RCL: Relation Contrastive Learning for Zero-Shot Relation Extraction

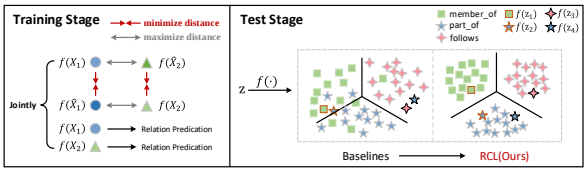
Shusen Wang, Bosen Zhang, Yajing Xu*, Yanan Wu, Bo Xiao
Pattern Recognition & Intelligent System Laboratory,
Beijing University of Posts and Telecommunications, Beijing, China
{shusenw,zhangbosen,xyj,yanan.wu,xiaobo}@bupt.edu.cn

Abstract

Zero-shot relation extraction aims to identify novel relations which cannot be observed at the training stage. However, it still faces some challenges since the unseen relations of instances are similar or the input sentences have similar entities, the unseen relation representations from different categories tend to overlap and lead to errors. In this paper, we propose a novel Relation Contrastive Learning framework (RCL) to mitigate above two types of similar problems: Similar Relations and Similar Entities. By jointly optimizing a contrastive instance loss with a relation classification loss on seen relations, RCL can learn subtle difference between instances and achieve better separation between different relation categories in the representation space simultaneously. Especially in contrastive instance learning, the dropout noise as data augmentation is adopted to amplify the semantic difference between similar instances without breaking relation representation, so as to promote model to learn more effective representations. Experiments conducted on two well-known datasets show that RCL can significantly outperform previous state-of-the-art methods. Moreover, if the seen relations are insufficient, RCL can also obtain comparable results with the model trained on the full training set, showing the robustness of our approach[1]

1 Introduction

Relation extraction is a fundamental problem in natural language processing, which aims to identify the semantic relation between a pair of entities mentioned in the text. Recent progress in supervised relation extraction has achieved great successes (Zeng et al., 2014; Zhou et al., 2016; Soares et al., 2019), but these approaches usually require large-scale labeled data. While in practice, human annotation is time-consuming and labor-intensive. To alleviate the human annotation efforts in relation extraction, some recent studies use distant supervision to generate labeled data for training (Mintz et al., 2009; Lin et al., 2016). However, in the real-world setting, the relations of instances are not always included in the training data, and existing supervised methods cannot well recognize unobserved relations due to weak generalization ability.



Error Type	Input Sentence	Ground Truth	Baseline Pred	RCL Pred(Ours)
Similar Relation	z_1 : In 1978, he replaced Thomas Erdelyi in the Ramones, assuming the name Marky Ramone.	member_of	member_of	member_of
	z_2 : The Doctor tries to restore the universe with the help of River and the alternative universe versions of his companions Amy Pond(Karen Gillan) and Rory Williams(Arthur Darvill).	part_of	member_of	part_of
Similar Entity	z_3 : In May 2015, Vienna hosted the Eurovision Song Contest following Austria 's victory in the 2014 contest.	follows	follows	follows
	z_4 : Thus, the song was succeeded as Romanian repre-sentative at the 2002 Contest by Monica Anghel & Marcel Pavel with "Tell Me Why".	part_of	follows	part_of

Figure 1: **Top:** Overview of the proposed RCL at the training and test stage. $f(\cdot)$ is a learnable projection function that projects the input sentence X_i to its corresponding relation representation $f(X_i)$. $f(\hat{X}_i)$ is the augmented view of $f(X_i)$ and Z is the whole test set. **Bottom:** Four examples at the test stage and their corresponding relation representations are shown in the right of Top. The entities are marked in orange.

To address the aforementioned limitations, zero-shot relation extraction has been proposed to extract relational facts where the target relations cannot be observed at the training stage. The challenge of zero-shot relation extraction models is how to learn effective representations based on seen relations at the training stage and well generalize to unseen relations at the test stage. Two studies (Levy et al., 2017; Obamuyide and Vlachos, 2018) treat zero-shot relation extraction as a different task (i.e., question answering and textual entailment), but they both need human annotation auxiliary information for input, i.e., pre-defining question templates and relation descriptions. ZS-BERT (Chen

*Yajing Xu is the corresponding author.
[1]https://github.com/ShusenWang/NAACL2022-RCL

Findings of the Association for Computational Linguistics: NAACL 2022, pages 2456 - 2468
July 10-15, 2022

and Li, 2021) predicts unseen relations with attribute representation learning. Despite promising improvements on directly predicting unseen relations, ZS-BERT still makes wrong predictions due to similar relations or similar entities. The same problem arises in supervised methods under the zero-shot settings.

As shown in Figure 1, there are two types of similar errors: **Similar Relations** and **Similar Entities**. For similar relations (see Z_1 and Z_2), existing methods predict wrongly results because the unseen relations possess similar semantics and data points belong to two relations in the representation space are overlapped. For similar entities (i.e., *2014 contest* and *2002 Contest*), since entities are the context of relation and relation representations are derived from entities, the relation representations containing similar entities are close (see $f(Z_3)$ and $f(Z_4)$) and baselines wrongly consider $f(Z_4)$ belongs to *follows* in the representation space, even if two unseen relations are not related. Recently, Instance-wise Contrastive Learning (Instance-CL) (He et al., 2020; Chen et al., 2020; Yan et al., 2021; Gao et al., 2021; Zhang et al., 2021) has achieved remarkable success in representation learning. Instance-CL is used to learn an effective representation by pulling together the instances from the same class, while pushing apart instances from different classes. Inspired by Instance-CL, we attempt to use Instance-CL on seen relations to learn the difference between similar relations and the divergence of relation representations derived from similar entities.

