ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230616

Siamese Representation Learning for Unsupervised Relation Extraction

Guangxin Zhang[a] **and Shu Chen**[a;*]

[a]School of IoT Engineering, Jiangnan University, China

Abstract. Unsupervised relation extraction (URE) aims at discovering underlying relations between named entity pairs from open-domain plain text without prior information on relational distribution. Existing URE models utilizing contrastive learning, which attract positive samples and repulse negative samples to promote better separation, have got decent effect. However, fine-grained relational semantic in relationship makes spurious negative samples, damaging the inherent hierarchical structure and hindering performances. To tackle this problem, we propose Siamese Representation Learning for Unsupervised Relation Extraction – a novel framework to simply leverage positive pairs to representation learning, possessing the capability to effectively optimize relation representation of instances and retain hierarchical information in relational feature space. Experimental results show that our model significantly advances the state-of-the-art results on two benchmark datasets and detailed analyses demonstrate the effectiveness and robustness of our proposed model on unsupervised relation extraction. We have released our code at https://github.com/gxxxzhang/siamese-ure.

1 Introduction

Relation Extraction (RE) is the task of extracting semantic relation between entity pair from raw text. For example, given the sentence *"ChatGPT is created by OpenAI, a research organization dedicated to creating and promoting friendly AI that benefits humanity"*, and the entity pair *(ChatGPT, OpenAI)*, RE model can predict the predefine relationship *"created_by"* and extract the corresponding triplet *(ChatGPT, created_by, OpenAI)* for downstream tasks, such as web search [35], knowledge base construction [1] and question answering [5]. Existing RE methods which are restricted to specific relation types have achieved good performance with annotated data. Nevertheless, with the rapid emergence of large, domain-specific text corpora (e.g., sports news, social media content, scientific publications) and new relation types in the real world, these methods face many challenges. On the one hand manually establishing and maintaining the ever-growing relation require expert knowledge and are time-consuming, on the other hand these methods are hard to scale up to newly emerged relations. Unsupervised relation extraction is promising and received widespread concern since it does not require prior information on relation distribution to reduce the reliance on labeled data and can discover new relation types in raw text.

Traditional unsupervised relation extraction approaches are based on variational autoencoder (VAE) architecture [20, 27, 28, 37]. These methods train the relation extraction model as an encoder that generates relation classifications. A decoder is trained along with the encoder to reconstruct the encoder input based on the encoder generated relation classifications. However, joint training for two networks (encoder and decoder) and requiring the exact number of relation classes during the training period of the encoder make model unstable. Different from treating relations as latent variables, the clustering-based approaches learn semantic relational representation from high-dimensional embeddings and adopt unsupervised clustering algorithms to recognize relation classes in feature space. In this process, the main challenge is how to learn semantic representation of instances in the relational feature space.

Elsahar et al. [10] extracts KB types and NER tags of entities as well as re-weighted word embeddings from sentences, then adopts Principal Component Analysis (PCA) to reduce feature dimensionality that can alleviate the problem of features sparsity, and finally uses Hierarchical Agglomerative Clustering (HAC) to cluster the feature representations. Because integrating word embeddings in a rule-based way, the method heavily rely on hand-craft features and make many simplifying assumptions and its feature space is lack of semantic information. Liu et al. [18] formulate URE using a structural causal model and conduct Element Intervention to eliminate spurious correlations, which intervenes on the context and entities respectively to obtain the underlying causal effects of them and learn the causal effects through instance-wise contrastive learning.

The core idea of instance-wise contrastive learning is to pull together the representations within positive instances while pushing apart negative ones. However, negative examples are commonly sampled from the batch or training data at random due to the lack of ground-truth annotations. In relation extraction, the relational semantic tend to be more fine-grained and based on a potential hierarchical structure [38]. For example, the relations *"per:stateorprovinces_of_residence"*, *"per:countries_of_residence"* and *"per:cities_of_residence"* in one benchmark dataset share the same parent semantic on */people/residence*, which means that they belong to the same semantic cluster from a hierarchical perspective. Naturally, instances may have highly similar semantics in a batch but contrastive learning pushes these representations apart as long as they are from different original instances, regardless of their semantic similarities. To alleviate the dilemma which instance-wise contrastive learning unreasonably pushes apart those sentence pairs that are semantically similar, Liu et al. [19] propose hierarchical exemplar contrastive learning. The model leverages Hierarchical Propagation Clustering to obtain hierarchical exemplars from relational feature space and further utilizes exemplars to hierarchically update relational features of sentences and is optimized by performing both instance and

