previously unseen data, owing the limited size of training data.

2.2 Knowledge Enhanced Text Understanding

The importance of background knowledge in text understanding has long been recognized [Minsky, 1974]. With the development of knowledge bases (KBs) — ranging from manually annotated networks like WordNet [Miller, 1995] to semi-automatically/automatically constructed knowledge graphs like DBPedia [Lehmann *et al.*, 2014] and ConceptNet [Speer *et al.*, 2017] — large amounts of knowledge become available. Many studies have investigated to leverage such knowledge to boost text understanding tasks. To name a few, Rahman and Ng [2011] studies knowledge-enhanced entity co-reference; Yang and Mitchell [2017] took advantage of external KBs to improve recurrent neural networks for entity recognition and event detection; Zhou *et al.* [2018] studied incorporating commonsense knowledge for conversation generation. But to the best of our knowledge, no work has studied introducing external knowledge for ECI.

3 Approach

Figure 2 schematically visualizes our approach. Specifically, we formulate ECI as a binary classification problem, following previous works [Mirza, 2014a; Ning *et al.*, 2018; Gao *et al.*, 2019] — for every pair of events in a sentence, we predict whether a causal relation holds. Our approach contains three major components:

- Knowledge-aware reasoner, which retrieves background knowledge from CONCEPTNET, and then integrate the knowledge with texts for reasoning (§ 3.1).
- Event masking reasoner, which masks event mentions in texts, aiming to learn event-agnostic, context-specific patterns for reasoning (§ 3.2).
- The attentive sentinel, which adopts an attention mechanism to balance the above two components for the final prediction (§ 3.3).

We illustrate each component in details in the followings.

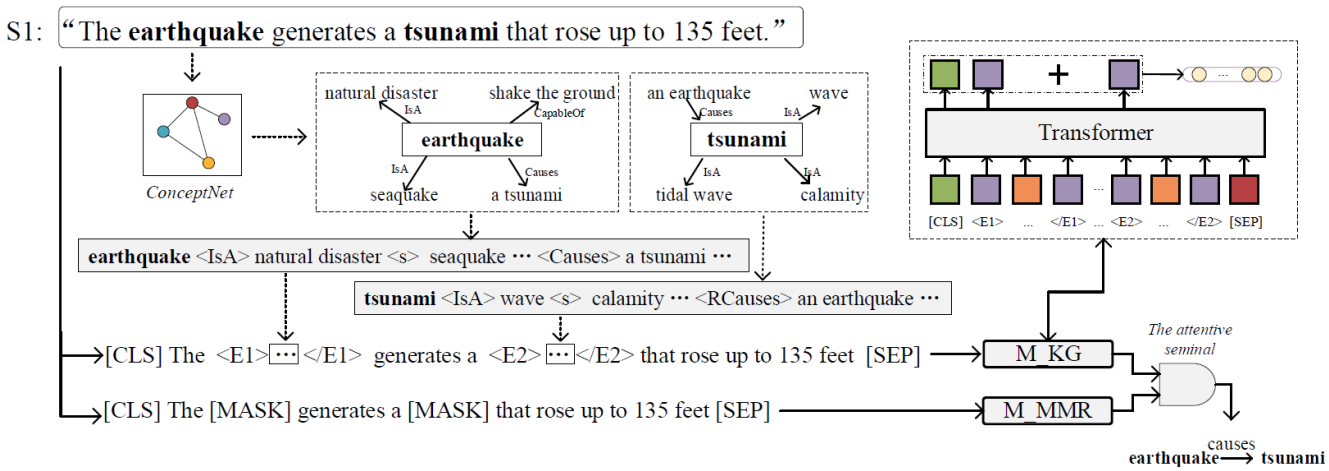

Figure 2: The overview of our approach, which consists of three major components: 1) M_KG, the knowledge aware reasoner; 2) M_MMR, the mention masking reasoner; and 3) the attentive sentinel trading off between the two modules.

