Unified Generative Model with Multi-Dimensional Prefix for Zero-Shot Event-Relational Reasoning

Zhengwei Tao[1] **Zhi Jin**[1*] **Haiyan Zhao**[1]
Chengfeng Dou[1] **Yongqiang Zhao**[1] **Tao Shen**[2] **Chongyang Tao**[3]
[1]Peking University, [2]FEIT, University of Technology Sydney, [3]Microsoft
{tttzw, yongqiangzhao}@stu.pku.edu.cn,
{zhijin,zhhy.sei,chengfengdou}@pku.edu.cn
tao.shen@uts.edu.au, chotao@microsoft.com

Abstract

Reasoning about events and their relations attracts surging research efforts since it is regarded as an indispensable ability to fulfill various event-centric or common-sense reasoning tasks. However, these tasks often suffer from limited data availability due to the labor-intensive nature of their annotations. Consequently, recent studies have explored knowledge transfer approaches within a multi-task learning framework to address this challenge. Although such methods have achieved acceptable results, such brute-force solutions struggle to effectively transfer event-relational knowledge due to the vast array of inter-event relations (e.g. temporal, causal, conditional) and reasoning formulations (e.g. discriminative, abductive, ending prediction). To enhance knowledge transfer and enable zero-shot generalization among various combinations, in this work we propose a novel unified framework, called UNIEVENT. Inspired by prefix-based multi-task learning, our approach organizes event relational reasoning tasks into a coordinate system with multiple axes, representing inter-event relations and reasoning formulations. We then train a unified text-to-text generative model that utilizes coordinate-assigning prefixes for each task. By leveraging our adapted prefixes, our unified model achieves state-of-the-art or competitive performance on both zero-shot and supervised reasoning tasks, as demonstrated in extensive experiments.

1 Introduction

An 'event' is defined as a semantic molecule to explain the states or actions of a person, entity, or thing (Zhou et al., 2022). In natural language literature, it is usually represented as a span in narrative text (e.g., sentences, paragraphs or documents), which is composed of an event trigger (e.g., predicate) and its arguments (e.g., subject, object, adverbial modifier). Based on the semantic unit at the event level, a broad spectrum of *event-relational reasoning* tasks was presented to learn various inter-event relations (e.g., temporal, causal, conditional) and thus enable commonsense or cognitive reasoning capabilities for advanced AI systems. The inherent event-relational reasoning logic has been formulated as tasks as relation extraction (Han et al., 2021b), question answering (Yang et al., 2022b; Han et al., 2021a), intend prediction (Rashkin et al., 2018), summarization (Daumé and Marcu, 2006) and knowledge base construc-

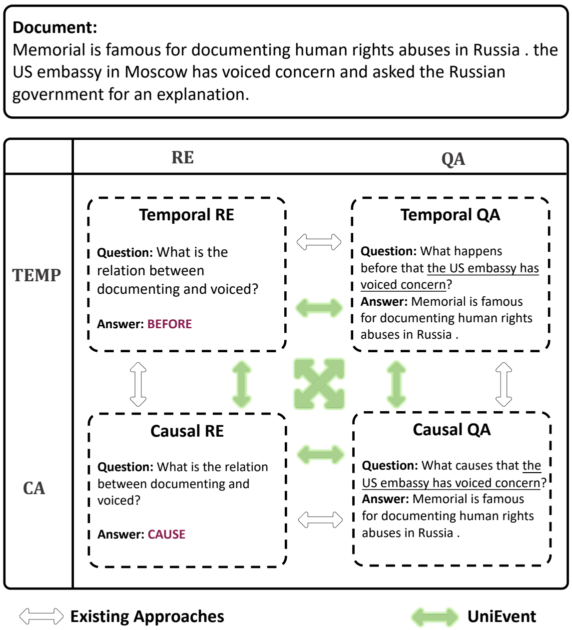

Figure 1: Illustration of knowledge transfer types across event-relational reasoning tasks. Existing approaches can only achieve inter-relation or inter-formulation transfer while UNIEVENT succeeds in all.

*Corresponding author.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7088–7102

tion (Sap et al., 2019; Li et al., 2020).

Attributed to recently advanced language models (e.g., BERT and GPT-3) pre-trained on raw corpora with billions of words in a self-supervised manner, data-driven methods via a *pre-training & fine-tuning* paradigm achieves acceptable performance on the event-relational reasoning tasks (Han et al., 2021b; Chen et al., 2022; Man et al., 2022a). Nonetheless, its inherently complex intra-event semantics and intricate inter-event relations inevitably increase the labor intensity of human annotation processes (e.g., experts-required, time-consuming, label-inconsistent). This limits the scale of human-labeled data for fine-tuning and thus affects the effectiveness of the data-driven methods on those tasks (Ning et al., 2018). For example, considering the event temporal question answering task, there are only 198 training instances in CIDER (Ghosal et al., 2021) among all datasets.

Therefore, such a data-scarcity issue necessitates knowledge transfer to an event-relational reasoning task. Besides task-specific heuristic pseudo labeling in a self- or semi-supervised framework to transfer from large-scale in-domain raw corpus, recent event-centric research works resort to supervised knowledge transfer due to its general learning methodology and superior fine-tuning performance. That is, transferring knowledge among supervised datasets under a variety of inter-event relations (e.g., temporal, causal) (Han et al., 2019; Wang et al., 2020) and reasoning formulations (e.g., event relation extraction, question answering) (Tang et al., 2021; Li et al., 2022b; Lourie et al., 2021). Despite their superior transfer performance, as shown in Figure 1, these works do not well consider knowledge transfer among a variety of both targeted relations and reasoning formulations in event-relational reasoning, and they usually fail to generalize to unseen event-relational reasoning tasks with distinct relations and/or formulations. For example, according to our empirical study shown in NT column of Table 3, unified training on T5 fails to transfer to tasks both unseen in formulation and relation.

To enhance knowledge transfer and empower zero-shot generalization among event-relational reasoning tasks, in this work we propose a brand-new unified framework UNIEVENT for zero-shot event-relational reasoning tasks transferring. We first categorize all event-relational reasoning tasks according to their original formulation types and event relation. We then construct generative formats for each task and convert them into generation forms. We train on adapted tasks based on a pretrained generation model (Raffel et al., 2020). Based on that, the proposed unified model enables implicit transfer across event-relational reasoning tasks. However, without explicitly discriminating the categorical coordination of the data, straightforward multi-task training may suffer from negative transfer (Liu et al., 2019) and intensive diversity of formulations and relations. Therefore, inspired by recent success of prompt tuning (Lester et al., 2021; Li and Liang, 2021; Liu et al., 2021b) where prompt instruction show great benefit in multi-task training (Sanh et al., 2021; Wei et al., 2021; Xu et al., 2022; Raffel et al., 2020), we propose to add prefix (Li and Liang, 2021) adapting to diversified formulations and relations. This multi-dimensional prefix additionally facilitates further transfer across tasks. We introduce to generate of these prefixes via the Adaptive Prefix Generators which allow sharing of flexible features among distinct dimensions. We then perform a contrastive regularization (Wu et al., 2020; Su et al., 2021) to learn to discriminate various task formulations and relations and enhance the representation.