* Corresponding Author. Email: kjcuse@jiangnan.edu.cn

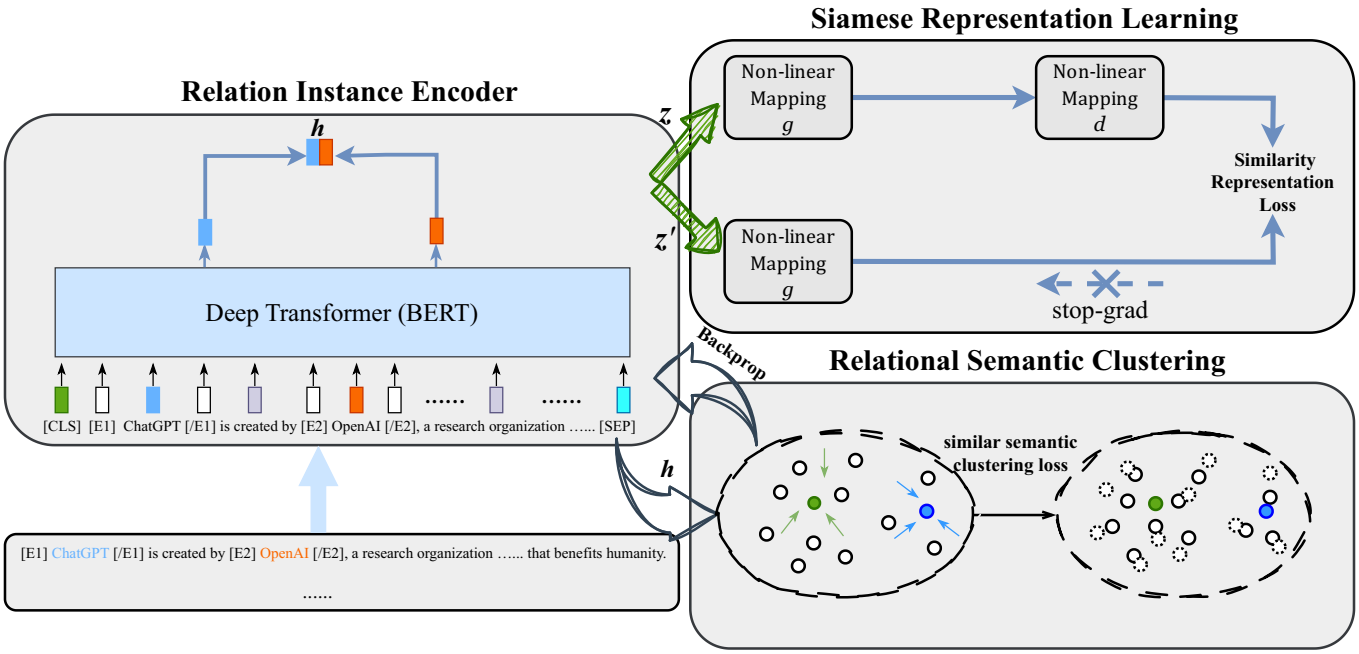

Figure 1: Overall architecture of our model. Instances will be fed into encoder with different dropout masks z and z' to obtain feature pairs, then transmitted into Non-linear Mapping $\boldsymbol{g}(\cdot)$ and $\boldsymbol{d}(\cdot)$ respectively. We use ones to predict other ones to optimize encoder. Besides, in relational semantic clustering, we mining nearest neighbors of each sample with semantic clustering loss.

exemplar-wise contrastive learning through Hierarchical Exemplar Contrastive Loss and propagation clustering iteratively. Nonetheless, the method still utilize traditional instance-wise contrastive learning loss to retain the local smoothness in relational feature space during training period and use an iterative way to obtain hierarchical exemplars will cause error accumulation [24].

In this paper, we attempt to propose a Siamese network architecture, only using positive pairs for representation learning, to eliminate adverse effect of spurious negative samples. Siamese networks are able to learn similarity metrics of relations from labeled data of predefined relations, and then transfer the relational knowledge to identify novel relations in unlabeled data [33]. Therefore, we intend to use Siamese architecture to construct positive pairs and learn relation representations in unsupervised setting. For the most part, owing to lack of labeled data, the network is easy to collapse (i.e., all outputs "collapsing" to a constant). To avoid that ,we introduce entity type, which provide a strong inductive bias for relation extraction [28], as prior information. Furthermore, followed by Chen and He [7], we use the analogous network architecture to further prevent to collapse. To recognize different relations, we through traditional unsupervised clustering algorithms (e.g., k-means clustering algorithm) to cluster learned representation in relational feature space. Nevertheless, naively applying these clustering algorithms on the obtained features can lead to cluster degeneracy [6]. We propose relational semantic clustering module that mining nearest neighbors of each instance in feature space. The module can support model learn more discriminative representations under semantic clustering loss, so that the learned representation is cluster-friendly and obtain better clustering performance.

Our main contributions are the following: (1) We propose a novel representation learning framework for unsupervised relation extraction. Furthermore, our model is much simpler than existing self-supervised learning models that apply empirical data augmentation and complicated network architecture. (2) We explore Relational Semantic Clustering module, encoding the relational semantic into the representations via unsupervised clustering. We efficiently conduct representation learning and unsupervised clustering in a unified framework. (3) We conduct extensive experiments on two datasets and achieves better performance than the existing state-of-the-art methods. Meanwhile, our ablation analysis shows the impacts of different modules in our framework.

2 Model

We aim at developing a joint model that leverages the beneficial properties of self-supervised learning to improve unsupervised relation extraction. As illustrated in Figure 1, our model consists of three modules: Relation Instance Encoder, Siamese Representation Learning and Relational Semantic Clustering. The encoder module uses instances as input which are composed of natural language sentences and entity pairs, and then employs the pre-trained model to output entity-level feature pair sets $\boldsymbol{H}$ and $\boldsymbol{H'}$ for all instances. The learning module is structured as Siamese architecture and takes pair sets respectively as input. We use similarity representation loss, which measure the cosine similarity of positive pairs, to enforces the relational feature of instances that have similar semantics to be more close in feature space. In clustering module, we mining nearest neighbors of each sample in relational feature space to learn discriminative representations under semantic clustering loss that can yield better clustering performance.

2.1 Relation Instance Encoder

The Relation Instance Encoder aims to obtain relational features from instances. We use instances $\boldsymbol{X}$ as inputs that each $\boldsymbol{x_i} \in \boldsymbol{X}$ is composed of a sentence $s_i = \{w_1, .., w_n\}$ with n words, $\mathrm{h_i} = (\mathrm{h_i^s}, \mathrm{h_i^e})$ representing the start and end position of head entity, and $\mathrm{t_i} = (\mathrm{t_i^s}, \mathrm{t_i^e})$

representing the start or end position of tail entity in the sentence. Same as the previous work, named entities in the sentences have been recognized in advance. We employ pre-trained BERT [8] model, which has strong performance on extracting contextual information, as encoder $\boldsymbol{f}$ to map relation instance x_i to embedding. However, BERT always induces a non-smooth anisotropic semantic space of sentences [16] , which is easier make model to collapse. To this end, we add entity types as prior information in head and tail entities as $[\mathrm{h_i^s}]$, $[\mathrm{h_i^e}]$, $[\mathrm{t_i^s}]$, $[\mathrm{t_i^e}]$ and inject them to each instance $\boldsymbol{x_i}$:

$$\boldsymbol{x_i} = [w_1, \ldots, [\mathrm{h_i^s}], \ldots, w_i, \ldots, [\mathrm{h_i^e}], \ldots, [\mathrm{t_i^s}], \ldots, [\mathrm{t_i^e}], \ldots, w_n] \tag{1}$$

then get the token embedding:

$$\mathbf{e}_1, ..., \mathbf{e}_n = \boldsymbol{f}_\theta(\mathbf{x_1}, ..., \mathbf{x_n}) \tag{2}$$

where θ is the learnable parameters in the encoder. We use the embeddings with position of $[\mathrm{h_i^s}]$ and position of $[\mathrm{t_i^s}]$ as outputs to obtain the entity-level feature $\boldsymbol{h}_i \in 2 \cdot \mathbb{R}^d$:

$$\boldsymbol{h}_i = \mathbf{e}_{head} \oplus \mathbf{e}_{tail} \tag{3}$$

We follow the **data augmentation** used in SimCSE [11] to construct positive pairs. Specifically, we only feed the same input $\boldsymbol{x_i}$ to the encoder twice with different dropout masks, which placed on fully-connected layers as well as attention probabilities, and we can obtain two different embeddings as positive pairs. We denote positive pair as $\boldsymbol{h_i} = \boldsymbol{f}_\theta(\mathbf{x_i};\, \mathbf{z})$ and $\boldsymbol{h_i'} = \boldsymbol{f}_\theta(\mathbf{x_i};\, \mathbf{z}')$, where $\mathbf{z}$ and $\mathbf{z}'$ is the different dropout masks.

2.2 Siamese Representation Learning

We use Siamese network which can naturally introduce inductive biases for modeling invariance to learn relational similarity metrics. However, Siamese networks suffers from the problem of model collapse, where the model converges to a constant value and the samples all mapped to a single point. Besides, it is difficult to learn a reasonable distance in feature space without negative samples. Followed by Chen and He [7], we attempt to use the similar approach to address this issue. As shown in Figure 1, positive feature pair $\boldsymbol{h_i}$ and $\boldsymbol{h_i'}$ of one instance are processed by the same encoder network $\boldsymbol{f}$ with different dropout masks $\mathbf{z}$ and $\mathbf{z}'$. Then we use the MLP $\boldsymbol{g}(\cdot)$ to map feature pair to g_i and g_i' respectively:

$$g_i, g_i' = \boldsymbol{g}_\phi\left(\boldsymbol{h_i},\ \boldsymbol{h_i'}\right) \tag{4}$$

where ϕ is the learnable parameters in the non-linear mapping network. And then, the non-linear mapping network $\boldsymbol{d}(\cdot)$ is applied on one side, and we denote the feature as $d_i' = \boldsymbol{d}_\psi(g_i')$, ψ is the learnable parameters. Meanwhile, a stop-gradient operation is applied on the other side. The model maximizes the similarity between both sides to learn relation representation under similarity representation loss.

Similarity Representation Loss

Given a training set $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}\}$ of n instances, Relation Instance Encoder can obtain two augmented relational features for each input sentences by feed the same input to the encoder twice with different dropout masks. In this process, we obtain feature sets $\boldsymbol{H} = \{\boldsymbol{h_1}, \boldsymbol{h_2}, \ldots, \boldsymbol{h_n}\}$ and $\boldsymbol{H'} = \{\boldsymbol{h_1'}, \boldsymbol{h_2'}, \ldots, \boldsymbol{h_n'}\}$. We minimize negative cosine similarity between positive pair:

$$\mathcal{L}_{Rl} = \frac{1}{2}\left[\mathcal{D}\big(d, \mathrm{stopgrad}(g')\big) + \mathcal{D}\big(d', \mathrm{stopgrad}(g)\big)\right] \tag{5}$$

where $\mathrm{stopgrad}(\cdot)$ is use stop-grad strategy that gradient does not back-propagate. The cost is described as a symmetrized form since pairs from the Siamese network. For each part of the loss, we minimize their negative cosine similarity:

$$\mathcal{D}(d, g') = -\frac{1}{n}\sum_{i=1}^{n}\frac{d_i}{\|d_i\|_2}\cdot\frac{g_i'}{\|g_i'\|_2} \tag{6}$$

for a mini-batch of N instances,where $i \in [1, N]$ and $\|\cdot\|_2$ is ℓ_2-norm.

2.3 Relational Semantic Clustering

One of the main obstacles for our model is difficult learn a discriminative representation without negative samples. Be enlightened by Van Gansbeke [30], we assume that in a excellent relational feature space, each sample with their nearest neighbors have similar relational semantic and belong to the same relation class. We propose Relational Semantic Clustering to learn discriminative relational representations and conduct clustering and representation learning in a unified framework. Specifically, for each instance $\mathbf{x_i}$, we mine its K nearest neighbors according to each representation $\boldsymbol{h_i}$ and define the set $\mathcal{N}_{x_i}$ as the output features of neighboring samples corresponding to each $\mathbf{x_i}$ in the dataset. Then, we use similar semantic clustering loss to attract instances and their neighboring samples to approach each other. Simultaneously, different instances are separated in feature space.

Similar Semantic Clustering Loss

Like adaptive clustering [34], we aim to learn a clustering function Φ_σ - parameterized by a neural network. The neural network classifies each instance $\mathbf{x_i}$ and its mined neighbors $\mathcal{N}_{x_i}$ together with the learnable parameters σ. The function Φ_σ terminates in a softmax function to perform a soft assignment over the clusters $\mathcal{C} = \{1, \ldots, C\}$, with $\Phi_\sigma(\mathbf{x_i}) \in [0,1]^C$. The probability of instance $\mathbf{x_i}$ being assigned to cluster c is denoted as $\Phi_\sigma^c(\mathbf{x_i})$. We learn the weights of Φ_σ by minimizing the following objective:

$$\mathcal{L}_{Cl} = -\frac{1}{|\mathbf{X}|}\sum_{\mathbf{x_i}\in\mathbf{X}}\sum_{k\in\mathcal{N}_{\mathbf{x_i}}}\log\langle\Phi_\sigma(\mathbf{x_i}), \Phi_\sigma(k)\rangle + \lambda\sum_{c\in\mathcal{C}}\Phi_\sigma'^c\log\Phi_\sigma'^c \tag{7}$$

where $\langle\cdot\rangle$ denotes the dot product operator. The first term in Equation 7 imposes Φ_σ to make consistent predictions for each instance $\mathbf{x_i}$ and its neighboring samples $\mathcal{N}_{\mathbf{x_i}}$. However, the dot product will be maximal when the predictions are one-hot (confident) and assigned to the same cluster (consistent). To avoid Φ_σ from assigning all samples to a single cluster, we include an entropy term(the second term in Equation 7):

$$\Phi_\sigma'^c = \frac{1}{|\mathbf{X}|}\sum_{\mathbf{x_i}\in\mathbf{X}}\Phi_\sigma^c(\mathbf{x_i}). \tag{8}$$

which spreads the predictions uniformly across the clusters $\mathcal{C}$ and encourages the classifier to scatter a set of instances into different classes.

2.4 Iterative Joint Training

In early training period, we only optimize $\mathcal{L}_{Rl}$ in Siamese Representation Learning module to drive instances with similar semantic get closer in feature space. After several warm-up epochs, instances have acquired reasonable semantic representations. We introduce Relational Semantic Clustering module to further refine the semantic representations and enable them to be more discriminative. In summary, our overall objective is:

$$\mathcal{L} = \mathcal{L}_{Rl} + \eta\mathcal{L}_{Cl} \tag{9}$$

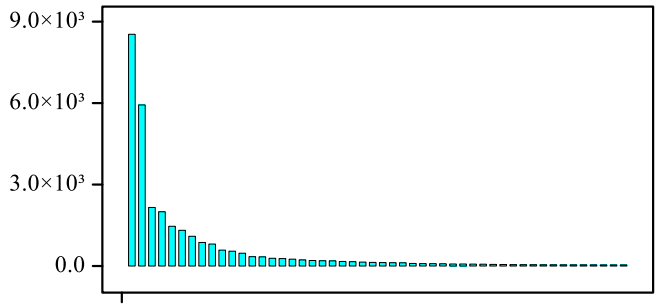

Figure 2: The relation distribution in the NYT+FB dataset. The ordinate represents the number of sentences in each relation type in the dataset.The x-axis is the relations sorted according to the number of sentences contained. For ease of observation, the x-axis label is omitted.

where η is a loss coefficient. Our approach involves the combined utilization of the Siamese Representation Learning module and the Relational Semantic Clustering module through an iterative procedure. This joint usage enables model to achieve a well-separated representation of distinct instances in the learned feature space while preserving local invariance for each individual instance.

3 Experiments

In this section, we first describe two relation extraction datasets for training and evaluating the proposed method, then detail the baseline models for comparison, and then expound the implementation details and hyperparameter configuration, finally we conduct a comprehensive and detailed analysis of our model.

3.1 Datasets and Evaluation Metrics

Datasets. Following previous work [28, 19], we conduct experiments on two relation extraction datasets – NYT+FB [20] and TACRED [39] with different constructing settings. The former is generated via distant supervision while the latter is manually annotated corpus, which is extremely challenging to model. The NYT+FB dataset is obtained by using Freebase to label the corpus of the New York Times corpus. That is, if the entity pair that appears in a sentence also appears in Freebase [4], then this sentence is automatically labeled as the relation stored by Freebase. After filtering out some sentences using syntactic patterns, there are 2 million sentences in the dataset, of which 41,000 are labeled with meaningful relations1. Of the 41,000 tagged sentences, 20% are used as validation set, and 80% are used as test set. The TACRED dataset is a large-scale crowd-sourced relation extraction dataset following the TAC KBP relation schema that covers 42 relation types. We remove the instances labeled as `no_relation` and use the remaining 21,773 instances including 41 relation types for training and evaluation.

Evaluation Metrics. B-cube (B^3) [3], V-measure [26] and Adjusted Rand Index (ARI) [13] are used as evaluation metrics for different models. Specifically, B^3 contains the precision and recall metrics to correspondingly measure the correct rate of putting each sentence in its cluster or clustering all samples into a single class, which are defined as follows:

$$\mathrm{B}^3_{\mathrm{Prec.}} = \mathop{\mathbb{E}}_{X,Y} P(g(X) = g(Y)|c(X) = c(Y))$$

$$\mathrm{B}^3_{\mathrm{Rec.}} = \mathop{\mathbb{E}}_{X,Y} P(c(X) = c(Y)|g(X) = g(Y))$$

Table 1: Hyper-parameter values used in our experiments.

Hyper-parameters	**value**
optimizer	*SGD*
learning rate	1e-5
weight_decay	1e-4
momentum	0.9
batch size	64
warm-up epochs L	5
dropout rate r	0.1
number of nearest neighbors K	20
loss coefficient η	0.5

Then B^3 F_1 is computed as the harmonic mean of the precision and recall.

V-measures contains the homogeneity and completeness, which is analogous to B^3 precision and recall. These two metrics penalize small impurities in a relatively "pure" cluster more harshly than in less "pure" ones:

$$\mathrm{V}_{\mathrm{Homo.}} = 1 - H(c(X)|g(X))/H(c(X))$$

$$\mathrm{V}_{\mathrm{Comp.}} = 1 - H(g(X)|c(X))/H(g(x))$$

ARI is a normalization of the Rand Index, which measures the agreement degree between the cluster and golden distribution. This metric ranges in [-1,1]. The larger the value, the more consistent the clustering result is with the real situation.

