Identify Event Causality with Knowledge and Analogy

Sifan Wu[1], **Ruihui Zhao**[2], **Yefeng Zheng**[2], **Jian Pei**[3], **Bang Liu**[1]*

[1] RALI & Mila, University of Montreal
[2] Tencent Jarvis Lab
[3] Duke University
sifan.wu@umontreal.ca, zachary@ruri.waseda.jp, yefengzheng@tencent.com,
j.pei@duke.edu, bang.liu@umontreal.ca

Abstract

Event causality identification (ECI) aims to identify the causal relationship between events, which plays a crucial role in deep text understanding. Due to the diversity of real-world causality events and difficulty in obtaining sufficient training data, existing ECI approaches have poor generalizability and struggle to identify the relation between seldom seen events. In this paper, we propose to utilize both external knowledge and internal analogy to improve ECI. On the one hand, we utilize a commonsense knowledge graph called ConceptNet to enrich the description of an event sample and reveal the commonalities or associations between different events. On the other hand, we retrieve similar events as analogy examples and glean useful experiences from such analogous neighbors to better identify the relationship between a new event pair. By better understanding different events through external knowledge and making an analogy with similar events, we can alleviate the data sparsity issue and improve model generalizability. Extensive evaluations on two benchmark datasets show that our model outperforms other baseline methods by around 18% on the F1-value on average.

Introduction

Event causality identification (ECI) is an important task in natural language processing (NLP) which aims to identity the causal relationships between events in text pieces, i.e., predict whether one event causes another one to happen. The term "event" is used as a cover term to refer to any situations that can happen, occur, or hold, which is a synonym to "eventuality" introduced by (Bach 1986) for covering both dynamic and static situations. With a better understanding of event causality, ECI can help with various NLP applications, such as question answering (Oh et al. 2016), machine reading comprehension (Berant et al. 2014), and logical reasoning (Ding et al. 2019; Hashimoto 2019).

Figure 1 shows an example to illustrate the task of ECI. Given two sentences "An *earthquake* ... *killing* 10 people, officials said." and "The U.S. ... a magnitude-6.1 *temblor*.", an ECI system needs to identify the causal relationships between mentioned events in the texts, such as "*killing*" and "*temblor*". Specifically, while most existing researches focus on sentence-level ECI which only predicts

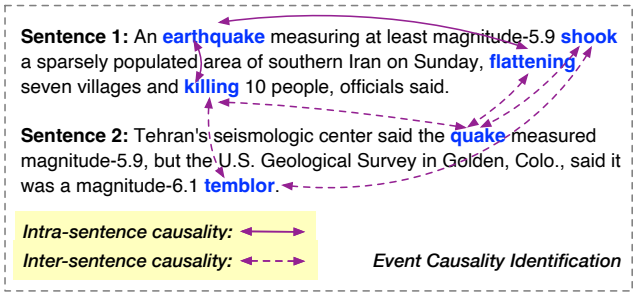

Figure 1: An example of ECI. Each double arrow line indicates that there is a causal relationship between the two noted events.

the intra-sentence causality between two events mentioned in the same sentence, here we aim to identify both sentence-level and document-level ECI (DECI) to predict both intra-sentence and inter-sentence event causality. It is also worth noting that although causal relationships are directed, we omit the causal directions following the same settings as prior research works (Zuo et al. 2021a; Cao et al. 2021).

Identifying event causality is challenging due to several reasons. First, existing datasets for ECI are relatively small and imbalanced. For example, the largest widely used public ECI dataset is the EventStoryLine Corpus (Caselli and Vossen 2017), which contains 258 documents consisting of 4,316 sentences, and only 1,770 out of 7,805 event pairs are annotated as causal relations. This situation poses challenges to existing data-hungry deep learning-based approaches for ECI tasks (Zuo et al. 2020; Cao et al. 2021; Liu, Chen, and Zhao 2020), which mainly utilize language models to model sentence context and treat the ECI task as a binary classification problem. Therefore, how to efficiently utilize limited data is one essential problem to be solved for ECI. Second, event mentions in texts are usually short and lack explicit definitions or descriptions, making it difficult to learn a good representation for an event. Third, as there are diverse and enormous amount of events in real-world, how to improve the generalizability of ECI models on unseen events is a critical problem.

We propose to exploit both external knowledge and internal analogy for improving the representation and generalization abilities of ECI models. On the one hand, by in-

*Corresponding author. Canada CIFAR AI Chair.

troducing knowledge or commonsense about an event from external knowledge bases, we can enrich the description of the event mention in the text and reveal the correlations between different events. For example, in Fig. 1, "temblor" is an uncommon word which comes from a Spanish word meaning a trembling. We can hardly realize there is causal relation between "tremblor" and "killing". But with the help of knowing "temblor" is a synonym of "earthquake", we can identify the causality more easily. On the other hand, manipulating concepts and making analogies between them is considered as a core aspect of human intelligence (Hofstadter 1995). For example, the analogy between DNA and zipper or source code allows us to better understand the paired structure of nucleotides and the ability of DNA to encode information. Similarly, to better modeling and representing an unseen event, we can recap similar seen events in our memory and make analogies between them to better understand new events or event pairs and predict the causal relations.

