Improving Event Representation via Simultaneous Weakly Supervised Contrastive Learning and Clustering

Jun Gao[1] **Wei Wang**[3] **Changlong Yu**[4] **Huan Zhao**[5] **Wilfred Ng**[4] **Ruifeng Xu**[1,2*]
[1]Harbin Institute of Technology (Shenzhen) [2]Peng Cheng Laboratory [3]Tsinghua University
imgaojun@gmail.com xuruifeng@hit.edu.cn weiwangorg@163.com
[4]HKUST, Hong Kong, China [5]4Paradigm. Inc.
{cyuaq,wilfred}@cse.ust.hk zhaohuan@4paradigm.com

Abstract

Representations of events described in text are important for various tasks. In this work, we present **SWCC**: a **S**imultaneous **W**eakly supervised **C**ontrastive learning and **C**lustering framework for event representation learning. SWCC learns event representations by making better use of co-occurrence information of events. Specifically, we introduce a weakly supervised contrastive learning method that allows us to consider multiple positives and multiple negatives, and a prototype-based clustering method that avoids semantically related events being pulled apart. For model training, SWCC learns representations by simultaneously performing weakly supervised contrastive learning and prototype-based clustering. Experimental results show that SWCC outperforms other baselines on `Hard Similarity` and `Transitive Sentence Similarity` tasks. In addition, a thorough analysis of the prototype-based clustering method demonstrates that the learned prototype vectors are able to implicitly capture various relations between events. Our code will be available at `https://github.com/gaojun4ever/SWCC4Event`.

1 Introduction

Distributed representations of events, are a common way to represent events in a machine-readable form and have shown to provide meaningful features for various tasks (Lee and Goldwasser, 2018; Rezaee and Ferraro, 2021; Deng et al., 2021; Martin et al., 2018; Chen et al., 2021). Obtaining effective event representations is challenging, as it requires representations to capture various relations between events. Figure 1 presents four pairs of events with different relations. Two events may share the same event attributes (e.g. event types and sentiments), and there may also be a causal or temporal relation between two events.

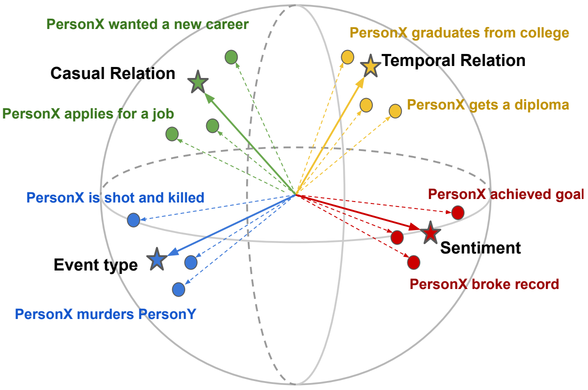


Figure 1: Four pairs of events with different relations. Stars represent prototypes and circles represent events.

Early works (Weber et al., 2018) exploit easily accessible co-occurrence relation of events to learn event representations. Although the use of co-occurrence relation works well, it is too coarse for deep understanding of events, which requires fine-grained knowledge (Lee and Goldwasser, 2019). Recent works focus on fine-grained knowledge, such as discourse relations (Lee and Goldwasser, 2019; Zheng et al., 2020) and commonsense knowledge (e.g. sentiments and intents) (Sap et al., 2019; Ding et al., 2019). Concretely, Lee and Goldwasser (2019) and Zheng et al. (2020) leverage 11 discourse relation types to model event script knowledge. Ding et al. (2019) incorporate manually labeled commonsense knowledge (intents and sentiments) into event representation learning. However, the types of fine-grained event knowledge are so diverse that we cannot enumerate all of them and currently adopted fine-grained knowledge fall under a small set of event knowledge. In addition, some manually labeled knowledge (Sap et al., 2019; Hwang et al., 2021) is costly and difficult to apply on large datasets.

In our work, we observe that there is a rich amount of information in co-occurring events, but previous works did not make good use of such information. Based on existing works on event relation

[*] Corresponding author
Jun Gao is currently a research intern at 4Paradigm.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 3036 - 3049
May 22-27, 2022 ©2022 Association for Computational Linguistics

extraction (Xue et al., 2016; Lee and Goldwasser, 2019; Zhang et al., 2020; Wang et al., 2020), we find that the co-occurrence relation, which refers to two events appearing in the same document, can be seen as a superset of currently defined explicit discourse relations. To be specific, these relations are often indicated by discourse markers (e.g., "because", capturing the casual relation) (Lee and Goldwasser, 2019). Therefore, two related events must exist in the same sentence or document. More than that, the co-occurrence relation also includes other implicit event knowledge. For example, events that occur in the same document may share the same topic and event type. To learn event representations, previous works (Granroth-Wilding and Clark, 2016; Weber et al., 2018) based on co-occurrence information usually exploit instance-wise contrastive learning approaches related to the margin loss, which consists of an anchor, positive, and negative sample, where the anchor is more similar to the positive than the negative. However, they share two common limitations: (1) such margin-based approaches struggle to capture the essential differences between events with different semantics, as they only consider one positive and one negative per anchor. (2) Randomly sampled negative samples may contain samples semantically related to the anchor, but are undesirably pushed apart in embedding space. This problem arises because these instance-wise contrastive learning approaches treat randomly selected events as negative samples, regardless of their semantic relevance.