3.2 Baselines

To evaluate the effectiveness of our method, we select the following unsupervised relation extraction models for comparison with standard evaluation metrics: 1) **rel-LDA** [36], a generative model that considers the unsupervised relation extraction as a topic model.We choose the full rel-LDA with a total number of 8 features for comparison in our experiment. 2) **March** [20], a VAE-based model learned by self-supervised signal of entity link predictor. 3) **UIE** [27],a discriminative model that adopts additional regularization to guide model learning. And it has different versions according to the choices of different relation encoding models (e.g., PCNN). We report the results of two versions—UIE-PCNN and UIE-BERT (i.e., using PCNN and BERT as the relation encoding models) with the highest performance. 4) **EType** [28], a simple and effective method relying only on entity types. The same link predictor as in **March** [20] is employed and two additional regularizers are used. 5) **SelfORE** [12],a self-supervised framework that bootstraps to learn a contextual relation representation through adaptive clustering and pseudo label. 6) **EIURE** [18], a contrastive learning framework that intervenes on the context and entities respectively to obtain the underlying causal effects of them. 7) **HiURE** [19], is the state-of-the-art method that derive hierarchical signals from relational feature space using cross hierarchy attention and effectively optimize relation representation of sentences under exemplar-wise contrastive learning.

3.3 Implementation Details

In order to do a fair comparison with baseline method, we adopted the setting by clustering all samples into 10 relation super-classes. In the process of training of our model, we used the development set to manually search part of the hyper-parameters, Table 1 shows our best parameter settings. In our implementation, we adopt the pre-trained `Bert-Base-Cased` model to initialize parameters for

Table 2: Main results on two relation extraction datasets.The results of all baseline are reproduced in liu et al. [19], MLPs refers to two mapping networks in Representation Learnin module and Semantic Clustering refers to Relational Semantic Cluster module in our model.

Dataset	Model	B^3			V-measure			ARI
		F1	Prec.	Rec.	F1	Hom.	Comp.	
NYT+FB	rel-LDA[36]	29.1±2.5	24.8±3.2	35.2±2.1	30.0±2.3	26.1±3.3	35.1±3.5	13.3±2.7
	March[20]	35.2±3.5	23.8±3.2	67.1±4.1	27.0±3.0	18.6 ±1.8	49.6±3.1	18.7±2.6
	UIE-PCNN[27]	37.5±2.9	31.1±3.0	47.4±2.8	38.7±3.2	32.6±3.3	47.8±2.9	27.6±2.5
	UIE-BERT[27]	38.7±2.8	32.2±2.4	48.5±2.9	37.8±2.1	32.3±2.9	45.7±3.1	29.4±2.3
	EType[28]	41.9±2.0	31.3±2.1	63.7±2.0	40.6±2.2	31.8±2.5	56.2±1.8	32.7±1.9
	SelfORE[12]	41.4±1.9	38.5±2.2	44.7±1.8	40.4±1.7	37.8±2.4	43.3±1.9	35.0±2.0
	EIURE[18]	43.1±1.8	48.4±1.9	38.8±1.8	42.7±1.6	37.7±1.5	49.2±1.6	34.5±1.4
	HiURE[19]	44.3±0.5	39.9±0.6	49.8±0.5	44.9±0.4	40.0±0.5	51.2±0.4	38.3±0.6
	Our w/o MLPs	40.5±0.6	35.9±0.5	46.5±0.8	43.3±0.4	38.2±0.2	50.0±0.5	31.0±0.5
	Our w/o Semantic Clustering	41.9±0.3	36.8±0.3	48.8±0.4	44.7±0.8	39.3±0.2	51.8±0.9	32.1±0.7
	Our	**44.9±0.4**	39.5±0.3	52.1±0.7	**45.7±0.6**	40.0±0.3	53.2±0.8	**39.6±0.3**
TACRED	rel-LDA[36]	35.6±2.6	32.9±2.5	38.8±3.1	38.0±3.5	33.7±2.6	43.6±3.7	21.9±2.6
	March[20]	38.8±2.9	35.5±2.8	42.7±3.2	40.6±3.1	36.1±2.7	46.5±3.2	25.3±2.7
	UIE-PCNN[27]	41.4±2.4	44.0±2.7	39.1±2.1	41.3±2.3	40.6±2.2	42.1±2.6	30.6±2.5
	UIE-BERT[27]	43.1±2.0	43.1±1.9	43.2±2.3	49.4±2.1	48.8±2.1	50.1±2.5	32.5±2.4
	EType[28]	49.3±1.9	51.9±2.1	47.0±1.8	53.6±2.2	52.5±2.1	54.8±1.9	35.7±2.1
	SelfORE[12]	47.6±1.7	51.6±2.0	44.2±1.9	52.1±2.2	51.3±2.0	52.9±2.3	36.1±2.0
	EIURE[18]	52.2±1.4	57.4±1.3	47.8±1.5	58.7±1.2	57.7±1.4	59.7±1.7	38.6±1.1
	HiURE[19]	55.8±0.4	57.8±0.3	54.0±0.5	59.7±0.6	57.6±0.5	61.9±0.6	40.5±0.4
	Our w/o MLPs	53.6±0.8	45.6±0.6	65.0±1.2	59.5±0.7	51.9±0.5	69.8±1.1	44.0±0.9
	Our w/o Semantic Clustering	56.1±0.3	47.7±0.2	68.1±0.8	64.1±0.7	56.2±0.8	74.6±1.3	48.4±0.6
	Our	**59.5±0.6**	49.4±0.4	74.9±0.8	**66.7±0.8**	58.1±0.6	78.5±0.9	**50.6±0.5**

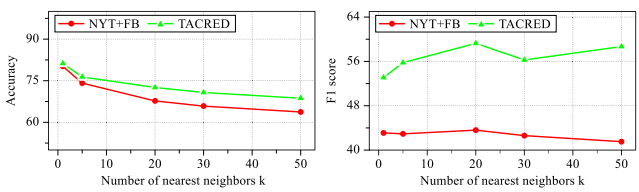

Figure 3: Influence of the used number of neighbors K.