Technically, we propose a two-stage Knowledge-Analogy Dual Enriched Representation (KADE) framework for ECI. In the first stage, we augment the event representations by retrieving relevant knowledge from ConceptNet (Speer, Chin, and Havasi 2017), a freely-available semantic network that include massive knowledge and common sense. By appending the relevant knowledge with the original texts and encode with BERT (Devlin et al. 2018), we can obtain knowledge-enriched representations of events. Such representations of the events are stored in a memory module during training for later usage. In the second stage, given an event, we compare its representation with other events in the memory to retrieve similar examples and making analogies between them. We also evaluate various ways to fuse the information of the analogy examples into the target event.

We conduct extensive experiments on the EventStoryLine dataset (Caselli and Vossen 2017) and the Causal-TimeBank dataset (Mirza and Tonelli 2014) to evaluate the performance of KADE and compare with baseline methods. The experimental results show that our method outperforms other SOTA baselines by at least 18% in terms of F1-value on both datasets. It is worth noting that our KADE framework is general and can be easily adapted to other NLP tasks. The code is open sourced to facilitate future research, which can be found here: **https://github.com/hihihihiwsf/KADE**.

Related Work

A wide range of approaches has been proposed for ECI. Early feature-based methods utilize different human-crafted features and resources to improve the performance, such as causality markers (Riaz and Girju 2014; Hidey and McKeown 2016), statistical co-occurrence of events (Beamer and Girju 2009; Hu, Rahimtoroghi, and Walker 2017), lexical patterns (Hashimoto 2019), or syntactic patterns (Mirza 2014). Such approaches rely on the domain knowledge of human. Deep learning-based approaches for ECI (Kadowaki et al. 2019; Zuo et al. 2020) leverage pretrained language models (e.g., BERT (Devlin et al. 2018)) and commonsense knowledge sources (e.g., ConceptNet (Speer, Chin, and Havasi 2017)) to improve the performance. To deal with implicit causal relations, Cao et al. (2021) conducts a descriptive graph induction module combining external knowledge and achieves promising results.

There are also research works aim to solve the data insufficiency problem. Zuo et al. (2021b) utilizes dual learning to generate task-related sentences for ECI. While data augmentation can alleviate data insufficiency to some extent, it still faces the data bias issue, which may cause the augmented data distribution be different from the original one. Besides, data augmentation-based ECI approaches usually enlarge the model size and make model less efficient. Document-level ECI (DECI) (Gao, Choubey, and Huang 2019; Phu and Nguyen 2021) further poses the new challenge of cross-sentence event causality identification. RichGCN (Phu and Nguyen 2021) constructs an interaction graph with heterogeneous edges from 6 information types, such as discourse-based and syntax-based edges. However, the structured representation is time-consuming to construct and contains redundant information which is useless for ECI tasks. The case-based model (Das et al. 2021) uses a neural retriever to retrieve other similar queries from a case memory, which can hardly solve ECI tasks. Overall, existing deep learning-based models cannot solve the data insufficiency problem of DECI efficiently. In this paper, we better utilize available datasets by analogy without generating noisy data.

K-nearest-neighbor (kNN) lookup is a widely-used technique for variety of machine learning tasks, especially combined with retrieval methods. Fan et al. (2021) retrieves related documents from external knowledge base to improve the performance for dialogue generation. Wu et al. (2022) integrates kNN module into Transformers to handle long context inputs. Memory-efficient Transformers (Gupta et al. 2021) replace dense attention with kNN lookup to increase speed and reduce memory usage. In our work, we retrieve analogy event examples with kNN to learn a better event representation for DECI.

Methodology

In this section, we formulate the task of ECI (including both sentence-level ECI and document-level DECI), and describe our proposed Knowledge-Analogy Dual Enriched Representation (KADE) framework for solving it.

Task definition. We formulate ECI as a binary classification problem following previous work (Phu and Nguyen 2021). Given two input sentences $S_1 = \{w_1, w_2, ..., w_{s_1}\}$, $S_2 = \{w_{s_1+1}, w_{s_1+2}, ..., w_{s_1+s_2}\}$ of length s_1 and s_2 respectively, the goal of DECI is to predict whether there exists a causal relationship between e_1 and e_2, where e_1 and e_2 represent two event mentions in the two sentences. For example, in Fig. 2, the two sentences contain e_1 *jolts* and e_2 *shook*, respectively.

KADE framework. As shown in Fig. 2, the framework of our proposed model mainly contains two stages: i) knowledge augmentation for incorporating commonsense knowledge into input sentences to better understand events; ii) analogy fusion for retrieve similar event examples from a memory of seen events and making analogies between them to better model unseen events. We will illustrate the two parts in detail in the following.