3.1 Knowledge-Aware Reasoner

Given a pair of events (denoted as e1 and e2), the knowledge aware reasoner first retrieves the related knowledge in CONCEPTNET, and then encodes the knowledge into contexts for reasoning.

Knowledge Retrieving. CONCEPTNET structures knowledge as graph, where each node corresponds a concept, and each edge corresponds to a semantic relation. For e1 and e2, we search their definitions in CONCEPTNET but we only consider 18 semantic relations that are potentially useful for ECI: CapableOf, IsA, HasProperty, Causes, MannerOf, CausesDesire, UsedFor, HasSubevent, HasPrerequisite, NotDesires, PartOf, HasA, Entails, ReceivesAction, UsedFor, CreatedBy, MadeOf, and Desires. Part of the knowledge related to **earthquake** and **tsunami** in S1 is shown in Figure 2.

Knowledge Encoding. To encode the knowledge and enrich representations of e1 and e2, we first conduct knowledge linearization, to transfer the discrete knowledge into a structured sequence, motivated by [Fan *et al.*, 2019]. As shown in Figure 2, for each semantic relation, one special marker (such as $\langle IsA \rangle$) is designed and following the marker is the related knowledge separated by a delimiter $\langle s \rangle$. Then, we adopt a BERT based encoder to encode the knowledge with the context texts jointly. Specifically, we first incorporate the linearized knowledge into the sentence; then we add additional event markers ⟨E1⟩, ⟨/E1⟩ and ⟨E2⟩ ⟨/E2⟩ to indicate boundaries of events (Two special tokens [CLS] and [SEP] are added at the beginning/ending of the sentence following BERT). Finally, after using BERT encoder to compute representations of the entire sequence, we concatenate representations of [CLS], ⟨E1⟩, and ⟨E2⟩ as the final representation regarding to (e1, e2), namely

$$F_{KG}^{(e1,e2)} = h_{[CLS]} \oplus h_{\langle E1 \rangle} \oplus h_{\langle E2 \rangle} \quad (1)$$

where $\oplus$ indicates the concatenation operator; $h_{[CLS]}$, $h_{\langle E1 \rangle}$, and $h_{\langle E2 \rangle}$ are representations of [CLS], ⟨E1⟩, and ⟨E2⟩ respectively. $F_{KG}^{(e1,e2)}$ is the knowledge-aware representation that would be used for further computation.

3.2 Mention Masking Reasoner

The mention masking reasoner aims to explore event-agnostic, context-specific patterns for reasoning. Specifically, e1 and e2 are firstly replaced with a special token '[MASK]' to exclude event information. Then, another BERT encoder is adopted to encode the mask-containing sentence ([CLS] and [SEP] are also added). Similar as in the knowledge aware reasoner, we regard $F_{MASK}^{(e1,e2)}$ as the masked representation of (e1, e2):

$$F_{MASK}^{(e1,e2)} = h_{[cls]} \oplus h_{[MASK]}^{e1} \oplus h_{[MASK]}^{e2} \quad (2)$$

where $h_{[MASK]}^{e1}$ and $h_{[MASK]}^{e2}$ are BERT representations of e1 and e2, which have been replaced by [MASK]. We train the mention masking reasoner with two different objectives:

Discrimination Learning. Our model is forced to predict whether e1 and e2 forms a causal relation based on the masked representation $F_{MASK}^{(e1,e2)}$. As $F_{MASK}^{(e1,e2)}$ does not contain any event-specific information, our model has to explore context-specific clues for reasoning, which would gain the ability to tackle unseen events.