We conduct extensive experiments on 16 datasets (3 for multi-task training, 13 for testing). Experiment results demonstrate that our method shows significant transferability and outperforms baselines in both zero-shot and full data multi-task settings We summarize our contributions as:

- We propose UNIEVENT for zero-shot event-relational reasoning tasks. We first categorize the event-relational reasoning tasks by the task formulation type and event relation. Then we unify the training datasets with event-relational reasoning targeted generative formats to enable knowledge transfer.
- We propose the Adaptive Prefix Generator to generate prefixes to guide the event-relational reasoning process. We also put up with a formulation- and relation-aware contrastive regularization to enhance further knowledge transfer across relations and formulations.
- We conduct extensive experiments to testify to our method. UNIEVENT outperforms the baselines on average of all datasets in both zero-shot and full-data settings.

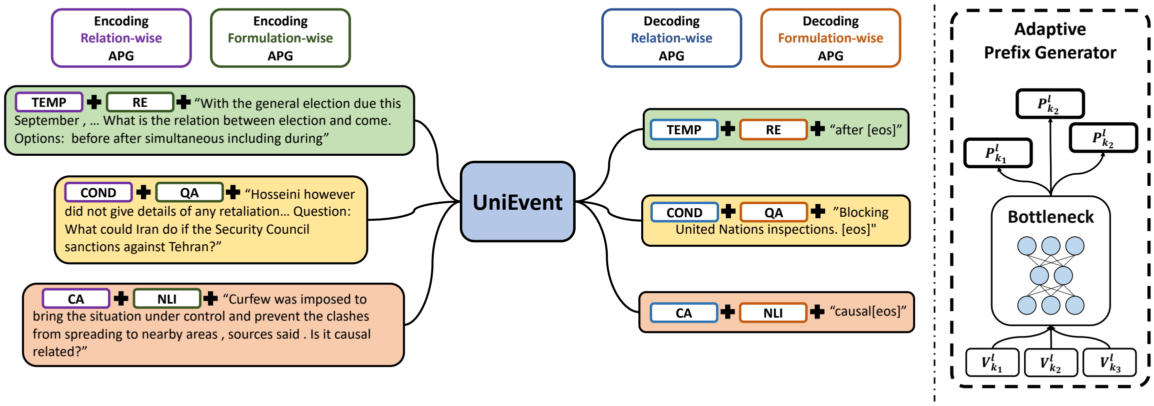

Figure 2: Overview of UNIEVENT. We propose to unify event-relational reasoning tasks with constructed generative formats. We use the Adaptive Prefix Generators to generate formulation-wise and relation-wise prefixes in both the encoder and decoder sides. In total, there are four Adaptive Prefix Generators of the same architecture. We illustrate the architecture on the right.

2 Method

Task Formulation. The objective of our study is to train a model using a combination of training datasets from different task formulations and event relations, enabling it to transfer its learning to a set of unseen datasets that were withheld during training. Formally, given a unified training dataset $\mathbb{T} = \bigcup \mathcal{T}_i$, we aim to train a model $P(\mathcal{Y}|\mathcal{X})$ on $\mathbb{T}$. Each data $(\mathcal{X}, \mathcal{Y}, \varsigma) \in \mathbb{T}$ consists of an input $\mathcal{X}$, an label $\mathcal{Y}$ and the original task formulation type ς. In summary, our framework encompasses relation extraction, natural language inference, question answering, and multiple-choice formulations., i.e. $\varsigma \in \{\texttt{RE}, \texttt{NLI}, \texttt{QA}, \texttt{MC}\}$. For all types of formulation, the inputs $\mathcal{X}$ and label $\mathcal{Y}$ are specifically:

$$(\mathcal{X}, \mathcal{Y}) = \begin{cases} ((\mathcal{D}, \mathcal{E}_0, \mathcal{E}_1, \gamma),\ \mathcal{L}), & \varsigma = RE \\ ((\mathcal{D},\ \gamma),\ \mathcal{L}), & \varsigma = NLI \\ ((\mathcal{D},\ \mathcal{Q},\ \gamma),\ \mathcal{A}), & \varsigma = QA \\ ((\mathcal{D},\ \mathcal{E}_0,\ \mathcal{E}_1,\ \mathcal{I},\ \gamma),\ \mathcal{A}), & \varsigma = MC \end{cases}$$

where $\mathcal{D}$ indicates the document, $\mathcal{E}_0$ and $\mathcal{E}_1$ are two queried events, $\mathcal{Q}$ is a question about events, $\mathcal{I}$ stands for the queried dimension (e.g. `cause` and `result` for event causality tasks), γ denotes for inherent event relation of that data, $\mathcal{L}$ is the gold label and $\mathcal{A}$ is the gold answer text. Then we transfer the models to the held-out unseen datasets $\mathbb{Z} = \bigcup \mathcal{Z}_z$ which is also composed by such four types of tasks. In this paper, we mainly consider four event relations which are temporal (`TEMP`), causal (`CA`), counterfactual (`COUNT`), and conditional (`COND`). We finally result in tasks taxonomy as shown in Table 1.

Form.	Rel.	Task
RE	TEMP	TBD (Chambers et al., 2014), MA (Ning et al., 2018) RED (O'Gorman et al., 2016), TM (Naik et al., 2019)
	CA	ESL (Caselli and Vossen, 2017), SCI (Li et al., 2021) CTB (Mirza and Tonelli, 2016)
NLI	CA	CNC (Tan et al., 2022a), ALT (Liang et al., 2022)
MC		ECA (Du et al., 2022)
QA	CA	EST (Han et al., 2021a), CQA (Yang et al., 2022c) RI (Poria et al., 2021), RD (Poria et al., 2021) CID (Ghosal et al., 2021)
	TEMP	CID (Ghosal et al., 2021)
	COUNT	EST (Han et al., 2021a), CQA (Yang et al., 2022c) SE (Yang et al., 2020), CID (Ghosal et al., 2021) EST (Han et al., 2021a), CQA (Yang et al., 2022c)
	COND	CID (Ghosal et al., 2021)

Table 1: Event-relational reasoning tasks taxonomy. We categorize these tasks according to their task formulations and event relations. Some of the tasks cover more than one relations such as CQA and EST.

Model Overview. Our model undergoes training on unified diverse datasets of task formulations and event relations, followed by evaluation on held-out test sets where it encounters zero-shot scenarios. We first convert all tasks into text-to-text generation based on our constructed generative formats as in Section 2.1. After that, UNIEVENT takes input with multi-dimensional prefix concatenated and generates output sequence. To improve knowledge transfer, we use the Adaptive Prefix Generators to generate the above prefixes according to the formulation and the relation of each data as in Section 2.2 and propose the formulation- and relation-aware contrastive regularizationas in Section 2.3. Finally, UNIEVENT perform unified multi-task training in Section 2.4. We depict an overview of UNIEVENT as shown in Figure 2.

