U-CORE: A Unified Deep Cluster-wise Contrastive Framework for Open Relation Extraction

Jie Zhou[◇] **Shenpo Dong**[◇] **Yunxin Huang**[†] **Meihan Wu**[◇]
Haili Li[◇] **Jingnan Wang**[◇] **Hongkui Tu**[◇*] **Xiaodong Wang**[◇]

[◇]College of Computer, National University of Defense Technology, China
[†]Technical Service Center for Professional Education,
National University of Defense Technology, China
{jiezhou, dsp, huangyunxin17, meihanwu20}@nudt.edu.cn
{lihaili20, wangjingnan17a, tuhongkui, xdwang}@nudt.edu.cn

Abstract

Within Open Relation Extraction (ORE) tasks, the Zero-shot ORE method is to generalize undefined relations from predefined relations, while the Unsupervised ORE method is to extract undefined relations without the need for annotations. However, despite the possibility of overlap between predefined and undefined relations in the training data, a unified framework for both Zero-shot and Unsupervised ORE has yet to be established. To address this gap, we propose **U-CORE**: A **U**nified Deep **C**luster-wise Contrastive Framework for both Zero-shot and Unsupervised **ORE**, by leveraging techniques from Contrastive Learning (CL) and Clustering.[1] U-CORE overcomes the limitations of CL-based Zero-shot ORE methods by employing Cluster-wise CL that preserves both local smoothness as well as global semantics. Additionally, we employ a deep-cluster-based updater that optimizes the cluster center, thus enhancing the accuracy and efficiency of the model. To increase the stability of the model, we adopt Adaptive Self-paced Learning that effectively addresses the data-shifting problems. Experimental results on three well-known datasets demonstrate that U-CORE significantly improves upon existing methods by showing an average improvement of 7.35% ARI on Zero-shot ORE tasks and 15.24% ARI on Unsupervised ORE tasks.

1 Introduction

Relation Extraction (RE) is a fundamental task in Natural Language Processing (NLP) that aims to extract the relationships between pairs of entities mentioned in a given text, such as identifying the Effect-Cause relation between ‘‘fire’’ and ‘‘fuel’’ in the sentence ‘‘The fire was caused by exploding fuel.’’ RE is an essential component of NLP systems that can facilitate diverse downstream tasks, including Question Answering (Soares and Parreiras, 2020), Knowledge Graphs (Ji et al., 2021), and Dialogue Systems (Chen et al., 2017).

While supervised methods have demonstrated great success in extracting predefined relations, in reality, new relations frequently arise, and it can be time-consuming and labor-intensive to define them manually. As a result, open-domain relation extraction has become a popular research topic. Based on prior studies, Open Relation Extraction (ORE) tasks can be categorized into two types, namely, Zero-shot Open Relation Extraction (ZORE) and Unsupervised Open Relation Extraction (UORE). ZORE aims to extract novel relational facts where the target relation types are not observed in the training set (Levy et al., 2017). On the other hand, UORE has the objective of extracting undefined relations without any annotation or prior knowledge (Elsahar et al., 2017).

In recent years, significant attention has been devoted to ORE tasks (Obamuyide and Vlachos, 2018; Hu et al., 2020; Chen and Li, 2021). Despite variations in complexity and annotations, both ZORE and UORE aim to develop an optimal encoder that can generate an appropriate relational representation using limited resources. Additionally, practical scenarios involving ZORE and UORE may overlap, such as when training data contain both predefined and undefined relations. Consequently, the previous ZORE and UORE methods, which concentrate on predefined or undefined relations, respectively, may yield

*Hongkui Tu is the corresponding author.

[1]The code can be found at: https://github.com/2kjiejie/U-CORE.

Transactions of the Association for Computational Linguistics, vol. 11, pp. 1301–1315, 2023. https://doi.org/10.1162/tacl_a_00604
Action Editor: Kristina Toutanova. Submission batch: 4/2023; Revision batch: 6/2023; Published 11/2023.

suboptimal results due to their inability to account for all relations. Given the similarities between ORE techniques, as well as their potential overlap in practical applications, we propose a unified framework that addresses diverse ORE tasks.

Recent studies have attempted to improve relation representation in ZORE by leveraging Contrastive Learning (CL) approaches (Chen and Li, 2021; Wang et al., 2022). These methods typically rely on instance-wise CL, which aims to bring together relations from the same instances while separating those from different instances. However, Li et al. (2021) have pointed out that instance-wise CL may treat instances with similar semantic information as negative pairs, leading to a drift in their representations and resulting in performance degradation. In the ORE task, instances that belong to the same relation type share ''similar semantic information'' and should not be treated as negative paris. To address this limitation, we propose employing a cluster-wise contrastive learning approach, which facilitates the alignment of relations within the same clusters and the separation of relations across different clusters.

Contrarily, in UORE tasks, the training process is unsupervised, and existing models tend to learn structural information from clustering techniques (Hu et al., 2020; Liu et al., 2021). However, the majority of existing UORE methods depend on conventional clustering algorithms, such as k-means, for defining clustering centers. Re-clustering at the end of each epoch is typically required, which may be time-consuming and computationally intensive. Consequently, we incorporate deep clustering into our ORE framework to eliminate the need for frequent re-clustering and enhance the clustering performance.

The combination of cluster-wise contrastive learning and deep clustering plays a crucial role in our unified ORE framework. Deep clustering enhances the performance of cluster-wise CL by providing more accurate clusters, while cluster-wise CL improves deep clustering by generating better relation representations. However, during the training process, while the encoder is updated in each minibatch, the clustering assignment is only updated at the end of each epoch. This inconsistency between the relation representations and cluster centers in the feature space is known as the data-shifting problem, which has been identified in previous works (Liu et al., 2022). This problem requires careful attention in our unified framework to avoid the misalignment between the relation representations and the cluster centers.

Based on the above analysis, we propose **U-CORE**: A **U**nified Deep **C**luster-wise Contrastive Framework for **O**pen **R**elation **E**xtraction in this article. The proposed framework aims to establish a unified approach to enhance the performance of both ZORE and UORE methods. Specifically, the Cluster-wise Contrastive Learning approach is employed in our ORE framework, which increases the inter-cluster spacing of clusters while minimizing intra-cluster spacing, enabling us to overcome the limitations of previous CL-based ZORE methods. To mitigate the need for regular re-clustering and enhance overall accuracy and efficiency, we introduce a deep Cluster Center Updater. Moreover, we propose the integration of Adaptive Self-paced Learning in the proposed U-CORE to address issues regarding data-shifting and produce a more stable model. The framework is capable of obtaining an effective representation of relations and able to handle both Zero-shot ORE and Unsupervised ORE tasks. Overall, our proposed U-CORE framework contributes towards the development of highly effective and versatile techniques for Open Relation Extraction. The architecture of our framework is shown in Figure 1.