In this paper, we propose a novel Relation Contrastive Learning framework (RCL) to solve the above-mentioned problems. Figure 1 depicts the overview of the proposed model, which consists of four steps: (i) The input for RCL is a batch of sentences containing the pair of target entities and each sentence is sent into input sentence encoder to generate the contextual sentence embeddings[2]. (ii) Taking the sentence embeddings as input, relation augmentation layer is designed to obtain the relation representations $f(X_i)$ and their corresponding augmented views $f(\hat{X}_i)$. (iii) By jointly optimizing a contrastive loss and a relation classification loss on seen relations, RCL can learn subtle difference between instances and achieve better separation between relations in the representation space simultaneously to obtain an effective projection function f. (iv) With the learned f, the whole test set Z can be projected for unseen relation representations in the representation space and zero-shot prediction is performed on unseen relation representations by K-Means.

To summarize, the major contributions of our work are as follows: (i) We propose a novel framework based on contrastive learning for zero-shot relation extraction. It effectively mitigates two types of similar problems: similar relations and similar entities by learning representations jointly optimized with contrastive loss and classification loss. (ii) We explore various data augmentation strategies in relation augmentation to minimize semantic impact for contrastive instance learning and experimental results show dropout noise as minimal data augmentation can help RCL learn the difference between similar instances better. (iii) We conduct experiments on two well-known datasets. Experimental results show that RCL can advance state-of-the-art performance by a large margin. Besides, even if the number of seen relations is insufficient, RCL can also achieve comparable results with the model trained on the full training set.

2 Related Work

Relation Extraction. Relation extraction aims at extracting relation between entities within a given sentence. Many relation extraction methods (Qian et al., 2008; Zeng et al., 2014; Zhou et al., 2016) are supervised model. Recently, some studies focus on pre-training language model (Devlin et al., 2019; Liu et al., 2019; Lan et al., 2020) because of its powerful capability of semantic representation. Wu and He (2019) propose R-BERT that uses BERT to extract relation features and incorporates entity information to perform relation extraction. Soares et al. (2019) propose a relation representation learning method based on BERT and have shown promising results. However, these models require labeled data. Unsupervised relation extraction (Yu et al., 2017; Saha and Mausam, 2018; Stanovsky et al., 2018) can discover semantic relation feature from data without human annotations. One representative work is Open relation extraction. Wu et al. (2019) propose a novel model to learn a similarity metric of relations from labeled data, and identify unseen relations by transferring knowledge learned from seen relations. While OpenRE method can identify novel relation without annotations and external resources, it cannot effectively

[2]The words, "embeddings", and "representations", are used interchangeably throughout this paper.

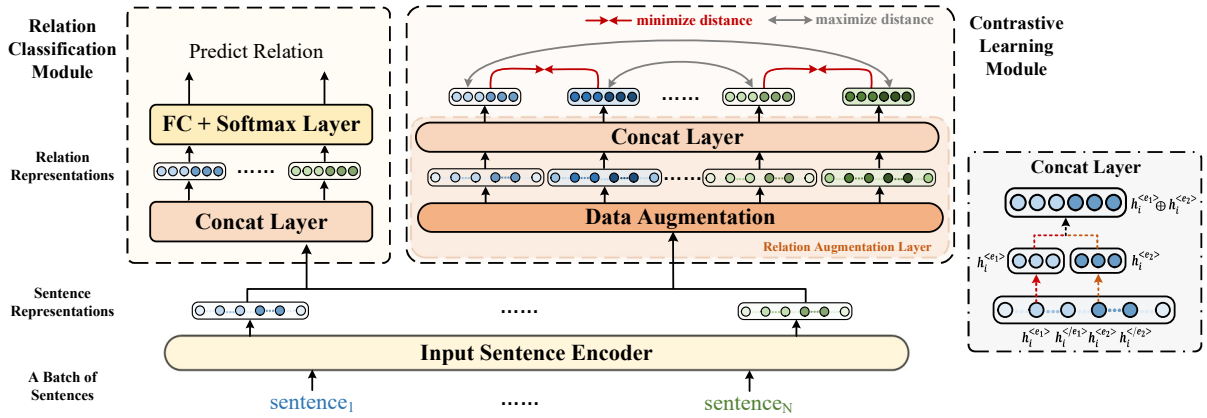


Figure 2: Architecture of the RCL, which consists of three components, and the details are introduced in Section 3. Note that relation augmentation layer contains data augmentation and a concat layer.

discard irrelevant information and severely suffers from the instability.

Zero-shot Relation Extraction. Zero-shot relation extraction aims to identify novel relation without training instances. Existing zero-shot relation extraction methods are few and most rely on human annotation auxiliary information for input. Levy et al. (2017) reduce zero-shot relation extraction to a question answering task. They use 10 predefining question templates to represent relations, and then train a reading comprehension model to infer which relation satisfies the given sentence and question. Obamuyide and Vlachos (2018) treat zero-shot relation extraction as a textual entailment task, which requires the model to input descriptions of relations. They train a textual entailment model to predict whether the input sentence containing two entities matches the description of a given relation, identifying novel relations by generalizing from the descriptions of seen relations at the training stage to those of unseen relations at test time. Chen and Li (2021) propose ZS-BERT to tackle zero-shot relation extraction task with attribute representation learning. ZS-BERT learns the representations of relations based on their descriptions during the training time, and generates the prediction of unseen relation for new sentence by nearest neighbor search. However, ZS-BERT suffers from similar relation error and similar entity error, and it needs human annotation auxiliary information for input, i.e., relation descriptions. In this paper, we do not require any human annotation auxiliary information for input.

