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

（ACL-2022)

Code： https://github.com/gaojun4ever/SWCC4Event

1. Introduction

2. Approach

3. Experiments

Introduction

在粒度上，事件介于词与句子之间：与词相比，事件通常包含多个词，用来描述事件的发生及事件的组成要素，是一种语义更完备的文本单元；与句子相比，事件更关注对现实世界中动作或变化的描述，是对现实世界一种更细粒度的刻画。

在形式上，事件的组成要素通常包括事件的触发词或类型、事件的参与者、事件发生的时间或地点等，与纯自然语言形式的文本相比，**事件是**现实世界中信息的一种更为**结构化**的表示形式。将结构化的事件信息表示为机器可以理解的形式对许多自然语言理解任务都十分必要，**例如脚本预测与故事生成**。

分布式语义表示将文本单元（如字、词等）嵌入到向量空间中，每个文本单元的语义信息由所有语义单元在向量空间中的位置共同决定。这种分布式的语义表示通常具有良好的性质，例如相关性较强（语义相近）的文本单元具有相似的向量表示，并且在很大程度上缓解了文本单元的稀疏性。

Introduction

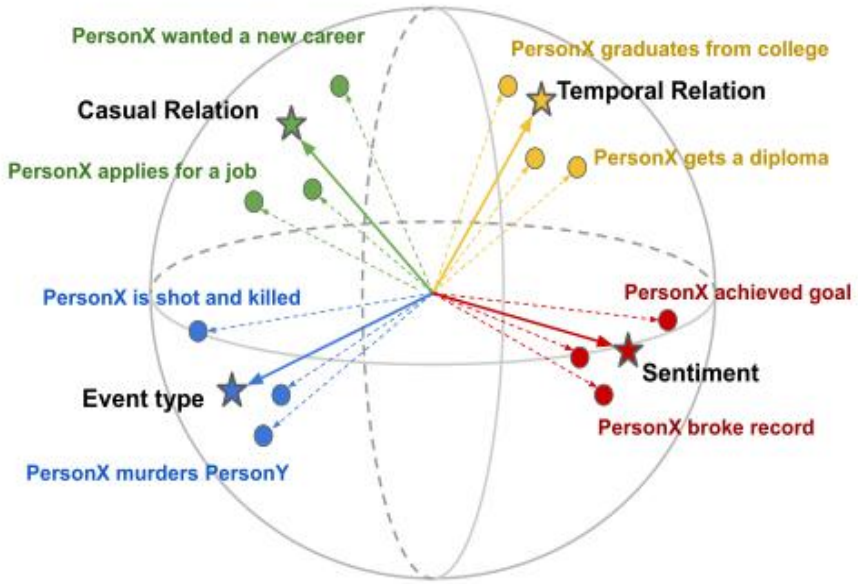

Figure 1: Four pairs of events with different relations. Stars represent prototypes and circles represent events.

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.