We briefly summarize our contribution as follows:

- We propose U-CORE, a novel deep cluster-wise contrastive framework that effectively addresses both zero-shot open relation extraction and unsupervised open relation extraction tasks.
- We introduce the Cluster-wise Contrastive Module for the ORE task, which combines instance-wise and cluster-wise Contrastive Learning to optimize relation representations both locally and globally.
- We introduce a deep-cluster-based Cluster Center Updater and Adaptive Self-paced Learning techniques, which enhance the efficiency and stability of our model.
- We conduct experiments on 3 well-known datasets. The results demonstrate that U-CORE outperforms existing state-of-the-art methods in both ZORE and UORE tasks.

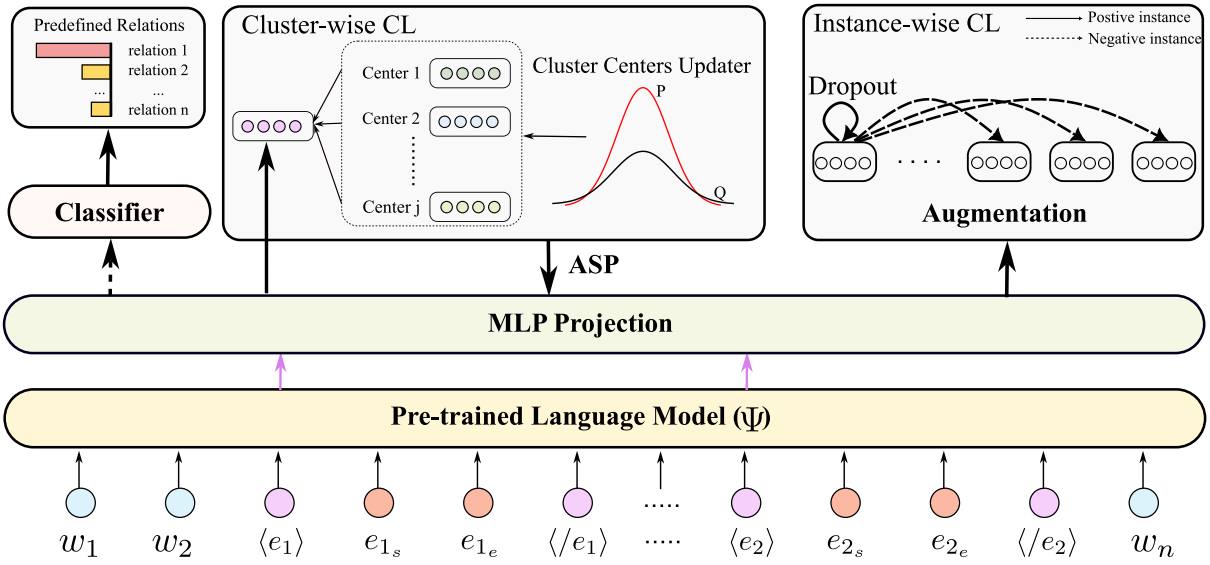


Figure 1: Framework of **U-CORE**. The Entity Markers representations generated by the PLM encoder Ψ and MLP projection will proceed through three parallel modules, namely, the Cluster-wise CL, Instance-wise CL, and Relation Prediction module. To further improve the efficiency and stability of U-CORE, the Cluster Center Updater and Adaptive Self-paced Learning (ASP) are incorporated.

2 Related Work

In this section, we survey the related work on Open Relation Extraction, Contrastive Learning, and Deep Clustering.

Open Relation Extraction In recent years, ORE has emerged as a significant research topic due to its practical applicability and downstream task potential. ORE tasks can be broadly classified as Zero-shot and Unsupervised open relation extraction. The former aims to distinguish novel relations without relying on prior training instances. Some researchers, such as Levy et al. (2017) and Obamuyide and Vlachos (2018), have drawn a parallel between this goal and reading comprehension or question answering. Zhao et al. (2021) have proposed a relation-oriented clustering method to solve the ZORE problem. In contrast, UORE is an unsupervised learning method that identifies semantic relation features from unannotated data. Several authors have pursued this strategy: Elsahar et al. (2017) have used re-weighted word embeddings for clustering free text, while Hu et al. (2020) have employed clustering-based techniques to generate pseudo-labels for new relation discovery. Both ZORE and UORE require robust representations of the relations. The key difference is that while ZORE is focused on extracting undefined relations from pre-existing ones, UORE optimizes representations from undefined relations themselves. Our proposed model U-CORE leverages supervised and self-supervised learning to optimize representation, making it well-suited for both ZORE and UORE tasks.

Contrastive Learning Contrastive Learning (CL) is a powerful strategy that extracts common attributes for each data class by contrasting samples while simultaneously identifying distinguishing characteristics. The effectiveness of Contrastive Learning was first widely recognized in the field of computer vision (He et al., 2020; Chen et al., 2020). This strategy has also seen successful applications in NLP. For instance, SimCSE (Gao et al., 2021) has been developed for NLP tasks. Recently, some researchers have attempted to apply CL for ZORE tasks. For example, Wang et al. (2022) employed instance-wise CL and a relational classification module in the ZORE task. These CL methods used in ZORE rely on instance-wise Contrastive Learning designed to emphasize the relationships among similar instances while distinguishing them from different instances. However, Li et al. (2021) has highlighted a critical issue with instance-wise CL, as it can often regard entity pairs with similar semantic information as negative pairs, resulting in a local smoothness but ignoring the global

semantics. Accordingly, in our ORE framework, U-CORE implements the cluster-wise contrastive learning approach, which aligns relations in the same clusters and separates relations in different ones. By doing so, we can avoid treating relations with similar semantics as negative pairs, thereby achieving improved performance.