Contrastive Learning. In the field of image and natural language processing, many recent successes are inspired by contrastive learning (He et al., 2020; Chen et al., 2020; Yan et al., 2021; Gao et al., 2021). Contrastive learning regards the input data and corresponding augmented views as an independent class. The goal of contrastive learning is to pull together representations from the same class, while keeping representations from different classes away. Therefore, the representations learned from contrastive learning are better separated and good for clustering. Gao et al. (2021) propose a novel sentence embeddings learning framework based on contrastive learning to produce superior sentence embeddings and show that dropout is an effective data augmentation. SCCL (Zhang et al., 2021) jointly optimizes a contrastive loss and a clustering loss to disperse overlap categories in the representation space. Inspired by contrastive learning, we leverage contrastive learning to help the model learn an effective representation.

3 Proposed Model

3.1 Model Overview

As illustrated in Figure 2, the proposed model RCL consists of three components: input sentence encoder, contrastive learning module and relation classification module. Given a batch of sentences containing two entities, the sentence representations are generated by input sentence encoder and then are sent to relation classification module and contrastive learning module. For contrastive learning module, the relation representations and their corresponding augmented views generated by a relation augmentation layer are used to perform contrastive instance learning to learn the difference between instances. For relation classification module, the relation representations generated by the concat layer are used to identify seen relations to achieve better separation between relations. We train RCL under a multi-task learning structure

with contrastive learning module and relation classification module to learn effective representations for unseen relations. At the test stage, we obtain the unseen relation representations by the input sentence encoder and concat layer, and then send them into K-Means to predict the unseen relations.

3.2 Input Sentence Encoder

Input Sentence Encoder aims to generate the contextual representation of each token. In this work, we assume entities contained in the sentence have been recognized before input. For a sentence $X_i = [x_i^1, .., x_i^L]$ where two entities e_1 and e_2 are mentioned, we use the ENTITY MARKERS (Soares et al., 2019) to augment X_i to better extract relation features from context. Specifically, we introduce four special tokens to mark the beginning and the end of each entity mentioned in the sentence. The input token sequence X_i for input sentence encoder is as follows:

$$X_i = [x_i^1, \ldots, <e_1>, x_i^k, \ldots, x_i^{l-1}, </e_1>, \ldots, <e_2>, x_i^p, \ldots, x_i^{z-1}, </e_2>, \ldots, x_i^L] \quad (1)$$

where $<e_1>, </e_1>, <e_2>, </e_2>$ are four special tokens to mark the beginning and the end of each entity mentioned in the sentence, L is the length of sentence. Then we use BERT (Devlin et al., 2019) to obtain the sentence embeddings $\boldsymbol{h}_i \in \mathbb{R}^{L \times d}$:

$$\mathbf{h_i} = [\mathbf{h_i^1}, \ldots, \mathbf{h_i^{<e_1>}}, \ldots, \mathbf{h_i^{</e_1>}}, \ldots, \mathbf{h_i^{<e_2>}}, \ldots, \mathbf{h_i^{</e_2>}}, \ldots, \mathbf{h_i^L}] \quad (2)$$

where d is the hidden dimension.

3.3 Contrastive Learning Module

Contrastive Learning Module aims at learning the difference between a batch of instances to better represent relations.

Contrastive Instance Learning. After we obtained $\boldsymbol{H} = \{\boldsymbol{h}_1, \ldots, \boldsymbol{h}_N\}$ from N input sentences using input sentence encoder, relation augmentation layer is used to generate relation representations and their augmented views. More specifically, the relation augmentation layer consists of data augmentation and a concat layer. For each sentence embeddings $\boldsymbol{h}_i$, a transformation $T(\cdot)$ is applied to generate its augmented view: $\hat{\boldsymbol{h}}_i = T(\boldsymbol{h}_i)$, where $\hat{\boldsymbol{h}}_i \in \mathbb{R}^{L \times d}$.

After obtaining sentence embeddings $\boldsymbol{h}_i$ and its augmentation $\hat{\boldsymbol{h}}_i$, we obtain relation representations and its augmentation by a concat layer. Specifically, we use the token embeddings corresponding to $<e_1>, <e_2>$ positions as the entity representation and concatenate them to derive a contextualized relation representation and its augmented view $\boldsymbol{r_i}, \hat{\boldsymbol{r}}_i \in \mathbb{R}^{2 \cdot d}$:

$$\begin{aligned} \mathbf{r_i} &= \mathbf{h_i^{<e_1>}} \oplus \mathbf{h_i^{<e_2>}} \\ \hat{\mathbf{r}}_\mathbf{i} &= \hat{\mathbf{h}}_\mathbf{i}^{<\mathbf{e_1}>} \oplus \hat{\mathbf{h}}_\mathbf{i}^{<\mathbf{e_2}>} \end{aligned} \quad (3)$$

where $\oplus$ is the concatenation operator and $\boldsymbol{r_i}, \hat{\boldsymbol{r}}_i$ are both fixed-length vector.

To better learn effective relation representations, we optimizes a contrastive objective, which disperses different relation of instances apart while implicitly bringing the same relation of instances together. Let $\boldsymbol{R} = \{\boldsymbol{r_1}, \ldots, \boldsymbol{r_N}\}$ and $\hat{\boldsymbol{R}} = \{\hat{\boldsymbol{r}}_\mathbf{1}, \ldots, \hat{\boldsymbol{r}}_\mathbf{N}\}$ denote a mini-batch of relation representations and its augmented views respectively. We regard $(\boldsymbol{r_i}, \hat{\boldsymbol{r}}_i)$ as a positive pair and other $N-1$ augmented views as negative instances. For a mini-batch with N pairs, we follow the contrastive framework in SimCSE (Gao et al., 2021) and take a cross-entropy objective with in-batch negatives (Chen et al., 2017). The training objective for $(\boldsymbol{r_i}, \hat{\boldsymbol{r}}_i)$ is:

$$\ell_i^{cl} = -\log \frac{e^{\mathrm{sim}(\mathbf{r_i}, \hat{\mathbf{r}}_\mathbf{i})/\tau}}{\sum_{j=1}^{N} e^{\mathrm{sim}(\mathbf{r_i}, \hat{\mathbf{r}}_\mathbf{j})/\tau}} \quad (4)$$

where $\boldsymbol{r_i} \in \boldsymbol{R}, \hat{\boldsymbol{r}}_i \in \hat{\boldsymbol{R}}$, $\mathrm{sim}(\mathbf{r_1}, \mathbf{r_2})$ is the cosine similarity $\frac{\mathbf{r}_1^\top \mathbf{r}_2}{\|\mathbf{r}_1\| \cdot \|\mathbf{r}_2\|}$,and τ is a temperature hyperparameter.

