Distill, Fuse, Pre-train: Towards Effective Event Causality Identification with Commonsense-Aware Pre-trained Model

Peixin Huang[1], **Xiang Zhao**[1], **Minghao Hu**[2]*, **Zhen Tan**[1], **Weidong Xiao**[1]
[1] National Key Laboratory of Information Systems Engineering,
National University of Defense Technology, Changsha, China
[2] Information Research Center of Military Science, Beijing, China
{huangpeixin15, xiangzhao, tanzhen08a, wdxiao}@nudt.edu.cn
huminghao16@gmail.com

Abstract

Event Causality Identification (ECI) aims to detect causal relations between events in unstructured texts. This task is challenged by the lack of data and explicit causal clues. Some methods incorporate explicit knowledge from external knowledge graphs (KGs) into Pre-trained Language Models (PLMs) to tackle these issues, achieving certain accomplishments. However, they ignore that existing KGs usually contain trivial knowledge which may prejudice the performance. Moreover, they simply integrate the concept triplets, underutilizing the deep interaction between the text and external graph. In this paper, we propose an effective pipeline DFP, i.e., Distill, Fuse and Pre-train, to build a commonsense-aware pre-trained model which integrates reliable task-specific knowledge from commonsense graphs. This pipeline works as follows: (1) To leverage the reliable knowledge, commonsense graph distillation is proposed to *distill* commonsense graphs and obtain the meta-graph which contain credible task-oriented knowledge. (2) To model the deep interaction between the text and external graph, heterogeneous information fusion is proposed to *fuse* them through a commonsense-aware memory network. (3) Continual pre-training is proposed to further align and fuse the text and the commonsense meta-graph with three continual *pre-training* tasks. Through extensive experiments on two benchmarks, we demonstrate the validity of our pipeline.

Keywords: event causality identification, commonsense graph, continual pre-training

1. Introduction

Event causality identification (ECI) is an important natural language processing task, which aims to identify causal relations of events in texts. As shown in Figure 1, an ECI model should identify the causalities between the mentioned events: (1) *quake* $\xrightarrow{cause}$ *tsunami* in S1; (2) *earthquake* $\xrightarrow{cause}$ *tsunami* in S2. ECI can support many NLP applications, including event forecasting (Hashimoto et al., 2014; Radinsky et al., 2012), why-question answering (Oh et al., 2016) and machine reading comprehension (Berant et al., 2014).

This task is often challenged by the lack of explicit causal clues. For example in Figure 1, there is no clue indicating the causality between "*earthquake*" and "*tsunami*" in S2, whereas causal clue words "*triggered*" explicitly indicate the causality existence in S1. Albeit that ECI models can use large amounts of annotated data to learn informative causal expressions, existing datasets are small (e.g., the largest ECI dataset only contains 258 documents (Caselli and Vossen, 2017)). Lately, pre-trained language models (PLMs) have achieved remarkable success on the ECI task (Liu et al., 2020a; Zuo et al., 2021a). They allow the model to learn useful information (i.e., semantic correlation in the latent space) from textual datasets, compensating the data lacking issue to some extent. However, knowledge from PLMs has inherent weakness of noisy and lacking in task related knowledge.

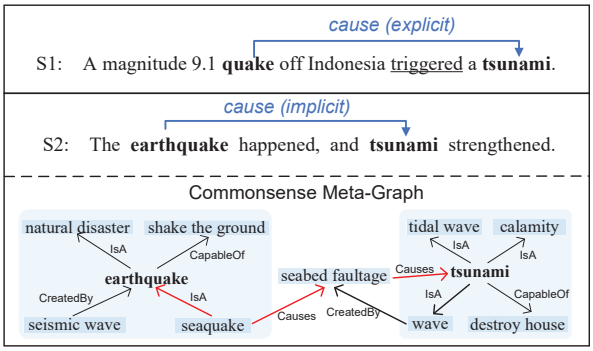

Figure 1: Examples of different causalities. S1 contains explicit causality between *quake* and *tsunami*. S2 contains implicit causality but it can leverage external knowledge from commonsense metagraph for inference.

Recently, several studies have shown that commonsense knowledge from knowledge graphs like ConceptNet (Speer et al., 2017) can benefit the PLMs-based ECI methods (Cao et al., 2021; Liu et al., 2023). For example in Figure 1, the one-hop definitions associated with *earthquake* include *earthquake* $\xrightarrow{IsA}$ *natural disas-*

* Corresponding authors

LREC-COLING 2024, pages 5029–5040
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

ter, *seaquake* $\xrightarrow{IsA}$ *earthquake* and so on, which can help the model understand the mentioned event. Besides, the three-hop path between two mentioned events through concept "seabed fault-age" can provide ample evidence for judging the causality. Despite achieving certain accomplishments, current methods are still not effective due to the following two issues: (1) They ignore that existing commonsense graphs are not constructed for this task. These commonsense graphs usually contain trivial concept triplets, which may prejudice the ECI performance. (2) They only introduce the concept triplet information through simply token concatenation or attention mechanism, underutilizing the informative relational knowledge contained in concept graphs and the deep interactions between the text and the external graph.

In view of this, we propose an effective pipeline, i.e., distill, fuse and pre-train (DFP) to build a commonsense-aware continual pre-trained model for ECI, tackling the aforementioned issues with the following three modules.

(1) Distill: We propose a *commonsense graph distillation* module. It includes commonsense graph pruning which prunes unreliable triplets from a given graph, and metagraph induction which extracts the task related knowledge, obtaining a event centered commonsense metagraph. In this way, trivial knowledge in the original commonsense graphs can be removed, and informative task specific knowledge ia retained. The first issue could be addressed through this module.