Deep Clustering Deep clustering (Ma et al., 2019; Guo et al., 2019) has demonstrated significant improvements over conventional algorithms in recent years. Subakti et al. (2022) has provided evidence that DEC (Xie et al., 2016) and IDEC (Guo et al., 2017) perform better than k-means in clustering sentence embeddings. In the past two years, several studies (Li et al., 2021; Caron et al., 2020; Liu et al., 2022) have attempted to integrate clustering into contrastive learning optimization. However, these works have primarily used conventional clustering algorithms such as k-means. These algorithms require re-clustering at the end of each epoch, which can significantly drain computational resources, particularly when processing massive datasets. Therefore, we have integrated deep clustering into our models. This approach results in two key benefits: 1) clustering centers can be refined in every epoch without excessive time or memory usage, and 2) the deep clustering approach has further improved the performance of cluster-wise contrastive learning.

3 Proposed Model

3.1 Relation Representation

Relation Representation is dedicated to generating features that represent each token in the input sentence. In this study, we presume that the entities included in the sentence have been recognized before inputting them. Based on the previous finding from Baldini Soares et al. (2019), we choose Entity Marker for relation representation. To include entities e_1 and e_2 in an input sentence $S = [w_1, \ldots, w_n]$, we augment S with four reserved word pieces to mark the start and end of each entity mention in the relation statement. We introduce $\langle e_1 \rangle$, $\langle /e_1 \rangle$, $\langle e_2 \rangle$, $\langle /e_2 \rangle$ and modify S to

$$\begin{aligned}\hat{S} = [w_1, \ldots, \langle e_1 \rangle, e_{1_s} \ldots, e_{1_e}, \langle /e_1 \rangle, \ldots, \\ \langle e_2 \rangle, e_{2_s}, \ldots, e_{2_e}, \langle /e_2 \rangle, \ldots, w_n]\end{aligned} \quad (1)$$

where $\hat{S}$ will be the input sentence for the encoder.

Then we choose BERT (Devlin et al., 2019) as the encoder to generate the sentence embedding $H \in \mathbb{R}^{n \times d}$:

$$\begin{aligned}H = [h_1, \ldots, h_{\langle e_1 \rangle}, h_i, \ldots, h_j, h_{\langle /e_1 \rangle}, \ldots, \\ h_{\langle e_2 \rangle}, h_k, \ldots, h_l, h_{\langle /e_2 \rangle}, \ldots, h_n]\end{aligned} \quad (2)$$

where d is the hidden dimension of BERT.

We utilize the concatenated results of the reserved word pieces to represent the relation r between $e1$ and $e2$:

$$r = MLP(h_{\langle e_1 \rangle} \oplus h_{\langle e_2 \rangle}) \quad (3)$$

where $\oplus$ is the concatenation operator and $r \in \mathbb{R}^{2 \times d}$. MLP is a a non-linearity projection from Chen et al. (2020) to generate enhanced representation, defined as $MLP(\cdot) = ReLU(W(\cdot))$.

3.2 Instance-wise Contrastive Module

3.2.1 Data Augmentation

In order to implement instance-wise contrastive learning, it is essential to utilize an appropriate augmentation method to generate positive pairs. Following the approach employed in prior research such as SimCSE (Gao et al., 2021), we employ Dropout noise as our data augmentation method. Specifically, the generated positive pairs consist of identical sentences with embeddings that differ only in terms of their dropout masks. For a relation r, its positive pair $\hat{r} = Dropout(r)$. Therefore, for a randomly sampled minibatch $B = \{r_i\}_{i=1}^{M}$, we use Dropout noise to generate a pair of augmentations for each relation instance in B. This results in an augmented batch $\hat{B}$ with double the size, represented as $\hat{B} = \{\hat{r_i}\}_{i=1}^{2M}$.

3.2.2 Instance-wise Contrastive Loss

Instance-wise Contrastive Loss aims to bring closer relations from the same instance and separate relations from different instances. For every minibatch B, the Instance-CL loss is determined by the augmented pairs in $\hat{B}$. A positive pair is represented by $r_i, \hat{r_i} \in \hat{B}$, while the remaining $(2M - 2)$ relations in $\hat{B}$ are considered as negative instances in relation to this positive pair. For the relation r_i, we aim to distinguish $\hat{r_i}$ from all negative instances in $\hat{B}$ by minimizing

the InfoNCE Loss (Oord et al., 2018) presented below:

$$\mathcal{L}_{\text{Inst}} = \sum_{i=1}^{M} -\log \frac{\exp(\text{sim}(r_i, \hat{r_i})/\tau)}{\sum_{j=1}^{2M} \mathbb{1}_{j \neq i} \exp(\text{sim}(r_i, \hat{r_j})/\tau)} \tag{4}$$

where $\mathbb{1}_{j \neq i}$ is an indicator function evaluating to 1 iff $j \neq i$ and τ is a temperature hyperparameter. $sim(r_i, \hat{r_i})$ indicates the cosine similarity between $r_i, \hat{r_i}$, which is ${r_i}^\top \hat{r_i} / \|r_i\| \|\hat{r_i}\|$.

3.3 Cluster-wise Contrastive Module

As described in Section 3.2, instance-wise contrastive learning results in an embedding space where each instance is distinctly separated and exhibits local smoothness. However, instance-wise contrastive learning treats two samples as a negative pair as long as they are from different instances, irrespective of whether they belong to the same relation type, causing alienation of the instances from the same type in the embedding space. To tackle this challenge, we integrate Cluster-wise Contrastive Learning.

3.3.1 Cluster Centers Initialization

At the beginning of training, we conduct k-means clustering on all representation vectors R, using varying numbers of clusters K, and assess the resultant clustering outcomes via the Davies–Bouldin index (DBI).

$$DBI = \frac{1}{K} \sum_{j=1}^{K} \max_{k \neq j} \left\{ \frac{\overline{D_j} + \overline{D_k}}{d(c_j, c_k)} \right\} \tag{5}$$

where c_j is the center of cluster j, $\overline{D_j}$ is the average distance from the data points in cluster i to its center, $d(c_j, c_k)$ is the distance between the centers of clusters j and k, and the maximum is taken over all pairs of clusters. The number of clusters with the lowest DBI value will be selected for utilization in our model.

3.3.2 Cluster-wise Contrastive Loss

Following Li et al. (2021), we take the same approach as NCE and use all negative cluster centers to calculate the normalization term:

$$\ell^i_{\text{Clus}} = -\log \frac{\exp(\text{sim}(r_i, c_j)/\phi_j)}{\sum_{k=1}^{K} \exp(\text{sim}(r_i, c_k)/\phi_k)} \tag{6}$$

where r_i belongs to cluster j. ϕ_j is the normalized density for cluster j, defined as:

$$\phi_j = \frac{\sum_i^J \|r_i - c_j\|^2}{J \log(J + a)}$$

where J is the number of relations in cluster j and a is a smooth parameter. The difference between τ in $\mathcal{L}_{Inst}$ and ϕ_j is that the value of ϕ_j varies between different clusters. A smaller value of ϕ_j indicates a better concentration of relations within the cluster.