Approach

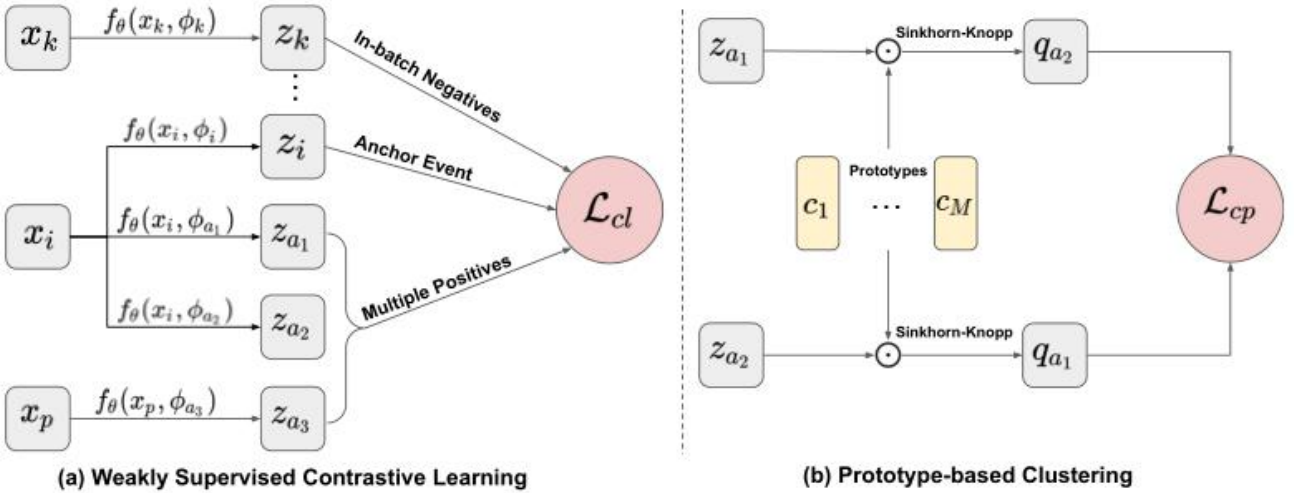

Figure 2: Architecture of the proposed framework, where the left part is the `Weakly Supervised Contrastive Learning` method and the right part is the `Prototype-based Clustering` method. Given an input event $\boldsymbol{x}_i$, we obtain three augmented representations $\boldsymbol{z}_i, \boldsymbol{z}_{a_1}$ 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.

Approach

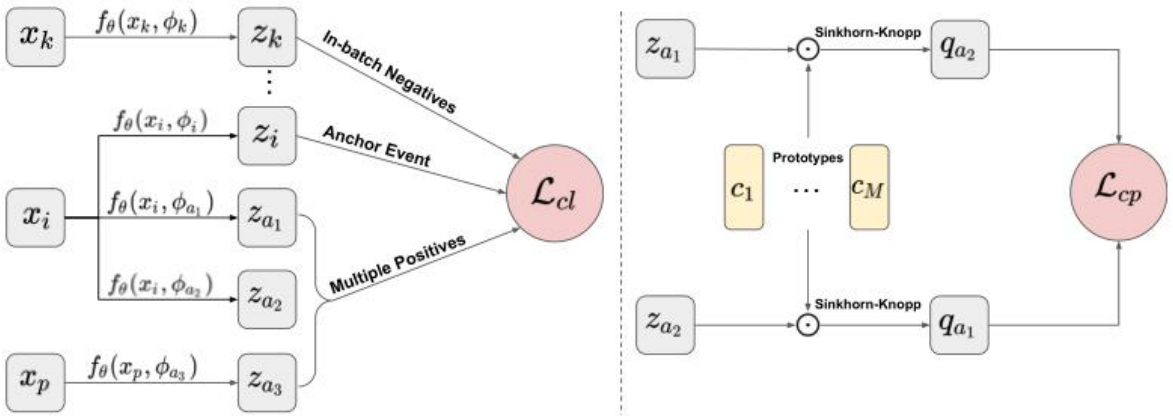

Figure 2: Architecture of the proposed framework, where the left part is the `Weakly Supervised Contrastive Learning` method and the right part is the `Prototype-based Clustering` method. Given an input event $\boldsymbol{x}_i$, we obtain three augmented representations $\boldsymbol{z}_i, \boldsymbol{z}_{a_1}$ 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.