2.1 Unified Generative Adaptation

We adapt all tasks into generation forms with constructed generative formats to enable unified generative training. However, there have been no available human-engineered prompts for event-related tasks so far. As is known to all, model performance is sensitive to the prompt and verbalizer designs (Shin et al., 2020). Such prompts from Prompt Source[1] are not directly suitable for event-relational reasoning tasks. Considering that, we construct the discrete generative formats from scratch. The generative format varies with task formulations and event relations, as listed in Figure 2. We mainly take RE as an example to explain the following process.

Input Adaptation. The adapted input is mainly a question "What is the relation between $\mathcal{E}_0$ and $\mathcal{E}_1$?". $\mathcal{E}_0$ and $\mathcal{E}_1$ are queried events. We prepend the document content placeholder $\mathcal{D}$ before the question. We also append optional $\mathcal{O}$ representing the candidate label set. For MC, there's another placeholder $\mathcal{I}$ which denotes for queried dimension(eg. `cause` and `effect` for the causal relation).

Ouput Adaptation. Conventionally in prompt tuning (Shin et al., 2020), we construct a verbalizer $\textsc{Verb}(\cdot)$ to map relation labels $\mathcal{L}$ to label words. As is shown in Figure 2, we show all verbalizers for all mentioned event relations. After, the generation output is the label word $\textsc{Verb}(\mathcal{L})$ appended by the $[eos]$ indicator. In QA and MC, we directly take the original answer $\mathcal{A}$ to compose the output. As a result, for data of any formulation and relation, we convert it into a text-to-text form with input $\mathcal{X}$ and linearized output sequence $\mathcal{Y}$.

Model Generation. Then given an input $\mathcal{X}$, $\mathcal{X} = (x_1, x_2, ..., x_n)$ where x_i is the i^{th} token of the input $\mathcal{X}$ and n is the sequence length, UniEvent output the prediction by generating the linearized answer $\mathcal{Y}$. The generation process is modeled by a pretrained encoder-decoder language model $\mathcal{M}$ such as BART (Lewis et al., 2019) and T5 (Raffel et al., 2020) which are pretrained on a large-scale corpus. After the generation adaptation, UniEvent first encode $\mathcal{X}$ by the encoder ***Enc*** of $\mathcal{M}$. Each encoder layer of $\mathcal{M}$ is a multi-head self-attention (Vaswani et al., 2017) block which take $\boldsymbol{H}^l \in \mathbb{R}^{n\times d}$ as input to compute input of next layer $\boldsymbol{H}^{l+1} = \boldsymbol{Enc}^l(\boldsymbol{H}^l; \theta_e^l)$. d is the hidden state dimension. UniEvent then generate answer $\mathcal{Y}$ with decoder of $\mathcal{M}$ in an auto-regressive generation process. We use $\theta_{\mathcal{M}} = (\theta_e, \theta_d)$ to denote both encoder and decoder parameters of $\mathcal{M}$

$$P(\mathcal{Y}|\mathcal{X}) = \prod_i \boldsymbol{Dec}(\mathcal{Y}_{<i}, \boldsymbol{H}; \theta_d). \quad (1)$$

2.2 Multi-Dimensional Prefix-Tuning

Straightforward unifying all tasks can impede a model's ability to recognize distinct formulations and relations, and could further result in negative transfer (Liu et al., 2019). To have UniEvent adapt to different tasks and relations while share basic information across them, we propose to use multi-dimensional prefix to instruct the generation. We generate formulation-wise prefix matrix $\boldsymbol{P}_{k^\varsigma}$ and relation-wise prefix matrix $\boldsymbol{P}_{k^\gamma}$ via our Adaptive Prefix Generators. To further train the Adaptive Prefix Generators and facilitate the discriminated representation, we propose the Task- and Relation-aware Contrastive Regularization.

2.2.1 Adaptive Prefix Generator (APG)

To better adapt UniEvent to different formulation types and relations, we utilize prepended layer-wise prefixes (Li and Liang, 2021) to guide the generation. Moreover, on account of sharing flexible features of various task formulations and event relations, we instead generate these prefixes via a novel Adaptive Prefix Generators.

We first introduce the learnable embeddings $\boldsymbol{V}_k^l \in \mathbb{R}^{s\times d^p}$ for various aspects in each layer, $k \in \mathbb{A}$. $\mathbb{A}$ can be any considering attributes which in this paper is the set of task formulations or event relations. d^p is the vector dimension, s is the length. $l \in [1, L]$ is the layer index. $\boldsymbol{V}_k^l$ can be randomly initialized or pretrained from other tasks before.

Given $\boldsymbol{V}_k^l$, our APG $g^l(\cdot)$ takes it as input and generates dimension-specific prefix $\boldsymbol{P}_k^l$. $g^l(\cdot)$ consists a trainable bottleneck layer which is a pair of down and up projections that firstly align different knowledge representations to the same semantic space and then project them to space of $\mathcal{M}$. Mathematically, given $\boldsymbol{V}_k^l$

$$\begin{aligned} \boldsymbol{P}_k^l[i,:] &:= g^l(\boldsymbol{V}_k^l[i,:]; \theta_g) \\ &= \boldsymbol{W}^{u^l}\ \boldsymbol{Tanh}\ (\boldsymbol{W}^{d^l} \boldsymbol{V}_k^l[i,:]), \quad (2) \\ \boldsymbol{P}_k^l[i,:] &\in \mathbb{R}^d, i \in [1, s], \end{aligned}$$

where $\boldsymbol{W}^{d^l} \in \mathbb{R}^{d^p\times d^m}$ and $\boldsymbol{W}^{u^l} \in \mathbb{R}^{d^m\times d}$. d^m is the mid dimension of the bottleneck layer. ***Tanh***

[1] https://github.com/bigscience-workshop/promptsource

Form.	Rel.	Generative Formats	Verbalizer
RE	TEMP	**Input:** $\mathcal{D}$ What is the relation between $\mathcal{E}_0$ and $\mathcal{E}_1$? Options: $\mathcal{O}$. Answer: **Output:** $\mathcal{L}$ $[eos]$	BEFORE: before AFTER: after INCLUDES: including VAGUE: vague IS_INCLUDED: during SIMULTANEOUS: simultaneous
	CA		CAUSAL: causal NONE: none
MC	CA	**Input:** What is the $\mathcal{I}$ of $\mathcal{D}$? Options: $\mathcal{E}_0$; $\mathcal{E}_1$. Answer: **Output:** $\mathcal{A}$ $[eos]$	-
NLI	CA	**Input:** $\mathcal{D}$ Is it causal related? Options: $\mathcal{O}$. Answer: **Output:** $\mathcal{L}$ $[eos]$	ENTAILMENT: causal CONTRADICTION: none
QA	*	**Input:** $\mathcal{D}$. $\mathcal{Q}$? **Output:** $\mathcal{A}$ $[eos]$	-

Table 2: Generative Formats and Verbalizers. We show inputs and outputs of all relations and formulations. $\mathcal{D}$, $\mathcal{E}$., $\mathcal{O}$ and $\mathcal{I}$ represents placeholders for document, queried events, options and queried dimension. $\mathcal{L}$ and $\mathcal{A}$ stands for answer label words and answer sequence.

is the hyperbolic tangent activation function. θ_g is the parameter for APG. The APG can apply to the both formulation and relation axis. Specifically, for formulation-wise APG, the attributes set $\mathbb{A}$ is:

$$\mathbb{A} = \{\texttt{RE}, \texttt{NLI}, \texttt{QA}, \texttt{MC}\}.$$

Turning to event relation:

$$\mathbb{A} = \{\texttt{TEMP}, \texttt{CA}, \texttt{COUNT}, \texttt{COND}\}.$$

The Adaptive Prefix Generator are learned end-to-end with the backbone transformer $\mathcal{M}$.