3.3.3 Cluster Centers Updater

Existing UORE methods that rely on clustering techniques frequently require the execution of conventional clustering algorithms at the end of each epoch, resulting in significant time investment. Furthermore, studies such as Subakti et al. (2022) demonstrate that deep clustering outperforms conventional clustering methods when analyzing high-dimensional data. Given the crucial role played by clustering quality in cluster-wise contrastive learning, we propose incorporating a deep-cluster-based updater for the cluster center. Following Xie et al. (2016) we use the Student's t-distribution as a kernel to measure the similarity between embedded relation r_i and center c_j :

$$q_{ij} = \frac{\left(1 + \|r_i - c_j\|^2 / \alpha\right)^{-\frac{\alpha+1}{2}}}{\sum_{j'} \left(1 + \|r_i - c_{j'}\|^2 / \alpha\right)^{-\frac{\alpha+1}{2}}} \tag{7}$$

where α are the degrees of freedom of the Student's t-distribution. q_{ij} can be interpreted as the probability of assigning relation r_i to cluster j (i.e., a soft assignment).

We proceed to construct a new distribution p_{ij}:

$$p_{ij} = \frac{q_{ij}^2 / f_j}{\sum_{j'} q_{ij'}^2 / f_{j'}} \tag{8}$$

where $f_j = \sum_j q_{ij}$ are soft cluster frequencies. Compared to q_{ij}, distribution p_{ij} is capable of placing greater emphasis on data points that have been assigned high confidence while ensuring that the feature space remains free from distortion caused by large clusters, owing to the normalization through cluster frequencies f_j.

Therefore, we consider P as the target distribution for Q and set our objective as the KL divergence loss between the soft assignments Q and the auxiliary distribution P, as outlined below:

$$\mathcal{L}_{\mathrm{KL}} = \mathrm{KL}\left[P \| Q\right] = \sum_{i} \sum_{j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \quad (9)$$

The clustering center will be optimized with higher confidence and reduce the impact of over-large clusters, thus improving the clustering accuracy and the Cluster-wise Contrastive Module's performance. Furthermore, the optimization process runs concurrently with the training process, obviating the necessity for a separate re-clustering procedure and yielding substantial reductions in both time and computational resources.

3.4 Adaptive Self-Paced Learning

During the training process, the relational feature r is updated in each batch, but the cluster assignment remains unchanged until the end of the epoch, leading to possible representation shift issues. This means that the representation of r_i and its corresponding cluster center c_j may not belong to the same feature space. To tackle this problem, previous research (Li et al., 2021; Liu et al., 2022) has employed a Momentum Encoder to ensure a consistent relational feature space. Nevertheless, with open relation extraction datasets that have fewer data and clusters, the use of Momentum Encoder may hinder convergence speed. Consequently, finding the right hyper-parameter θ for the moving average that balances both representation shift and convergence speed poses a significant challenge. Therefore, following Guo et al. (2019), we propose an Adaptive Self-Paced Learning to the cluster-wise contrastive loss as:

$$\mathcal{L}_{\text{Clus}} = \sum_{i=1}^{M} v_i \ell^i_{\text{Clus}} - \lambda v_i \quad (10)$$

where

$$v_i = \begin{cases} 0, & \text{if } \ell^i_{\text{Clus}} < \lambda \\ 1, & \text{otherwise} \end{cases}$$

To have an adaptive λ, we define it as:

$$\lambda = \mu(\mathcal{L}^t_{\text{Clus}}) + \frac{t}{T}\sigma(\mathcal{L}^t_{\text{Clus}}) \quad (11)$$

where $\mathcal{L}^t_{Clus}$ denotes the $\mathcal{L}_{Clus}$ in t-th epochs, T is the number of max epochs. $\mu(\cdot)$ and $\sigma(\cdot)$ are the means and standard deviation. λ is adaptive to the losses of the cluster-wise contrastive module, not an independent hyper-parameter like regular self-paced learning. Then the loss of all the unsupervised modules above can be formulated as:

$$\mathcal{L}_{\text{unsup}} = \mathcal{L}_{\text{Inst}} + \mathcal{L}_{\text{Clus}} + \eta \mathcal{L}_{\text{KL}} \quad (12)$$

where η is used to balance the KL loss and Contrastive loss.

3.5 Labeled Relation Prediction

All the modules we proposed above can be executed without the need for labeled relations. Nonetheless, incorporating those predefined data can enhance the encoder's performance in terms of generalization. Following this, we adopt a Feed-forward layer as the classifier to facilitate the prediction process. For a given input sentence S_i, the final output for the labeled entity pair (e1, e2) can be expressed as:

$$\hat{y}_i = \text{Softmax}(W(\text{ReLU}(r_i) + b) \quad (13)$$

We apply cross-entropy to compute the classification loss for the labeled data:

$$\mathcal{L}_{\text{label}} = -\frac{1}{N} \sum_{i=1}^{N} y^*_i \log(\hat{y}_i) \quad (14)$$

where N is the number of labeled relations and y^*_i is the ground truth.

If there are labeled data in the training set, then the Labeled Relation Prediction loss will be added to the final loss:

$$\mathcal{L} = \mathcal{L}_{\text{label}} + \gamma * \mathcal{L}_{\text{unsup}} \quad (15)$$

where γ is a hyper-parameter to balance two objectives.

4 Experiment

4.1 Datasets

Following previous studies, we adopt two datasets to evaluate zero-shot ORE: SemEval2010 Task8

and FewRel. **SemEval2010 Task8** (Hendrickx et al., 2019) was designed to classify a set of semantic relations between pairs of concepts, such as cause-effect or instrument-agency. It contains 9 relations and an ''Other'' relation. Each relation possesses a distinct direction (e.g., ''son of'' and ''father of''). Following previous works (Wang et al., 2022), we do not consider the direction of the 9 relations or use the ''Other'' relation in experiments. We combine the instances of the training set and testing set for each relation to obtain the overall instances. This collection consists of 10,717 instances, with different numbers of instances allocated to each relation. **FewRel** (Han et al., 2018), a publicly available dataset that utilizes data from Wikipedia, is specifically designed to evaluate the model's performance in carrying out few-shot relation extraction tasks. Unlike SemEval2010 Task8, FewRel is a balanced dataset comprising 80 relations, with 700 instances for each relation. Although FewRel is primarily utilized for a few-shot learning approach, it can also be effective for zero-shot learning if the relation labels between the training and testing data are distinct.