Data Augmentation Strategies. To amplify the semantic difference between similar instances without breaking the semantic of relation representations, we explore five different data augmentations $T(\cdot)$ for contrastive instance learning, including feature cutoff (Shen et al., 2020), random mask (Hinton et al., 2012), dropout (Gao et al., 2021), composition of dropout and feature cutoff and composition of dropout and random mask.

Feature cutoff is a simple and efficient data augmentation strategy to introduce minimal semantic impact for relation instances. Specifically, we randomly erase some feature dimensions in the sentence embeddings produced by input sentence encoder.

Random mask is proved its effectiveness as an augmentation strategy (Yan et al., 2021). In our experiments, we randomly drop elements in the sentence embeddings by a specific probability and sets their values to zero.

Dropout has been shown its effectiveness as minimal data augmentation by SimCSE (Gao et al., 2021). Thus, similar to SimCSE, we augment sentence embeddings by feeding the same input sentence to BERT again.

Composition of augmentations is an effective strategy in image domain (Chen et al., 2020). Based on dropout, we explore two strategies of composition of data augmentations. Composition of dropout and feature cutoff is a strategy that we first use dropout to obtain augmented view and then send it into feature cutoff to obtain the final augmented view. Similarly, composition of dropout and random mask is a strategy that dropout first and then random mask. We present the experimental results of these strategies and analyze their effects for contrastive learning in Section 4.4.

3.4 Relation Classification Module

Relation Classification Module aims to identify seen relations. With sentence embeddings $\boldsymbol{h_i}$ from input sentence encoder, we obtain relation representation $\boldsymbol{r_i}$ by the concat layer, following the way same as Equation (3). Let n denotes the number of seen relations and Y_s denotes the set of seen relations. By transforming the relation representation $\boldsymbol{r_i}$, along with a softmax layer, we generate the n-dimensional classification probability distribution of the i-th sample over seen relations:

$$p\left(y_i \mid X_i, \theta\right) = \text{softmax}\left(W\left(\tanh\left(\mathbf{r_i}\right)\right) + b\right) \tag{5}$$

where X_i is the i-th input sentence containing two entities, $y_i \in Y_s$ is the seen relation, θ is the model parameter, $W \in \mathbb{R}^{n \times 2 \cdot d}$, and $b \in \mathbb{R}^n$. Note that we use the relation representation $\boldsymbol{r_i}$ produced intermediately for predicting unseen relations under the zero-shot settings instead of the probability distribution. For each data point X_i, we use cross-entropy to calculate classification loss:

$$\ell_i^{rc} = CrossEntropy(p(y_i \mid X_i, \theta), \hat{y}_i) \tag{6}$$

where $\hat{y}_i$ is the ground-truth label of the i-th sample.

3.5 Train and Test

At the training stage, We train the model with two objectives under the multi-task learning structure. The first is to minimize the distance between the relation representation and its augmented view, while keeping the relation representation distant from other augmented relation representations in a mini-batch. The second objective is to bring high prediction accuracy of seen relations. For a mini-batch of input sentences, the training objective of RCL is as follows:

$$\begin{aligned} \mathcal{L}_{\text{CL}} &= -\frac{1}{N}\sum_{i=1}^{N}\ell_i^{cl}, \mathcal{L}_{\text{RC}} = -\frac{1}{N}\sum_{i=1}^{N}\ell_i^{rc} \\ \mathcal{L}_{\text{joint}} &= \mathcal{L}_{\text{RC}} + \alpha\mathcal{L}_{\text{CL}} \end{aligned} \tag{7}$$

where N is the number of input sentences, α is a hyper-parameter to balance two objectives.

At the test stage, we send the new-coming sentences into the input sentence encoder and concat layer to generate unseen relation representations, and the prediction on unseen relations can be achieved by K-Means.

4 Experiments

4.1 Datasets

Two datasets are used to evaluate our model: SemEval2010 Task8 (Hendrickx et al., 2010) and FewRel (Han et al., 2018). **SemEval2010 Task8** has been widely used in relation extraction task, which contains 9 relations and an Other relation. There are 10,717 instances in the dataset and the number of instances of each relation is not equal. Each relation has direction in the dataset, but in our experiments, we do not consider the direction of 9 relations and not use the Other relation. For each relation, we combine the instances of training set with instances of test set to obtain overall instances of each relation. **FewRel** is a public dataset based on Wikipedia, and it contains 80 types of relations, each with 700 instances. Although FewRel is widely used in few-shot learning setting, it is also suitable for zero-shot learning as long as we disjoint the relation labels within training and test data. The statistics of the two datasets are shown in Appendix A.

4.2 Evaluation Settings

Zero-shot Learning Settings. Let m denotes the number of unseen relations, and Y_u denotes the set of unseen relations. We randomly select m relations as unseen relations and the rest of relations n as seen relations. Note that $Y_s \cap Y_u = \emptyset$. Then we split the whole dataset into training and test data. The training data only contains the instances of seen relations, in contrast to test data only with the instances of unseen relations. We repeat experiments 5 times on SemEval2010 Task8 and FewRel, and then report the average results. As for implementation details for RCL, we implement

our model based on `Transformers` package[3] (Wolf et al., 2020). And we use an Adam optimizer (Kingma and Ba, 2014), in which the learning rate is 5e-5. Please refer to the Appendix B for more implementation details.