2.2.2 Prefix Instructed Generation

To instruct the accomplishment of a task and induce considering task formulation and relational knowledge from the model. We prepend generated formulation-wise and relation-wise prefix matrix $\boldsymbol{P}^l_{k^\varsigma}$ and $\boldsymbol{P}^l_{k^\gamma}$ to inputs of each encoder layer of $\mathcal{M}$:

$$\boldsymbol{H}^{l+1} = \boldsymbol{Enc}^l([\boldsymbol{P}^l_{k^\varsigma}; \boldsymbol{P}^l_{k^\gamma}; \boldsymbol{H}^l]; \theta^l_e), \quad (3)$$
$$\boldsymbol{H}^l \in \mathbb{R}^{n\times d}, \boldsymbol{P}^l_{k^\varsigma} \in \mathbb{R}^{s^\varsigma \times d}, \boldsymbol{P}^l_{k^\gamma} \in \mathbb{R}^{s^\gamma \times d},$$

$\boldsymbol{H}$ is hidden states of the l^{th} layer. s^ς and s^γ are lengths of formulation-wise and relation-wise prefix respectively. $[;]$ is the concatenation operation.

We also add non-identical prefixes generated by another two APGs to each layer of decoder. Therefore, in total, we have four APGs in UNIEVENT.

2.3 Formulation- and Relation-aware Contrastive Regularization (TRC)

When trained solely on the supervised multi-task loss, a model has a tendency to undergo shortcuts of neglecting the prefixes. If this happens, UNIEVENT degrades to normal multi-task training on $\mathcal{M}$. To avoid such a dilemma and further adapt UNIEVENT to various dimensions, we add an additional contrastive regularization (Wu et al., 2020; Su et al., 2021). We take the vector $\boldsymbol{H}_{[bos]}$ of the first token $[bos]$ after all prefixes from the last layer's hidden states as the representation. Then we map $\boldsymbol{H}_{[bos]}$ to another space via a feed-forward layer $f(\cdot)$:

$$\boldsymbol{u}_\mathcal{X} := f(\boldsymbol{H}_{[bos]}; \theta_c) = \boldsymbol{Tanh}(\boldsymbol{W}^c \boldsymbol{H}_{[\text{bos}]}), \quad (4)$$

$\boldsymbol{u}_\mathcal{X} \in \mathbb{R}^d$. $\boldsymbol{W}^c \in \mathbb{R}^{d\times d^c}$. θ_c represents the parameters. Then we take $\boldsymbol{u}_\mathcal{X}$ as the representation of $\mathcal{X}$. For a data point $\mathcal{X}$ with its formulation type $\varsigma_\mathcal{X}$ and event relation $\gamma_\mathcal{X}$, we sample a subset $\mathbb{K}_\mathcal{X}$ from the whole training set. Then we conduct contrastive learning on $\mathcal{X}$ with $\mathbb{K}_\mathcal{X}$:

$$\varphi_\mathcal{X} = \sum_{\mathcal{X}_p \in \mathbb{K}^+_\mathcal{X}} \log \frac{\exp(\boldsymbol{u}_\mathcal{X} \cdot \boldsymbol{u}_{\mathcal{X}_p}/\tau)}{\sum_{\mathcal{X}_a \in \mathbb{K}_\mathcal{X}} \exp(\boldsymbol{u}_\mathcal{X} \cdot \boldsymbol{u}_{\mathcal{X}_a}/\tau)},$$
$$\mathcal{L}^C = -\sum_{\mathcal{X} \in \mathbb{T}} \frac{1}{|\mathbb{K}_\mathcal{X}|} \varphi_\mathcal{X}, \quad (5)$$
$$\mathbb{K}^+_\mathcal{X} = \{\mathcal{X}_p | \mathcal{X}_p \in \mathbb{K}_\mathcal{X}, \varsigma_{\mathcal{X}_p} = \varsigma_\mathcal{X} \wedge \gamma_{\mathcal{X}_p} = \gamma_\mathcal{X}\},$$

where τ is the temperature parameter and $\cdot$ is vector inner production.

2.4 Multi-Task Training

To train UNIEVENT, we perform multi-task training on $\mathbb{T}$. We shuffle all data of $\mathbb{T}$ which ends up with a mixed-up training batch composed of data from various datasets. Then we acquire the final training loss with scaling factor α:

$$\mathcal{L}^E = -\sum_{(\mathcal{X},\mathcal{Y}) \in \mathbb{T}} \log P(\mathcal{Y}|\mathcal{X}; \theta_\mathcal{M}, \theta_g, \theta_c), \quad (6)$$
$$\mathcal{L} = \mathcal{L}^E + \alpha \times \mathcal{L}^C,$$

	AVG							
	NT	**TEMP**	**CA**	**NLU**	**QA-F1**	**QA-EM**	**QA**	**ALL**
T5-zero (Raffel et al., 2020)	22.62	19.99	17.09	36.93	3.01	0.31	2.00	17.11
T0-3B (Sanh et al., 2021)	34.77	29.28	30.85	47.89	31.08	5.92	19.90	30.91
T5-unified (Raffel et al., 2020)	28.36	29.36	27.90	42.72	39.95	9.20	24.57	30.19
UniEvent (**Ours**)	**42.89**	**29.94**	**37.07**	**48.37**	**46.87**	**11.79**	**29.33**	**37.43**
	Ablation							
UniEvent - r (**Ours**)	30.89	29.29	31.83	46.76	**45.33**	**11.86**	25.32	34.85
UniEvent - t (**Ours**)	**38.92**	**33.39**	**36.59**	**47.96**	43.69	10.85	**27.27**	**36.02**
UniEvent - c (**Ours**)	38.76	28.11	36.13	47.09	40.81	10.07	25.44	35.07

Table 3: The main results for average zero-shot performance on event-relational reasoning tasks. Bold numbers are best scores for each average metrics. **ALL** averages scores of all metrics of all zero-shot dataset. **TEMP**, **CA**, **NLU**, **QA** average all metrics of temporal, causal, NLU and QA datasets respectively. **QA-F1** and **QA-EM** evaluate **F1** and **EM** of all QA datasets. **NT** denotes for all datasets of those there are no training datasets with both the same formulation type and event relation, namely ESL, CTB, SCI, ECA and temporal part of CID.