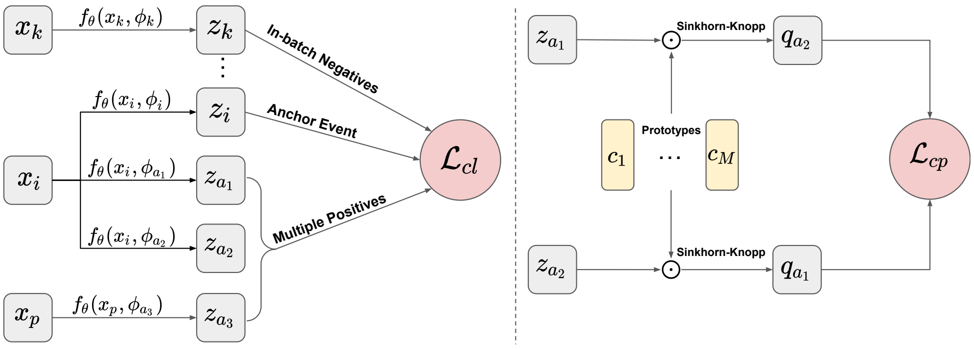
We are motivated to address the above issues with the goal of making better use of co-occurrence information of events. To this end, we present **SWCC**: a **S**imultaneous **W**eakly supervised **C**ontrastive learning and **C**lustering framework for event representation learning, where we exploit document-level co-occurrence information of events as weak supervision and learn event representations by simultaneously performing weakly supervised contrastive learning and prototype-based clustering. To address the first issue, we build our approach on the contrastive framework with the InfoNCE objective (van den Oord et al., 2019), which is a self-supervised contrastive learning method that uses one positive and multiple negatives. Further, we extend the InfoNCE to a weakly supervised contrastive learning setting, allowing us to consider multiple positives and multiple negatives per anchor (as opposed to the previous works which use only one positive and one negative). Co-occurring events are then incorporated as additional positives, weighted by a normalized co-occurrence frequency. To address the second issue, we introduce a prototype-based clustering method to avoid semantically related events being pulled apart. Specifically, we impose a prototype for each cluster, which is a representative embedding for a group of semantically related events. Then we cluster the data while enforce consistency between cluster assignments produced for different augmented representations of an event. Unlike the instance-wise contrastive learning, our clustering method focuses on the cluster-level semantic concepts by contrasting between representations of events and clusters. Overall, we make the following contributions:

- We propose a simple and effective framework (**SWCC**) that learns event representations by making better use of co-occurrence information of events. Experimental results show that our approach outperforms previous approaches on several event related tasks.
- We introduce a weakly supervised contrastive learning method that allows us to consider multiple positives and multiple negatives, and a prototype-based clustering method that avoids semantically related events being pulled apart.
- We provide a thorough analysis of the prototype-based clustering method to demonstrate that the learned prototype vectors are able to implicitly capture various relations between events.

2 Preliminaries

Event representation model. In the early works (Weber et al., 2018; Ding et al., 2019), Neural Tensor Networks (NTNs) (Socher et al., 2013b,a) are widely adopted to compose the representation of event constitutions, i.e., `(subject, predicate, object)`. However, such methods introduced strong compositional inductive bias and can not extend to events with more additional arguments, such as time, location etc. Several recent works (Zheng et al., 2020; Vijayaraghavan and Roy, 2021) replaced static word vector compositions with powerful pretrained language models, such as BERT (Devlin et al., 2019), for flexible event representations and achieved better performance. Following them, we also take the BERT as the backbone model.

The BERT encoder can take as input a free-form



(a) Weakly Supervised Contrastive Learning

(b) Prototype-based Clustering

Figure 2: Architecture of the proposed framework, where the left part is the `Weakly Supervised Contrastive Learning` method and the right part is the `Prototype-based Clustering` method. Given an input event $\boldsymbol{x}_i$, we obtain three augmented representations $\boldsymbol{z}_i, \boldsymbol{z}_{a_1}$ and $\boldsymbol{z}_{a_2}$ of the same event $\boldsymbol{x}_i$ using the BERT model with different dropout masks. Using the same approach, we obtain the representation set $\{\boldsymbol{z}_k\}_{k \in \mathcal{N}(i)}$ of in-batch negatives and the representation $\boldsymbol{z}_{a_3}$ of its co-occurrence event.

event text, which contains a sequence of tokens and the input format can be represented as follows:

$$[\text{CLS}], pred, subj, obj, [\text{SEP}]. \tag{1}$$

Define $\boldsymbol{x} = [x_0, x_1, \cdots, x_L]$ to be the input sequence of length L, where x_0 and x_L are the [CLS] token and the [SEP] token respectively. Given $\boldsymbol{x}$, the BERT returns a sequence of contextualized vectors:

$$[\boldsymbol{v}_{[\text{CLS}]}, \boldsymbol{v}_{x_1}, \cdots, \boldsymbol{v}_{x_L}] = \text{BERT}(\boldsymbol{x}), \tag{2}$$

where $\boldsymbol{v}_{[\text{CLS}]}$ is the representation for the [CLS] token. In the default case, the final vector representation $\boldsymbol{z}$ of the event is the output representation of the [CLS] token: $\boldsymbol{z} = \boldsymbol{v}_{[\text{CLS}]}$.

Instance-wise contrastive learning. Event representation models learn representations with contrastive learning, which aims to pull related events together and push apart unrelated events. Margin loss (Schroff et al., 2015) is a widely used contrastive loss in most of the existing works on event representation learning (Weber et al., 2018; Ding et al., 2019; Zheng et al., 2020). Most recently, an alternative contrastive loss function, called InfoNCE (van den Oord et al., 2019), has been proposed and shown effective in various contrastive learning tasks (He et al., 2020; Hu et al., 2021; Gao et al., 2021). Chen et al. (2020a) further demonstrate that InfoNCE works better than the Margin loss. In this work, we explore the use of InfoNCE to train our event representation model.

Formally, given a set of N paired events $\mathcal{D} = \{\boldsymbol{x}_i, \boldsymbol{x}_i^+\}_{i=1}^N$, where $\boldsymbol{x}_i^+$ is a positive sample for $\boldsymbol{x}_i$, the InfoNCE objective for $(\boldsymbol{x}_i, \boldsymbol{x}_i^+)$ is presented in a softmax form with in-batch negatives (Chen et al., 2020a; Gao et al., 2021):

$$\mathcal{L} = -\log \frac{g(\boldsymbol{z}_i, \boldsymbol{z}_i^+)}{g(\boldsymbol{z}_i, \boldsymbol{z}_i^+) + \sum_{k \in \mathcal{N}(i)} g(\boldsymbol{z}_i, \boldsymbol{z}_k)}, \tag{3}$$

where $\boldsymbol{z}_i$ and $\boldsymbol{z}_i^+$ are the augmented representations of $\boldsymbol{x}_i$ and $\boldsymbol{x}_i^+$ obtained through a representation model , $k \in \mathcal{N}(i)$ is the index of in-batch negatives. and g is a function: $g(\boldsymbol{z}_i, \boldsymbol{z}_k) = \exp(\boldsymbol{z}_i^\top \boldsymbol{z}_k / \tau)$, where $\tau \in \mathbb{R}^+$ is a positive value of temperature.