(2) Fuse: We propose a *heterogeneous information fusion* module. It introduces a commonsense-aware memory network which associates the representation spaces of the text and metagraph through memory read and write operations. Through this module, the relational knowledge contained in the metagraph, not just the discrete triplet knowledge could be fused with the text knowledge. The deep interactions between the text and the external graph could be captured. The second issue is alleviated.

(3) Pre-train: We introduce a *continual pre-training* module and design three pre-training tasks to further fuse the text and the graph, including the masked language model and concept triplet completion tasks to improve the understanding of events in the text and triplet correlations in the metagraph, respectively, and the text-metagraph contrastive learning task to align and unify the representations of text and graph. Through this module, knowledge from the text and metagraph can further boost each other to achieve a better ECI performance.

Our contributions are summarized as follows:

- We propose an effective pipeline named DFP for ECI, which follows distill, fuse and pre-train, to build a commonsense-aware pre-trained model which incorporates reliable task-specific knowledge from existing commonsense graphs.
- We design a commonsense graph distillation module. It distills a credible task-specific metagraph through commonsense graph pruning and metagraph induction.
- We devise a commonsense-aware memory network and three continual pre-training tasks to capture the information interactions between the text and the external commonsense metagraph, which are heterogeneous.
- Extensive experimental results demonstrate that the proposed DFP can achieve a new state-of-the-art (SOTA) performance.

2. Related Work

Studies related to our work are mainly discussed from the following two aspects.

2.1. Event Causality Identification

ECI has attracted much attention to date. Early studies usually rely on predefined patterns or linguistic rules to identify causal relations. For example, some studies utilize linguistic and syntactic features tailored for causal expressions (Riaz and Girju, 2013, 2014). Some incorporate explicit causal markers or clues (Riaz and Girju, 2010; Do et al., 2011). And some others pay attention to statistical information (Hashimoto et al., 2014).

Later, several ECI datasets are released, e.g., Causal-TimeBank (Mirza et al., 2014), EventStoryLine (Caselli and Vossen, 2017), BECAUSE (Dunietz et al., 2017) and CNC (Tan et al., 2022). Based on these benchmarks, supervised learning methods are applied and achieve the better performance (Kruengkrai et al., 2017; Hu et al., 2017, 2023). However, the scale of these datasets is relatively small. To ease this issue, some studies utilize weakly supervised methods to generate labeled training data (Hashimoto, 2019; Zuo et al., 2020). Some recent methods introduce external knowledge graphs to enhance the abilities of the ECI models (Zuo et al., 2021b; Cao et al., 2021).

Recently, PLMs-based ECI methods have achieved the better performance, as PLMs can generate high-quality text representations (Kadowaki et al., 2019; Zuo et al., 2021a; Shen et al., 2022). Although PLMs can capture the associations among tokens in the texts, the implicit associations between events hinder the performance. To empower PLMs with the task-specific knowledge, we propose a commonsense-aware pre-trained model, which thoughtfully incorporates reliable commonsense knowledge.

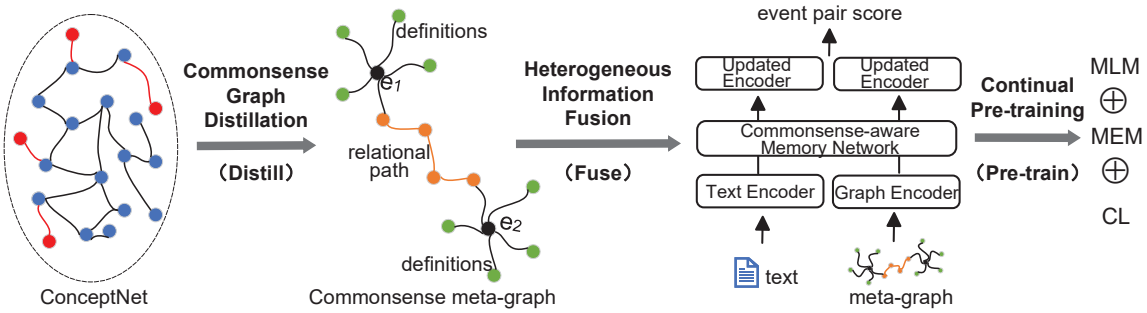

Figure 2: Overview of DFP pipeline. (1) Commonsense graph distillation to distill a reliable task-specific commonsense metagraph. (2) Heterogeneous information fusion to fuse the heterogeneous information from the external graph and the text. (3) Continual pre-training and fine-tuning to further fuse the text and metagraph, and make final predictions.

2.2. Knowledge Graph Enhanced PLMs

Knowledge graph enhanced PLMs (KG-PLMs) leverage the rich structured knowledge of a knowledge graph (KG) and integrate it with PLMs to enhance the performance of natural language understanding tasks. Existing KG-PLMs can be grouped by the forms of knowledge incorporated from KG, i.e., semantic triplets and knowledge subgraphs. To incorporate semantic triplets from KG, K-BERT (Liu et al., 2020b) and ERNIE 3.0 (Sun et al., 2021) turn to append these triplets to the specific position in the text. Liu et al. (2022) directly concatenate triplets embeddings with text embeddings. Recently, some studies turn to integrate more sophisticated knowledge, i.e., knowledge subgraphs. QA-GNN (Yasunaga et al., 2021) designs interaction nodes to integrate knowledge from the subgraph space and the text space. GreaseLM (Zhang et al., 2022) introduces a fusion layer to fuse the information from different modalities. In this work, we first distill metagraphs through graph pruning and induction, which is task-specific subgraphs from existing commonsense graphs. Then we construct a commonsense-aware memory network to associate the text space and the subgraph space, integrating knowledge from them through the memory network and continual pre-training.