Distributional Similarity Learning. In distributional similarity learning, we assume causal statements may share similar representations in some ways, by taking in pair of mask-containing statements as inputs and encouraging their representations to be similar if both of them express causal relations. Assume A and B are two pairs of events; F_{MASK}^{A} and F_{MASK}^{B} are their masked representation. We optimize the following loss to achieve distributional similarity:

$$\begin{aligned} L = &- \delta_{A,B} * log(p(l=1|A,B)) \\ &+ (1-\delta_{A,B}) * log(1-p(l=1|A,B)) \end{aligned} \quad (3)$$

where $\delta_{A,B}$ is the Kronecker function that take the values of 1 when both of A and B express a causal relation and 0 otherwise. $p(l=1|A,B) = \frac{1}{1+exp({F_{MASK}^{A}}^{T} F_{MASK}^{B})}$ defines the distributional similarity score. In practice, we alternately adopt discrimination learning and distributional similarity learning to train the mention masking reasoner.

3.3 The Attentive Sentinel

The attentive sentinel aims to learn a trade-off between the knowledge aware reasoner and the mention masking reasoner, by learning an attentive gate as their combination weights, namely:

$$g_{e_1,e_2} = \sigma(W(F_{KG}^{(e1,e2)} \oplus F_{MASK}^{(e1,e2)}) + b) \quad (4)$$

where W and b are model parameters; $\oplus$ denotes the concatenation operator. Then it adopts an weighted summation to integrate $F_{KG}^{(e1,e2)}$ and $F_{MASK}^{(e1,e2)}$ as the final feature for (e1, e2), namely:

$$F_{e_1,e_2} = g_{e_1,e_2} * F_{KG}^{(e1,e2)} + (1 - g_{e_1,e_2}) * F_{MASK}^{(e1,e2)} \quad (5)$$

The attentive sentinel allows to balance the knowledge aware reasoner and the mention masking reasoner to make the final prediction.

3.4 Model Prediction and Training

To make the final prediction, we perform a binary classification by taking F_{e_1,e_2} as input:

$$o_{e_1,e_2} = \sigma(W_o F_{e_1,e_2} + b_o) \quad (6)$$

where o_{e_1,e_2} denotes the probability of e1 $\xrightarrow{\text{cause}}$ e2; W_o and b_o are model parameters. For training, we adopt cross-entropy as the loss function:

$$J(\Theta) = -\sum_{s} \sum_{\substack{e_i,e_j \in E_s \\ e_i \neq e_j}} y_{e_i,e_j} \log(o_{e_i,e_j}) + \quad (7)$$

$$(1 - y_{e_i,e_j}) \log(1 - o_{e_i,e_j}) \quad (8)$$

where Θ denotes the parameter set of our model; s ranges over each sentence in the training set; e_i, and e_j ranges over each events in s. We adopt the Adam [Kingma and Ba, 2015] algorithm to optimize model parameters.

4 Experiments

4.1 Experimental Setups

Datasets and Evaluations. Our experiments are conducted on three benchmark datasets, including: a) EventStroryLine [Caselli and Vossen, 2017], which contains 258 documents in 22 topics, 5,334 events in total, and 1,770 of 7,805 event pairs are causally related; b) Causal-TimeBank [Mirza *et al.*, 2014], which contains 184 documents, 6,813 events, and 318 of 7,608 event pairs are causally related. c) EventCausality [Do *et al.*, 2011; Ning *et al.*, 2018], which contains 25 documents, 1,134 events, and 414 of 887 event pairs are causally related. For evaluation, we adopt Precision (P), Recall (R) and F1-score (F1) as evaluation metrics, same as previous methods to ensure comparability. Significant test is conducted using paired t-test at a significance level of 0.05.

Implementations. In our implementations[1], both the knowledge aware reasoner and the mention masking reasoner are implemented as BERT-Large architecture, which has 24-layer, 1024-hidden, and 16-heads. We use CONCEPTNET 5.0 as the KB. Regarding hyper-parameters, the batch size is set as 10, and the learning rate is initialized as 5×10^{-5} with a linear decay. We also adopt a negative sampling rate of 0.5 for training, owing to the sparseness of positive examples.