Data augmentation. One critical question in contrastive learning is how to obtain $\boldsymbol{z}_i^+$. In language representation, $\boldsymbol{z}_i^+$ are often obtained by first applying data augmentation in the form of word deletion, reordering, or substitution on $\boldsymbol{x}_i$ and then feeding it into the event representation model. Several recent works (Gao et al., 2021; Liang et al., 2021) exploit dropout noise as data augmentation for NLP tasks and find that this data augmentation technique performs much better than common data augmentation techniques. Specifically, given an input event $\boldsymbol{x}_i$, we obtain $\boldsymbol{z}_i$ and $\boldsymbol{z}_i^+$ by feeding the same input to the BERT encoder with the parametric weights θ twice, and each time we apply a different dropout mask:

$$\boldsymbol{z}_i = f_\theta(\boldsymbol{x}_i, \boldsymbol{\phi}_1), \boldsymbol{z}_i^+ = f_\theta(\boldsymbol{x}_i, \boldsymbol{\phi}_2), \tag{4}$$

where $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ are two different random masks for dropout. As described in Sec.3.1, given an anchor event $\boldsymbol{z}_i$, we generate 3 positive samples $\boldsymbol{z}_{a_1}$, $\boldsymbol{z}_{a_2}$ and $\boldsymbol{z}_{a_3}$ with different dropout masks.

3 The Proposed Approach

In this section, we will present technical details of our proposed approach and our goal is to learn event representations by making better use of co-occurrence information of events. Figure 2 presents an overview of our proposed approach, which contains two parts: the weakly-supervised contrastive learning method (left) and the prototype-based clustering method (right). In the following sections, we will introduce both methods separately.

3.1 Weakly Supervised Contrastive Learning

We build our approach on the contrastive framework with the InfoNCE objective (Eq.3) instead of the margin loss. To incorporate co-occurrence information into event representation learning, a straightforward way is to consider the co-occurring event of each input event as an additional positive sample, that is, the positive augmented representations of $\boldsymbol{x}_i$ come not only from itself but also from its co-occurring event denoted as $\boldsymbol{x}_p$. However, The original InfoNCE objective cannot handle the case where there exists multiple positive samples. Inspired by Khosla et al. (2020), we take a similar formulation to tackle this problem. More than that, we also introduce a weighting mechanism to consider co-occurrence frequency of two events, which indicates the strength of the connection between two events.

Co-occurrence as weak supervision. Formally, for each input pair $(\boldsymbol{x}_i, \boldsymbol{x}_p)$, where $\boldsymbol{x}_i$ and $\boldsymbol{x}_p$ refer to the input event and one of its co-occurring events, we first compute an augmented representation $\boldsymbol{z}_i$ of $\boldsymbol{x}_i$ as an anchor event, through the event representation model mentioned in § 2. How the method differs from InfoNCE is in the construction of the positive set $\mathcal{A}(i)$ for $\boldsymbol{x}_i$. In InfoNCE, $\mathcal{A}(i)$ only contains one positive. In our method, we generalize Eq. 3 to support multiple positives learning:

$$\mathcal{L} = \sum_{a \in \mathcal{A}(i)} -\log \frac{g(\boldsymbol{z}_i, \boldsymbol{z}_a)}{g(\boldsymbol{z}_i, \boldsymbol{z}_a) + \sum_{k \in \mathcal{N}(i)} g(\boldsymbol{z}_i, \boldsymbol{z}_k)}, \tag{5}$$

where $\mathcal{A}(i)$ and $\mathcal{N}(i)$ refer to the positive set and the negative set for the event $\boldsymbol{x}_i$. Note that we support arbitrary number of positives here. In our work, considering the limited GPU memory, we use $\mathcal{A}(i) = \{\boldsymbol{z}_{a_1}, \boldsymbol{z}_{a_2}, \boldsymbol{z}_{a_3}\}$, where $\boldsymbol{z}_{a_1}$ and $\boldsymbol{z}_{a_2}$ are two augmented representations of the same event $\boldsymbol{x}_i$, obtained with different dropout masks, and $\boldsymbol{z}_{a_3}$ is an augmented representation of its co-occurring event. Here $\boldsymbol{z}_{a_1}$ and $\boldsymbol{z}_{a_2}$ will then be used in the prototype-based clustering method (See Fig. 2 for example) as detailed later (§ 3.2).

Incorporating co-occurrence frequency. The co-occurrence frequency indicates the strength of the connection between two events. To make better use of data, we introduce a weighting mechanism to exploit the co-occurrence frequency between events as instance weights and rewrite the Eq. 5:

$$\mathcal{L}_{cl} = \sum_{a \in \mathcal{A}(i)} -\log \frac{\varepsilon_a \cdot g(\boldsymbol{z}_i, \boldsymbol{z}_a)}{g(\boldsymbol{z}_i, \boldsymbol{z}_a) + \sum_{k \in \mathcal{N}(i)} g(\boldsymbol{z}_i, \boldsymbol{z}_k)}. \tag{6}$$

Here ε_a is a weight for the positive sample $\boldsymbol{z}_a$. In our work, the two weights ε_{a_1} and ε_{a_2} of the positive samples ($\boldsymbol{z}_{a_1}$ and $\boldsymbol{z}_{a_2}$) obtained from the input event, are set as $\varepsilon_{a_1} = \varepsilon_{a_2} = \frac{1}{|\mathcal{A}(i)|-1}$, where $|\mathcal{A}(i)|$ is its cardinality. To obtain the weight ε_{a_3} for the augmented representation $\boldsymbol{z}_{a_3}$ of the co-occurring event, we create a co–occurrence matrix, $\boldsymbol{V}$ with each entry corresponding to the co-occurrence frequency of two distinct events. Then $\boldsymbol{V}$ is normalized to $\hat{\boldsymbol{V}}$ with the `Min-Max` normalization method, and we take the entry in $\hat{\boldsymbol{V}}$ as the weight ε_{a_3} for the co-occurrence event. In this way, the model draws the input events closer to the events with higher co-occurrence frequency, as each entry in $\hat{\boldsymbol{V}}$ indicates the strength of the connection between two events.