Relation Instance Encoder and set dropout rate $r = 0.1$ to generate positive pairs. The output entity-level features $\mathbf{h}_i$ and $\mathbf{h}'_i$ possess the dimension of $2 \cdot \mathbb{R}^d$, where $\mathbb{R}^d = 768$. For Siamese Representation Learning, we use Non-linear Mapping $\boldsymbol{g}(\cdot)$ and $\boldsymbol{d}(\cdot)$ in our network and use SGD with 1e-5 learning rate to optimize the loss. The $\boldsymbol{g}(\cdot)$ has layer normalization (LN) [2] applied to each fully-connected (fc) layer, including its output fc. Its output fc has no ReLU. This MLP has 3 layers. The $\boldsymbol{d}(\cdot)$ has LN applied to its hidden fc layers. Its output fc does not have LN or ReLU. This MLP has 2 layers. For Relation Semantic Clustering, we set warm-up epochs $L = 5$ and number of nearest neighbors of each instance $K = 20$. In the evaluation period, we simply adopt the pre-trained models for representation extraction, then cluster the evaluation instances based on these representations.

3.4 Results

We summarize the performances of our method and seven baseline models in Table 2. All of these models are evaluated on identical test set to show their performance. From the experimental results, we can see that our method significantly outperforms baselines. For NYT+FB dataset, compared with the previous SOTA model, our method improves V-measure F1 by 1.6%, and ARI by 5.7%, but the B^3 F1 score is only by 0.6%. The reason why the performance gain is minuscule in NYT+FB is that the dataset contains numerous wrongly labeled instances in the train and test sets. These instances are unable to reflect the real performance of the model, and the number of samples in different relations is very unbalanced. As shown in Figure 2, the relation distribution is similar to a long-tailed distribution. Besides, most relations only have several samples that make a lot of noisy nearest neighbors to hurt performance when use mining nearest neighbors in Relation Semantic Clustering. For TACRED, compared with the previous SOTA model, our method improves B^3 F1 by 3.7%, V-measure F1 by 7.0%, and ARI by 10.1%. It is worth noting that the score of precision and recall in B^3 seems to be extremely disequilibrium. By definition, precision measure the correct rate of putting each sample in its cluster and recall measure the correct rate of clustering all samples into a single class. Therefore, the results indicate that most of the samples from corresponding relations are clustered in the same cluster.

Ablation Study. To study the contribution of different components in the proposed method, we conduct an ablation study on each component. For fair comparisons, the other settings remain the same as the main model. From Table 2, we can see that in both NYT+FB and TACRED, the model's performance is degraded if any component is removed, indicating that both modules are important for the final model performance.

3.5 Detailed Analysis

Hyperparameter Analysis. On account of the importance of two hyperparameters dropout rate r and number of nearest neighbors K which is in the encoder module and cluster module respectively, we conduct a detail analysis on them. Firstly, to further study the role of dropout rate in relation instance encoder for data augmentation, we try out different rates and report the performance of B^3 F1 on NYT+FB and TACRED. As shown in Table 3, we observe that all the variants underperform the default dropout rate $r = 0.1$ of Transformers [31]. Using small dropout rate will introduce small divergence so that it is difficult for our model to learn discriminative representation,

Table 3: Effects of different dropout rates r on the NYT+FB and TACRED development sets.

Dataset / r	0.01	0.05	0.1	0.15	0.2	0.5
NYT+FB	41.3	42.1	**44.3**	44.1	43.8	41.5
TACRED	50.4	52.1	**58.7**	57.6	57.1	51.1

while large dropout rate will make more noise and prejudice similar semantic information. Secondly, we study the influence of number of nearest neighbors K in cluster module for mining nearest neighbors and report the accuracy and F1 score on two datasets. As shown in Figure 3, the accuracy (left),which is the correct rate of that nearest neighbors with their corresponding samples are all come from same relation, is gradually decrease with the increase of K. However, the result of B^3 F1 score (right) is not very sensitive to the value and even remain perform well when increasing K to 50, despite the increasing value will introduce more noise. This is beneficial, since we do not have to fine-tune the value on new raw text.

Visualization of Relation Representations. In this experiment, to intuitively show the effectiveness of our model to learn representations in relational feature space, we visualize the representations of the instances in TACRED datasets with t-SNE [29] and randomly select 4 relations from the test set. As shown in Figure 4, we color each instance according to its ground-truth relation label and we can observe that the proposed model without Relation Semantic Clustering (left) gives general results and does not provide discriminative cluster assignments. For example, the instances with black and red colors may have similar syntactic or surface features and clustering them directly will lead to a poor result. When we use clustering module in our model, model with full module (right) can learn more discrimintaive features and each relation is mostly separate from others.

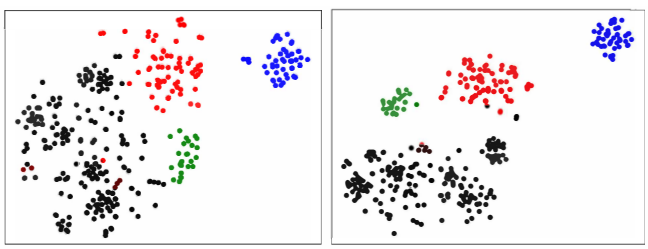

Figure 4: Visualizing contextualized entity-level features after t-SNE dimension reduction on TACRED.

Analysis on Clustering Results. Due to the number of clusters is lower than the number of true relations, different relations are likely to be clustered into same relation class. We attempt to have a detailed analysis on clustering results from TACRED to further verify whether different relation types in same cluster group have similar semantic or not. Specifically, we select the largest and smallest cluster group, the two most typical groups, which contain the most and least number of samples, to conduct detailed analysis. We find the top 5 real relations that appear most frequently in each of these two cluster groups. The 5 relation types in the former are: *"per: cities_of_residence, per: countries_of_residence, per: origin, per: stateorprovinces_of_residence, per: city_of_death"*; The 5 relation types in the latter are: *"org: country_of_headquarters, org: members, org: stateorprovince_of_headquarters, org: member_of, org: parents*. The findings of the analysis have led to the conclusion that these relation types in the same cluster have analogous relational semantics and a potential hierarchical structure. Furthermore, we count the

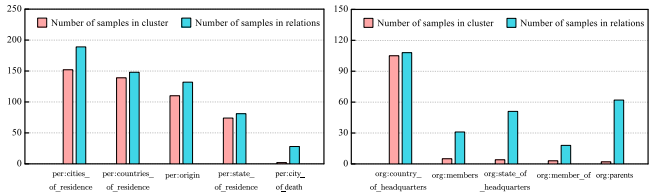

Figure 5: Statistics of samples in clusters and classes from the largest and smallest group.

exact number of samples based on their true relation types in each cluster group. As shown in Figure 5, in the largest cluster group (left), the number of samples from different relation types is nearly and these samples occupy the vast majority of their corresponding relation types, which indicate that the cluster is a fine super-class. On the contrary, the most frequent relation type dominants the smallest cluster group (right), while other types only have one or few samples in this cluster, which indicate that the cluster is very pure and have less small impurities.