3 Experiments

3.1 Event-Relational Reasoning Datasets

In total, we assess the performance of UNIEVENT across 16 datasets that involve event-relational reasoning . Datasets can be divided by their original formulation types and event relation. Datasets we use are TB-Dense (TBD) (Chambers et al., 2014), MATRES (MA) (Ning et al., 2018), RED (O'Gorman et al., 2016), TD-DMan (TM) (Naik et al., 2019) which are temporal relation extraction; ESL (Caselli and Vossen, 2017)[2], SCITE (SCI) (Li et al., 2021), CTB (CTB) (Mirza and Tonelli, 2016) which are event causality identification; CNC (CNC) (Tan et al., 2022b), ALTLEX (ALT) (Liang et al., 2022) which are causal natural language inference; ESTER (EST) (Lester et al., 2021), CQA (Yang et al., 2022c) and CIDER (CID) (Ghosal et al., 2021) are multi-relational question extractive answering datasets which cover causal, counterfactual and conditional event relation. RECCON-IE (RI) and RECCON-DD (RD) (Poria et al., 2021) are causal QA tasks. SE2020-EQA (SE) (Yang et al., 2020) which is a counterfactual question answering task. ECARE (ECA) (Du et al., 2022) is a causal multiple choice task. To better show the results, in the following part, we organize RD, RI, SE, CQA, CID as QA part and leave the rest as NLU part. We summarize data statistics in Figure 8 and state the details of each dataset in Appendix A. We select TBD, CNC, and EST as train sets and leave others as held-out unseen test datasets.

[2]In this paper, we don't perform 5-folds cross-validation and instead split each dataset into 8 :1 :1 for training, validation and test.

3.2 Evaluation Metrics

All evaluation metrics follow previous researches on each dataset. We use micro-F1 score to evaluate all relation extraction tasks. Since causal NLI only has two labels(entailment and contradiction), we evaluate them by binary-F1 score. We denote both micro-F1 and binary-F1 as **F1**. We use F1-score (**F1**), **EM** to measure QA task. **F1** measures the correctness of uni-grams in generated sentence comparing those in ground truth sentences. **EM** score measures the exactly matches of uni-grams. In ESTER dataset, previous works also evaluate by HIT@1 which measures whether the event trigger words are generated in the sentences. Multiple choice tasks are measured by accuracy.

3.3 Parameters

We choose T5-base (Raffel et al., 2020) as the backbone of UNIEVENT. We set both formulation-wise and relational knowledge-wise prefix length s^{ς} and s^{γ} as 200. For all experiments, we use batch size 32, learning rate 5e-5 on AdamW optimizer. For contrastive learning, we set $\tau = 0.07$, $\alpha = 0.05$ and $|\mathbb{K}| = 512$. We don't use any optimization tricks like label smoothing and randomly initialize all parameters our Adaptive Prefix Generators. We train till 15 epochs for all model and select best performing checkpoint on average score of all validation sets. We use deepspeed[3] framework and train on two Tesla V-100 GPUs.

[3]https://www.deepspeed.ai/

Dataset	RD		RI		SE		CQA		CID	
Metric	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM
T5-zero (Raffel et al., 2020)	5.55	0.21	3.88	0.37	2.23	0.36	0.12	0.00	3.25	0.00
T0-3B (Sanh et al., 2021)	36.57	**8.55**	30.75	**7.77**	37.66	0.97	40.68	6.37	9.76	0.00
T5-unified (Raffel et al., 2020)	23.48	0.58	23.97	0.45	64.72	7.53	69.60	**28.99**	**17.98**	**8.44**
UniEvent (Ours)	**38.40**	3.27	**34.32**	2.22	**72.03**	**20.46**	**72.25**	28.54	17.36	4.45

Table 4: Results of QA tasks. Bold numbers are highest scores of the columns.

Dataset	TM	MA	RED	SCI	ESL	CTB	ALT	ECA
Metric	F1	F1	F1	F1	F1	F1	F1	ACC
T5-zero (Raffel et al., 2020)	13.80	38.94	38.78	49.89	31.40	3.49	67.90	51.27
T0-3B (Sanh et al., 2021)	25.27	**55.46**	39.61	49.87	**72.21**	4.39	**68.03**	**68.25**
T5-unified (Raffel et al., 2020)	28.93	35.57	**44.87**	51.87	31.91	0.00	56.95	48.97
UniEvent (Ours)	**30.66**	35.29	42.11	**82.78**	70.64	**8.95**	62.50	54.03

Table 5: Results of NLU tasks. Bold numbers are highest scores of the columns.

3.4 Baselines

- **T0-3B**(Sanh et al., 2021) This is the strongest baseline which is trained on a massive corpus of hundreds of general datasets. And more, this model is 10× bigger than our model.
- **T5-zero** (Raffel et al., 2020). We directly test on T5 without any training.
- **T5-unified** This is the baseline that only conducts multi-task training on T5-base without multi-dimensional prefix-tuning.
- **UniEvent-r** This is the ablated model of UNIEVENT without relation-wise prefixes.
- **UniEvent-t** This is the ablated model of UNIEVENT without formulation-wise prefixes.
- **UniEvent-c** This is the ablated model of UNIEVENT without formulation- and relation-aware contrastive regularization.

3.5 Zero-Shot Results

We list models' average performances on all zero-shot test datasets in Table 3. We find UNIEVENT outperforms strong baseline T0-3B on average 6.52 scores of all tasks in column **ALL**. This demonstrate the effectiveness of transferability on zero-shot event-relational reasoning tasks. The multi-dimensional prefixes with task- and relation-aware contrastive loss further boost the model to transfer across tasks. We also find T5-unified achieves comparable performance with T0-3B which is 10 × larger than it. All above findings testify our motivations that transfer knowledge via task formulation and relation axis is promising. Moreover, our multi-dimensional prefix-tuning ensures the knowledge transfer.

We list average score of `QA` tasks of all models in columns **QA-F1** (i.e. average of f1-scores.), **QA-Em** (i.e. average of exactly match scores.) and **QA** of Table 3 and show score of each dataset in Table 4. In Table 3, we find UNIEVENT outperforms T0-3B with average 9.43 scores on **QA** which is average scores of all both F1 and EM. This reveals UNIEVENT works encouragingly on `QA` reasoning. We show average score of NLU tasks in column **NLU** of Table 3 and results of each dataset in Table 5. We find UNIEVENT exceeds 0.48 scores on average which indicates the effectiveness of UNIEVENT on NLU part of datasets. As we can find, UNIEVENT performs not that well on NLU as on QA, we believe this is probably due to the pretrained generation backbone $\mathcal{M}$ is more suitable for generation tasks and T0-3B are trained on massive NLU datasets.

We also conduct experiments to evaluate cross-formulation and cross-relation transfer. Results are listed in **NT** column in Table 3 which are average scores of all datasets without training data in the same coordination in Figure 1. We surprisingly find that UNIEVENT exceeds T0-3B on a large margin, i.e. 8.12 average scores. These results indicate promising transferability of UNIEVENT since those tested dataset can only be completed by transfering from other datasets.