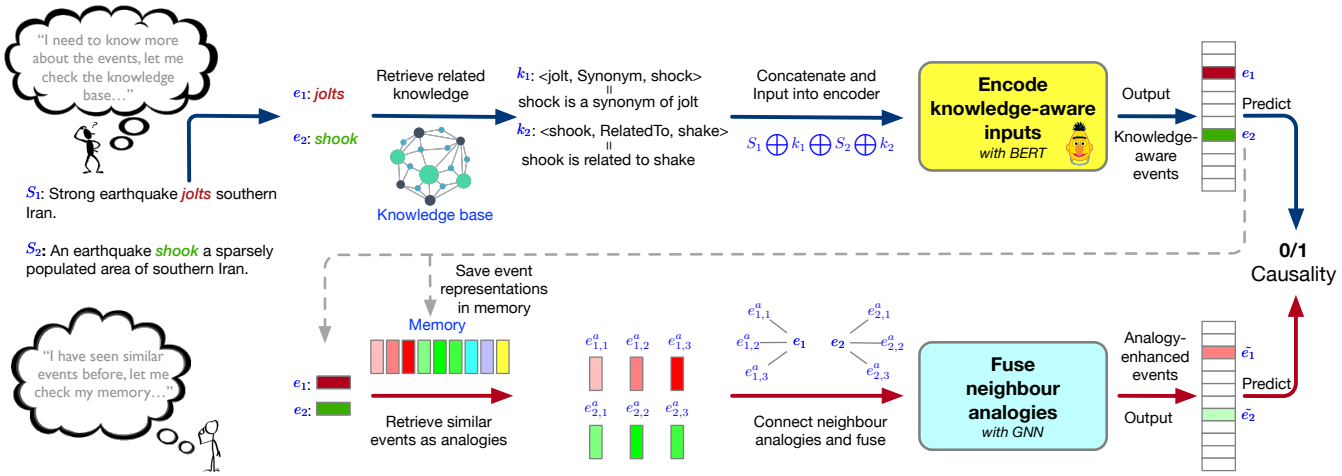

Figure 2: Illustration of the KADE framework for event causality identification. It consists of knowledge augmentation stage (upper part) and analogy fusion stage (lower part).

Knowledge-Enriched Event Representation

Human can identify event causalities not only by reading the input sentences but also by leveraging commonsense knowledge, which is important for ECI (Liu, Chen, and Zhao 2020). In our work, we exploit the relevant knowledge of events from ConceptNet (Speer, Chin, and Havasi 2017), which contains plentiful commonsense knowledge of various concepts. Specifically, to we only pay attention to 19 useful semantic relations for ECI which is the same as (Liu, Chen, and Zhao 2020).

To give an example, in Fig. 2, the input sentence S_1 "Strong earthquake *jolts* southern Iran." contains an event e_1 "*jolts*". We first extract the structural knowledge of e_1 from ConceptNet, then we construct a structured sequence to linearize the extracted knowledge, like k_1 "Shock is a synonym of jolt". After obtaining the linearized concept sentences k_1 and k_2 from ConceptNet for the event pair e_1 and e_2, we concatenate them to form knowledge-enriched input representation:

$$S_1^k = S_1 \bigoplus k_1, S_2^k = S_2 \bigoplus k_2, S_{1,2}^k = \quad S_1^k \bigoplus S_2^k, \tag{1}$$

where S_1^k and S_2^k denote knowledge-enriched sentences of S_1 and S_2, respectively. $S_{1,2}^k$ is the concatenation of them.

After obtaining knowledge-enriched input sentences $S_{1,2}^k$, we encode the input by BERT (Devlin et al. 2018) to learn a knowledge-aware representation for each event. For the convenience of notation, we still denote the knowledge-aware representation of events from BERT as e_1 and e_2.

Our KADE framework consists of two-stage training procedures. In the first training stage, we directly concatenate a classifier module on top of the BERT encoder to classify whether there is a causal relationship between e_1 and e_2. We train the model by the following cross-entropy loss:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_i [y_i \cdot log(p_i) + (1-y_i) \cdot log(1-p_i)], \tag{2}$$

where N is the number of event pairs, p_i is the prediction output. y_i is the ground-truth of i-th event pair, where $y_i = 0$ means there are no causality between e_1 and e_2, and vice versa.

After obtaining the knowledge-aware event representations from the first-stage training, we save the event representations in a memory module $\mathcal{M} = \{e_1, e_2, ..., e_{|\mathcal{M}|}\}$, which will play a key role in the second-stage training. Note that we only save the events from the training dataset.

Analogy-Enhanced Event Refinement

Although the knowledge-aware representations can better characterize the input events, they are still insufficient to model rare events and make full utilization of the limited training data. Therefore, in the second-stage, we further refine the event representations by retrieving similar seen events from the memory $\mathcal{M}$ and making analogies between a target event and the retrieved analogy examples.

Specifically, after we learned the representation of input event pairs (e_1, e_2) from the encoder, we look up the top-n similar representations from memory $\mathcal{M}$. We denote the retrieved analogy examples as $(e_{1,1}^a, ..., e_{1,n}^a)$ and $(e_{2,1}^a, ..., e_{2,n}^a)$, where $e_{1,i}^a$ means the i-th analogy example of e_1. Similarly for $e_{2,i}^a$. In our implementation, we utilize kNN to retrieve the similar events from memory and we set the number of neighbours as $n = 3$.

Next, we fuse the information of analogy examples into the knowledge-aware event representations to futher refine it. This can be done in various ways, and we consider two strategies in this work.

The first strategy is *mean fusion*. We can make an average of the n representations of the retrieved analogy examples and fuse it as follows:

$$\begin{aligned} \tilde{e_1} &= \alpha \cdot e_1 + (1-\alpha) \cdot Mean(e_{1,1}^a, ..., e_{1,n}^a), \\ \tilde{e_2} &= \alpha \cdot e_2 + (1-\alpha) \cdot Mean(e_{2,1}^a, ..., e_{2,n}^a), \end{aligned} \tag{3}$$

where α is a hyper-parameter.

The second strategy is *graph fusion*. Considering n similar events have similar semantics, we can construct a local graph between the retrieved events and the target event.