3.2 Prototype-based Clustering

To avoid semantically related events being pulled apart, we draw inspiration from the recent approach (Caron et al., 2020) in the computer vision domain and introduce a prototype-based clustering method, where we impose a prototype, which is a representative embedding for a group of semantically related events for each cluster. Then we cluster the data while enforce consistency between cluster assignments produced for different augmented representations of an event. These prototypes essentially serve as the center of data representation clusters for a group of semantically related events (See Figure 1 for example). Unlike the instance-wise contrastive learning, our clustering method focuses on the cluster-level semantic concepts by contrasting between representations of events and clusters.

Cluster prediction. This method works by comparing two different augmented representations of the same event using their intermediate cluster assignments. The motivation is that if these two representations capture the same information, it should be possible to predict the cluster assignment of one augmented representation from another augmented representation. In detail, we consider a set of M prototypes, each associated with a learnable vector $\boldsymbol{c}_i$, where $i \in [\![M]\!]$. Given an input event, we first transform the event into two augmented representations with two different dropout masks. Here we use the two augmented representations $\boldsymbol{z}_{a_1}$ and $\boldsymbol{z}_{a_2}$ of the event $\boldsymbol{x}_i$. We compute their cluster assignments $\boldsymbol{q}_{a_1}$ and $\boldsymbol{q}_{a_2}$ by matching the two augmented representations to the set of M prototypes. The cluster assignments are then swapped between the two augmented representations: the cluster assignment $\boldsymbol{q}_{a_1}$ of the augmented representation $\boldsymbol{z}_{a_1}$ should be predicted from the augmented representation $\boldsymbol{z}_{a_2}$, and vice-versa. Formally, the cluster prediction loss is defined as:

$$\mathcal{L}_{cp} = \ell(\boldsymbol{z}_{a_1}, \boldsymbol{q}_{a_2}) + \ell(\boldsymbol{z}_{a_2}, \boldsymbol{q}_{a_1}), \qquad (7)$$

where function $\ell(\boldsymbol{z}, \boldsymbol{q})$ measures the fit between the representation $\boldsymbol{z}$ and the cluster assignment $\boldsymbol{q}$, as defined by: $\ell(\boldsymbol{z}, \boldsymbol{q}) = -\boldsymbol{q} \log \boldsymbol{p}$. Here $\boldsymbol{p}$ is a probability vector over the M prototypes whose components are:

$$p^{(j)} = \frac{\exp(\boldsymbol{z}^\top \boldsymbol{c}_j / \tau)}{\sum_{k=1}^{M} \exp(\exp(\boldsymbol{z}^\top \boldsymbol{c}_k / \tau)}, \qquad (8)$$

where τ is a temperature hyperparameter. Intuitively, this cluster prediction method links representations $\boldsymbol{z}_{a_1}$ and $\boldsymbol{z}_{a_2}$ using the intermediate cluster assignments $\boldsymbol{q}_{a_1}$ and $\boldsymbol{q}_{a_2}$.

Computing cluster assignments. We compute the cluster assignments using an Optimal Transport solver. This solver ensures equal partitioning of the prototypes or clusters across all augmented representations, avoiding trivial solutions where all representations are mapped to a unique prototype. In particular, we employ the Sinkhorn-Knopp algorithm (Cuturi, 2013). The algorithm first begins with a matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{M \times N}$ with each element initialized to $\boldsymbol{z}_b^\top \boldsymbol{c}_m$, where $b \in [\![N]\!]$ is the index of each column. It then iteratively produces a doubly-normalized matrix, the columns of which comprise $\boldsymbol{q}$ for the minibatch.

3.3 Model Training

Our approach learns event representations by simultaneously performing weakly supervised contrastive learning and prototype-based clustering. The overall training objective has three terms:

$$\mathcal{L}_{overall} = \mathcal{L}_{cl} + \beta \mathcal{L}_{cp} + \gamma \mathcal{L}_{mlm}, \qquad (9)$$

where β and γ are hyperparameters. The first term is the weakly supervised contrastive learning loss that allows us to effectively incorporate co-occurrence information into event representation learning. The second term is the prototype-based clustering loss, whose goal is to cluster the events while enforcing consistency between cluster assignments produced for different augmented representations of the input event. Lastly, we introduce the masked language modeling (MLM) objective (Devlin et al., 2019) as an auxiliary loss to avoid forgetting of token-level knowledge.

4 Experiments

Following common practice in event representation learning (Weber et al., 2018; Ding et al., 2019; Zheng et al., 2020), we analyze the event representations learned by our approach on two event similarity tasks (§ 4.2) and one transfer task (§ 4.4).

4.1 Dataset and Implementation Details

The event triples we use for the training data are extracted from the `New York Times Gigaword Corpus` using the Open Information Extraction system Ollie (Mausam et al., 2012). We filtered the events with frequencies less than 3 and ended up with 4,029,877 distinct events. We use the MCNC dataset adopted in Lee and Goldwasser (2019)[1] for the transfer task.

Our event representation model is implemented using the Texar-PyTorch package (Hu et al., 2019). The model starts from the pre-trained checkpoint of `BERT-based-uncased` (Devlin et al., 2019) and we use the [CLS] token representation as the event representation. We train our model with a batch size of 256 using an Adam optimizer. The learning rate is set as 2e-7 for the event representation model and 2e-5 for the prototype memory. We adopt the temperature $\tau = 0.3$ and the numbers of prototypes used in our experiment is 10.

4.2 Event Similarity Tasks

Similarity task is a common way to measure the quality of vector representations. Weber et al. (2018) introduce two event related similarity tasks: (1) `Hard Similarity Task` and (2) `Transitive Sentence Similarity`.