We also carry out our unsupervised open relation extraction experiments on **TACRED** (Zhang et al., 2017), which is one of the largest and most widely used datasets for relation classification. TACRED is a comprehensive supervised relation extraction dataset that focuses on Text Analysis Conference's Knowledge Base Population (TAC KBP) relations. The dataset contains an extensive collection of 21,773 positive examples sourced through crowdsourcing, encompassing a wide range of relationships.

4.2 Evaluation Settings

Zero-shot ORE Settings Following Wang et al. (2022), we randomly select m relations as the undefined relation set R_{test}, and n relations as the predefined relation set R_{train}. Note that $(m + n)$ equals to the whole numbers of relations in the dataset and $R_{\text{train}} \cap R_{\text{test}} = \varnothing$. The training data only contains the instances of predefined relations while the testing data only contains undefined relations. We repeat experiments 10 times on SemEval2010 Task8 and FewRel, then report the average clustering results on k-means. To show an appropriate clustering result, the clustering number is set to m.

Unsupervised ORE Settings The TACRED[2] dataset has been officially split into the training, validation, and testing sets. Following Tran et al. (2020) and Liu et al. (2022), we train the unsupervised models on the training set and report the clustering results of the testing set. For our U-CORE model, we suppose the training and testing set have the same relation types, so we directly use the clustering centers generated in the training process to assign cluster labels.

Evaluation Metrics To evaluate the effectiveness of clustering, we choose three commonly used metrics as our evaluation criteria, namely, B^3 (Bagga and Baldwin, 1998), Normalized Mutual Information (NMI), and Adjusted Rand Index (ARI) (Hubert and Arabie, 1985). B^3 utilizes both precision and recall to accurately assess the rate of correctly assigning data points to their respective cluster or clustering all points into a singular class. Then B^3 F1 score is computed as the harmonic mean of the B^3 precision and recall:

$$\text{B}^3 \text{ prec} = \mathop{\mathbb{E}}_{X,Y} P(g(X) = g(Y) \mid c(X) = c(Y))$$
$$\text{B}^3 \text{ recall} = \mathop{\mathbb{E}}_{X,Y} P(c(X) = c(Y) \mid g(X) = g(Y))$$

where $g(X)$ and $g(Y)$ are the predicted labels of two data points X and Y, respectively. $c(X)$ and $c(Y)$ are the true labels of X and Y, respectively. The NMI score quantifies the amount of information shared between the predicted label and the ground truth. A perfect partition of data results in an NMI score of 1, while an independent prediction and ground truth yield a score of 0.

$$\text{NMI}(Y^*, Y) = \frac{2I(Y^*, Y)}{H(Y^*) + H(Y)}$$

where Y^* and Y are predicted labels and the ground truth, respectively. $I(Y^*, Y)$ is the mutual information between Y^* and Y, and $H(Y^*)$ and $H(Y)$ are the entropies of Y^* and Y, respectively.

The ARI metric gauges the level of conformity between the cluster and golden distribution, ranging from -1 to 1. A high score indicates

[2] https://nlp.stanford.edu/projects/tacred/.

	FewRel									SemEval		
	m = 5			m = 10			m = 15			m = 4		
Model	F1	NMI	ARI	F1	NMI	ARI	F1	NMI	ARI	F1	NMI	ARI
CNN	74.47	68.51	66.31	60.87	64.59	53.79	55.3	62.35	49.87	38.42	17.06	15.43
Att-BiLSTM	82.75	79.36	76.63	75.89	79.10	71.46	69.84	75.94	66.03	41.6	21.45	19.97
Supervised RSN	73.33	67.89	64.49	59.11	64.96	48.66	50.99	59.98	39.74	38.41	11.98	10.96
ZS-BERT	74.51	69.24	66.96	70.63	74.10	65.23	63.33	70.7	59.24	35.03	12.47	9.53
MTB	88.06	85.32	84.03	82.7	84.16	79.19	76.72	77.66	71.65	44.35	25.25	20.59
RCL	89.69	87.12	85.69	85.61	86.59	80.36	81.48	85.64	78.18	68.02	55.91	54.71
U-CORE	**96.38**	**95.04**	**95.33**	**90.37**	**90.08**	**82.45**	**83.35**	**89.03**	**79.55**	**78.83**	**66.79**	**70.88**

Table 1: Experiment results(%) on FewRel and SemEval in terms of B^3 F1, NMI, and ARI. m denotes the number of undefined relation types. The best results among these models are represented in bold.

greater consistency between the two distributions. The formula for ARI is as follows:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}}$$

where n is the total number of samples, a_i is the number of samples in cluster C_i, b_j is the number of samples in golden relation r_j, and n_{ij} is the number of samples assigned to both cluster C_i and relation r_j.

Baselines To conduct the Zero-shot experiment, we conduct a comparative analysis of U-CORE against two distinct sets of models. The first set comprises supervised relation extraction models, including *CNN* (Zeng et al., 2014), *Attention-BiLSTM* (Zhou et al., 2016), and *MTB* (Baldini Soares et al., 2019). These models have demonstrated remarkable efficacy in supervised learning settings; however, their effectiveness in the zero-shot environment remains untested. The second set includes three zero-shot relation extraction models, namely, *Supervised RSN* (Wu et al., 2019), *ZS-BERT* (Chen and Li, 2021), and *RCL* (Wang et al., 2022). Following the previous setting (Wang et al., 2022), we have modified the supervised relational extraction models to fit the zero-shot experiments. These models' outputs will be replaced with vectors that have the same dimensionality as the U-CORE. Subsequently, we utilize the k-means algorithm to predict undefined relations in our sample data.