4 Related Work

Self-supervised Learning Self-supervised learning enables AI systems to learn from orders of magnitude more data, which is important to recognize and understand patterns of more subtle, less common representations of the world. Specifically, self-supervised learning tries to learn an encoder that extracts generic feature representations from unlabeled datasets. Early work focuses on solving different artificially designed pretext tasks that does not require any supervision and can be easily constructed on the dataset, such as predicting neighbor words [21], generating neighbor sentences [14] for textual data, and denoising [32], colorization [15], adversarial generative models [9] for image data. Nevertheless, the feature representations are tailored to the specific pretext tasks with limited generalization.

Recent self-supervised learning algorithms mainly solve an instance discrimination task. In these algorithms, instance-wise contrastive learning with InfoNCE loss function [23] is prominent. Instance-CL treats each instance in the dataset and its augmentations as an independent pair and tries to pull together the representations within each pair while pushing apart different pairs . Consequently, different instances are well-separated in the learned feature space with local invariance being preserved for each instance. Although Instance-CL may implicitly group similar instances together, it pushes representations apart as long as they are from different original instances, regardless of their semantic similarities. Thereby, the implicit grouping effect of Instance-CL is less stable and more data-dependent, giving rise to worse representations in some cases [17, 25].

Open Relation Extraction Open relation extraction has received more attention in recent years and many efforts have been undertaken to exploring methods for it, due to the ability to extract new emerging relation types. The first line of research is Open Information Extraction , in which relation phrases are extracted directly to represent different relation types. However, using surface forms to represent relations results in an associated lack of generality since many surface forms can express the same relation. Another exploration is Relation Discovery, aims at discovering unseen relation types from open-domain text. Relation discovery can be divided into two different approaches: 1) cluster the relation representations learned from the instances, or 2) make more assumptions as learning signals to discover better relational representations.

The variational autoencoder (VAE) based approaches are under unsupervised setting. Marcheggiani and Titov [20] first propose the variational autoencoder method on unsupervised relation extraction. The model utilize the encoder extracts the semantic relation from hand-crafted features of the sentence and the decoder tries to predict one of the two entities given the relation and the other entity with a general triplet scoring function [22]. However, Simon et al. [27] point out that the aforementioned method severely suffer from the instability, and they also propose two regularizers to guide the learning procedure. But the fundamental cause of the instability is still undiscovered. On this basis, Tran et al. [28] demonstrate that by using only named entities to induce relation types can achieve better performance. Yuan et al. [37] assume that these classifications are a latent variable so they are required to follow a pre-defined prior distribution which results in unstable training and overcome this limitation by using the classifications as an intermediate variable instead of a latent variable. In clustering-based approaches, the supervised learning model [33, 38, 40] are restricted by labeled data despite achieving good performance. In unsupervised setting, Yao et al. [36] proposed Rel-LDA model ,using a generative model inspired by LDA to cluster sentences: each relation defines a distribution over a high-level handcrafted set of features describing the relationship between the two entities in the text (e.g. the dependency path). Hu et al. [12] proposed SelfORE which encodes relational feature space in a self-supervised method that bootstraps relational feature signals by leveraging adaptive clustering and classification iteratively.

5 Conclusion

In this work, we investigate the deficiencies of the contrastive learning on unsupervised relation extraction and propose a similarity-based representation learning method, which can learn well semantic of instances to effectively improve the performance with unsupervised clustering. In addition, we further obtain discriminative feature representations through relational semantic clustering. Owing to the fact that our model is straightforward and efficient, we believe that our approach easily admits extensions to different open-domain texts. However, our model still has many shortcomings. Similar to other unsupervised relation extraction methods, our model is unable to handle instances where entity pairs appearing in a sentence do not exhibit any relation. We leave these problems as future work and look forward to seeking possible solutions from a broader perspective.

Acknowledgements

We would like to thank the referees for their comments, which helped improve this paper considerably.

References

[1] Rabah A Al-Zaidy and C Lee Giles, 'Extracting semantic relations for scholarly knowledge base construction', in *2018 IEEE 12th international conference on semantic computing (ICSC)*, pp. 56–63. IEEE, (2018).

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton, 'Layer normalization', *arXiv preprint arXiv:1607.06450*, (2016).

[3] Amit Bagga and Breck Baldwin, 'Entity-based cross-document coreferencing using the vector space model', in *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*, (1998).

[4] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor, 'Freebase: a collaboratively created graph database for structuring human knowledge', in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pp. 1247–1250, (2008).

[5] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt, 'Discovering latent knowledge in language models without supervision', *arXiv preprint arXiv:2212.03827*, (2022).

[6] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze, 'Deep clustering for unsupervised learning of visual features', in *Proceedings of the European conference on computer vision (ECCV)*, pp. 132–149, (2018).

[7] Xinlei Chen and Kaiming He, 'Exploring simple siamese representation learning', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15750–15758, (June 2021).

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, '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)*, pp. 4171–4186, Minneapolis, Minnesota, (June 2019). Association for Computational Linguistics.

[9] Jeff Donahue and Karen Simonyan, 'Large scale adversarial representation learning', in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 10542–10552, (2019).