3. Approach

We formulate ECI as a binary classification problem. For a pair of events (e_1, e_2) in a sentence S, we predict whether a causal relation holds. Figure 2 schematically visualizes our approach DFP, which are elaborated by the following subsections.

3.1. Commonsense Graph Distillation

In this paper, we harness ConceptNet (Speer et al., 2017) as the external commonsense graph, which contains abundant background knowledge. Directly introducing ConceptNet is unsuitable, as it contains trivial triplets and some knowledge might not be useful. In view of this, we adopt a pipeline of commonsense graph pruning and metagraph induction to obtain the task-oriented knowledge, forming a commonsense metagraph G_{meta}.

3.1.1. Commonsense Graph Pruning

Given a commonsense graph G, where the nodes correspond to concepts, and edges correspond to semantic relations, we harness TransE (Bordes et al., 2013) to measure the confidence of a given triplet (n_i, r, n_j) in G. Intuitively, TransE models latent knowledge distribution in a graph, allowing it to distinguish valuable knowledge from less informative knowledge. We first calculate the distance between two linked concepts as follows:

$$D(n_i, n_j) = \frac{1}{\mathbf{n}_i\mathbf{r} + \mathbf{n}_i\mathbf{n}_j + \mathbf{r}\mathbf{n}_j} \tag{1}$$

where $\mathbf{n}$ and $\mathbf{r}$ are the TransE embeddings of concept and relation. We regard triplets with small distance values as informative ones.

Then, given each node n_i in G, we keep the top-K neighboring nodes $N(n_i)$:

$$N(n_i) = \bigcup_{k=1}^{K} \{n_j^k\}, where D(n_i, n_j^k) \leq D(n_i, n_j^{k+1}) \tag{2}$$

Thus, the pruned commonsense graph $\overline{G}$ is:

$$\overline{G} = \{(n_i, r, n_j) | n_j \in N(n_i)\} \tag{3}$$

3.1.2. Metagraph Induction

Given each event pair (e_1, e_2), the aim of metagraph induction is to construct a corresponding task-oriented commonsense metagraph from $\overline{G}$.

To obtain the definition knowledge, we first match the event mention of e_1, e_2 with concept tokens in $\overline{G}$ through matching rules, i.e., soft matching

and stop word filtering. Then we search the one-hop definitions of the matched concepts from the pruned graph $\overline{G}$.

To obtain the relational knowledge, we perform Breadth First Search to discover the multi-hop path between the matched concept pairs of e_1 and e_2 from $\overline{G}$. Note that we only keep the shortest relational path that might be the most informative, as the shorter path between two nodes indicates a stronger relevance of intermediate nodes on the path. Moreover, if there are multiple shortest relational paths, we randomly choose one of them.

Finally, the metagraph G_{meta} is built with the one-hop definitions of e_1, e_2 and the multi-hop path discovered between e_1 and e_2. Figure 1 shows an example of the metagraph, which contains rich background knowledge of two events, and prior information of the relevance between them.

3.2. Heterogeneous Information Fusion

After obtaining G_{meta}, we propose a commonsense-aware memory network to deeply fuse the explicit knowledge in G_{meta} with the text knowledge. In what follows, we first present the base models for encoding the text and the metagraph. Then we introduce the devised commonsense-aware memory network.

3.2.1. Text Encoder

We use BERT (Devlin et al., 2019) as PLMs to encode the input sentence $S = \{t_1, t_2, ..., t_L\}$. Specifically, PLMs first map the tokens into corresponding embeddings. Then a stack of Transformer layers encodes the embeddings to generate the l-th layer token representations $\mathbf{H}^{(l)} = \{\mathbf{h}_1^{(l)}, \mathbf{h}_2^{(l)}, ..., \mathbf{h}_L^{(l)}\}$.

3.2.2. Metagraph Encoder

We harness the graph attention network (GAT) (Velickovic et al., 2018) to encode the commonsense metagraph. Given G_{meta} with N nodes, GAT first initializes the node embeddings by TransE, obtaining a set of embeddings $\{\mathbf{n}_1, \mathbf{n}_2, ..., \mathbf{n}_N\}$. Then the node representation is updated as:

$$\mathbf{n}_i^{(l+1)} = \Big\Vert_{k=1}^{K} \sigma\Big(\sum_{j \in N_i} \alpha_{ij}^k \mathbf{W}_k^{(l)} \mathbf{n}_j^{(l)}\Big) \quad (4)$$

where $\mathbf{n}_i^{(l+1)}$ is the representation of node n_i in the $l+1$ layer, $\|$ represents the concatenation operation, N_i is the neighbor set of node i in G_{meta}, K is the number of attention heads, α_{ij}^k is the attention value of node i to j in attention head k and $\mathbf{W}_k^{(l)}$ is a learnable weight matrix.

3.2.3. Commonsense-aware Memory Network

To deeply aggregate the text and metagraph, we design k memory networks between the last k layers of PLMs and GAT. They allow the text knowledge and the commonsense knowledge to interact and mutually boost each other. The illustration of our devised commonsense-aware memory network is shown in Figure 3.