Algorithm 1: Two-stage Training of KADE

Input: Two sentences S_1 and S_2, event pairs $(e1, e2)$, ground truth label y, ConeptNet knowledge graph $\mathcal{KG}$. A memory $\mathcal{M}$. A pre-trained encoder $\mathcal{E}$, classifier $\mathcal{C}$, graph encoder $\mathcal{G}$, and the number of neighbors n to be retrieved.

Stage 1: Training the encoder and classifier with knowledge.
Enrich sentences S_1, S_2 to S_1^k, S_2^k by commonsense knowledge from $\mathcal{KG}$ as Equation (1).
for each event pair (S_1^k, S_2^k, e_1, e_2) in batch **do**
- Optimize the encoder $\mathcal{E}$ and the classifier $\mathcal{C}$ with loss as Equation (2).
- Save representation embeddings of the two events e_1 and e_2 to the memory $\mathcal{M}$.

end for
Stage 2: Fine-tuning the classifier with kNN-GCN analogy
for each event pair (S_1^k, S_2^k, e_1, e_2) in batch **do**
- Encoder $\mathcal{E}$ outputs representation embeddings for two events e_1 and e_2. Then retrieve the nearest n embeddings of e_1 and e_2 from memory $\mathcal{M}$ as $e^a_{1,1}, ..., e^a_{1,n}$ and $e^a_{2,1}$, ..., $e^a_{2,n}$
- Compute the refined event representations $\tilde{e_1}$ and $\tilde{e_2}$ as Equation (7) with $\mathcal{G}$.
- Use classifier $\mathcal{C}$ to predict the causality probability $\tilde{p_i}$ between $\tilde{e_1}$ and $\tilde{e_2}$.
- Update graph encoder $\mathcal{G}$ and classifier $\mathcal{C}$ with loss as equation (8).

end for

The target event and the n nearest neighbour events are the nodes, and the similarities between the retrieved events and the target event are the edge weights. Therefore, the graphs are represented as:

$$\begin{aligned}
&\boldsymbol{V}_1 = \{e_1, e^a_{1,1}, ..., e^a_{1,n}\},\ \boldsymbol{E}_1 = \{(e_1, e^a_{1,1}), ..., (e_1, e^a_{1,n})\},\\
&\boldsymbol{V}_2 = \{e_2, e^a_{2,1}, ..., e^a_{2,n}\},\ \boldsymbol{E}_2 = \{(e_2, e^a_{2,1}), ..., (e_2, e^a_{2,n})\},\\
&\boldsymbol{G}_1 = (\boldsymbol{V}_1, \boldsymbol{E}_1),\ \boldsymbol{G}_2 = (\boldsymbol{V}_2, \boldsymbol{E}_2),
\end{aligned} \tag{4}$$

where $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ are the event nodes. $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$ are weighted edges from the target event to the retrieved similar events whose weights are computed as the cosine similarity between the event embeddings. Formally, the weight of edge between the target node e_i and the corresponding retrieved event node e_j is defined as:

$$A_{ij} = \begin{cases} f_{\text{Cosine}}(e_i, e_j), & e_j \text{ is the retrieved events,} \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

The cosine similarity is computed as:

$$f_{\text{Cosine}}(x, y) = \frac{x^T y}{||x|| \cdot ||y||}, \tag{6}$$

where $||x|| = \sqrt{\sum_i x_i^2}$ and $||y|| = \sqrt{\sum_i y_i^2}$ are the length of the vectors of x and y.

Given the constructed event graphs, we can learn how to propagate the neighbour information to the target node with a Graph Convolutional Network (GCN) (Kipf and Welling 2016). The refined representation of each input event after two layers GCN can be calculated as:

$$\begin{aligned}
\tilde{e_1} =& \tilde{A}_1 ReLU(\tilde{A}_1 X W_0) W_1, \\
\tilde{e_2} =& \tilde{A}_2 ReLU(\tilde{A}_2 X W_0) W_1,
\end{aligned} \tag{7}$$

where $\tilde{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ is the normalized symmetric adjacency matrix, D is the degree matrix of A with $D_{ii} = \sum_j A_{ij}$, W_0 and W_1 are the weight matrices, and $ReLU(x) = max(0, x)$ is the activation function.

Two-stage Training of KADE

We briefly describe the training process of KADE in Alg. 1, including the first-stage knowledge-enrich training and the second-stage analogy-fusion training. As shown in Alg. 1, we firstly optimize the BERT encoder and the knowledge-enrich stage classifier by Equation (2). Then, in the analogy-fusion stage, we train the GCN for analogy fusion and fine-tune the classifier $\mathcal{C}$ with adjusted representation event embeddings as Equation (7). Our optimization function is simply the cross-entropy loss on both training stages. Thus, the second stage training loss is:

$$\mathcal{L}_{CE} = -\frac{1}{N} \sum_i [y_i \cdot log(\tilde{p_i}) + (1 - y_i) \cdot log(1 - \tilde{p_i})] \tag{8}$$

In our experiments, we separately train the first stages and second stages for 40 epochs.

Experiments

Datasets and Evaluation Metrics

Following prior works (Zuo et al. 2021b; Liu, Chen, and Zhao 2020), we evaluate our methods on two benchmark datasets for ECI, i.e., EventStoryLine v0.9 (Caselli and Vossen 2017) and Causal-TimeBank (Mirza and Tonelli 2014).