$$[\text{CLS}], pred, subj, obj, [\text{SEP}]. \quad (1)$$

$$\boldsymbol{x} = [x_0, x_1, \cdots, x_L]$$

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

$$\boldsymbol{z} = \boldsymbol{v}_{[\text{CLS}]}.$$

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

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

Approach

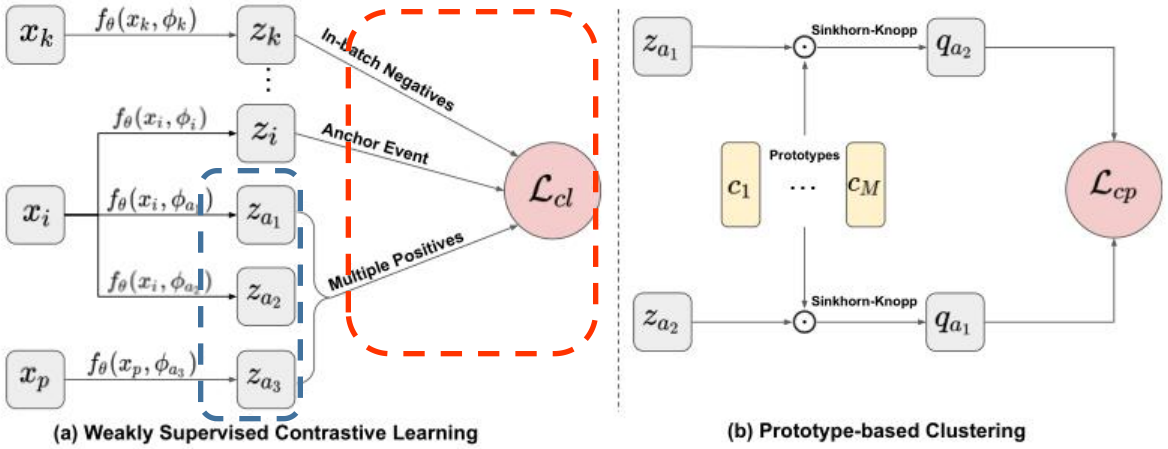

Figure 2: Architecture of the proposed framework, where the left part is the `Weakly Supervised Contrastive Learning` method and the right part is the `Prototype-based Clustering` method. Given an input event $\boldsymbol{x}_i$, we obtain three augmented representations $\boldsymbol{z}_i, \boldsymbol{z}_{a_1}$ 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.

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

$$\mathcal{A}(i) = \{\boldsymbol{z}_{a_1}, \boldsymbol{z}_{a_2}, \boldsymbol{z}_{a_3}\}$$