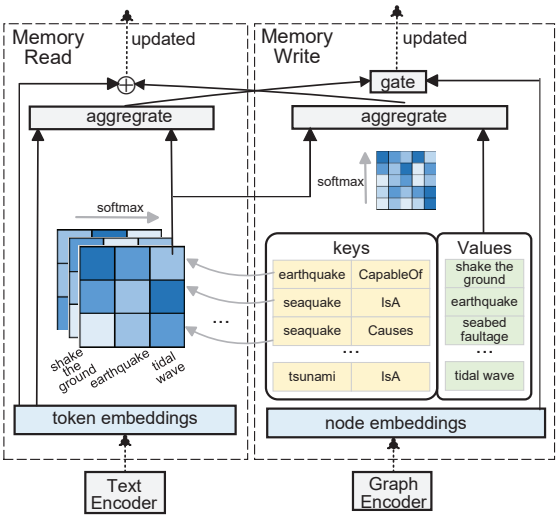

Figure 3: An illustration of the commonsense-aware memory network.

Specifically, given the concept triplets (n_i, r, n_j) in G_{meta}, we regard the concept n_i and the relation r as the key, and the concept n_j as the value. The representations of n_i and n_j are from the metagraph encoder. The relation representations are initialized from TransE and will be optimized through continual pre-training. Thus the representation matrices of keys and values are:

$$\mathbf{K}^{(l)} = \{[\mathbf{n}_{i_1}^l; \mathbf{r}_1], [\mathbf{n}_{i_2}^l; \mathbf{r}_2], ..., [\mathbf{n}_{i_N}^l; \mathbf{r}_N]\} \quad (5)$$

$$\mathbf{V}^{(l)} = \{\mathbf{n}_{j_1}^l, \mathbf{n}_{j_2}^l, ..., \mathbf{n}_{j_N}^l\} \quad (6)$$

Memory Read Operation reads important prior information within the commonsense-aware memory to update the token representations from the text encoder. Specifically, we first calculate multi-head similarity matrices $\mathbf{S}_i$ between tokens and keys:

$$\mathbf{S}_i = \mathbf{H}^{(l)} \mathbf{W}_i^S \mathbf{K}^{(l)\mathsf{T}} \quad (7)$$

where $\mathbf{S}_i$ is the similarity of head i, and $\mathbf{W}_i^S$ is a learnable matrix. We aggregate values to update the token representations as:

$$\hat{\mathbf{H}}^{(l)} = \mathbf{H}^{(l)} + [\alpha_1 \mathbf{V}^{(l)}; \alpha_2 \mathbf{V}^{(l)}; ...; \alpha_h \mathbf{V}^{(l)}] \mathbf{W}^r \quad (8)$$

$$\alpha_i = \text{softmax}(\mathbf{S}_i) \quad (9)$$

where α_i is the attention score distribution along the key dimension and $\mathbf{W}^r$ is a learnable matrix. Through this, the token representations can be enriched by the concept representations from the

metagraph. Later, the updated token representations $\hat{\mathbf{H}}^{(l)}$ are fed into the next layer of PLMs.

Memory Write Operation updates the node representations from the metagraph encoder. Given the similarity matrices $\mathbf{S}_i$, we aggregate the token representations as:

$$\widetilde{\mathbf{V}}^{(l)} = [\beta_1\mathbf{H}^{(l)}; \beta_2\mathbf{H}^{(l)}; ...; \beta_h\mathbf{H}^{(l)}]\mathbf{W}^w \quad (10)$$

$$\beta_i = \text{softmax}(\mathbf{S}_i^{\mathsf{T}}) \quad (11)$$

where β_i is the attention score distribution along the token dimension and $\mathbf{W}^w$ is a learnable matrix. Then we design a gate to update the value representations:

$$g = \sigma(\widetilde{\mathbf{V}}^{(l)}\mathbf{W}^{new} + \mathbf{V}^{(l)}\mathbf{W}^{old}) \quad (12)$$

$$\hat{\mathbf{V}}^{(l)} = g\widetilde{\mathbf{V}}^{(l)} + (1-g)\mathbf{V}^{(l)} \quad (13)$$

where $\mathbf{W}^{new}$ and $\mathbf{W}^{ori}$ are learnable matrices. The gate mechanism allows for selective information flow from the original node representations and the updated ones. The updated node representations are later fed into the next layer of GAT.

3.3. Continual Pre-training and Fine-tuning

3.3.1. Continual Pre-training

We design the following three continual pre-training tasks to further enhance and fuse the text and the commonsense metagraph.

Masked Language Model (MLM). The event texts contain a number of event mentions and arguments, which may facilitate the understanding of event relations. Thus, we follow the masking mechanism in Devlin et al. (2019) and harness the MLM task to learn task-related semantic information at the lexical and sentence levels. The objective is to predict the masked tokens with cross-entropy loss.

Concept Triplet Completion (CTC). In the commonsense metagraph, the semantic correlations within concept triplets contributes to the understanding of event causalities. Thus, we design the concept triplet completion task. Given a triplet (n_i, r, n_j) in G_{meta}, the objective is:

$$L_{CTC} = \max(\gamma + d(\mathbf{n}_i + \mathbf{r}, \mathbf{n}_j) - d(\mathbf{n}_i + \mathbf{r}', \mathbf{n}_j)) \quad (14)$$

where $\gamma > 0$ is the margin, $d(*,*)$ denotes the euclidean distance and $\mathbf{r}'$ is the sampled negative relation embedding.

Text-Metagraph Contrastive Learning (CL). To further unify the representations of the texts and metagraphs, we adopt contrastive learning to bring those of the same event pairs together while separating the negatives samples. We use in-batch negatives and the training objective is:

$$L_{CL} = -\log \frac{exp(f(\mathbf{h}_{S_i}, \mathbf{n}_{G_i})/\tau)}{\sum_{i \neq j} exp(f(\mathbf{h}_{S_i}, \mathbf{n}_{G_j})/\tau)} \quad (15)$$

where S_i and G_i denote the text and metagraph of an event pair, $\mathbf{h}_{S_i}$ is the [CLS] token representation of S_i, $\mathbf{n}_{G_i}$ is the mean pooling of the node representations of G_i, $f(\bullet)$ denotes the cosine similarity and τ is the temperature hyperparameter.