EventStoryLine v0.9 comes from (Caselli and Vossen 2017), which involves 258 documents, 22 topics, 4,316 sentences, 5,334 event mentions, 7,805 intra-sentence and 46,521 inter-sentence event mention pairs (1,779 and 3,855 are annotated with a causal relation, respectively). Following (Liu, Chen, and Zhao 2020), we use the documents of the last two topics as the development set while the documents of the remaining 20 topics are employed for a 5-fold cross-validation evaluation, using the same data split of (Liu, Chen, and Zhao 2020).

Causal-TimeBank (Mirza and Tonelli 2014) contains 184 documents, 6,813 events, and 318 of 7,608 event mention pairs annotated with causal relation. As the number of inter-sentence event mention pairs with the causal relation is very small (i.e., only 18 pairs), we only evaluate the ECI performance for intra-sentence events in Causal-TimeBank. Following (Liu, Chen, and Zhao 2020), we perform 10-fold cross-validation evaluation for Causal-TimeBank.

For evaluation, we consider Precision (P), Recall (R), and F1-score (F1) as evaluation metrics, same to previous methods to ensure comparability.

Model	Intra-sentence			Inter-sentence			Intra+Inter		
	P	R	F1	P	R	F1	P	R	F1
OP (Caselli and Vossen 2017)	22.5	**98.6**	36.6	8.4	99.5	15.6	10.5	99.2	19.0
LR+ (Gao, Choubey, and Huang 2019)	37.0	45.2	40.7	25.2	48.1	33.1	27.9	47.2	35.1
LIP (Gao, Choubey, and Huang 2019)	38.8	52.4	44.6	35.1	48.2	40.6	36.2	49.5	41.9
KMMG (Liu, Chen, and Zhao 2020)	41.9	62.5	50.1	-	-	-	-	-	-
KnowDis (Zuo et al. 2020)	39.7	66.5	49.7	-	-	-	-	-	-
RichGCN (Phu and Nguyen 2021)	49.2	63.0	55.2	39.2	45.7	42.2	42.6	51.3	46.6
LearnDA (Zuo et al. 2021b)	42.2	69.8	52.6	-	-	-	-	-	-
BERT (Our implement)	47.3	55.8	51.2	22.3	29.2	25.3	27.3	35.3	30.8
BERT$_{kg}$	44.7	57.4	50.3	39.2	63.2	40.8	47.3	55.3	43.8
KADE$_{mANA}$	58.5	**78.6**	**67.1**	37.1	67.7	47.9	42.3	**72.3**	53.5
KADE$_{full}$	**61.5**	73.2	66.8	**51.2**	**74.2**	**60.5**	**51.9**	70.6	**59.8**

Table 1: Compare different methods on EventStoryLine. The best results are in **bold** and the second-best results are underlined. Overall, our proposed KADE outperforms other SOTA models on Precision, Recall and F1.

Model	P	R	F1
KMMG	36.6	55.6	44.1
knowDis	42.3	60.5	49.8
LearnDA	41.9	68.0	51.9
CauseRL	43.6	68.1	53.2
RichGCN	39.7	56.5	46.7
BERT	45.2	50.1	47.5
BERT$_{kg}$	21.6	44.7	27.4
KADE$_{mANA}$	**60.7**	69.2	64.8
KADE$_{full}$	56.8	**70.6**	**66.7**

Table 2: Compare different methods on CausalTimeBank.

Parameter Settings

We implement our method based on PyTorch (Paszke et al. 2019). We use uncased BERT-base (Devlin et al. 2018) as the encoder like previous works (Zuo et al. 2021b; Liu, Chen, and Zhao 2020), with 12 layers, embedding dimensions of 768, and 12 heads. We employ feed forward network for the classifier. For analogy enhancement, we use $k = 3$ most similar entities for all our experiments and show the impact of k. For the optimizer, we use BertAdam (Zhang et al. 2020) and train the model for 40 epochs during the first-stage training, with 1×10^{-6} as learning rate and 1×10^{-4} as weight decay. For the second stage of training, we only fine-tune the classifier for 40 epochs. The batch size is set to 16 for both training stages. We also adopt a negative sampling rate of 0.6 for the first step training, owing to the sparseness of positive examples of ECI datasets.

Compared Baselines

We choose both state-of-the-art deep learning-based models and feature-based models for comparison: 1) **OP** (Caselli and Vossen 2017): a dummy model assigns a causal relation to every pair of event mentions; 2) **LR+** and **LIP** (Gao, Choubey, and Huang 2019): the current SOTA for inter-sentence ECI with a document structure-based model; 3) **KMMG** (Liu, Chen, and Zhao 2020): a mention masked generalization method using external knowledge databases; 4) **KnowDis** (Zuo et al. 2020): a model utilizing both original sentence and event mention masking sentence; 5) **LSIN** (Cao et al. 2021), the current SOTA for intra-sentence ECI with a descriptive graph base model. 6) **LearnDA** (Zuo et al. 2021b): a model used knowledge bases to augment training data; 7) **CauSeRL** (Zuo et al. 2021a): a model which can extract causal patterns from external causal statements; 8) **RichGCN** (Phu and Nguyen 2021): a GCN based model to use document-level interaction graph, which is the current SOTA for inter-sentence ECI.