For the Unsupervised Clustering experiment, we choose five representative models. 1) *RAE* (Marcheggiani and Titov, 2016) proposes a reconstruction-based method for ORE by reconstructing entities from pairing entities and predicted relations. 2) *RW-HAC* (Elsahar et al., 2017) involves re-weighting word vectors based on the sentence's dependency parse tree. 3) *EType+* (Tran et al., 2020) incorporates entity type knowledge into the relation extraction task. 4) *SelfORE* (Hu et al., 2020) leverages a pre-trained language model to detect weak self-supervised signals and group contextualized relational features into clusters. 5) *HiURE* (Liu et al., 2022) introduces a contrastive learning framework that utilizes cross-hierarchy attention to derive hierarchical signals from relational feature space. Note that in the studies of EType+ (Tran et al., 2020) and HiURE (Liu et al., 2022), their models were trained on the NYT-FB dataset (Marcheggiani and Titov, 2016) and tested on TACRED. However, we fail to obtain the NYT-FB dataset as it is private. Thus, we train and test EType+ and HiURE on TACRED. In order to ensure a fair comparison, the number of clusters for each baseline model has been set to 16, following previous work (Tran et al., 2020).

4.3 Results

Results on Zero-shot Open Relation Tasks Table 1 displays the results of our experiments on ZORE tasks. Our proposed method U-CORE outperforms other state-of-the-art models on FewRel and SemEval datasets. U-CORE effectively learns the relation representations from both predefined relations and global semantics. A decrease in performance is observed with an increase in undefined relation set R_{test} for all models. Moreover, our evaluation shows that SemEval is a

TACRED					
Model	F1	P	R	NMI	ARI
RAE	40.82	34.70	49.55	33.51	26.42
RW-HAC	50.94	42.61	63.33	51.67	28.15
EType+	49.91	41.05	63.65	45.51	31.85
SelfORE	54.16	51.06	57.64	61.91	44.70
HiURE	57.12	54.13	60.46	63.03	46.16
U-CORE	**63.74**	**59.61**	**68.49**	**75.77**	**61.40**

Table 2: Experiment results(%) on TACRED in terms of B^3 precision, B^3 recall, B^3 F1, NMI, and ARI.

Dataset	Model	F1	NMI	ARI
SemEval	w/o CCM	68.85	53.17	54.17
	w/o Updater	75.39	65.61	68.16
	w/o ASP	73.44	60.35	62.71
	U-CORE	78.83	66.79	70.88
	w self-training	80.28	70.24	70.94
TACRED	w/o CCM	60.17	71.43	54.22
	w/o Updater	61.07	73.25	58.19
	w/o ASP	62.49	74.11	60.39
	U-CORE	63.74	75.77	61.40
	w self-training	67.33	78.41	62.25

Table 3: Effectiveness of each U-CORE operation.

more challenging dataset with the lower performance of all models, attributable to its imbalanced data and limited relationship with the general domains on pre-trained BERT, as also observed by Wang et al. (2022). Directly using pre-trained BERT for clustering only yields a 5.73% ARI.

The results of CNN, Att-BiLSTM, and Supervised-RSN are relatively low without the performance boost provided by Pre-trained Language Models (PLMs). Although ZS-BERT can achieve impressive ZORE performance, as demonstrated in the original paper, it relies on a manual description of novel relations, resulting in decreased clustering performance. While MTB can capture information from predefined relations effectively, its ability to generalize on undefined relations is insufficient. RCL, the previous state-of-the-art method, uses instance-wise CL to enhance performance, but it tends to separate similar semantics and only preserve the local smoothness of instances. The visualization of RCL in Section 4.9 reveals its failure to differentiate some similar relations. The performance of U-CORE proves that it can optimize the encoder both locally and globally to generate a better relation representation.

Results on Unsupervised Open Relation Tasks Table 2 displays the performance of various models on the UORE tasks. The challenge of TACRED is extracting undefined relations without annotations. TACRED has 41 relations, yet we used only 16 clusters based on previous work, resulting in a higher value of B^3 recall than B^3 precision. Our proposed method, U-CORE, outperforms state-of-the-art models on TACRED datasets with remarkable improvements of 6.62% B^3 F1, 12.74% NMI, and 15.24% ARI. The proposed cluster-wise contrastive module of U-CORE minimizes intra-cluster distances while maximizing inter-cluster distances, leading to a more accurate clustering distribution closer to the actual distribution. This has led to substantial improvements, especially in the ARI value. The performance on the UORE task shows that U-CORE excels in self-training and can effectively learn relation representations from global semantics.

4.4 Ablation Study

Effect of Cluster-wise Contrastive Module We have introduced a Cluster-wise Contrastive Module (CCM) to prevent the identification of instances with similar semantics as negative pairs. As shown in Table 3, the performance of U-CORE without CCM has a significant decrease. Additionally, the performance of U-CORE without CCM on the SemEval dataset is similar to that of RCL, which utilizes instance-wise contrastive learning. This highlights the ability of our cluster-wise contrastive module to capture global semantic structures and effectively generalize over undefined relations.

Effect of Cluster Center Updater Our proposed Cluster Center Updater is a deep-cluster-based mechanism that enables U-CORE to update cluster centers in parallel with the training process. We show in Table 3 that U-CORE without Center Updater, which utilizes k-means to update the centers, results in an average performance loss of 3.06% B^3 F1, 1.85% NMI, and 2.97% ARI compared to U-CORE. The experimental

results demonstrate that the proposed module significantly improves the accuracy of clustering. Furthermore, the Center Updater also improves efficiency, which will be discussed further in Section 4.6.

Effect of Adaptive Self-paced Learning The main objective of introducing the Adaptive Self-paced Learning (ASP) module is to enhance training stability by apprising the model of the optimal timing for learning. Our preceding experimental analyses reveal that SemEval represents a demanding dataset with unsatisfactory clustering outcomes in the absence of training. Additionally, the feature space undergoes rapid changes due to predefined relations, leading to significant data-shifting problems. The results presented in Table 3 demonstrate that U-CORE without ASP performs considerably worse, exhibiting a loss of 5.29% in B^3 F1, 6.34% in NMI, and 8.11% in ARI. In comparison, the severity of the data-shifting problem in TACRED is relatively lower due to the self-supervised nature of UORE and consequently results in relatively smaller performance degradation in the absence of ASP.

4.5 Effect of Self-training on Testing Set

It is worth noting that U-CORE with self-training represents a special case of our proposed model. In Table 3, we present the results of conducting self-training on U-CORE with testing data, which yields improved performance over U-CORE in both SemEval and TACRED, as it can optimize relation representations without requiring any human annotations. This aspect is not featured in the baseline comparison section, as no other baseline in ZORE is capable of self-training in the absence of predefined relations. Furthermore, our analysis reveals that the scenarios of ZORE and UORE may converge in situations where both predefined and undefined relations are present. As a unified framework, U-CORE facilitates supervised training on predefined data and self-training on both predefined and undefined relations, leading to an enhanced performance by optimizing global semantics.