Evaluation Metrics. We follow the setting in the previous work (Simon et al., 2019) to convert pseudo labels predicted by clustering to relation labels. In each cluster, the relation label with the largest proportion among the cluster is assigned to all samples as the prediction label. For evaluation metrics, we adopt three commonly-used metrics (Wu et al., 2019; Hu et al., 2020; Zhang et al., 2021) to measure the effectiveness of clustering : B^3 (Bagga and Baldwin, 1998), Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI). For B^3, B^3 precision and recall correspondingly measure the correct rate of putting each sentence in its cluster or clustering all samples into a single class. Then B^3 F_1 is computed as the harmonic mean of the B^3 precision and recall:

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

NMI measures the information shared between the predicted label and the ground truth. When data are partitioned perfectly, the NMI score is 1, while it becomes 0 when prediction and ground truth are independent. ARI is a metric to measure the degree of agreement between the cluster and golden distribution, which ranges in [-1,1]. The more consistent two distributions, the higher the score.

Baselines. We compare RCL to previous methods consisting of CNN (Zeng et al., 2014), Attention BiLSTM (Zhou et al., 2016), RSNs (Wu et al., 2019), MTB (Soares et al., 2019), ZS-BERT (Chen and Li, 2021). For CNN, Attention BiLSTM and MTB, these methods have great success in supervised relation extraction (SRE) but fail to perform zero-shot prediction. Specifically, we consider two variations of MTB which only differ in the backbone (MTB-BERT and MTB-RoBERTa). For fair comparison and zero-shot prediction, we make the relation representation from encoder become the output of the SRE model, instead of originally outputting a probability vector whose dimension is equal to the seen relations. The dimension of output vector is same as RCL. The K-Means is applied over output vector to generate zero-shot prediction. Although RSNs is a open relation extraction method, its Supervised RSN model also meets the setting of zero-shot. For ZS-BERT, the original relation descriptions are used for FewRel and we collect the descriptions of relations for SemEval2010 Task8 from open resources. Then we use the sentence embeddings for K-Means to predict unseen relations. Note that we set the dimension of sentence embeddings same as RCL for fair comparison.

[3] https://github.com/huggingface/transformers

SemEval2010 Task8					
Model	**P**	**R**	**F1**	**NMI**	**ARI**
CNN	38.37	38.49	38.42	17.06	15.43
Att-BiLSTM	41.46	41.79	41.6	21.45	19.97
Supervised RSN	33.14	47.06	38.41	11.98	10.96
MTB-BERT	45.1	46.35	45.71	28.12	23.69
MTB-RoBERTa	42.71	44.84	43.71	24.52	21.01
ZS-BERT	33.86	36.33	35.03	12.47	9.53
RCL w/o RC	50.31	54.87	52.45	34.55	28.97
RCL	**68.1**	**67.95**	**68.02**	**55.91**	**54.71**

Table 1: Experimental results(%) on SemEval2010 Task8 in terms of B^3 precision, B^3 recall, B^3 F1, NMI, ARI. We also report the standard F1 score results in Table 5.

4.3 Experimental Results

Results on SemEval2010 Task8. Table 1 show the comparison results on SemEval2010 Task8. RCL achieves the best performance, significantly outperforming the previous state-of-the-art with 22.31% F1, 27.79% NMI and 31.02% ARI improvements. Due to the relations of SemEval2010 Task8 dataset with high similarity, baseline models severely suffer from similar errors and the performances of baselines are poor. Another reason why baselines perform poorly is that small number of seen relations and class imbalance are more challenging for model. Moreover, SemEval2010 Task8 is much less related to the general domains on which the transformers are pretrained. However, comparing with baselines, experimental results show RCL can effectively mitigates similar problems and better use the general knowledge of the pre-training language model.

Results on FewRel. For FewRel, the experimental results are shown in Table 2. From the results, we observe that our model RCL outperforms existing baselines on FewRel when targeting at different numbers of unseen relations m. Specifically, RCL achieves an average of 2.87% F1, 1.98% NMI and 2.98% ARI improvements compared to previous best results. Since relations on FewRel are class balance and sufficient, MTB-BERT and MTB-

	FewRel																	
	m=5			m=10			m=15			m=30			m=40			Avg.		
Model	**F1**	**NMI**	**ARI**	**F1**	**NMI**	**ARI**	**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	39.15	54.61	35.49	34.09	53.46	30.37	52.78	60.7	47.17
Att-BiLSTM	82.75	79.36	76.63	75.89	79.1	71.46	69.84	75.94	66.03	50.76	66.99	47.64	45.01	64.66	42.23	64.85	73.21	60.8
Supervised RSN	73.33	67.89	64.49	59.11	64.96	48.66	50.99	59.98	39.74	26.01	44.31	18.71	23.55	48.26	18.08	46.6	57.08	37.94
MTB-BERT	88.06	85.32	84.03	81.08	83.95	76.22	78.62	83.57	74.83	63.51	76.61	59.98	60.35	75.9	54.54	74.32	81.07	69.92
MTB-RoBERTa	90.14	87.12	86.7	82.39	84.77	78.03	79.78	84.35	76.82	62.98	75.91	58.83	60.58	75.99	55.08	75.17	81.63	71.15
ZS-BERT	74.51	69.24	66.96	70.63	74.1	65.23	63.33	70.7	59.24	46.43	61.66	42.94	45.68	64.43	42.68	60.12	68.03	55.41
RCL w/o RC	73.58	68.23	64.5	70.52	74.28	59.53	58.02	64.67	51.74	39.89	54.62	33.7	33.09	50.15	28.57	55.02	62.39	47.61
RCL	**90.73**	**87.41**	**86.72**	**84.52**	**86.73**	**80.23**	**81.48**	**85.64**	**78.18**	**67.75**	**79.21**	**64.43**	**65.74**	**79.09**	**61.1**	**78.04**	**83.61**	**74.13**

Table 2: Experimental results(%) produced by the baseline models and the proposed model RCL on FewRel dataset in terms of B^3 F-score, NMI, ARI. m is the number of unseen relations, and we vary m in [5, 10, 15, 30, 40] to examine how performance is affected. RCL w/o RC means RCL without relation classification module. In addition, we also report the standard F1 score results in Table 6.