We also develop several BERT-based methods to evaluate the effectiveness of knowledge enhancement and kNN-GNN analogy (i.e., analogy with kNN retrieval and GNN-based graph fusion): 1) **BERT(our implement)**: a baseline method that takes the embedding vectors from BERT and performs classification for ECI; 2) **BERT**$_{kg}$: a BERT-based model with knowledge-aware inputs; 3) **KADE**$_{mANA}$: mean fusion analogy with knowledge-aware inputs; 4) **KADE**$_{full}$: GNN-based graph fusion analogy with knowledge-aware inputs. The **KADE**$_{full}$ is our full model shown in Fig. 2.

Main Results

Since we only evaluate intra-sentence ECI on the CausalTimeBank, the baselines used for EventStoryLine and Causal-TimeBank are different. The experimental results for EventStoryLine and CausalTimeBank are summarized in Table 1 and Tabel 2, respectively. We make the following observations.

First, our models outperform baselines by a large margin. From the results, we can see that our proposed $\mathrm{KADE_{mANA}}$ and $\mathrm{KADE_{full}}$ significantly outperform all baseline methods and achieve the best performance in terms of the three metrics on both intra-sentence and inter-sentence ECI. Our $\mathrm{KADE_{mANA}}$ outperforms other deep learning-based baselines by $18.9\%, 12.6\%, 21.6\%$ for precision, recall and F1 on intra-sentence ECI on EventStoryLine, and $27.5\%, 1.6\%, 21.8\%$ compared with CauseRL on CausalTimeBank, which justifies the effectiveness of our proposed method.$\mathrm{KADE_{full}}$ outperforms SOTA by

Model	P	R	F1
BERT	47.3	55.8	51.2
BERT$_{kg}$	44.7	57.4	50.2
BERT$_{mANA}$	58.9	<u>74.0</u>	65.6
BERT$_{gANA}$	<u>60.6</u>	69.5	64.8
KADE$_{mANA}$	58.5	**78.6**	**67.1**
KADE$_{full}$	**61.4**	73.2	<u>66.8</u>

Table 3: Ablation results on intra-sentence EventStoryLine dataset. The best results are in **bold** and the second-best results are <u>underlined</u>. BERT denote the input of BERT model is enhanced by external knowledge as descriped in Section Knowledge enhancement.

Model	P	R	F1
BERT	44.1	39.4	41.6
BERT$_{kg}$	21.6	44.7	27.4
BERT$_{mANA}$	42.8	53.5	57.8
BERT$_{gANA}$	50.6	53.4	56.5
KADE$_{mANA}$	**60.7**	<u>69.2</u>	<u>64.7</u>
KADE$_{full}$	<u>56.8</u>	**70.5**	**66.7**

Table 4: Ablation results on Causal-TimeBank dataset.

$21.8\%, 26.5\%, 28\%$ compared with RichGCN for precision, recall and F1 on intra+inter-sentence ECI on EventStoryLine, and $25.7\%, 3.6\%, 25.3\%$ compared with CauseRL on CausalTimeBank. Especially, our proposed $\mathrm{KADE_{mANA}}$ and $\mathrm{KADE_{full}}$ show remarkable ability to solve inter-sentence ECI, which is a large challenge for previous ECI methods.

Second, knowledge is helpful for ECI. Our implemented BERT achieves comparable performance with previous work. Compare $\mathrm{BERT_{kg}}$ with BERT, we can see the knowledge enrichment method can improve the performance of ECI. That is because commonsense knowledge is essential for understanding event causality. We also note that $\mathrm{BERT_{kg}}$ performs worse than RichGCN and LearnDA on some cases, especially on CausalTimeBank dataset. That may because the enriched commonsense knowledge can also introduce noise, which may disturb the attention of the BERT model. When equipped with analogy module, the model has stronger ability to distinguish the important information for event pairs. $\mathrm{KADE_{mANA}}$ improves a lot over $\mathrm{BERT_{kg}}$, which shows that the similar events retrieved by kNN analogy can largely help the model learning a better representation of event mentions.

Third, graph fusion performs better than mean fusion. Compared to $\mathrm{KADE_{mANA}}$, $\mathrm{KADE_{full}}$ improves F1 by $3.24\%, 11.7\%$ for intra- and inter-sentence ECI on EventStoryLine dataset, respectively, as well as improves F1 by 2.93% on CausalTimeBank. This shows that GCN can better capture effective information between the target event and the retrieved analogy event examples. The learnable parameters of GCN also enables more flexibility than mean fusion-based analogy.

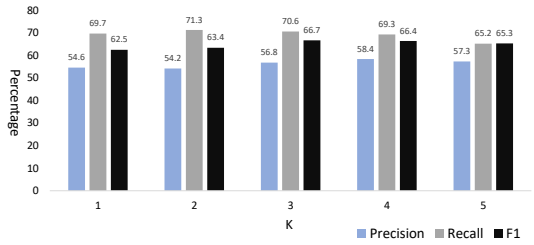

Figure 3: Impact of the number of similar entities K in analogy on Causal-TimeBank.

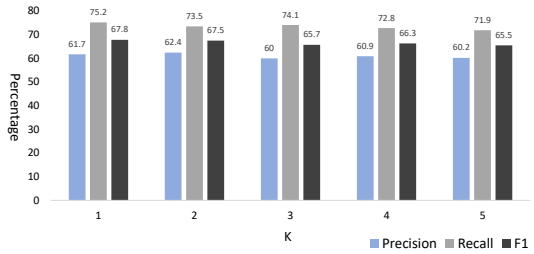

Figure 4: Impact of the number of similar entities K in analogy on EventStoryLine.