Hard Similarity Task. The hard similarity task tests whether the event representation model can

[1] `https://github.com/doug919/multi_relational_script_learning`

Model	Hard similarity (Accuracy %) Original	Hard similarity (Accuracy %) Extended	Transitive sentence similarity (ρ)
Event-comp (Weber et al., 2018)*	33.9	18.7	0.57
Predicate Tensor (Weber et al., 2018)*	41.0	25.6	0.63
Role-factor Tensor (Weber et al., 2018)*	43.5	20.7	0.64
KGEB (Ding et al., 2016)*	52.6	49.8	0.61
NTN-IntSent (Ding et al., 2019)*	77.4	62.8	0.74
SAM-Net (Lv et al., 2019)*	51.3	45.2	0.59
FEEL (Lee and Goldwasser, 2018)*	58.7	50.7	0.67
UniFA-S (Zheng et al., 2020)*	78.3	64.1	0.75
SWCC	**80.9**	**72.1**	**0.82**

Table 1: Evaluation performance on the similarity tasks. Best results are bold. *: results reported in the original papers.

push away representations of dissimilar events while pulling together those of similar events. Weber et al. (2018) created a dataset (denoted as "Original"), where each sample has two types of event pairs: one with events that should be close to each other but have very little lexical overlap, and another with events that should be farther apart but have high overlap. This dataset contains 230 event pairs. After that, Ding et al. (2019) extended this dataset to 1,000 event pairs (denoted as "Extended"). For this task, we use Accuracy as the evaluation metric, which measures the percentage of cases where the similar pair receives a higher cosine value than the dissimilar pair.

Transitive Sentence Similarity. The transitive sentence similarity dataset (Kartsaklis and Sadrzadeh, 2014) contains 108 pairs of transitive sentences that contain a single subject, object, and verb (e.g., `agent sell property`) and each pair in this dataset is manually annotated by a similarity score from 1 to 7. A larger score indicates that the two events are more similar. Following previous work (Weber et al., 2018; Ding et al., 2019; Zheng et al., 2020), we evaluate using the Spearman's correlation of the cosine similarity predicted by each method and the annotated similarity score.

4.3 Comparison methods.

We compare our proposed approach with a variety of baselines. These methods can be categorized into three types:

(1) **Co-occurrence**: **Event-comp** (Weber et al., 2018), **Role-factor Tensor** (Weber et al., 2018) and **Predicate Tensor** (Weber et al., 2018) are models that use tensors to learn the interactions between the predicate and its arguments and are trained using co-occurring events as supervision.

(2) **Discourse Relations**: This line of work exploits discourse relations. **SAM-Net** (Lv et al., 2019) explores event segment relations, **FEEL** (Lee and Goldwasser, 2018) and **UniFA-S** (Zheng et al., 2020) adopt discourse relations.

(3) **Commonsense Knowledge**: Several works have shown the effectiveness of using commonsense knowledge. **KGEB** (Ding et al., 2016) incorporates knowledge graph information. **NTN-IntSent** (Ding et al., 2019) leverages external commonsense knowledge about the intent and sentiment of the event.

Results. Table 1 reports the performance of different methods on the hard similarity tasks and the transitive sentence similarity task. The result shows that the proposed SWCC achieves the best performance among the compared methods. It not only outperforms the Role-factor Tensor method that based on co-occurrence information, but also has better performance than the methods trained with additional annotations and commonsense knowledge, e.g. NTN-IntSent and UniFA-S. This implies the co-occurrence information of events is effective but underutilized by previous works, and the proposed SWCC makes better use of the co-occurrence information.

Ablation study. To investigate the effect of each component in our approach, we conduct an ablation study as reported in Table 2. We remove a certain component of SWCC and examine the corresponding performance of the incomplete SWCC on the similarity tasks. We first explore the impact of our prototype-based clustering method by removing the loss term $\mathcal{L}_{cp}$ in Eq. 9. We find that this component has a significant impact on the transitive sentence similarity task. Removing this component causes a 0.05 (maximum) point drop in performance on the transitive sentence similarity task. And for the weakly supervised contrastive learning method, we find that it has a strong impact on both hard simi-

Model	Hard similarity (Accuracy %)		Transitive sentence similarity (ρ)
	Original	Extended	
SWCC	**80.9**	**72.1**	**0.82**
w/o Prototype-based Clustering	77.4 (-3.5)	67.4 (-4.7)	0.77 (-0.05)
w/o Weakly Supervised CL	75.7 (-5.2)	65.1 (-7.0)	0.78 (-0.04)
w/o MLM	77.4 (-3.5)	70.4 (-1.7)	0.80 (-0.02)
BERT (InfoNCE)	72.1	63.4	0.75
BERT (Margin)	43.5	51.4	0.67

Table 2: Ablation study for several methods evaluated on the similarity tasks.

larity tasks, especially the extended hard similarity task. Removing this component causes an 7.0 point drop in performance of the model. We also study the impact of the MLM auxiliary objective. As shown in Table 2 the token-level MLM objective improves the performance on the extended hard similarity task modestly, it does not help much for the transitive sentence similarity task.

Next, we compare the InfoNCE against the margin loss in Table 2. For a fair comparison, the BERT (InfoNCE) is trained using the InfoNCE objective only, with co-occurring events as positives and other samples in the minibatch as negatives, and the BERT (Margin) is trained using the margin loss, with co-occurring events as positives and randomly sampled events as negatives. Obviously, BERT (InfoNCE) achieves much competitive results on all tasks, suggesting that the InfoNCE with adjustable temperature works better than the margin loss. This can be explained by the fact that the InfoNCE weighs multiple different negatives, and an appropriate temperature can help the model learn from hard negatives, while the margin loss uses only one negative and can not weigh the negatives by their relative hardness.