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

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

Approach

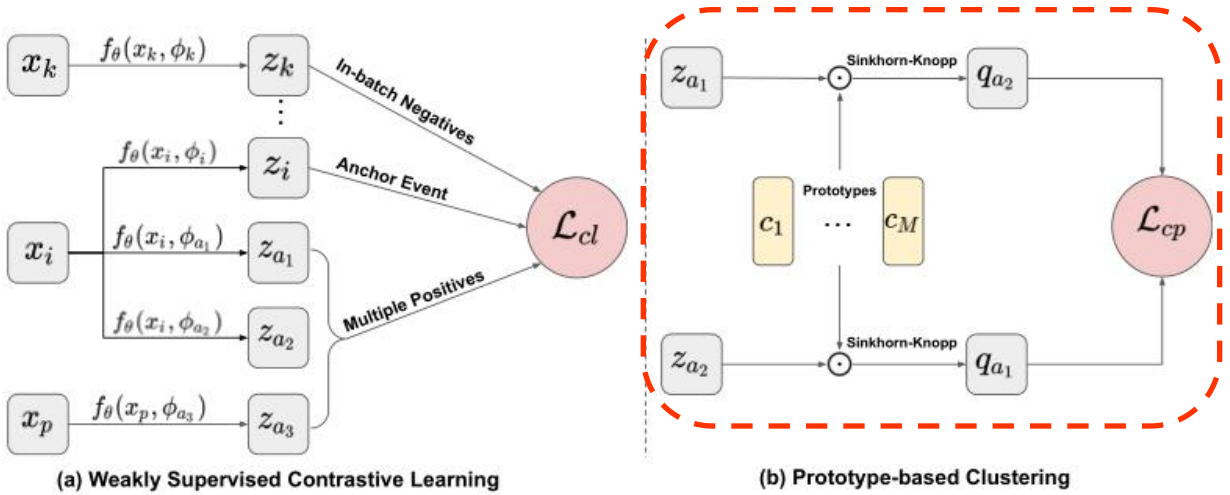

Figure 2: Architecture of the proposed framework, where the left part is the `Weakly Supervised Contrastive Learning` method and the right part is the `Prototype-based Clustering` method. Given an input event $\boldsymbol{x}_i$, we obtain three augmented representations $\boldsymbol{z}_i, \boldsymbol{z}_{a_1}$ 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.

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

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:

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

$\boldsymbol{z}_{a_1}$ 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.

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

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.

Experiments

Model	Hard similarity (Accuracy %)		Transitive sentence similarity (ρ)
	Original	Extended	
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.

Experiments

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.

Experiments

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).

Experiments

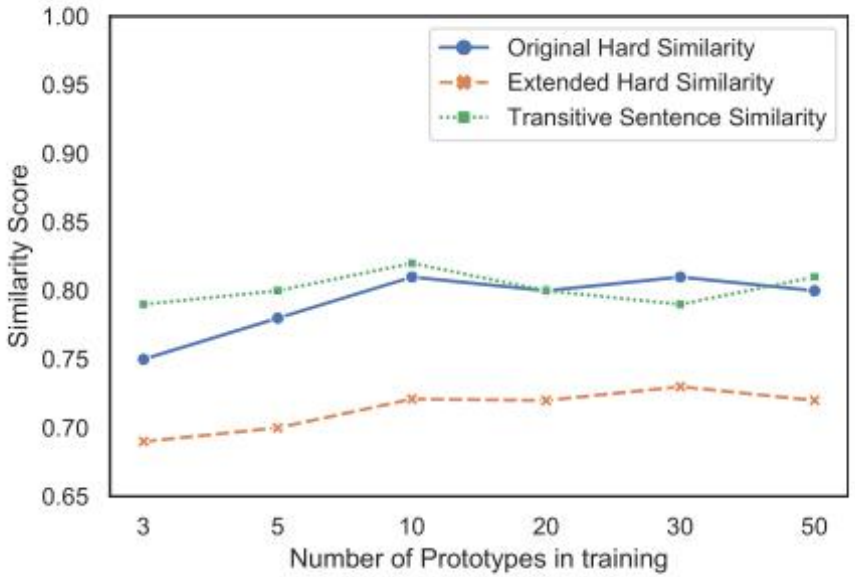

Figure 3: Impact of # of Prototypes

Experiments

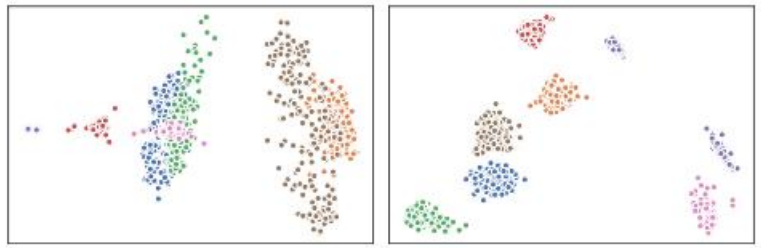

Figure 4: 2D visualizations of the event representation spaces learned by BERT (InfoNCE) (left) and SWCC (right), respectively. Each event is denoted by a color indicating a prototype.

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.

Experiments

SWCC	Hard similarity (Acc. %)		Transitive sentence similarity (ρ)
	Original	Extended	
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).

SWCC	Hard similarity (Acc. %)		Transitive sentence similarity (ρ)
	Original	Extended	
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.

Experiments

SWCC	Hard similarity (Acc. %)		Transitive sentence similarity (ρ)
	Original	Extended	
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.

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

Table 8: Example events of different prototypes.

Thank you !