Ablation Study of KADE Components

To analyze the effect of different designs in KADE, in Table 3 and Table 4, we compare the following methods: 1) **BERT**, basic BERT model implemented by ourselves; 2) **BERT**$_{kg}$, BERT with knowledge-enriched inputs; 3) **BERT**$_{mANA}$, mean analogy model without knowledge-enriched inputs; 4) **BERT**$_{gANA}$, GCN analogy model without knowledge-enriched inputs; 5) **KADE**$_{mANA}$ and **KADE**$_{full}$ are the same with Table 1.

From the results, we have the following observations: First, the external knowledge only improves recall a little bit with precision and F1 decrease on both dataset. This illustrates that the external knowledge may improve the performance by introducing commonsense knowledge but also incurs noise, which influences the model performance, especially for a small dataset like Causal-TimeBank. Second, mean fusion-based analogy such as **BERT**$_{mANA}$ leads to substantial gains on both datasets for all metrics, especially for Causal-TimeBank. This may be related to the characteristic of the Causal-TimeBank, which is very small and has similarity and commonality between samples. Therefore, mean analogy can effectively utilize the data and enhance the generalizability of model. Third, for intra-sentence ECI, graph fusion-based analogy performs better than mean fusion-based analogy.

Effect of Different k

To validate the effect of different k values in kNN lookup for ECI, we test **KADE**$_{full}$ on EventStoryLine with $k \in \{1, 2, 3, 4, 5\}$ while fixing other settings. The results are summarized in Fig. 3 and Fig. 4. The results show that when $k = 3$, $KADE_{full}$ achieves the best performance for both datasets, which shows that too small k can hardly learn

Sentence of target event	Sentence of retrieved event	Patterns
Strong earthquake jolts southern Iran.	Large Riot Breaks Out In Brooklyn During Vigil For Teen Shot 11 Times By Police.	Retrieved event "Breaks Out" is the synonym of target event "jolts".
On Qeshm island, between half and two-thirds of homes in five villages had been damaged, officials said.	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 .	Retrieved event "destroyed" has similar meaning with the target event "damaged".
The fire started when demonstrators hurled Molotov cocktail fire bombs at the Bank .	After the tragic death of the three workers made the round of Athens, new clashes started to spread in the Greek capital, with a large crowd gathered outside the burned bank when Martin's boss tried to visit the site.	Retrieved event "started" are the same events with the target event, but in different sentences.
Convicted of second-degree murder and assault in the first degree , Lopez , 20 , faces a potential sentence of life in prison when he is sentenced by Justice Vincent Del Giudice .	Prosecutors say Andrew Lopez, 20, an alleged 8 Block gang member fired the shots meant for rival gang members from the Howard Projects while his brother Jonathan Carrasquillo, 24, gave the orders.	Retrieved event "orders" has similar meanings with the target event "sentenced", which are in different sentences.
First came the shooting : an armed teenager killed by police officers on a darkened Brooklyn street .	An earthquake measuring at least magnitude-5.9 shook a sparsely populated area of southern Iran on Sunday , flattening seven villages and killing 10 people, officials said.	Retrieved event "earthquake" has high-level latent similarity with target event "shooting", such as similar sentence structure.

Figure 5: Samples of retrieved events and their corresponding sentence.

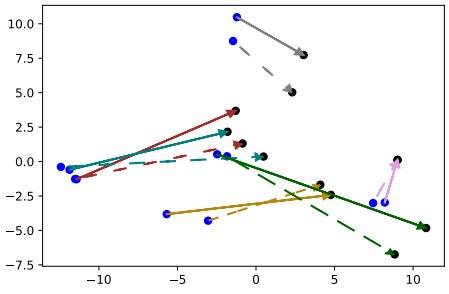

Figure 6: Two-dimensional PCA projection of the causality event pairs embeddings before and after analogy. The dotted lines denote embeddings before analogy.

enough analogy information, and too large k could introduce noise which may deteriorate the performance.

Case Study

We conduct a qualitative study of how the model actually benefits from analogy by showing what analogy event examples the kNN really retrieved. A few examples we analyzed are shown in Fig. 5, from which we can see that kNN lookup can find related and general event mentions that can help the model to focus on more general information. For example, in Fig. 5, the model retrieved 'earthquake' with the target event of 'shooting'. The information from similar sentence structure can help the representation expand the information boundary.

Based our analysis, we classify the retrieved events into three categories:

- Retrieved events are the same events to the target event, but in different sentences. In this situation, the refined event representation incorporates information of another sentence, which can help the model understand the meaning of event more comprehensively.
- Retrieved events are the events with similar meanings, which can be in either the same sentence or a different sentence. This is an analogy to the situation when we search for the meaning of one word in dictionary, we not only need to check the meaning of this word, but also the meaning of similar words.
- Retrieved events are the different events with similar semantics, which can still improve the generalization ability of the model, especially when the model is evaluated on unseen data.

To better visualize the effect of analogy, we further perform two-dimensional PCA projection for the representation embeddings of events before and after GCN-based analogy. As shown in Fig. 6, we can discover that after analogy, the extent of orthogonality between causality event embedding pairs have decreased. Also, some arrows tend to be parallel with other arrows (e.g., the grey arrow and the yellow arrow). These observations confirm that after analogy, event pairs with causal relationships have more common features than before analogy.

