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] and Weidong Xiao[1]
1 National University of Defense Technology, Changsha, China
2 Information Research Center of Military Science, Beijing, China

Introduction

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.

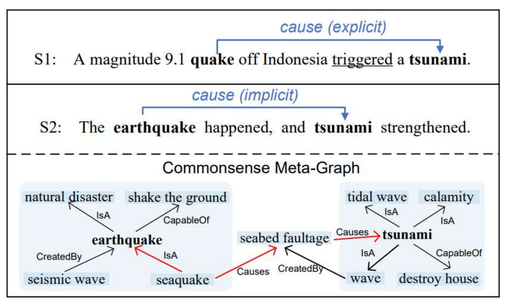

Figure 1: Examples of different causalities.

Methods

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.

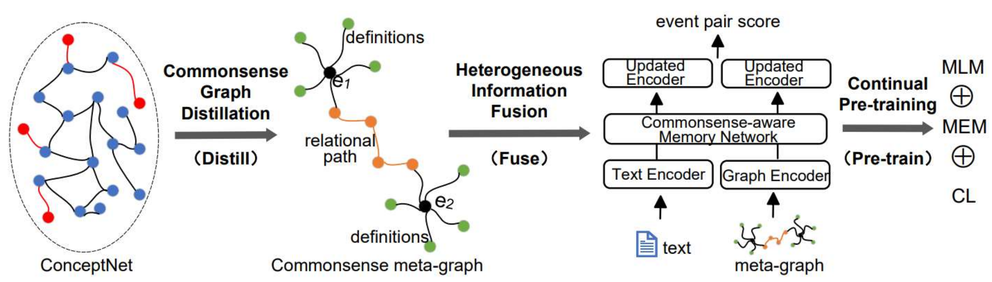

Figure 2: Overview of DFP pipeline.

Commonsense Graph Distillation

We harness ConceptNet as the external commonsense graph. Directly introducing ConceptNet is unsuitable, as itcontains 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.

Commonsense Graph Pruning Given a commonsense graph G, we harness TransE to measure the confidence of a given triplet in G. We first calculate the distance between two linked concepts as follows: $D(n_i, n_j) = \frac{1}{\mathbf{n}_i\mathbf{r} + \mathbf{n}_i\mathbf{n}_j + \mathbf{r}\mathbf{n}_j}$

$\mathbf{n}$ and $\mathbf{r}$ are the TransE embeddings of concept and relation.

Then, given each node in G, we keep the top-K neighboring nodes:

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

Thus, the pruned commonsense graph is $\overline{G} = \{(n_i, r, n_j) | n_j \in N(n_i)\}$

Metagraph Induction Given an event pair (e_1, e_2), the aim of metagraph induction is to construct a task-oriented commonsense metagraph. To obtain the definition knowledge, we first match the event mention with concept tokens in G through matching rules. Then we search the onehop 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 from $\overline{G}$. 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 and the multi-hop path.

Heterogeneous Information Fusion

We propose a commonsense-aware memory network to deeply fuse the explicit knowledge in G_{meta} with the text knowledge.

Text Encoder We use BERT as PLMs to encode the input sentence, obtaining token representations $\mathbf{H}^{(l)} = \{\mathbf{h}_1^{(l)}, \mathbf{h}_2^{(l)}, ..., \mathbf{h}_L^{(l)}\}$.

Metagraph Encoder We harness the graph attention network (GAT) to encode the commonsense metagraph. GAT first initializes the node embeddings by TransE. Then the node representation is updated as:

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

Commonsense-aware Memory Network The input (n_i, r, n_j), the representation matrices of keys and values are

$$\mathbf{K}^{(l)} = \{[\mathbf{n}_{i_1}^l; \mathbf{r}_1], [\mathbf{n}_{i_2}^l; \mathbf{r}_2], ..., [\mathbf{n}_{i_N}^l; \mathbf{r}_N]\}$$
$$\mathbf{V}^{(l)} = \{\mathbf{n}_{j_1}^l, \mathbf{n}_{j_2}^l, ..., \mathbf{n}_{j_N}^l\}$$

Memory Read operation

$$\mathbf{S}_i = \mathbf{H}^{(l)}\mathbf{W}_i^S \mathbf{K}^{(l)\top}$$
$$\hat{\mathbf{H}}^{(l)} = \mathbf{H}^{(l)} + [\alpha_1\mathbf{V}^{(l)}; \alpha_2\mathbf{V}^{(l)}; ...; \alpha_h\mathbf{V}^{(l)}]\mathbf{W}^r$$
$$\alpha_i = \text{softmax}(\mathbf{S}_i)$$

Memory Write operation

$$\widetilde{\mathbf{V}}^{(l)} = [\beta_1\mathbf{H}^{(l)}; \beta_2\mathbf{H}^{(l)}; ...; \beta_h\mathbf{H}^{(l)}]\mathbf{W}^w$$
$$\beta_i = \text{softmax}(\mathbf{S}_i^\top)$$
$$g = \sigma(\widetilde{\mathbf{V}}^{(l)}\mathbf{W}^{new} + \mathbf{V}^{(l)}\mathbf{W}^{old})$$
$$\hat{\mathbf{V}}^{(l)} = g\widetilde{\mathbf{V}}^{(l)} + (1-g)\mathbf{V}^{(l)}$$

Continual Pre-training and Fine-tuning

Continual Pre-training Masked Language Model (MLM), Concept Triplet Completion (CTC), Text-Metagraph Contrastive Learning (CL).

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

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

Continual pre-training loss: $L_{MLM} + L_{MEM} + L_{CL}$

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

Results

We evaluate DFP model on two benchmark datasets, including EventStoryLine v0.9(ESC), which contains and Causal-TimeBank (CTB).

Main results

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 (%).

Ablation Experiment

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 (%).

Case Study

Samples	w/o. memory network	w/o. continual pre-training	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.