RoBERTa perform well among competing models but their performance is still lower than RCL. The reason is that their approaches cannot well deal with similar problems. ZS-BERT performs worse than most competing models because ZS-BERT severely relies on the unseen relation descriptions for prediction, while our approach can perform well without external resources. In addition, we find that the improvement of RCL gets larger when m is larger, especially when $m = 40$. It is obvious that it becomes more difficult for prediction since the number of unseen relations increases leading to more seriously similar problems.

Ablation Study. To better validate our model, we conduct an ablation study on each module by correspondingly ablating one. Note that MTB-BERT is the version of RCL without contrastive learning module. From Table 1 and Table 2, we can see that combining these two modules can result in a noticeable performance gain over two datasets. Especially in SemEval2010 Task8, RCL w/o RC outperforms existing baselines by all evaluation metrics, which prove the effectiveness of contrastive learning. However, our proposed RCL significantly outperforms RCL w/o RC with 15.57% F1, 21.36% NMI and 25.74% ARI improvements. It demonstrates that these two modules are complementary on relation representation learning: contrastive learning focuses on learning the difference between instances and implicitly obtaining some knowledge about the difference between relations while relation classification can explicitly learn the difference between relations by identifying the relations but cannot learn the difference between similar instances and suffers from similar problems. When the number of unseen relations increases on FewRel, RCL w/o RC performs worse than competing methods due to without effectively learning relation difference, which also shows that both two modules are important to final model performance.

	SemEval2010 Task8	FewRel
Data augmentation	**F1**	**F1**
None	58.14	86.95
Random Mask	60.25	87.42
Feature Cutoff	59.92	88.46
Dropout	**68.02**	**90.73**
Dropout+Random Mask	67.53	89.84
Dropout+Feature Cutoff	65.51	89.23

Table 3: Experimental results(%) with different data augmentation strategies over two datasets in term of B^3 F1 score. For FewRel, we report the results on $m = 5$.

4.4 Qualitative Analysis

Effect of Data Augmentations. To study the effect of data augmentations, we consider six different data augmentation strategies for contrastive learning in our experiments, including None (i.e. doing nothing), Random Mask, Feature Cutoff, Dropout, Composition of Dropout and Feature Cutoff (Dropout+Feature Cutoff) and Composition of Dropout and Random Mask (Dropout+Random Mask).

The results are shown in Table 3. We can make the following observations: (a) Dropout is the most effective strategy, outperforming all competing strategies. It demonstrate that Dropout essentially acts as minimal data augmentation (Gao et al., 2021) and the noise produced by Dropout can make model learn the difference between similar instances better. (b) When compared with None, Random Mask and Feature Cutoff also improve performance across two datasets. Moreover, Dropout+Random Mask and Dropout+Feature Cutoff significantly outperform Random Mask and Feature Cutoff with roughly 6 and 2 points gain respectively while Dropout still outperforms these two composition of augmentations. It shows that different from the image domain (Chen et al., 2020), composition of augmentations is not always effective for the text domain. (c) We find that our model can improve performance on two datasets

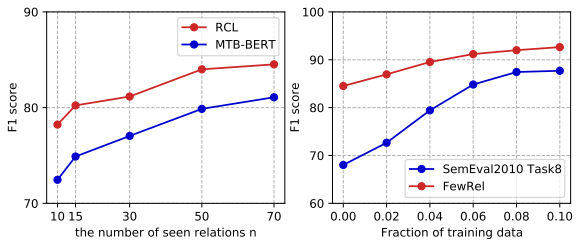


Figure 3: Left: The results of RCL and MTB-BERT with different numbers of seen relations. Right: The performance of RCL with different fractions of unseen instances available for training. The number of unseen relations is set to 10 on FewRel.

even without any data augmentation (None), especially for SemEval2010 Task8 (from 45.71 to 58.14). This is because None tunes the representation space by keeping each representation away from others, even if it has no effect on minimizing the distance between instance and its augmented view since the embeddings of augmented view are same with original instance. It also demonstrates that the effectiveness of the contrastive learning without external resources.

Effect of Number of Seen Relations. In this section, we study the effect of the number of seen relations on FewRel which contains sufficient relations. In our experiment, we vary the number of seen relations n from 10 (insufficient) to 70 (sufficient) and consistently set the number of unseen relations m to 10. Experimental results are presented in Figure 3. As the number of seen relations increases, RCL continuously outperforms MTB-BERT, which shows the effectiveness of our approach. More specifically, when n is set to 10, RCL can achieve 90% F1 score of the model trained on the full seen relations. In addition, the performance of RCL declines more slighter and smoother than MTB-BERT when seen relations gradually become insufficient (from 30 to 10), showing the robustness of our approach.

Capability under Few-shot Settings. In this section, we conduct the experiment of few-shot prediction by following the setting of Chen and Li (2021) to understand the capability of RCL. We move a small fraction of sentences of each unseen relation from test data to training data. Experimental results are shown in Figure 3. As expected on two datasets, RCL achieves more F1 score improvement with more unseen relation instances available at the training stage. When the fraction is set to 4%, RCL can achieve 90% F1 score on FewRel and 80% F1 score on SemEval2010 Task8. It shows the capability of few-shot learning for RCL.