Baseline Systems. We prefer different baseline systems to compare for different datasets. For EventStoryLine, we prefer OP [Caselli and Vossen, 2017], a dummy model assigns causal relation to every event pair; 2) LSTM [Cheng and Miyao, 2017], a dependency path based sequential model that models the context between events to identify causality; 3) Seq [Choubey and Huang, 2017], a sequence model explore complex human designed features for the task. 4) LR+, and 5) LIP [Gao *et al.*, 2019], state-of-the art ECI system that adopts document structure for the task. For Causal-TimeBank, we prefer 1) RB, a rule-based system; 2) ML, a machine learning based model, and 3) HB, a hybrid method combine rules with features for comparison. These models are designed by [Mirza, 2014a; Mirza and Tonelli, 2016] for ECI. For EventCausality, we prefer PMI, ECD, and CEA [Do *et al.*, 2011], which adopt different co-occurrence patterns for the task as baselines systems. For each dataset, we add a BERT based model as baseline.

In our approach, we use M_{KG} to denote the knowledge-aware reasoner, which adopts $F_{KG}^{(e1,e2)}$ for prediction; we denote M_{MMR} as the mention masking reasoner, which adopts $F_{MASK}^{(e1,e2)}$ for prediction. M_{FULL} indicates our full model.

4.2 Experimental Results

Experimental results on the three benchmark datasets follow.

EventStoryLine. Table 1 shows the results on EventStoryLine, where we use the last two topics as development set, and conduct a 5-fold cross-validation on the rest 20 topics, as suggested by [Gao *et al.*, 2019]. From the results, our full model M_{FULL} outperforms all baseline methods and achieves the best performance (50.1% on F1 score), outperforming the state-of-the-art model LIP by a margin of 5.5%, which justifies its effectiveness. Comparing with M_{KG} with BERT, we note adding external knowledge improves the performance by 3.6% in F1 score. Moreover, the mention masking reasoner (M_{MMR}) is more effective than the knowledge aware reasoner (M_{KG}) (43.9% v.s. 41.8%). This may imply that, for small dataset, generalization knowledge is more important than discrimination knowledge.

Causal-TimeBank. Table 2 shows results on Causal-TimeBank, where we adopt two settings: 1) 10-fold cross-validation (CV) as in [Mirza, 2014a], and 2) evaluating on an additional TemporalEval-3 datasets as in [Mirza and Tonelli, 2016]. Our models show a consistence performance as in EventStoryLine, which achieves the best performance (44.1% for CV, and 66.7% for TE). Moreover, M_{FULL} and M_{MASK} demonstrate high recall value, which is benefited from their generalization ability.

EventCausality. Table 3 shows results on EventCausality. Note this is an extremely tiny datasets, and [Do *et al.*, 2011] adopts a weakly-supervised methods to retrieve additional examples for training. Nevertheless, in our approach, we use

[1] https://github.com/jianliu-ml/EventCausalityIdentification

METHODS	PRE.	REC.	F1
OP [Caselli and Vossen, 2017]	22.5	98.6	36.6
LSTM [Cheng and Miyao, 2017]	34.0	41.5	37.4
Seq [Choubey and Huang, 2017]	32.7	44.9	37.8
LR+ [Gao *et al.*, 2019]	37.0	45.2	40.7
LIP [Gao *et al.*, 2019]	38.8	52.4	44.6
BERT	37.9	38.5	38.2
M_{KG} (Ours)	**44.5**	39.3	41.8
M_{MMR} (Ours)	37.6	52.6	43.9
M_{FULL} (Ours)	41.9	**62.5**	**50.1***

Table 1: Results on EventStoryLine. Pre., Rec. and F1 indicate precision (%), recall (%) and F1-score (%) respectively; Bold denotes best results; * denotes a significant test at the level of 0.05.