We conduct continual pre-training on the BECAUSE corpus (Dunietz et al., 2017). The continual pre-training loss is $L_{MLM} + L_{CTC} + L_{CL}$.

3.3.2. Knowledge-enhanced Fine-tuning

We concatenate the representations of [CLS] token, e_1 and e_2 as the contextual representation of the given event pairs:

$$\mathbf{F}_{\mathsf{T}} = \mathbf{h}_S \oplus \mathbf{h}_{e_1} \oplus \mathbf{h}_{e_2} \quad (16)$$

We concatenate the corresponding node representations of e_1 and e_2 in G_{meta} as the commonsense enhanced event pair representation:

$$\mathbf{F}_{\mathsf{C}} = \mathbf{n}_{e_1} \oplus \mathbf{n}_{e_2} \quad (17)$$

Then we make the final prediction through binary classification:

$$\mathbf{p}_{(e_1,e_2)} = \text{softmax}(\mathbf{W}_o(\mathbf{F}_{\mathsf{T}} \oplus \mathbf{F}_{\mathsf{C}}) + \mathbf{b}_o) \quad (18)$$

where $\mathbf{W}_o$ and $\mathbf{b}_o$ are learnable parameters.

We use cross-entropy loss to fine-tune our pre-trained DFP model:

$$L_{DFP}(\Theta) = -\sum_s \sum_{\substack{e_i, e_j \in E_s \\ e_i \neq e_j}} \mathbf{y}_{(e_i,e_j)} \log(\mathbf{p}_{(e_1,e_2)}) \quad (19)$$

where Θ is the parameter set, s ranges over each sentence in the training set and e_i, e_j range over events in s.

4. Experiments

4.1. Datasets and Evaluation Metrics

We evaluate DFP model on two benchmark datasets, including EventStoryLine v0.9 (ESC) (Caselli and Vossen, 2017), which contains 258 documents, 4,316 sentences, 5,334 events in total, and 1,770 of 7,805 event pairs are causally related; Causal-TimeBank (CTB) (Mirza et al., 2014), which contains 184 documents, 6,813 events, and 318 of 7,608 event pairs are causally related. Same as previous work (Shen et al., 2022; Zuo et al., 2021a), we set the last two topics of ESC as the development set for two datasets. Notably, CNC (Tan et al., 2022) is a corpus of annotating event sentences instead of event pairs in sentences with causal labels, for which the task format is inconsistent with our setting. As a result, we do not consider it during experiments.

To ensure fairness, we conduct 5-fold and 10-fold cross-validation on ESC and CTB, respectively. Models are evaluated by Precision (P), Recall (R), and macro-F1 (F1). All the results are the average of three independent experiments.

Methods	P	R	F1
LSTM	34.0	41.5	37.4
Seq	32.7	44.9	37.8
ILP	37.4	55.8	44.7
BERT-base	36.9	56.0	44.5
KnowDis	39.7	66.5	49.7
LearnDA	42.2	69.8	52.6
CauSeRL	41.9	69.0	52.1
LSIN	47.9	58.1	52.5
KEPT	50.0	68.8	57.9
SemSIn	50.5	63.0	56.1
DFP	**55.9**	**69.8**	**62.1***

Table 1: Experimental results on ESC (%).

Methods	P	R	F1
RULE	36.8	12.3	18.4
DD	67.3	22.6	33.9
VerbRule	**69.0**	31.5	43.2
BERT-base	38.8	44.1	41.3
KnowDis	42.3	60.5	49.8
LearnDA	41.9	68.0	51.9
CauSeRL	43.6	**68.1**	53.2
LSIN	51.5	56.2	53.7
KEPT	48.2	60.0	53.5
SemSIn	52.3	65.8	58.3
DFP	53.7	64.2	**58.5**

Table 2: Experimental results on CTB (%).

4.2. Implementation Details

Our implementation uses HuggingFace Transformers [1] (Wolf et al., 2020) and PyTorch (Paszke et al., 2019). We harness ConceptNet 5.0[2] as the external commonsense graph. When pruning the commonsense graph, we perform grid search for $K \in [500, 1000, 2000, 4000]$ and set it to 1000. We use uncased BERT-base (Devlin et al., 2019) as the text encoder. For the metagraph encoder GAT, we set the number of layers, attention heads and hidden states to 6, 12 and 64, respectively. For the commonsense-aware memory network, we search $k \in [1, 2, 3, 4, 5]$ and set it to 2. In the continual pre-training stage, we pre-train the parameters with a total of 32 batch size for 200 steps. The max length of input sequences is set to 512. We use AdamW (Loshchilov and Hutter, 2019) optimizer and learning rate is set to 2e-4. During fine-tuning stage, we use AdamW with the same setting as pre-training. The batch size is set to 32, and the learning rate is set to 3e-5.

4.3. Baselines

We compare DFP with two types of SOTA methods, i.e., feature-based ones and PLMs-based ones. For ESC, we select: **LSTM** (Cheng and Miyao, 2017) is a sequential model based on the dependency path; **Seq** (Choubey and Huang, 2017) is a sequential model with handcrafted features; **ILP** (Gao et al., 2019) models causal structures with integer linear programming. For CTB, we choose: **RULE** (Mirza and Tonelli, 2014) is a rule-based framework; **DD** (Mirza and Tonelli, 2014) is a data driven supervised model; **VerbRule** (Mirza, 2016) enhances ECI with verb rule and gold causal signals.

we also compare with methods based on PLMs: **BERT-base** (Zuo et al., 2021b) is a BERT- based baseline with a linear classifier; **KnowDis** (Zuo et al., 2020) is a distantly supervised data augmentation method; **LearnDA** (Zuo et al., 2021b) is a knowledge-guided data augmentation model; **CauSeRL** (Zuo et al., 2021a) is a self-supervised framework which learns context-specific causal patterns; **LSIN** (Cao et al., 2021) is a knowledge-enriched latent structure induction model; **KEPT** (Liu et al., 2023) is a knowledge enhanced model which converts external triples into textual descriptions; **SemSIn** (Hu et al., 2023) leverage semantic structures in the context.