We report performances on `TEMP` datasets (MA, RED, TM, temporal part of CID) and `CA` datasets(ESL, SCI, CTB, ALT, ECA, RD, RI, causality parts of CQA and CID) of all models as well. Results are illustrated in **TP** and **CA** columns

Dataset	TBD	CNC	EST			AVG
Metric	F1	F1	F1	HIT@1	EM	
T5-base	41.94	73.79	58.23	**79.73**	21.93	55.12
T5-unified	29.90	78.05	60.76	78.41	**23.59**	54.14
UniEvent	**42.97**	**78.69**	**61.44**	79.07	21.59	**56.75**

Table 6: Performances on training set of all models.

	AVG				
	NLU	QA-F1	QA-EM	QA	ALL
T5-zero	38.32	4.38	0.25	4.38	18.05
T0-3B	**49.93**	30.49	3.86	20.60	33.33
T5-unified	44.19	30.96	6.38	22.74	31.78
UniEvent	38.50	**41.20**	**12.51**	**27.72**	**33.65**

Table 7: Trainset substituted by TM, ALT and CQA.

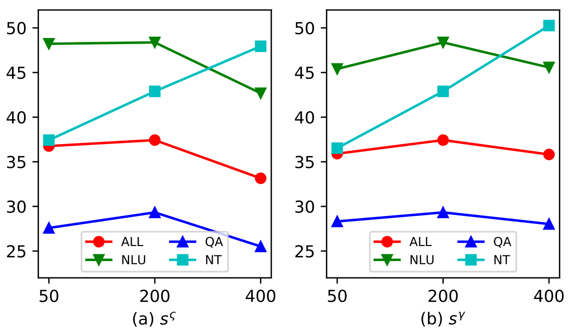

Figure 3: Prefix length analysis. (a) Formulation-wise prefix length s^ς under $s^\gamma = 200$, (b) Relation-wise prefix length s^γ under $s^\varsigma = 200$

in Table 3. Firstly, we find UNIEVENT performs well on CA datasets. However, we find formulation-wise prefixes harms performances of TEMP tasks which is probably due to most of TEMP datasets are RE.

3.6 Multi-Task Training Results

We also report multi-task training results on three trainsets. We find scores of trainsets can still increase if we continue training after the 10^{th} epoch while zero-shot performance would drop. Therefore, for fair comparison, we report best results within 10^{th} epochs for all models. As shown in the Table 6, UNIEVENT exceeds T5-unified. **T5-base** is a model finetuned on T5 base model in single task. Results demonstrate that our unified model can even transfer knowledge in full data setting. We believe our multi-dimensional prefix-tuning can reduce notorious negative transfer to some degree.

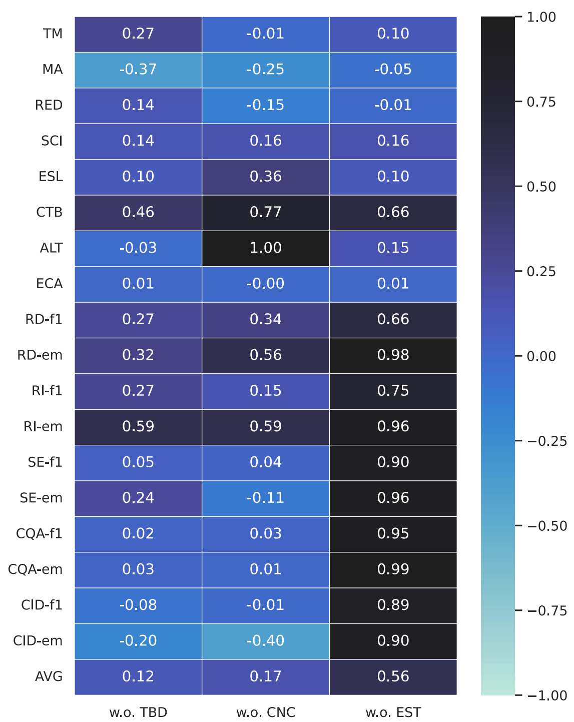

Figure 4: Dataset ablation study. Each score is computed by $\frac{x-\hat{x}}{x}$, where x is the score of UNIEVENT, $\hat{x}$ is the score with a dataset ablated.

3.7 Ablation Study

Model Ablation. We conduct model ablation studies. The results are detailed in Table 3. We find both formulation-wise and relation-wise prefixes effect. UniEvent outperforms UniEvent-c, which indicates task- and relation-aware contrastive regularization is crucial since it discriminates all sorts of dimensions in the unified training.

Dataset Ablation. In order to inspect the transferability and quantify the amount, we conduct dataset ablation studies. We complete three experiments, each with one of three training set ablated. Then we compute the transfer ratio of each trainset on all metrics as $\frac{x-\hat{x}}{x}$, where x is the score of UniEvent, $\hat{x}$ is the score with a dataset ablated. We detail the results in Figure 4. Basically, these experimental results are consistent with our motivation. EST contributes to all QA datasets. Causal part of EST transfer to CTB. CNC transfers causality knowledge to SCI, ESL, CTB, ALT and QA datasets as RD and RI. TBD can transfer to most of the RE dataset except MA. We believe MA suffers from negative transfer of all training sets. We surprisingly find TBD contributes to RD and RI. In sum, all training sets can transfer to other datasets on average (AVG row of Figure 4).

3.8 Prefix Length

In this part, we study influence of prefix length. In `UniEvent`, there are two types of prefix, i.e. P_{k^ς} and P_{k^γ}. We illustrate the results in Figure 3. Specifically, in Figure 3(a), we fix length of P_{k^γ} to 200, and vary length of P_{k^ς} (i.e. s^ς) from 50 to 400. We find almost all average metrics increase with s^ς varying from 50 to 400 except from temporal relation average performance. The results show that formulation-wise prefix length should reach to a scale to guarantee zero-shot performance.

On the other hand, we also analysis the length of P_{k^γ} under fixed $s^\varsigma = 200$. Results are depicted in Figure 3(b). Results are similar with s^ς, s^γ should reach a critical scale to make P_{k^γ} work.

We also find a interesting phenomena that **NT** metrics are still increasing in both experiments which indicates prefix length should be large for both formulation and relation unseen tasks.

3.9 Dataset Substitution

We substitute training set with TM, ALT and CQA. Results are shown in Table 7. We find UNIEVENT outperforms all baselines with dataset substituted. It indicates that UNIEVENT can transfer knowledge in various datasets permutations.

4 Related Work

Unified Training To fulfill knowledge transfer, sorts of brute-force solutions known as multi-task learning trains parameter-sharing neural models (Raffel et al., 2020; Sanh et al., 2021; Xu et al., 2022; Wei et al., 2021; Li et al., 2022a). However, learning out-of-domain and -formulation data could diminish the model efficacy on the targeted tasks, not to mention domain/formulation varying significantly in event-relational reasoning. Built upon a multi-task learning framework recent works are dedicated to integrating knowledge by unifying massive tasks (Lourie et al., 2021; Zhong et al., 2022; Xie et al., 2022; Lu et al., 2022; Khashabi et al., 2020). Via unified task formulations (e.g., text-to-text generation) and advanced training strategies, these works excel single task finetuning in conventional multi-task learning.