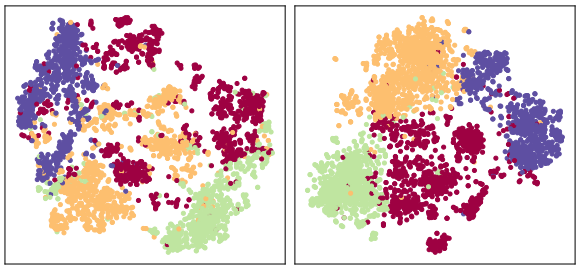


Figure 4: t-SNE visualization of unseen relation representations learned by MTB-BERT and RCL on SemEval2010 Task8 dataset.

Visualization of Relation Representations. To intuitively show how our approach helps to learn better relation representations on seen relations, we visualize the representations of unseen relations by using t-SNE (Van der Maaten and Hinton, 2008) to reduce the dimension to 2. We randomly choose 4 relations as unseen relation from SemEval2010 Task8 and the visualization results are shown in Figure 4. In each figure, relation instances are colored according to their ground-truth labels.

As we can see from Figure 4(a), the data points from MTB-BERT are mingled with different clusters, especially for red points. The reason is these instances possess similar relations or similar entities, and MTB-BERT has not learned the corresponding knowledge to deal with similar problems. However, as illustrated in Figure 4(b), RCL effectively mitigates these two types of similar problems since our approach can learn the difference between instances and the difference between seen relations. It again exhibits the effectiveness of the contrastive loss and multi-task learning structure. We also provide a case study in the Appendix F.

5 Conclusion

In this paper, we propose a jointly framework for zero-shot relation extraction to mitigate two types of similar errors: Similar Relations and Similar Entities. Different from conventional zero-shot relation extraction models which require external resources for training and test, our model does not require external resources. We demonstrate the effectiveness of our framework on two datasets, and our method achieves new state-of-the-art performance. Furthermore, we compare various data augmentation strategies for contrastive learning and provide fine-grained analysis for interpreting how our approach works.

Acknowledgements

We thank all anonymous reviewers for their valuable suggestions. This work was supported by the National Natural Science Foundation of China (NSFC No.62076031).

References

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 79–85.

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.

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.

Ting Chen, Yizhou Sun, Yue Shi, and Liangjie Hong. 2017. On sampling strategies for neural network-based collaborative filtering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 767–776.

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.

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.

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.

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.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38.

Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

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.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

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.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2124–2133, Berlin, Germany. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

Longhua Qian, Guodong Zhou, Fang Kong, Qiaoming Zhu, and Peide Qian. 2008. Exploiting constituent dependencies for tree kernel-based semantic relation extraction. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 697–704, Manchester, UK. Coling 2008 Organizing Committee.

Swarnadeep Saha and Mausam. 2018. Open information extraction from conjunctive sentences. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2288–2299, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

David Sculley. 2010. Web-scale k-means clustering. In *Proceedings of the 19th international conference on World wide web*, pages 1177–1178.

Dinghan Shen, Mingzhi Zheng, Yelong Shen, Yanru Qu, and Weizhu Chen. 2020. A simple but tough-to-beat data augmentation approach for natural language understanding and generation. *arXiv preprint arXiv:2009.13818*.

Étienne Simon, Vincent Guigue, and Benjamin Piwowarski. 2019. Unsupervised information extraction: Regularizing discriminative approaches with relation distribution losses. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1378–1387, Florence, Italy. Association for Computational Linguistics.

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.

Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 885–895, New Orleans, Louisiana. Association for Computational Linguistics.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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.

Shanchan Wu and Yifan He. 2019. Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 2361–2364.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A contrastive framework for self-supervised sentence representation transfer. 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 5065–5075, Online. Association for Computational Linguistics.

Dian Yu, Lifu Huang, and Heng Ji. 2017. Open relation extraction and grounding. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 854–864, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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.

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

Tian Zhang, Raghu Ramakrishnan, and Miron Livny. 1996. Birch: an efficient data clustering method for very large databases. *ACM sigmod record*, 25(2):103–114.

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.

A Statistics of Datasets

The statistics of SemEval2010 Task8 and FewRel are shown in Table 4. For SemEval2010 Task8, we use 9 relations except the Other relaiton. Because of small number of relations, class imbalance and relations with high similarity, the experiment on SemEval2010 Task8 is more challenging and close to real world setting. For FewRel, we use the train and valid split but not test split, because the test split is not publicly available.

	#Instances	#Entities	#Relations	Avg.Len.
SemEval2010 Task8	10,717	7,984	10	18.84
FewRel	56,000	72,954	80	24.95

Table 4: Statistics of two datasets. "Avg.Len." means the average length of sentences.

B Implementation Details

We implement RCL based on `Transformers` package[4] (Wolf et al., 2020), where we use *Bert-base-uncased* as backbone. We set the maximum input length to 96 for SemEval2010 Task8 and 80 for FewRel. The epoch is set to 6 for training and we use an Adam optimizer (Kingma and Ba, 2014) with a batch size of 32. The learning rate is set to 5e-5 and the weight decay is set to 0.1. Same as SimCSE (Gao et al., 2021), the dropout probability of data augmentation is set to 0.1. The temperature τ is set to 0.05 across two datasets and we set α to 0.4 and 0.6 on SemEval2010 Task8 and FewRel respectively. The hidden size of fully-connected layer is set to 1536. All experiments are conducted by using a GeForce RTX 3090Ti with 24 GB memory.

C Standard F1 score Results on Two Datasets

To comprehensively compare the performance of baselines and our method, we report the standard F1 score results in Table 5 and Table 6. We follow the setting in the previous work (Simon et al., 2019) to convert pseudo labels predicted by clustering to relation labels. In each cluster, the relation label with the largest proportion among the cluster is assigned to all samples as the prediction label. From the standard F1 score results, we can see that the performance of our method is much better than the baselines, especially in SemEval2010 Task8. For FewRel, we find that the improvement of RCL gets larger when m is larger, especially when $m = 40$. It is obvious that it becomes more difficult for prediction since the number of unseen relations increases leading to more seriously similar problems. It proves that our method can effectively mitigates two types of similar problems.