We directly adopt the best parameter setup reported in papers that originally introduced the methods listed. $\star$ denotes a paired t-test at a significance level of 0.05.

4.4. Main Results

Table 1 and Table 2 show the results of ECI on ESC and CTB datasets, respectively. From these results, we can find that:

(1) Overall, our method significantly or comparably outperforms all baselines in terms of the F1-score on both datasets. Compared with these methods, DFP achieves at least 4.2% and 0.2% F1-score improvements on the ESC and CTB datasets. This demonstrates the effectiveness of DFP. Moreover, the usage of PLMs boosts the performance. On both datasets, PLMs-based baselines comparably or consistently outperform feature-based ones.

(2) The experimental results of KnowDis, LearnDA, LSIN and KEPT show that the introduction of different external knowledge and the method of introducing external knowledge can affect the performance of the ECI model. We note that the performance of DFP is higher than those of other knowledge-enhanced methods. It indicates that DFP is more advantageous in leveraging external knowledge for the ECI task, by deeply integrating external knowledge graphs into PLMs.

(3) DFP significantly outperforms KEPT which utilizes descriptive and relational information from ConceptNet. The reason may be that KEPT simply utilizes the knowledge of concept triplets for

[1] https://huggingface.co/transformers/

[2] https://github.com/commonsense/conceptnet5

Methods	P	R	F1
w/o. graph pruning	55.0	69.7	61.5(-0.6)
w/o. one-hop definition	54.5	68.5	60.7(-1.4)
w/o. multi-hop path	52.3	62.9	57.1(-5.0)
w/o. memory network	48.9	65.0	55.8(-6.3)
w/o. continual pre-training	51.5	66.9	58.2 (-3.9)
w/o. MLM	53.4	67.9	59.8(-2.3)
w/o. CTC	54.9	68.6	61.0(-1.1)
w/o. CL	54.4	68.2	60.5 (-1.6)
DFP	**55.9**	**69.8**	**62.1**

Table 3: Ablation results on ESC (%).

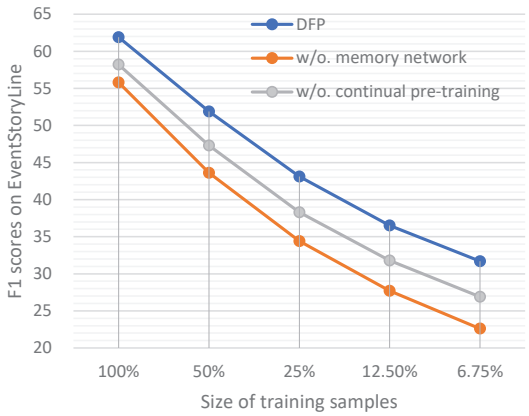

Figure 4: Effect of varying training sample size on ESC.

event causality inference, neglecting the information interaction between the text and the external graph. Whereas our DFP deeply fuses the heterogeneous information with commonsense-aware memory network and continual pre-training, obtaining more informative event pair embeddings for causality inference.

(4) DFP significantly outperforms SemSIn by 6% F1-score on the ESC dataset, which only 0.2% on the CTB dataset. The reason may be that ESC is an event story line dataset, in which sentences are higher related than CTB which is from news. SemSIn digs the information contained in the independent sentence with the AMR parser, which may be less effective for ESC than our model. Comparatively, the enhanced PLMs of DFP might be more effective for the ESC dataset which contains texts with a certain degree of causal continuity.

4.5. Ablation Experiment

To elucidate the effect of main components, we set up ablation experiments and design eight internal baselines: **w/o. graph pruning** removes the commonsense graph pruning process and directly induces metagraph from ConceptNet; **w/o. one-hop definition** removes one-hop definition knowledge of the event pair from the original commonsense metagraph; **w/o. multi-hop path** removes multi-hop relational path knowledge of the event pair from the original commonsense metagraph; **w/o. memory network** removes commonsense-aware memory network module and directly concatenates event pair representations from the PLMs and the GAT for fine-tuning and inference; **w/o. continual pre-training** removes all continual pre-training tasks and only conducts knowledge-enhanced fine-tuning; **w/o. MLM**, **w/o. CTC** and **w/o. CL** remove masked language model, concept triplet completion and text-metagraph contrastive learning tasks respectively.

Results of internal baselines on the ESC dataset are shown in Table 3. As seen:

(1) DFP significantly outperforms all internal baselines on ESC. Compared to DFP, w/o. graph pruning drops 0.6% F1-score. The reason may be that the original commonsense graph exists noisy data which may affect the performance.

(2) Experimental results of w/o. one-hop definition and w/o. multi-hop path demonstrate that, after removing either definition knowledge or relational path knowledge, the F1-scores go down. It indicates that these two kinds of knowledge both contribute to our model. Simultaneously using two kinds of knowledge further improves the overall performance, which reveals that the external structured knowledge in the commonsense graphs is effective for causality inference.

(3) w/o. memory network shows huge performance drop, with a decrease of 6.3% F1-score on ESC. This result demonstrates that the deep information interaction between the text and the graph contributes to more informative event representations, impelling better event pair relation inference.