4.4 Transfer Task

We test the generalization of the event representations by transferring to a downstream event related tasks, the Multiple Choice Narrative Cloze (MCNC) task (Granroth-Wilding and Clark, 2016), which was proposed to evaluate script knowledge. In particular, given an event chain which is a series of events, this task requires a reasoning system to distinguish the next event from a small set of randomly drawn events. We evaluate our methods with several methods based on unsupervised learning: (1) **Random** picks a candidate at random uniformly; (2) **PPMI** (Chambers and Jurafsky, 2008) uses co-occurrence information and calculates Positive PMI for event pairs; (3) **BiGram** (Jans et al., 2012) calculates bi-gram conditional probabilities based on event term frequencies; (4) **Word2Vec** (Mikolov et al., 2013) uses the word embeddings trained by Skipgram algorithm and event representations are the summation of word embeddings of predicates and arguments. Note that we did not compare with supervised methods (Bai et al., 2021; Zhou et al., 2021; Lv et al., 2020) since unsupervised ones are more suitable for purely evaluating event representations.

Results. Table 3 reports the performance of different methods on the MCNC task. As shown in the table, SWCC achieves the best accuracy on the MCNC task under the zero-shot transfer setting, suggesting the proposed SWCC has better generalizability to the downstream tasks than other compared methods.

Model	Accuracy (%)
Random	20.00
PPMI*	30.52
BiGram*	29.67
Word2Vec*	37.39
BERT (Margin)	36.50
BERT (InfoNCE)	39.23
SWCC	**44.50**

Table 3: Evaluation performance on the MCNC task. Best results are bold. *: results reported in the previous work (Lee and Goldwasser, 2019).

5 Analysis and Visualization

In this section, we further analyze the prototype-based clustering method.

Number of prototypes. Figure 3 displays the impact of the number of prototypes in training. As shown in the figure, the performance increases as the number M increases, but it will not further increase after 10. We speculate that because these evaluation data are too small and contain too few types of relations, a larger number of prototypes would not help much in performance improvement.

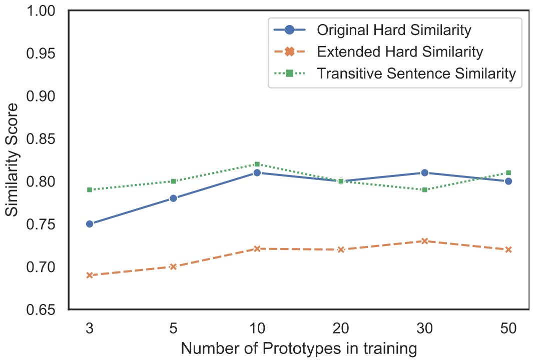


Figure 3: Impact of # of Prototypes

Visualization of learned representation. We randomly sample 3000 events and embed the event representations learned by BERT (InfoNCE) and SWCC in 2D using the PCA method. The cluster label of each event is determined by matching its representation to the set of M prototypes. The resulting visualizations are given in Figure 4. It shows that the proposed SWCC yields significantly better clustering performance than the BERT (InfoNCE), which means, to a certain extent, the prototype-based clustering method can help the event representation model capture various relations of events. Overall, the class separation in the visualizations qualitatively agrees with the performance in Table 1.

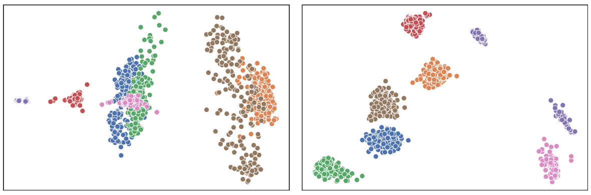


Figure 4: 2D visualizations of the event representation spaces learned by BERT (InfoNCE) (left) and SWCC (right), respectively. Each event is denoted by a color indicating a prototype.

Case study. We also present sampled events from two different prototypes in Table 4 (see Appendix for more examples), to further demonstrate the ability of SWCC to capture various relations of events. We can see that the events belonging to "Prototype1" mainly describe financial stuff, for example, "earnings be reduced", while the events belonging to "Prototype2" are mainly related to politics. Clearly, the events in the same cluster have the same topic. And we also find that there are also causal and temporal relations between some of these events. For example, "earnings be reduced" led to "company cut costs".

Prototype1	Prototype2
loans be sell in market	president asked senate
earnings be reduced	he deal with congress
company cut costs	senate reject it
earnings be flat	council gave approval
banks earn fees	council rejected bill

Table 4: Example events of two different prototypes.

6 Related Work

Event representation learning. Effectively representing events and their relations (casual, temporal, entailment (Ning et al., 2018; Yu et al., 2020)) becomes important for various downstream tasks, such as event schema induction (Li et al., 2020), event narrative modeling (Chambers and Jurafsky, 2008; Li et al., 2018; Lee and Goldwasser, 2019), event knowledge graph construction (Sap et al., 2019; Zhang et al., 2020) etc. Many efforts have been devoted into learning distributed event representation. Though driven by various motivations, the main idea of these methods is to exploit explicit relations of events as supervision signals and these supervision signals can be roughly categorized into three types: (1) discourse relations (e.g. casual and temporal relations) obtained with automatic annotation tools (Zheng et al., 2020); (2) manually annotated external knowledge (e.g. sentiments and intents) (Lee and Goldwasser, 2018; Ding et al., 2019) and (3) co-occurrence information (Weber et al., 2018). Existing work has focused on the first two supervision signals, with less research on how to better utilize co-occurrence information. Though, discourse relations and external knowledge are fine-grained relations that can provide more accurate knowledge, the current explicitly defined fine-grained relations fall under a small set of event relations. Co-occurrence information is easily accessible but underutilized. Our work focus on exploiting document-level co-occurrence information of events to learn event representations, without any additional annotations.

Instance-wise contrastive learning. Recently, a number of instance-wise contrastive learning methods have emerged to greatly improve the performance of unsupervised visual and text representations (He et al., 2020; Chen et al., 2020b,a; Chen and He, 2021; Grill et al., 2020; Zbontar et al., 2021; Chen et al., 2020a; Hu et al., 2021; Gao et al., 2021; Yang et al., 2021). This line of work aims at learning an embedding space where samples from the same instance are pulled closer and samples from different instances are pushed apart, and usually adopt InfoNCE (van den Oord et al., 2019) objective for training their models. Unlike the margin loss using one positive example and one negative example, the InfoNCE can handle the case where there exists multiple negative samples. In our work, we extend the InfoNCE, which is a self-supervised contrastive learning approach, to a weakly supervised contrastive learning setting, allowing us to effectively leverage co-occurrence information.