	METHOD	PRE.	REC.	F1
CV	Rule-based [Mirza, 2014b]	36.8	12.3	18.4
	Data-driven [Mirza, 2014a]	**67.3**	22.6	33.9
	BERT	30.3	41.1	34.9
	M_{KG} (Ours)	38.7	44.4	41.3
	M_{MMR} (Ours)	31.1	51.9	38.8
	M_{FULL} (Ours)	36.6	**55.6**	**44.1***
TE	RB [Mirza and Tonelli, 2016]	**91.7**	42.3	57.9
	ML [Mirza and Tonelli, 2016]	42.9	11.5	18.2
	HB [Mirza and Tonelli, 2016]	73.7	53.8	62.2
	BERT	56.6	42.3	48.4
	M_{KG} (Ours)	50.0	57.7	53.5
	M_{MMR} (Ours)	52.8	**73.1**	61.2
	M_{FULL} (Ours)	61.3	**73.1**	**66.7***

Table 2: Results on Causal-TimeBank. CV denotes 10-fold cross-validation. TE denotes evaluating on a manually Temporal Eval-3 datasets. Pre., Rec. and F1 indicate precision (%), recall (%) and F1-score (%) respectively. Bold denotes best results; * denotes a significant test at the level of 0.05.

only 10 documents ([Do *et al.*, 2011] use them as seed documents) for training. From the results, the superior performance of M_{FULL} (45.4% on F1) demonstrates the applicability of our approach for small datasets.

5 Model Generalization Evaluation

Generalization refers to a model's ability to adapt to new, previously unseen data. We conduct 1) cross-topic adaptation, 2) unseen predicates, and 3) cross-task adaptation to estimate the generalization ability of our model.

5.1 Cross-Topic Adaptation

Different topics usually involve different events. In our cross-topic adaptation, we train our model on a source topic, but test our model on other topics. We use EventStoryLine to conduct our experiments. Specifically, we first randomly select a topic as the source topic (for model training and tuning), and then we rank the remaining topics based on their similarities with the source topic (The similarity of two topics $t1$ and $t2$ is defined as $\frac{E_{t1} \cap E_{t2}}{E_{t1} \cup E_{t2}}$, where E_t denotes the event set of t). Finally, we test how our model performs on topics with the lowest, medium, and highest similarity value with the source topic. We re-implement previous state-of-the-art system LIP to compare with our models.

METHODS	PRE.	REC.	F1
PMI [Do *et al.*, 2011]	26.6	20.8	23.3
ECD PMI [Do *et al.*, 2011]	40.9	23.5	29.9
CEA [Do *et al.*, 2011]	**62.2**	28.0	38.6
BERT	16.8	30.7	21.7
M_{BERT} (Ours)	17.2	68.2	27.5
M_{MMR} (Ours)	20.7	**77.3**	32.6
M_{FULL} (Ours)	34.1	68.2	**45.4***

Table 3: Results on EventCausality datasets. Pre., Rec. and F1 indicate precision (%), recall (%) and F1-score (%) respectively. Bold denotes best results; * denotes a significant test at the level of 0.05.

SET.	ST → TT (δ)	LIP	M_{KG}	M_{MMR}	M_F
Low	T8 → T35 (0.0%)	2.8	17.6	29.7	44.7
	T13 → T12 (0.0%)	-	6.0	20.4	25.1
	T18 → T30 (0.0%)	-	-	10.3	19.5
Med.	T8 → T3 (1.7%)	6.7	22.1	24.9	30.9
	T13 → T41 (0.1%)	4.5	12.1	20.7	28.6
	T18 → T35 (2.8%)	17.1	40.4	38.4	44.5
High	T8 → T19 (12.4%)	19.4	42.7	42.8	45.1
	T13 → T14 (17.1%)	27.4	44.4	42.7	46.0
	T18 → T33 (27.2%)	32.2	45.3	44.1	49.0

Table 4: Results (F1 score (%)) of cross-topic adaptation. ST → TT (δ) denotes that the source topic is ST, and the target topic is TT, and their similarity is δ. M_F denotes our full model.