(4) When the continual pre-training is removed, the F1-score decreases, indicating the significance of designing continual pre-training tasks to further inject task-specific knowledge into the PLMs. Besides, removing MLM task also results in a large performance drop, since it further injects knowledge of events into our model.

4.6. Analysis of Training Data

To validate the reliability of DFP under the data lacking scenarios, we test DFP, w/o. memory network and w/o. continual pre-training on randomly sub-sampled labeled data of ESC training set. As shown in Figure 4, the performance of three models drops with the decline of training sample size. However, our full model performs consistently better than two internal baselines. Moreover, we observe from the vertical line that the degree of performance drop of the full model is less than that of the other two internal baselines. In sum, the above observations demonstrate that DFP is capable of leveraging the data more effectively with the help of the commonsense-aware memory network and

Samples	w/o. memory network	w/o. continual pre-training	SemSIn	DFP
1) Iraq said it **invaded** Kuwait <u>because of</u> **disputes** over oil …	√	√	√	√
2) The **fights** erupted in Flatbush, and 46 were **arrested** at Wednesday …	×	√	√	√
3) more traditional groups are also **opening** new chapters, <u>thanks</u> in part to their ability to **use** new technologies …	√	×	×	√

Figure 5: Results of case study where bold denotes target events, and underlined words indicate causal clues.

continual pre-training tasks.

4.7. Case Study

We conduct a case study to further visually demonstrate the effectiveness of DFP. Figure 5 shows three instances as well as the identification results of DFP, two internal baselines and previous SOTA method SemSIn. Instance 1 is a simple sample of explicit causality, and all methods can make correct predictions. Instance 2 is an implicit sample without any clue. However, there is a relational path between two events in the commonsense graph, i.e., *fight* $\xrightarrow{HasSubevent}$ *hurt someone* $\xrightarrow{HasSubevent}$ *get arrested*. Methods with the memory network can elicit relational knowledge to make correct predictions. In instance 3, there is no explicit semantic relationship between two events, and the clue word "thanks" did not appear in the training set. Nonetheless, methods with continual pre-training can correctly identify it, indicating that the continual pre-training can elicit PLMs' ability to identify causal clues.

4.8. Compared with In-Context Learning

To illustrate that ECI task is still not well-solved in the era of Large Language Models (LLMs), we set such baseline instructing LLMs to identify event causalities by the means of in-context learning (ICL). In this work, we adopt ChatGPT (gpt-3.5-turbo-0301) provided by OpenAI APIs [3] for in-context learning. We design the prompt to simulate the chatting history between the user and the model. Specifically, it contains three parts: (1) the instruction telling LLMs the task purposes and the output format, (2) the demonstration giving an input-output pair to teach LLMs and (3) some test instances. We feed the prompt into LLMs and expect them to generate answers.

Figure 6 presents two examples which have been correctly identified by DFP. In instance 2, ChatGPT is confused and wrongly identifies the causal relation between "*escaped*" and "*transporting*", because if the inmate was not transported, he would not have a chance to escape. However, in this context, "*transporting*" does not directly lead to "*escape*". From the test instances, we find that even a powerful tool like ChatGPT still face challenges in dealing with some ambiguous causalities. A series of recent work (Jin et al., 2023; Kiciman et al., 2023) observe the similar results as ours, that off-the-shelf LLMs perform poorly on inferring causation from correlation. Moreover, it is important to consider that ChatGPT relies on large-scale corpora and high-performance hardwares, while for most research environments, such conditions are almost impossible. In contrast, DFP requires relatively fewer conditions, is easier to implement, and is more reliable in dealing with some ambiguous causalities.

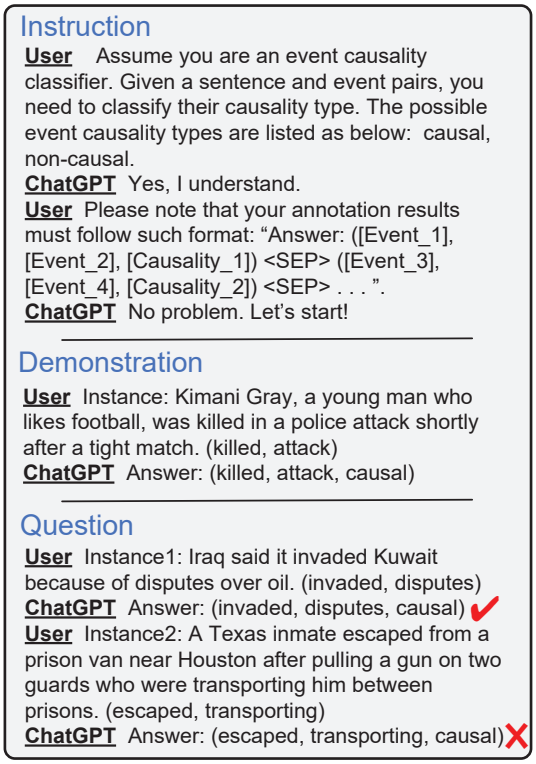

Figure 6: The instruction, demonstration and test questions of in-context learning.

[3] https://platform.openai.com/docs/api-reference

5. Conclusion

In this paper, we propose an effective pipeline, i.e., distill, fuse and pre-train (DFP) to integrate reliable task-specific knowledge from existing commonsense graph for ECI task. DFP includes a *commonsense graph distillation* module which aims to distill task-oriented metagraph, a *heterogeneous information fusion* module which adopts commonsense-aware memory network to fuse the text and metagraph, and a *continual pre-training* module which further injects task-specific knowledge into the model. Extensive experiments demonstrate the effectiveness of DFP over existing baselines.