[10] Hady Elsahar, Elena Demidova, Simon Gottschalk, Christophe Gravier, and Frederique Laforest, 'Unsupervised open relation extraction', in *The Semantic Web: ESWC 2017 Satellite Events: ESWC 2017 Satellite Events, Portorož, Slovenia, May 28–June 1, 2017, Revised Selected Papers 14*, pp. 12–16. Springer, (2017).

[11] Tianyu Gao, Xingcheng Yao, and Danqi Chen, 'SimCSE: Simple contrastive learning of sentence embeddings', in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6894–6910, Online and Punta Cana, Dominican Republic, (November 2021). Association for Computational Linguistics.

[12] Xuming Hu, Lijie Wen, Yusong Xu, Chenwei Zhang, and S Yu Philip, 'Selfore: Self-supervised relational feature learning for open relation extraction', in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3673–3682, (2020).

[13] Lawrence Hubert and Phipps Arabie, 'Comparing partitions', *Journal of classification*, **2**, 193–218, (1985).

[14] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler, 'Skip-thought vectors', in *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*, pp. 3294–3302, (2015).

[15] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich, 'Learning representations for automatic colorization', in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pp. 577–593. Springer, (2016).

[16] Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li, 'On the sentence embeddings from pre-trained language models', in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9119–9130, (2020).

[17] Junnan Li, Pan Zhou, Caiming Xiong, and Steven CH Hoi, 'Prototypical contrastive learning of unsupervised representations', *arXiv preprint arXiv:2005.04966*, (2020).

[18] Fangchao Liu, Lingyong Yan, Hongyu Lin, Xianpei Han, and Le Sun, 'Element intervention for open 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)*, pp. 4683–4693, Online, (August 2021). Association for Computational Linguistics.

[19] Shuliang Liu, Xuming Hu, Chenwei Zhang, Shu'ang Li, Lijie Wen, and Philip Yu, 'HiURE: Hierarchical exemplar contrastive learning for unsupervised relation extraction', in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5970–5980, Seattle, United States, (July 2022). Association for Computational Linguistics.

[20] Diego Marcheggiani and Ivan Titov, 'Discrete-state variational autoencoders for joint discovery and factorization of relations', *Transactions of the Association for Computational Linguistics*, **4**, 231–244, (2016).

[21] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean, 'Distributed representations of words and phrases and their compositionality', in *Proceedings of the 26th International Conference on Neural Information Processing Systems-Volume 2*, pp. 3111–3119, (2013).

[22] Maximilian Nickel, Volker Tresp, and Hans-Peter Kriegel, 'A three-way model for collective learning on multi-relational data', in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pp. 809–816, (2011).

[23] Aaron van den Oord, Yazhe Li, and Oriol Vinyals, 'Representation learning with contrastive predictive coding', *arXiv preprint arXiv:1807.03748*, (2018).
[24] Ashokkumar Palanivinayagam and Sureshkumar Nagarajan, 'An optimized iterative clustering framework for recognizing speech', *International Journal of Speech Technology*, **23**, 767–777, (2020).
[25] Senthil Purushwalkam and Abhinav Gupta, 'Demystifying contrastive self-supervised learning: Invariances, augmentations and dataset biases', *Advances in Neural Information Processing Systems*, **33**, 3407–3418, (2020).
[26] Andrew Rosenberg and Julia Hirschberg, 'V-measure: A conditional entropy-based external cluster evaluation measure', in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 410–420, Prague, Czech Republic, (June 2007). Association for Computational Linguistics.
[27] Étienne Simon, Vincent Guigue, and Benjamin Piwowarski, 'Unsupervised information extraction: Regularizing discriminative approaches with relation distribution losses', in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1378–1387, (2019).
[28] Thy Thy Tran, Phong Le, and Sophia Ananiadou, 'Revisiting unsupervised relation extraction', in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7498–7505, Online, (July 2020). Association for Computational Linguistics.
[29] Laurens van der Maaten and Geoffrey Hinton, 'Visualizing data using t-sne', *Journal of Machine Learning Research*, **9**, 2579–2605, (2008).
[30] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool, 'Scan: Learning to classify images without labels', in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X*, pp. 268–285. Springer, (2020).
[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin, 'Attention is all you need', in *Advances in Neural Information Processing Systems*, eds., I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, volume 30. Curran Associates, Inc., (2017).
[32] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol, 'Extracting and composing robust features with denoising autoencoders', in *Proceedings of the 25th international conference on Machine learning*, pp. 1096–1103, (2008).
[33] Ruidong Wu, Yuan Yao, Xu Han, Ruobing Xie, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun, 'Open relation extraction: Relational knowledge transfer from supervised data to unsupervised data', 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)*, pp. 219–228, Hong Kong, China, (November 2019). Association for Computational Linguistics.
[34] Junyuan Xie, Ross Girshick, and Ali Farhadi, 'Unsupervised deep embedding for clustering analysis', in *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pp. 478–487, (2016).
[35] Chenyan Xiong, Russell Power, and Jamie Callan, 'Explicit semantic ranking for academic search via knowledge graph embedding', in *Proceedings of the 26th international conference on world wide web*, pp. 1271–1279, (2017).
[36] Limin Yao, Aria Haghighi, Sebastian Riedel, and Andrew McCallum, 'Structured relation discovery using generative models', in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 1456–1466, (2011).
[37] Chenhan Yuan and Hoda Eldardiry, 'Unsupervised relation extraction: A variational autoencoder approach', in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1929–1938, Online and Punta Cana, Dominican Republic, (November 2021). Association for Computational Linguistics.
[38] Kai Zhang, Yuan Yao, Ruobing Xie, Xu Han, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun, 'Open hierarchical relation extraction', in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5682–5693, Online, (June 2021). Association for Computational Linguistics.
[39] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning, 'Position-aware attention and supervised data improve slot filling', in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 35–45, Copenhagen, Denmark, (September 2017). Association for Computational Linguistics.
[40] Jun Zhao, Tao Gui, Qi Zhang, and Yaqian Zhou, 'A relation-oriented clustering method for open relation extraction', in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9707–9718, (2021).