From the results in Table 4, the performance of LIP is highly depended on the similarity of source and target topics. It achieves relative good performance when the target and source topics are of high-similarity, but behaves extremely poorly when the target and source topics are of low-similarity. While our approach, especially M_{MMR} and M_{FULL}, are robust in cross-topic adaptation, which achieve superior performance even in low-similarity cases.

5.2 Unseen Predicates

To further test the generalization ability of our model, we conduct experiments to explore unseen predicates. For the EventStoryLine corpus, we first randomly select 1/3 of documents as the training set. Then, we divide the remaining corpus as 1) 'Both Seen' set, where both of events exist in the training data (with a size of 3,464); 2) 'One Unseen' set, where only one of the events exists in the training data (with a size of 4,381); 3) 'Both Unseen' set, where both events are unobserved during training (with a size of 1,891). From the results in Figure 3, 1) LIP behaves relatively good on 'Both Seen', but poorly on 'One Unseen' and 'Both Unseen' (only 11.3% in F1). 2) Our full model achieves the best performance on all of the three sets. 3) M_{MMR} achieves better performance than M_{KG} on 'One Unseen' and 'Both Unseen'.

5.3 Cross-Task Adaptation

Finally, we investigate cross-task adaptation, where we train our model on ECI datasets but test the performance of our

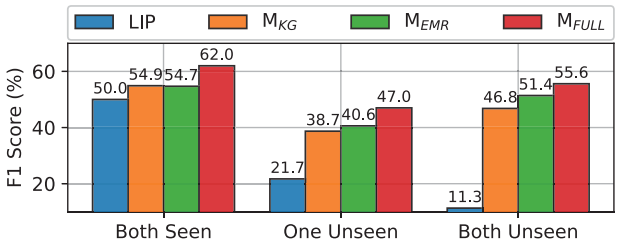

Figure 3: Results (F1 score (%)) of unseen predicates. 'Both Seen' indicates that both of events exist in the training data; 'One Unseen' indicates that only one of the events exists in the training data; 'Both Unseen' indicates that both events are unobserved during training.

DATASETS	METHODS	PRE.	REC.	F1
SemEval	LIP [Gao *et al.*, 2019]	24.6	21.1	22.8
	$\mathrm{M_{KG}}$ (Ours)	**63.5**	55.2	59.1
	$\mathrm{M_{MMR}}$ (Ours)	37.7	**89.3**	52.9
	$\mathrm{M_{FULL}}$ (Ours)	59.4	75.0	**66.0**
FrameNet	LIP [Gao *et al.*, 2019]	10.5	11.8	11.1
	$\mathrm{M_{KG}}$ (Ours)	64.6	13.5	22.0
	$\mathrm{M_{MMR}}$ (Ours)	**85.9**	57.0	68.5
	$\mathrm{M_{FULL}}$ (Ours)	84.4	**60.3**	**70.3**

Table 5: Results of cross-task adaptation. The model is trained/tuned on EventStoryLine. Pre., Rec. and F1 indicate precision (%), recall (%) and F1-score (%) respectively.

model in generalizing to other tasks. Specifically, we train our model on EventStoryLine, but we test our model on identifying causal relations in SemEval-8 (which focuses on causal relations between entities) and FrameNet (which focuses on causal relations between frame elements). From the results in Table 5, LIP performs relatively poor in cross-task adaptation. The reason might be that features adopted by LIP are not applied to entities and frame elements. $\mathrm{M_{KG}}$ behaves better than $\mathrm{M_{MMR}}$ in SemEval, but much worse than $\mathrm{M_{MMR}}$ in FrameNet. The reason is that, SemEval focus on relations between entities, which are more likely to have definitions on a KG. But FrameNet focuses on frame elements, which can be any span of the sentence, and do likely to have definitions on a KG. Our full model achieves the best performance among all models regarding cross-task adaptation.