Deep unsupervised clustering. Clustering based methods have been proposed for representation learning (Caron et al., 2018; Zhan et al., 2020; Caron et al., 2020; Li et al., 2021; Zhang et al., 2021). Caron et al. (2018) use k-means assignments pseudo-labels to learn visual representations. Later, Asano et al. (2020) and Caron et al. (2020) cast the pseudo-label assignment problem as an instance of the optimal transport problem. Inspired by Caron et al. (2020), we leverage a similar formulation to map event representations to prototype vectors. Different from Caron et al. (2020), we simultaneously perform weakly supervised contrastive learning and prototype-based clustering.

7 Conclusion

In this work, we propose a simple and effective framework (**SWCC**) that learns event representations by making better use of co-occurrence information of events, without any addition annotations. In particular, we introduce a weakly supervised contrastive learning method that allows us to consider multiple positives and multiple negatives, and a prototype-based clustering method that avoids semantically related events being pulled apart. Our experiments indicate that our approach not only outperforms other baselines on several event related tasks, but has a good clustering performance on events. We also provide a thorough analysis of the prototype-based clustering method to demonstrate that the learned prototype vectors are able to implicitly capture various relations between events.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (61876053, 62006062, 62176076), the Shenzhen Foundational Research Funding (JCYJ20200109113441941, JCYJ20210324115614039), Joint Lab of Lab of HITSZ and China Merchants Securities.

References

Yuki Markus Asano, Christian Rupprecht, and Andrea Vedaldi. 2020. Self-labelling via simultaneous clustering and representation learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Long Bai, Saiping Guan, Jiafeng Guo, Zixuan Li, Xiaolong Jin, and Xueqi Cheng. 2021. Integrating deep event-level and script-level information for script event prediction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9869–9878, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. 2018. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 132–149.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

Hong Chen, Raphael Shu, Hiroya Takamura, and Hideki Nakayama. 2021. Graphplan: Story generation by planning with event graph. *ArXiv preprint*, abs/2102.02977.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020a. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020b. Improved baselines with momentum contrastive learning. *ArXiv preprint*, abs/2003.04297.

Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. 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 2292–2300.

Shumin Deng, Ningyu Zhang, Luoqiu Li, Chen Hui, Tou Huaixiao, Mosha Chen, Fei Huang, and Huajun Chen. 2021. OntoED: Low-resource event detection with ontology embedding. 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 2828–2839, Online. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiao Ding, Kuo Liao, Ting Liu, Zhongyang Li, and Junwen Duan. 2019. Event representation learning enhanced with external commonsense 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)*, pages 4894–4903, Hong Kong, China. Association for Computational Linguistics.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2016. Knowledge-driven event embedding for stock prediction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2133–2142, Osaka, Japan. The COLING 2016 Organizing Committee.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *ArXiv preprint*, abs/2104.08821.

Mark Granroth-Wilding and Stephen Clark. 2016. What happens next? event prediction using a compositional neural network model. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 2727–2733. AAAI Press.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. 2020. Bootstrap your own latent - A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9726–9735. IEEE.

Qianjiang Hu, Xiao Wang, Wei Hu, and Guo-Jun Qi. 2021. Adco: Adversarial contrast for efficient learning of unsupervised representations from self-trained negative adversaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1074–1083.

Zhiting Hu, Haoran Shi, Bowen Tan, Wentao Wang, Zichao Yang, Tiancheng Zhao, Junxian He, Lianhui Qin, Di Wang, Xuezhe Ma, Zhengzhong Liu, Xiaodan Liang, Wanrong Zhu, Devendra Sachan, and Eric Xing. 2019. Texar: A modularized, versatile, and extensible toolkit for text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 159–164, Florence, Italy. Association for Computational Linguistics.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. In *AAAI*.

Bram Jans, Steven Bethard, Ivan Vulić, and Marie Francine Moens. 2012. Skip n-grams and ranking functions for predicting script events. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 336–344, Avignon, France. Association for Computational Linguistics.

Dimitri Kartsaklis and Mehrnoosh Sadrzadeh. 2014. A study of entanglement in a categorical framework of natural language. *ArXiv preprint*, abs/1405.2874.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

I-Ta Lee and Dan Goldwasser. 2018. FEEL: featured event embedding learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4840–4847. AAAI Press.

I-Ta Lee and Dan Goldwasser. 2019. Multi-relational script learning for discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4214–4226, Florence, Italy. Association for Computational Linguistics.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven C. H. Hoi. 2021. Prototypical contrastive learning of unsupervised representations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Manling Li, Qi Zeng, Ying Lin, Kyunghyun Cho, Heng Ji, Jonathan May, Nathanael Chambers, and Clare Voss. 2020. Connecting the dots: Event graph schema induction with path language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 684–695, Online. Association for Computational Linguistics.

Zhongyang Li, Xiao Ding, and Ting Liu. 2018. Constructing narrative event evolutionary graph for script event prediction. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4201–4207. ijcai.org.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. 2021. R-drop: Regularized dropout for neural networks. *ArXiv preprint*, abs/2106.14448.

Shangwen Lv, Wanhui Qian, Longtao Huang, Jizhong Han, and Songlin Hu. 2019. Sam-net: Integrating event-level and chain-level attentions to predict what happens next. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6802–6809. AAAI Press.

Shangwen Lv, Fuqing Zhu, and Songlin Hu. 2020. Integrating external event knowledge for script learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 306–315, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Lara J. Martin, Prithviraj Ammanabrolu, Xinyu Wang, William Hancock, Shruti Singh, Brent Harrison, and Mark O. Riedl. 2018. Event representations for automated story generation with deep neural nets. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 868–875. AAAI Press.

Mausam, Michael Schmitz, Stephen Soderland, Robert Bart, and Oren Etzioni. 2012. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 523–534, Jeju Island, Korea. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *ArXiv preprint*, abs/1301.3781.