Conclusion

In this paper, we emphasize the importance of both knowledge and analogy in event causality identification task, which is similar to human intelligence. Motivated by this insight, we propose the KADE framework that exploits both knowledge from ConceptNet and analogy from kNN retrieved similar events. By comparing our KADE model and its variants to a series of baseline methods, we see that KADE outperforms existing methods by a large margin, demonstrating the significant effect of knowledge and analogy. Our KADE framework is flexible and general: the different components can be easily replaced by other models, and the idea of making use of both knowledge and analogy examples can be easily extended to other NLP tasks. In the future, we plan to explore analogy-based framework on a wider range of tasks, and utilize global and heterogeneous graph for better graph fusion.

Acknowledgments

This work was supported by the FRQNT Établissement de la relève professorale 2022-2023 under Grant No. 313315 and the Canada CIFAR AI Chair Program.

References

Bach, E. 1986. The algebra of events. *Linguistics and philosophy*, 5–16.

Beamer, B.; and Girju, R. 2009. Using a bigram event model to predict causal potential. In *International Conference on Intelligent Text Processing and Computational Linguistics*, 430–441. Springer.

Berant, J.; Srikumar, V.; Chen, P.-C.; Vander Linden, A.; Harding, B.; Huang, B.; Clark, P.; and Manning, C. D. 2014. Modeling biological processes for reading comprehension. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1499–1510.

Cao, P.; Zuo, X.; Chen, Y.; Liu, K.; Zhao, J.; Chen, Y.; and Peng, W. 2021. Knowledge-enriched event causality identification via latent structure induction networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4862–4872.

Caselli, T.; and Vossen, P. 2017. The event storyline corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, 77–86.

Das, R.; Zaheer, M.; Thai, D.; Godbole, A.; Perez, E.; Lee, J.-Y.; Tan, L.; Polymenakos, L.; and McCallum, A. 2021. Case-based reasoning for natural language queries over knowledge bases. *arXiv preprint arXiv:2104.08762*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ding, X.; Li, Z.; Liu, T.; and Liao, K. 2019. ELG: an event logic graph. *arXiv preprint arXiv:1907.08015*.

Fan, A.; Gardent, C.; Braud, C.; and Bordes, A. 2021. Augmenting transformers with KNN-based composite memory for dialog. *Transactions of the Association for Computational Linguistics*, 9: 82–99.

Gao, L.; Choubey, P. K.; and Huang, R. 2019. Modeling document-level causal structures for event causal relation identification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Gupta, A.; Dar, G.; Goodman, S.; Ciprut, D.; and Berant, J. 2021. Memory-efficient Transformers via Top-k Attention. In *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, 39–52.

Hashimoto, C. 2019. Weakly supervised multilingual causality extraction from Wikipedia. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2988–2999.

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

Hofstadter, D. R. 1995. *Fluid concepts and creative analogies: Computer models of the fundamental mechanisms of thought.* Basic books.

Hu, Z.; Rahimtoroghi, E.; and Walker, M. A. 2017. Inference of fine-grained event causality from blogs and films. *arXiv preprint arXiv:1708.09453*.

Kadowaki, K.; Iida, R.; Torisawa, K.; Oh, J.-H.; and Kloetzer, J. 2019. Event causality recognition exploiting multiple annotators' judgments and background knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5816–5822.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Liu, J.; Chen, Y.; and Zhao, J. 2020. Knowledge Enhanced Event Causality Identification with Mention Masking Generalizations. In *IJCAI*, 3608–3614.

Mirza, P. 2014. Extracting temporal and causal relations between events. In *Proceedings of the ACL 2014 Student Research Workshop*, 10–17.

Mirza, P.; and Tonelli, S. 2014. An analysis of causality between events and its relation to temporal information. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2097–2106.

Oh, J.-H.; Torisawa, K.; Hashimoto, C.; Iida, R.; Tanaka, M.; and Kloetzer, J. 2016. A semi-supervised learning approach to why-question answering. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Phu, M. T.; and Nguyen, T. H. 2021. Graph convolutional networks for event causality identification with rich document-level structures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3480–3490.

Riaz, M.; and Girju, R. 2014. In-depth exploitation of noun and verb semantics to identify causation in verb-noun pairs. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 161–170.

Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.

Wu, Y.; Rabe, M. N.; Hutchins, D.; and Szegedy, C. 2022. Memorizing transformers. *arXiv preprint arXiv:2203.08913*.

Zhang, T.; Wu, F.; Katiyar, A.; Weinberger, K. Q.; and Artzi, Y. 2020. Revisiting few-sample BERT fine-tuning. *arXiv preprint arXiv:2006.05987*.

Zuo, X.; Cao, P.; Chen, Y.; Liu, K.; Zhao, J.; Peng, W.; and Chen, Y. 2021a. Improving Event Causality Identification via Self-Supervised Representation Learning on External Causal Statement. *arXiv preprint arXiv:2106.01654*.

Zuo, X.; Cao, P.; Chen, Y.; Liu, K.; Zhao, J.; Peng, W.; and Chen, Y. 2021b. LearnDA: Learnable knowledge-guided data augmentation for event causality identification. *arXiv preprint arXiv:2106.01649*.

Zuo, X.; Chen, Y.; Liu, K.; and Zhao, J. 2020. KnowDis: Knowledge enhanced data augmentation for event causality detection via distant supervision. *arXiv preprint arXiv:2010.10833*.