6 Further Discussion

Inductive Bias. To further explore the effectiveness of our model, we investigate the prediction bias of $\mathrm{M_{KG}}$ and $\mathrm{M_{MMR}}$ by inspecting their outputs. Accordingly, there are 685 cause relations only predicted by $\mathrm{M_{KG}}$, 655 relations only by $\mathrm{M_{MMR}}$ and 382 relations by both of them in the experiments shown in Table 1 (for a specific fold). The values change to 102, 132 and 58 in the experiment of cross-topic adaptation (T18→T33). The relatively less of their common predictions indicate that $\mathrm{M_{KG}}$ and $\mathrm{M_{MMR}}$ focus on different aspects of features to identify the cause relations and they share complementary effects. This provides explanation for the good performance of our full model.

EXAMPLES	$\mathrm{M_{KG}}$	$\mathrm{M_{MMR}}$	$\mathrm{M_F}$
a) ... has confessed to **killing** a pregnant mom, who **died** on ...	✓	×	✓
b) his half-brother, ..., is also **on trial** for **murder**.	✓	✓	✓
c) A gang member was **convicted** Tuesday for **claiming the life** of a mother of ...	×	✓	✓
d) Horton was **struck** by a stray bullet as lopez **targeted** gang members ...	×	✓	✓
e) ... Carrasquillo allegedly **ordered** Lopez to **shoot** ...	×	×	✓

Table 6: Results of case study where bold denotes the two event pair. ✓and × denote a correct and incorrect prediction respectively.

Case Study. We conduct case study to further investigate the effectiveness of our model. To simplify the discussion, we limit the experiments to a specific cross-topic adaptation, i.e., T18→T33 adaptation. Table 6 shows several cases showing the outputs of $\mathrm{M_{KG}}$ and $\mathrm{M_{MMR}}$. Basically, $\mathrm{M_{KG}}$ is good at finding commonsense causality that is usually context-independent, such as **killing** $\xrightarrow{\text{cause}}$ **die** in a), and **murder** $\xrightarrow{\text{cause}}$ **on trial** in b), but cannot handle context depended cases as in c), d), and e). While $\mathrm{M_{MMR}}$ is completely opposite. The full model can take advantage of $\mathrm{M_{KG}}$ and $\mathrm{M_{MMR}}$ to make a more accurate prediction.

7 Conclusion and Future Work

In this paper, we propose a new approach for event causality identification. Our approach on the one hand can leverage background knowledge to enhance the reasoning, on the other hand cam mine event-agnostic context-specific patterns for reasoning, which greatly enhances its generalization ability. The effectiveness of our model is verified on three datasets with diverse settings. In the future, we would like to apply our model to other NLP tasks such as relation classification, event temporal relation extraction and others.

Acknowledgments

This work is supported by the Natural Key R&D Program of China (No.2018YFB1005100), the National Natural Science Foundation of China (No.61922085, No.U1936207, No.61976211, No.61806201) and the Key Research Program of the Chinese Academy of Sciences (Grant NO. ZDBS-SSW-JSC006). This work is also supported by Beijing Academy of Artificial Intelligence (BAAI2019QN0301), CCF-Tencent Open Research Fund and independent research project of National Laboratory of Pattern Recognition.

References

[Beamer and Girju, 2009] Brandon Beamer and Roxana Girju. Using a bigram event model to predict causal potential. In *COLING*, pages 430–441, 2009.

[Berant *et al.*, 2014] Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D. Manning. Modeling biological processes for reading comprehension. In *EMNLP*, pages 1499–1510, 2014.

[Caselli and Vossen, 2017] Tommaso Caselli and Piek Vossen. The event StoryLine corpus: A new benchmark for causal and temporal relation extraction. In *ACL Workshop*, pages 77–86, 2017.

[Cheng and Miyao, 2017] Fei Cheng and Yusuke Miyao. Classifying temporal relations by bidirectional LSTM over dependency paths. In *ACL*, pages 1–6, 2017.

[Choubey and Huang, 2017] Prafulla Kumar Choubey and Ruihong Huang. A sequential model for classifying temporal relations between intra-sentence events. In *EMNLP*, pages 1796–1802, 2017.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019.