Prompting Transfer Yang et al. (2022a); Liu et al. (2022); Gu et al. (2022); Asai et al. (2022); Vu et al. (2021) transfer knowledge from pretrained tasks to downstream ones via prompting. In this work, we don't acquire prior knowledge from other tasks while enhance generalization across tasks.

Event-Relational Reasoning Zuo et al. (2020); Liu et al. (2021a); Zuo et al. (2021a); Cao et al. (2021); Zuo et al. (2021b); Chen et al. (2022); Phu and Nguyen (2021); Man et al. (2022b) identify event causality between two event trigger mentions. Zuo et al. (2020); Liu et al. (2021a); Zuo et al. (2021a) utilize external knowledge. Chen et al. (2022); Phu and Nguyen (2021) develop novel graph neural networks to capture structural information. Tan et al. (2022b); Liang et al. (2022) obtain event causality via natural language inference formulation.

Mathur et al. (2021); Zhou et al. (2020, 2021); Han et al. (2021b); Zhang et al. (2021); Hwang et al. (2022); Man et al. (2022a) extract temporal relations of events from documents or sentences. Zhou et al. (2020, 2021); Han et al. (2021b) learn from unsupervised or distant supervision.

Yang et al. (2020) asks for counterfactual statements. Du et al. (2022) aims to choose correct cause or effect from choices. Poria et al. (2021); Han et al. (2021a); Yang et al. (2022c) question about diversified event relations. Among all methods, we are the first to study the unification across these relations and formulations.

5 Conclusion

In this work, we propose UNIEVENT to transfer knowledge for unseen event-relational reasoning tasks. We first categorize these tasks. Then we construct generative formats and then unify them with generated multi-dimensional prefixes. UNIEVENT outperforms all baselines in both zero-shot and full-data settings.

6 Acknowledgement

Our work is supported by the National Key Research and Development Program of China (Project Number: 2020AAA0109400). we kindly appreciate all the researchers who provide valuable insights, discussions, and comments on this work.

Limilations

The current UniEvent is limited to performing event-relational reasoning tasks in a textual modality. It is unable to transfer knowledge between tasks of different modalities. However, combining event knowledge from different modalities may have more interactions and further enhance performance. As this is beyond the scope of our current work, we leave it to future research.

References

Akari Asai, Mohammadreza Salehi, Matthew E Peters, and Hannaneh Hajishirzi. 2022. Attentional mixtures of soft prompt tuning for parameter-efficient multi-task knowledge sharing. *arXiv preprint arXiv:2205.11961*.

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 (Volume 1: Long Papers)*, pages 4862–4872.

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

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284.

Meiqi Chen, Yixin Cao, Kunquan Deng, Mukai Li, Kun Wang, Jing Shao, and Yan Zhang. 2022. Ergo: Event relational graph transformer for document-level event causality identification. *arXiv preprint arXiv:2204.07434*.

Hal Daumé and Daniel Marcu. 2006. Bayesian query-focused summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, page 305–312, USA. Association for Computational Linguistics.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-care: a new dataset for exploring explainable causal reasoning. *arXiv preprint arXiv:2205.05849*.

Deepanway Ghosal, Pengfei Hong, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2021. Cider: Commonsense inference for dialogue explanation and reasoning. *arXiv preprint arXiv:2106.00510*.

Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. 2022. PPT: Pre-trained prompt tuning for few-shot learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8410–8423, Dublin, Ireland. Association for Computational Linguistics.

Rujun Han, I-Hung Hsu, Jiao Sun, Julia Baylon, Qiang Ning, Dan Roth, and Nanyun Peng. 2021a. Ester: A machine reading comprehension dataset for reasoning about event semantic relations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7543–7559.

Rujun Han, Qiang Ning, and Nanyun Peng. 2019. Joint event and temporal relation extraction with shared representations and structured prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 434–444, Hong Kong, China. Association for Computational Linguistics.

Rujun Han, Xiang Ren, and Nanyun Peng. 2021b. Econet: Effective continual pretraining of language models for event temporal reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5367–5380.

EunJeong Hwang, Jay-Yoon Lee, Tianyi Yang, Dhruvesh Patel, Dongxu Zhang, and Andrew McCallum. 2022. Event-event relation extraction using probabilistic box embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 235–244.

Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. Unifiedqa: Crossing format boundaries with a single qa system. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1896–1907.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Jia Li, Yuyuan Zhao, Zhi Jin, Ge Li, Tao Shen, Zhengwei Tao, and Chongyang Tao. 2022a. Sk2: Integrating implicit sentiment knowledge and explicit syntax knowledge for aspect-based sentiment analysis. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1114–1123.

Qintong Li, Piji Li, Wei Bi, Zhaochun Ren, Yuxuan Lai, and Lingpeng Kong. 2022b. Event transition planning for open-ended text generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3412–3426.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.

Zhaoning Li, Qi Li, Xiaotian Zou, and Jiangtao Ren. 2021. Causality extraction based on self-attentive bilstm-crf with transferred embeddings. *Neurocomputing*, 423:207–219.

Zhongyang Li, Xiao Ding, Ting Liu, J. Edward Hu, and Benjamin Van Durme. 2020. Guided generation of cause and effect. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3629–3636.

Shining Liang, Wanli Zuo, Zhenkun Shi, Sen Wang, Junhu Wang, and Xianglin Zuo. 2022. A multi-level neural network for implicit causality detection in web texts. *Neurocomputing*, 481:121–132.

Jian Liu, Yubo Chen, and Jun Zhao. 2021a. Knowledge enhanced event causality identification with mention masking generalizations. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3608–3614.

Shengchao Liu, Yingyu Liang, and Anthony Gitter. 2019. Loss-balanced task weighting to reduce negative transfer in multi-task learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9977–9978.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021b. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*.

Xiaochen Liu, Yu Bai, Jiawei Li, Yinan Hu, and Yang Gao. 2022. Psp: Pre-trained soft prompts for few-shot abstractive summarization. *arXiv preprint arXiv:2204.04413*.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13480–13488.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. *arXiv preprint arXiv:2203.12277*.

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

Hieu Man, Minh Van Nguyen, and Thien Huu Nguyen. 2022b. Event causality identification via generation of important context words. In *The 11th Joint Conference on Lexical and Computational Semantics*, page 323.

Puneet Mathur, Rajiv Jain, Franck Dernoncourt, Vlad Morariu, Quan Hung Tran, and Dinesh Manocha. 2021. Timers: document-level temporal relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 524–533.

Paramita Mirza and Sara Tonelli. 2016. Catena: Causal and temporal relation extraction from natural language texts. In *The 26th international conference on computational linguistics*, pages 64–75. ACL.

Aakanksha Naik, Luke Breitfeller, and Carolyn Rose. 2019. Tddiscourse: A dataset for discourse-level temporal ordering of events. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 239–249.

Qiang Ning, Hao Wu, and Dan Roth. 2018. A multi-axis annotation scheme for event temporal relations. *arXiv preprint arXiv:1804.07828*.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56.