6. Acknowledgments

We would like to thank the PC chairs, and the anonymous reviewers for their invaluable suggestions and feedback. This work is partially supported by National Key RD Program of China No. 2022YFB3103600, NSFC under grants Nos. U23A20296, 72371245, 62376284, and the Science and Technology Innovation Program of Hunan Province No. 2023RC1007.

7. Bibliographical References

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 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*. ACL.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795.

Pengfei Cao, Xinyu Zuo, Yubo Chen, Kang Liu, Jun Zhao, Yuguang Chen, and Weihua Peng. 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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4862–4872. Association for Computational Linguistics.

Fei Cheng and Yusuke Miyao. 2017. Classifying temporal relations by bidirectional LSTM over dependency paths. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 1–6. 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, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1796–1802. 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 *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

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, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 294–303. ACL.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1808–1817. 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 2019, Hong Kong, China, November 3-7, 2019*,

pages 2986–2997. Association for Computational Linguistics.

Chikara Hashimoto, Kentaro Torisawa, Julien Kloetzer, Motoki Sano, István Varga, Jong-Hoon Oh, and Yutaka Kidawara. 2014. Toward future scenario generation: Extracting event causality exploiting semantic relation, context, and association features. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pages 987–997. The Association for Computer Linguistics.

Zhichao Hu, Elahe Rahimtoroghi, and Marilyn A. Walker. 2017. Inference of fine-grained event causality from blogs and films. In *Proceedings of the Events and Stories in the News Workshop@ACL 2017, Vancouver, Canada, August 4, 2017*, pages 52–58. 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), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 10901–10913. Association for Computational Linguistics.

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona T. Diab, and Bernhard Schölkopf. 2023. Can large language models infer causation from correlation? *CoRR*, abs/2306.05836.

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 2019, Hong Kong, China, November 3-7, 2019*, pages 5815–5821. Association for Computational Linguistics.

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *CoRR*, abs/2305.00050.

Canasai Kruengkrai, Kentaro Torisawa, Chikara Hashimoto, Julien Kloetzer, Jong-Hoon Oh, and Masahiro Tanaka. 2017. Improving event causality recognition with multiple background knowledge sources using multi-column convolutional neural networks. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3466–3473. AAAI Press.

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

Jintao Liu, Zequn Zhang, Zhi Guo, Li Jin, Xiaoyu Li, Kaiwen Wei, and Xian Sun. 2023. KEPT: knowledge enhanced prompt tuning for event causality identification. *Knowl. Based Syst.*, 259:110064.

Qi Liu, Dani Yogatama, and Phil Blunsom. 2022. Relational memory-augmented language models. *Trans. Assoc. Comput. Linguistics*, 10:555–572.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020b. K-BERT: enabling language representation with knowledge graph. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 2901–2908. AAAI Press.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Paramita Mirza. 2016. *Extracting Temporal and Causal Relations between Events*. Ph.D. thesis, University of Trento, Italy.

Paramita Mirza and Sara Tonelli. 2014. An analysis of causality between events and its relation to temporal information. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 2097–2106. ACL.

Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Ryu Iida, Masahiro Tanaka, and Julien Kloetzer. 2016. A semi-supervised learning approach to why-question answering. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 3022–3029. AAAI Press.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. 2012. Learning causality for news events prediction. In *Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16-20, 2012*, pages 909–918. ACM.

Mehwish Riaz and Roxana Girju. 2010. Another look at causality: Discovering scenario-specific contingency relationships with no supervision. In *Proceedings of the 4th IEEE International Conference on Semantic Computing (ICSC 2010), September 22-24, 2010, Carnegie Mellon University, Pittsburgh, PA, USA*, pages 361–368. IEEE Computer Society.

Mehwish Riaz and Roxana Girju. 2013. Toward a better understanding of causality between verbal events: Extraction and analysis of the causal power of verb-verb associations. In *Proceedings of the SIGDIAL 2013 Conference, The 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 22-24 August 2013, SUPELEC, Metz, France*, pages 21–30. The Association for Computer Linguistics.

Mehwish Riaz and Roxana Girju. 2014. In-depth exploitation of noun and verb semantics to identify causation in verb-noun pairs. In *Proceedings of the SIGDIAL 2014 Conference, The 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 18-20 June 2014, Philadelphia, PA, USA*, pages 161–170. The Association for Computer Linguistics.

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, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 2288–2299. 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 Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *CoRR*, abs/2107.02137.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 535–546. Association for Computational Linguistics.

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D. Manning, and Jure Leskovec. 2022. Greaselm: Graph reasoning enhanced language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

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, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2162–2172. 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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3558–3571. 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, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 1544–1550. International Committee on Computational Linguistics.

8. Language Resource References

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@ACL 2017, Vancouver, Canada, August 4, 2017*, pages 77–86. Association for Computational Linguistics.

Jesse Dunietz, Lori S. Levin, and Jaime G. Carbonell. 2017. The because corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the 11th Linguistic Annotation Workshop, LAW@EACL 2017, Valencia, Spain, April 3, 2017*, pages 95–104. Association for Computational Linguistics.

P. Mirza, R. Sprugnoli, S. Tonelli, and M. Speranza. 2014. Annotating causality in the tempeval-3 corpus. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*.

Fiona Anting Tan, Ali Hürriyetoglu, Tommaso Caselli, Nelleke Oostdijk, Tadashi Nomoto, Hansi Hettiarachchi, Iqra Ameer, Onur Uca, Farhana Ferdousi Liza, and Tiancheng Hu. 2022. The causal news corpus: Annotating causal relations in event sentences from news. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 2298–2310. European Language Resources Association.