[Do *et al.*, 2011] Quang Do, Yee Seng Chan, and Dan Roth. Minimally supervised event causality identification. In *EMNLP*, pages 294–303, 2011.

[Fan *et al.*, 2019] Angela Fan, Claire Gardent, Chloé Braud, and Antoine Bordes. Using local knowledge graph construction to scale Seq2Seq models to multi-document inputs. In *EMNLP*, pages 4186–4196, 2019.

[Gao *et al.*, 2019] Lei Gao, Prafulla Kumar Choubey, and Ruihong Huang. Modeling document-level causal structures for event causal relation identification. In *NAACL*, pages 1808–1817, 2019.

[Girju, 2003] Roxana Girju. Automatic detection of causal relations for question answering. In *ACL Workshop*, pages 76–83, 2003.

[Hashimoto *et al.*, 2014] Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, Motoki Sano, István Varga, Jong-Hoon Oh, and Yutaka Kidawara. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In *ACL*, pages 987–997, 2014.

[Hashimoto, 2019] Chikara Hashimoto. Weakly supervised multilingual causality extraction from Wikipedia. In *EMNLP*, pages 2988–2999, 2019.

[Hu *et al.*, 2017] Zhichao Hu, Elahe Rahimtoroghi, and Marilyn Walker. Inference of fine-grained event causality from blogs and films. In *ACL Workshop*, pages 52–58, 2017.

[Kadowaki *et al.*, 2019] Kazuma Kadowaki, Ryu Iida, Kentaro Torisawa, Jong-Hoon Oh, and Julien Kloetzer. Event causality recognition exploiting multiple annotators' judgments and background knowledge. In *EMNLP*, pages 5816–5822, 2019.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[Lehmann *et al.*, 2014] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, (2):167–195, 2014.

[Miller, 1995] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 1995.

[Min *et al.*, 2013] Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek. Distant supervision for relation extraction with an incomplete knowledge base. In *NAACL*, pages 777–782, 2013.

[Minsky, 1974] Marvin Minsky. A framework for representing knowledge. Technical report, USA, 1974.

[Mirza and Tonelli, 2016] Paramita Mirza and Sara Tonelli. CATENA: CAusal and TEmporal relation extraction from NAtural language texts. In *COLING*, pages 64–75, 2016.

[Mirza *et al.*, 2014] Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. Annotating causality in the TempEval-3 corpus. In *EACL Workshop*, pages 10–19, 2014.

[Mirza, 2014a] Paramita Mirza. Extracting temporal and causal relations between events. In *ACL Workshop*, pages 10–17, 2014.

[Mirza, 2014b] Paramita Mirza. Fbk-hlt-time : a complete italian temporal processing system for eventi-evalita 2014. In *EVALITA 2014*, pages 44–49, 2014.

[Ning *et al.*, 2018] Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. Joint reasoning for temporal and causal relations. In *ACL*, pages 2278–2288, 2018.

[Oh *et al.*, 2016] Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Ryu Iida, Masahiro Tanaka, and Julien Kloetzer. A semi-supervised learning approach to why-question answering. In *AAAI*, page 3022–3029, 2016.

[Rahman and Ng, 2011] Altaf Rahman and Vincent Ng. Coreference resolution with world knowledge. In *ACL*, pages 814–824, 2011.

[Riaz and Girju, 2014] Mehwish Riaz and Roxana Girju. In-depth exploitation of noun and verb semantics to identify causation in verb-noun pairs. In *SIGDIAL*, pages 161–170, 2014.

[Speer *et al.*, 2017] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, page 4444–4451, 2017.

[Yang and Mitchell, 2017] Bishan Yang and Tom Mitchell. Leveraging knowledge bases in LSTMs for improving machine reading. In *ACL*, pages 1436–1446, July 2017.

[Zhou *et al.*, 2018] Hao Zhou, Tom Yang, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI-ECAI*, pages 4623–4629, 2018.