Qiang Ning, Hao Wu, and Dan Roth. 2018. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1318–1328, Melbourne, Australia. Association for Computational Linguistics.

Mehdi Rezaee and Francis Ferraro. 2021. Event representation with sequential, semi-supervised discrete variables. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4701–4716, Online. Association for Computational Linguistics.

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 *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3027–3035. AAAI Press.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 815–823. IEEE Computer Society.

Richard Socher, Danqi Chen, Christopher D. Manning, and Andrew Y. Ng. 2013a. Reasoning with neural tensor networks for knowledge base completion. 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 926–934.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013b. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2019. Representation learning with contrastive predictive coding.

Prashanth Vijayaraghavan and Deb Roy. 2021. Lifelong knowledge-enriched social event representation learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3624–3635, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Noah Weber, Niranjan Balasubramanian, and Nathanael Chambers. 2018. Event representations with tensor-based compositions. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4946–4953. AAAI Press.

Nianwen Xue, Hwee Tou Ng, Sameer Pradhan, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. 2016. CoNLL 2016 shared task on multilingual shallow discourse parsing. In *Proceedings of the CoNLL-16 shared task*, pages 1–19, Berlin, Germany. Association for Computational Linguistics.

Haoran Yang, Wai Lam, and Piji Li. 2021. Contrastive representation learning for exemplar-guided paraphrase generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4754–4761, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Changlong Yu, Hongming Zhang, Yangqiu Song, Wilfred Ng, and Lifeng Shang. 2020. Enriching large-scale eventuality knowledge graph with entailment relations. In *Automated Knowledge Base Construction*.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. 2021. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12310–12320. PMLR.

Xiaohang Zhan, Jiahao Xie, Ziwei Liu, Yew-Soon Ong, and Chen Change Loy. 2020. Online deep clustering for unsupervised representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 6687–6696. IEEE.

Dejiao Zhang, Feng Nan, Xiaokai Wei, Shang-Wen Li, Henghui Zhu, Kathleen McKeown, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. 2021. Supporting clustering with contrastive learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5419–5430, Online. Association for Computational Linguistics.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020. ASER: A large-scale eventuality knowledge graph. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 201–211. ACM / IW3C2.

Jianming Zheng, Fei Cai, and Honghui Chen. 2020. Incorporating scenario knowledge into A unified fine-tuning architecture for event representation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 249–258. ACM.

Yucheng Zhou, Xiubo Geng, Tao Shen, Jian Pei, Wenqiang Zhang, and Daxin Jiang. 2021. Modeling event-pair relations in external knowledge graphs for script reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4586–4596, Online. Association for Computational Linguistics.

A Appendix

A.1 Model Analysis

Impact of Temperature. We study the impact of the temperature by trying out different temperature rates in Table 5 and observe that all the variants underperform the $\tau = 0.3$.

SWCC	Hard similarity (Acc. %)		Transitive sentence
	Original	Extended	similarity (ρ)
with Temperature			
$\tau = 0.2$	80.0	71.0	0.80
$\tau = 0.3$	**80.9**	**71.3**	**0.82**
$\tau = 0.5$	77.4	68.7	0.78
$\tau = 0.7$	72.2	50.5	0.75
$\tau = 1.0$	48.7	22.9	0.67

Table 5: Impact of Temperature (τ).

Impact of the MLM objective with different γ. Table 6 presents the results obtained with different γ. As can be seen in the table, larger or smaller values of gamma can harm the performance of the model. $\gamma = 1.0$ gives a better overall performance of the model.

SWCC	Hard similarity (Acc. %)		Transitive sentence
	Original	Extended	similarity (ρ)
with MLM			
$\gamma = 0.1$	76.5	70.9	0.80
$\gamma = 0.5$	79.1	71.1	0.81
$\gamma = 1.0$	**80.9**	**72.1**	**0.82**
$\gamma = 1.5$	**80.9**	71.9	0.81
$\gamma = 2.0$	**80.9**	**72.1**	0.80

Table 6: Impact of the MLM objective with different γ.

Impact of the prototype-based clustering objective with different β. Finally, we study the impact of the prototype-based clustering objective with different β. As can be seen in the Table 7, the larger the $beta$, the better the performance of the model on the hard similarity task.

SWCC	Hard similarity (Acc. %)		Transitive sentence
	Original	Extended	similarity (ρ)
with $\mathcal{L}_{pc}$			
$\beta = 0.01$	78.3	71.6	0.80
$\beta = 0.05$	76.5	71.6	0.80
$\beta = 0.1$	**80.9**	72.1	**0.82**
$\beta = 0.3$	**80.9**	71.3	**0.82**
$\beta = 0.5$	**80.9**	**73.1**	0.80
$\beta = 0.7$	**80.9**	72.8	0.80
$\beta = 1.0$	**80.9**	72.1	0.80

Table 7: Impact of the prototype-based clustering objective with different β.

A.2 Case Study

Case study. We present sampled events from six different prototypes in Table 8 to further demonstrate the ability of SWCC to capture various relations of events. We can see that the events belonging to "Prototype1" mainly describe financial stuff, for example, "earnings be reduced", while the events belonging to "Prototype2" are mainly related to politics. Clearly, the events in the same cluster have the same topic. And we also find that there are also causal and temporal relations between some of these events. For example, "earnings be reduced" leads to "company cut costs".

Prototype1	Prototype2	Prototype3
loans be sell in market earnings be reduced company cut costs earnings be flat banks earn fees	president asked senate he deal with congress senate reject it council gave approval council rejected bill	he be known as director Wright be president of NBC Cook be chairman of ARCO Bernardo be manager for Chamber Philbin be manager of Board
Prototype4	**Prototype5**	**Prototype6**
he be encouraged by things I be content they be motivated by part they be meaningful he be ideal	kind is essential it be approach to life we respect desire thing be do for ourselves it be goal of people	Dorsey said to James Gephardt said to Richard Pherson said to Kathy Stone said to Professor Stiles said to Thomas

Table 8: Example events of different prototypes.