4.6 Efficiency Analysis

As previously discussed, U-CORE's cluster center updater is more efficient compared to conventional clustering algorithms. To provide a comparison, we employed *HiURE*, which uses re-clustering with a k-means-based approach after every epoch. Results are presented in Table 4, indicating the average epoch time and epoch interval of both methods on the TACRED dataset. Despite having similar epoch time, U-CORE's epoch interval is only a quarter of HiURE's.

	Epoch time	Epoch Interval
HiURE	70.87s	40.83s
U-CORE	72.25s	**10.87s**

Table 4: Model's epoch time and epoch interval.

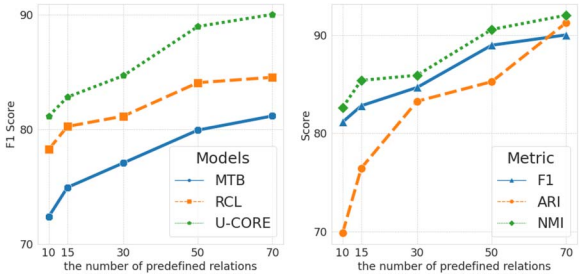


Figure 2: Left: The results of models with different numbers of predefined relations. Right: The different metric scores of U-CORE with different numbers predefined relations.

4.7 Effect of Predefined Relations Numbers

This section investigates the impact of predefined relation quantity on model performance on the FewRel dataset. To this end, we selected 10 undefined relations for the testing set and varied the number of predefined relations (n) in the training set from 10 to 70. The experimental outcomes are depicted in Figure 2 (Left). The results reveal that U-CORE displays significant performance improvements with increasing numbers of predefined relations, achieving nearly a 10% lead over both RCL and MTB models when trained on equivalent quantities of data. It is noteworthy, however, that the benefits of U-CORE are less pronounced in settings with a small number of relations, such as $n = 5$, highlighting potential areas for future research.

Moreover, as illustrated in Figure 2 (Right), the ARI score experiences a prominent upswing within the range of $n = 10$ to $n = 30$, indicating an intensifying impact of the cluster-wise contrastive loss, which is relation-based. This

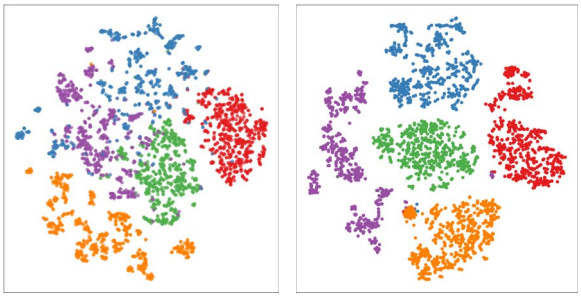
	w/ Negative Relation			Mixed Test Set		
Model	F1	NMI	ARI	F1	NMI	ARI
MTB	32.03	16.10	11.87	30.06	14.02	10.22
RCL	55.52	44.64	39.55	53.52	46.86	35.58
U-CORE	**64.63**	**55.57**	**52.23**	**58.17**	**54.68**	**49.54**

Table 5: Experiment results(%) on SemEval with additional complex real-world settings in terms of B3 F1, NMI and ARI. ''w/ negative relation'' means ''no relation'' type is added to train and test set. ''Mixed Test Set'' means the test set contains both predefined and undefined relation types.

phenomenon can be attributed to the larger vocabulary of relational knowledge that emerges as the number of predefined relations rises, significantly amplifying the effect of this loss on the model.

4.8 Effect of Additional Complex Settings

In our previous experiments, we follow mainstream work to design our evaluation settings for fair comparisons. In this section, we delve deeper into assessing the robustness of U-CORE by exploring additional complex real-world scenarios. In realistic scenarios, the ''no relation'' type may appear in the dataset, and the test set may contain both predefined and undefined relation types. We present the experimental results for these two challenging real-world settings on the SemEval dataset in Table 5. Additionally, we include the results of the two best-performing models from our previous experiments, RCL and MTB. In the ''w/ negative'' setting, we add the ''no relation'' type based on the proportion of train and test sets. In the ''Mixed Test Set'' setting, we randomly allocated 20% of the data in predefined relation types to the test set. The experimental results demonstrate that the ''Mixed Test Set'' setting presents a greater challenge, as it involves reduced training data and an increased number of relation types in the test set. Consequently, all models experience a significant performance loss in this scenario. However, even under these more complex real-world conditions, U-CORE consistently outperforms the other models and achieves the best performance. This highlights the robustness and effectiveness of U-CORE in handling these intricate settings.



(a) RCL (b) U-CORE

Figure 3: t-SNE visualization of RCL and U-CORE on FewRel dataset ($m = 5$).

4.9 Visualization

To visually illustrate how our method enhances the understanding of undefined relations, we employ t-SNE (Van der Maaten and Hinton, 2008) to visualize the representation by mapping relation representation to a low-dimensional space. We choose undefined categories ($m = 5$) for the zero-shot experiment on the FewRel dataset. In each figure, the relation instances are colored according to their ground truth labels. As depicted in Figure 3a, the RCL struggles to differentiate between the five relationship types effectively. Due to Instance-wise CL implementation, blue dots representing the same relation are pushed away from each other. In contrast, U-CORE has effectively separated and categorized these five types, exhibiting a noteworthy capability in identifying differences. This success may be attributed to the cluster-wise contrastive module that collaborates with Adaptive Self-paced learning to optimize relation clustering performance by expanding inter-cluster spacing while minimizing intra-cluster spacing.

5 Conclusion

In this paper, we present a unified deep cluster-wise contrastive framework, U-CORE, for Open Relation Extraction tasks. Our proposed framework can tackle various ORE tasks and overcome the limitations of previous instance-wise CL-based methods. Furthermore, we introduce the cluster center updater and adaptive self-paced learning to enhance the stability and efficiency of our model. The results of our experiments on three datasets provide evidence of the effectiveness of our framework, achieving new state-of-the-art performance. Recently, Large Language

Models (LLMs) like ChatGPT[3] have demonstrated remarkable performance in various NLP tasks, but Han et al. (2023) and Li et al. (2023) indicate that LLMs exhibit subpar performance in ORE tasks. From our aspect, we believe that LLMs have the potential to address ORE tasks. In light of this, our feature work is to further explore the potential of LLMs in ORE tasks.