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

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, et al. 2021. Recognizing emotion cause in conversations. *Cognitive Computation*, 13:1317–1332.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2Mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473, Melbourne, Australia. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Yixuan Su, Fangyu Liu, Zaiqiao Meng, Tian Lan, Lei Shu, Ehsan Shareghi, and Nigel Collier. 2021. Tacl: Improving bert pre-training with token-aware contrastive learning. *arXiv preprint arXiv:2111.04198*.

Fiona Anting Tan, Ali Hürriyetoğlu, Tommaso Caselli, Nelleke Oostdijk, Tadashi Nomoto, Hansi Hettiarachchi, Iqra Ameer, Onur Uca, Farhana Ferdousi Liza, and Tiancheng Hu. 2022a. The causal news corpus: Annotating causal relations in event sentences from news. *arXiv preprint arXiv:2204.11714*.

Fiona Anting Tan, Ali Hürriyetoğlu, Tommaso Caselli, Nelleke Oostdijk, Tadashi Nomoto, Hansi Hettiarachchi, Iqra Ameer, Onur Uca, Farhana Ferdousi Liza, and Tiancheng Hu. 2022b. The causal news corpus: Annotating causal relations in event sentences from news. *arXiv preprint arXiv:2204.11714*.

Jialong Tang, Hongyu Lin, Meng Liao, Yaojie Lu, Xianpei Han, Le Sun, Weijian Xie, and Jin Xu. 2021. From discourse to narrative: Knowledge projection for event relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 732–742.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Tu Vu, Brian Lester, Noah Constant, Rami Al-Rfou, and Daniel Cer. 2021. Spot: Better frozen model adaptation through soft prompt transfer. *arXiv preprint arXiv:2110.07904*.

Haoyu Wang, Muhao Chen, Hongming Zhang, and Dan Roth. 2020. Joint constrained learning for event-event relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 696–706.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Sen Wu, Hongyang R Zhang, and Christopher Ré. 2020. Understanding and improving information transfer in multi-task learning. *arXiv preprint arXiv:2005.00944*.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I. Wang, Victor Zhong, Bailin Wang, Chengzu Li, Connor Boyle, Ansong Ni, Ziyu Yao, Dragomir Radev, Caiming Xiong, Lingpeng Kong, Rui Zhang, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2022. Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. *arXiv preprint arXiv:2201.05966*.

Hanwei Xu, Yujun Chen, Yulun Du, Nan Shao, Yanggang Wang, Haiyu Li, and Zhilin Yang. 2022. Zeroprompt: Scaling prompt-based pretraining to 1,000 tasks improves zero-shot generalization. *arXiv preprint arXiv:2201.06910*.

Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Mingfeng Xue, Boxing Chen, and Jun Xie. 2022a. Tailor: A prompt-based approach to attribute-based controlled text generation. *arXiv preprint arXiv:2204.13362*.

Linyi Yang, Zhen Wang, Yuxiang Wu, Jie Yang, and Yue Zhang. 2022b. Towards fine-grained causal reasoning and qa. *arXiv preprint arXiv:2204.07408*.

Linyi Yang, Zhen Wang, Yuxiang Wu, Jie Yang, and Yue Zhang. 2022c. Towards fine-grained causal reasoning and qa. *arXiv preprint arXiv:2204.07408*.

Xiaoyu Yang, Stephen Obadinma, Huasha Zhao, Qiong Zhang, Stan Matwin, and Xiaodan Zhu. 2020. Semeval-2020 task 5: Counterfactual recognition. *arXiv preprint arXiv:2008.00563*.

Shuaicheng Zhang, Lifu Huang, and Qiang Ning. 2021. Extracting temporal event relation with syntactic-guided temporal graph transformer. *arXiv preprint arXiv:2104.09570*.

Wanjun Zhong, Yifan Gao, Ning Ding, Yujia Qin, Zhiyuan Liu, Ming Zhou, Jiahai Wang, Jian Yin, and Nan Duan. 2022. ProQA: Structural prompt-based pre-training for unified question answering. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4230–4243, Seattle, United States. Association for Computational Linguistics.

Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. 2020. Temporal common sense acquisition with minimal supervision. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7579–7589.

Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2021. Temporal reasoning on implicit events from distant supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1361–1371.

Yucheng Zhou, Tao Shen, Xiubo Geng, Guodong Long, and Daxin Jiang. 2022. Claret: Pre-training a correlation-aware context-to-event transformer for event-centric generation and classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2559–2575.

Xinyu Zuo, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Weihua Peng, and Yuguang Chen. 2021a. Improving event causality identification via self-supervised representation learning on external causal statement. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2162–2172.

Xinyu Zuo, Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Weihua Peng, and Yuguang Chen. 2021b. LearnDA: Learnable knowledge-guided data augmentation for event causality identification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3558–3571, Online. Association for Computational Linguistics.

Xinyu Zuo, Yubo Chen, Kang Liu, and Jun Zhao. 2020. KnowDis: Knowledge enhanced data augmentation for event causality detection via distant supervision. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1544–1550, Barcelona, Spain (Online). International Committee on Computational Linguistics.

A Dataset Details

Dataset	Train	Validation	Test
TBD	4,032	629	1,427
MA	5,412	920	827
TM	3,987	650	1500
RED	2,609	303	361
SCI	4,936	-	891
ESL	4,611	499	492
CTB	1,212	845	846
CNC	2,632	293	293
ALT	100,744	488	611
ECA	14,928	2,132	2,132
RD	7,271	347	1,894
RI	-	-	1,080
SE	3,551	-	1,950
EST	4,547	301	301
CQA	19,588	2,449	2,449
CID	1,938	237	225

Table 8: Dataset statistics. There are no validation set in SCI and SE. RI only have test set.

In this section, we state processing details of all datasets. We show dataset statistics in Table 8.

Considering temporal event relation extraction, we strictly follow settings in Han et al. (2021b) for MATRES, TBD, RED and setting in (Naik et al., 2019) for TM.

For event causality identification, in ESL, CTB, we don't perform 5-folds cross validation as in Zuo et al. (2021b) and instead split each dataset into 8:1:1 for train, validation and test. We follow Li et al. (2021) for SCI.

We follow CNC in Tan et al. (2022b) and ALT in Liang et al. (2022) respectively for causal NLI.

In view of question answering datasets, we follow Han et al. (2021a), Yang et al. (2022c), Ghosal et al. (2021) and Yang et al. (2020) for EST, CQA, CID and SE. RD and RI are the same with Poria et al. (2021).

Lastly, the setting for ECA is the same with Du et al. (2022).

There are no validation set for SCI, RI, SE, so when compute average score in validation, we don't consider these three datasets.

ACL 2023 Responsible NLP Checklist

A For every submission:

☐ A1. Did you describe the limitations of your work?
Left blank.

☐ A2. Did you discuss any potential risks of your work?
Left blank.

☐ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☐ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☐ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Left blank.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C ☐ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Left blank.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Left blank.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Left blank.

D ☐ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Left blank.