References

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*. https://doi.org/10.3115/980451.980859

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1279

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924.

Chih-Yao Chen and Cheng-Te Li. 2021. ZS-BERT: Towards zero-shot relation extraction with attribute representation learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3470–3479, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.272

Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter*, 19(2):25–35. https://doi.org/10.1145/3166054.3166058

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR.

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. https://doi.org/10.18653/v1/N19-1423

Hady Elsahar, Elena Demidova, Simon Gottschalk, Christophe Gravier, and Frederique Laforest. 2017. 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*, pages 12–16. Springer. https://doi.org/10.1007/978-3-319-70407-4_3

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.552

Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin. 2017. Improved deep embedded clustering with local structure preservation. In *IJCAI*, pages 1753–1759. https://doi.org/10.24963/ijcai.2017/243

Xifeng Guo, Xinwang Liu, En Zhu, Xinzhong Zhu, Miaomiao Li, Xin Xu, and Jianping Yin. 2019. Adaptive self-paced deep clustering with data augmentation. *IEEE Transactions on Knowledge and Data Engineering*, 32(9):1680–1693.

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023. Is information extraction solved by ChatGPT?

[3] https://chat.openai.com/.

An analysis of performance, evaluation criteria, robustness and errors. *arXiv preprint arXiv:2305.14450*.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2018. FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1514

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738. https://doi.org/10.1109/CVPR42600.2020.00975

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid O. Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2019. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. *arXiv preprint arXiv:1911.10422*.

Xuming Hu, Lijie Wen, Yusong Xu, Chenwei Zhang, and Philip Yu. 2020. SelfORE: Self-supervised relational feature learning for open relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3673–3682, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.299

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification*, 2:193–218. https://doi.org/10.1007/BF01908075

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and S. Yu Philip. 2021. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2):494–514. https://doi.org/10.1109/TNNLS.2021.3070843, PubMed: 33900922

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 333–342, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/K17-1034

Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. 2023. Evaluating ChatGPT's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *arXiv preprint arXiv:2304.11633*.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven Hoi. 2021. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations*.

Fangchao Liu, Lingyong Yan, Hongyu Lin, Xianpei Han, and Le Sun. 2021. 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)*, pages 4683–4693, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.361

Shuliang Liu, Xuming Hu, Chenwei Zhang, Shu'ang Li, Lijie Wen, and Philip Yu. 2022. 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*, pages 5970–5980, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.437

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Qianli Ma, Jiawei Zheng, Sen Li, and Gary W. Cottrell. 2019. Learning representations for time series clustering. *Advances in Neural Information Processing Systems*, 32.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11).

Diego Marcheggiani and Ivan Titov. 2016. Discrete-state variational autoencoders for joint

discovery and factorization of relations. *Transactions of the Association for Computational Linguistics*, 4:231–244. https://doi.org/10.1162/tacl_a_00095

Abiola Obamuyide and Andreas Vlachos. 2018. Zero-shot relation classification as textual entailment. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 72–78. https://doi.org/10.18653/v1/W18-5511

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Marco Antonio Calijorne Soares and Fernando Silva Parreiras. 2020. A literature review on question answering techniques, paradigms and systems. *Journal of King Saud University-Computer and Information Sciences*, 32(6):635–646. https://doi.org/10.1016/j.jksuci.2018.08.005

Alvin Subakti, Hendri Murfi, and Nora Hariadi. 2022. The performance of BERT as data representation of text clustering. *Journal of Big Data*, 9(1):1–21. https://doi.org/10.1186/s40537-022-00564-9, PubMed: 35194542

Thy Thy Tran, Phong Le, and Sophia Ananiadou. 2020. Revisiting unsupervised relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7498–7505, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.669

Shusen Wang, Bosen Zhang, Yajing Xu, Yanan Wu, and Bo Xiao. 2022. RCL: Relation contrastive learning for zero-shot relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2456–2468, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-naacl.188

Ruidong Wu, Yuan Yao, Xu Han, Ruobing Xie, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun. 2019. 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)*, pages 219–228, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1021

Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *International Conference on Machine Learning*, pages 478–487. PMLR.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/D17-1004

Jun Zhao, Tao Gui, Qi Zhang, and Yaqian Zhou. 2021. A relation-oriented clustering method for open relation extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9707–9718, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.765

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 207–212, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-2034

Models	F1	NMI	ARI
MTB	55.40	46.41	40.11
RCL	73.97	64.53	61.18
U-CORE	**85.77**	**79.94**	**76.86**

Table 6: Experiment results (%) on SemEval when setting popular relation types as predefined.

A Appendix

A.1 Implement Details

In the U-CORE model, the encoder utilized is *BERT-base-uncased*, and it undergoes 10 epochs of training with an AdamW optimizer (Loshchilov and Hutter, 2019) set to a learning rate of 1e-5, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of 0.01. Additionally, the values of τ and γ are set to 0.05 and 0.6, respectively. The value of η is set to 10. Following Gao et al. (2021), the dropout rate in data augmentation is 0.1. The training process utilizes double NVIDIA RTX 3090 with 24 GB memory, and the batch size is 128.

A.2 Popular Relations only as Predefined

In this section, we conduct an experiment considering only popular relation types as predefined, and the corresponding results are presented in Table 6. In this scenario, each model exhibits even better performance due to the availability of a larger training dataset.

Dataset	Model	F1	NMI	ARI
SemEval	w/o CCI	78.25	66.34	68.70
	w/o ϕ_j	77.82	64.55	69.72
	U-CORE	78.83	66.79	70.88
TACRED	w/o CCI	61.97	74.08	60.51
	w/o ϕ_j	60.52	73.27	58.53
	U-CORE	63.74	75.77	61.40

Table 7: Effectiveness of CCI and ϕ_j.

A.3 Additional Ablation Studies

In this section, we have included two additional ablations. The first ablation replaces the Cluster Centers Initialization method in Section 3.3.1 by manually setting the number of clusters to match the number of predefined relation types during the training process. The second ablation replaces ϕ_j in Equation (6) with a fixed value, which is equivalent to τ in Equation (4). Table 7 presents the results of these two ablations. Note that in some cases CCI has no effect on the performance of U-CORE, as the number of clusters generated by CCI may align with the number of predefined relation types. CCI (Cluster Centers Initialization) and ϕ_j were introduced to avoid the artificial setting of two hyperparameters: the number of clusters and the temperature of cluster-wise contrastive loss.