[4] https://github.com/huggingface/transformers

SemEval2010 Task8			
Model	**P**	**R**	**F1**
CNN	46.83	48.53	47.52
Att-BiLSTM	47.46	51.87	49.41
Supervised RSN	32.87	40.12	36.07
MTB-BERT	54.42	58.93	56.29
MTB-RoBERTa	53.28	55.09	54.13
ZS-BERT	34.0	41.88	37.25
RCL w/o RC	61.41	60.91	61.13
RCL	**79.6**	**79.91**	**79.72**

Table 5: Experimental results(%) on SemEval2010 Task8 in terms of standard precision, recall, F1 score.

FewRel						
Model	**m=5**	**m=10**	**m=15**	**m=30**	**m=40**	**Avg.**
CNN	83.88	68.19	67.97	49.3	45.51	62.97
Att-BiLSTM	88.11	80.10	77.19	59.29	54.5	71.84
Supervised RSN	83.28	49.43	44.24	23.22	14.68	42.97
MTB-BERT	92.82	84.76	83.43	68.99	65.46	79.09
MTB-RoBERTa	**94.57**	86.25	85.21	69.66	66.49	80.43
ZS-BERT	82.83	77.57	70.93	53.68	55.75	68.15
RCL w/o RC	78.63	72.6	67.63	49.71	41.5	62.02
RCL	94.08	**87.46**	**85.95**	**75.2**	**72.64**	**83.07**

Table 6: Experimental results(%) on FewRel in terms of standard F1 score.

D More Ablation Studies

The effects of hyper-parameters are shown in Table 7 and Table 8. For hyper-parameter α, we vary α in the list of [0.0, 0.2, 0.4, 0.6, 0.8, 1.0] and find RCL can achieve the best performance when α is set to 0.4 on SemEval2010 Task8 or 0.6 on Fewrel. For temperature hyper-parameter τ, we vary τ in the list of [0.001, 0.01, 0.05, 0.1, 1.0] and find $\tau = 0.05$ can achieve the best performance across two datasets.

E Different Clustering methods for Zero-shot Prediction

Figure 5 shows the results of different clustering methods for RCL, including Mini-Batch K-Means (Sculley, 2010), Gaussian Mixture Model

α	0.0	0.2	0.4	0.6	0.8	1.0
SemEval2010 Task8	45.71	65.45	**68.02**	67.08	66.00	64.46
FewRel	81.08	82.32	82.94	**84.52**	83.35	83.11

Table 7: Experimental results(%) with different α in term of B^3 F1 score. For FewRel, we report the results on the unseen relation number $m = 10$.

τ	0.001	0.01	0.05	0.1	1.0
SemEval2010 Task8	44.99	47.18	**68.02**	61.85	42.23
FewRel	82.23	83.51	**84.52**	83.11	77.29

Table 8: Experimental results(%) with different temperatures over two datasets in term of B^3 F1 score. For FewRel, we report the results on the unseen relation number $m = 10$.

(GMM), Hierarchical Agglomerative Clustering (HAC), Birch (Zhang et al., 1996), K-Means. We can find that the performance of K-Means is much better than other clustering methods on two datasets. Moreover, Mini-Batch K-Means still outperforms MTB-BERT on SemEval2010 Task8, even its performance is worse than other clustering methods, showing the effectiveness of our model.

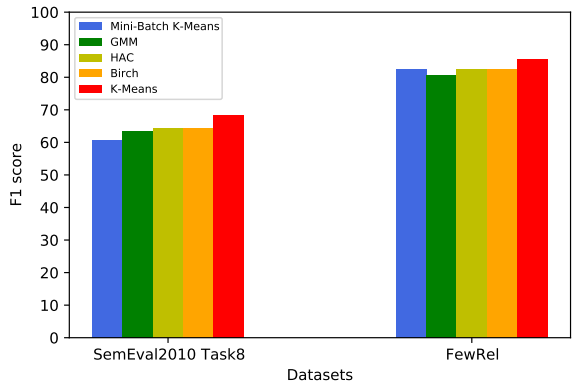


Figure 5: Different clustering methods for our proposed RCL model on two datasets. For FewRel, the number of unseen relations is set to 10.

F Case Study

To intuitively show how RCL helps to solve two types of similar problems (similar relations and similar entities), we conduct some case studies on two datasets. As shown in Figure 6, it is clear to see that RCL effectively solves these two problems under the multi-task learning structure. Specifically, RCL can better represent two sentences which have similar relations or similar entities, and then make their euclidean distance closer to the cluster corresponding to their ground truth.

SemEval 2010 Task8	Sentence	Label	Cluster Center	Euclidean Distance		
				ZS-BERT	MTB-BERT	RCL
Similar Relations	The envelope contained an important intelligence discovery of the war.	Content-Container	**Member-Collection**	**12.08**	**9.92**	8.55
			Content-Container	12.18	12.42	**8.36**
	The kitchen holds patient drinks and snacks.	Content-Container	**Member-Collection**	**10.41**	**8.68**	8.52
			Content-Container	11.71	10.66	**8.41**
Similar Entities	***Group1:*** China has tested Barack Obama early in his presidency, with a flotilla of naval vessels surrounding and harassing a US spy ship in the South China Sea.	Member-Collection	Member-Collection	**11.59**	**7.02**	**6.83**
			Instrument-Agency	11.83	12.41	8.30
	Group1: Until 1864 vessels in the service of certain UK public offices defaced the Red Ensign with the badge of their office.	Instrument-Agency	**Member-Collection**	**11.84**	**8.62**	8.20
			Instrument-Agency	13.74	11.02	**7.70**
	Group2: The puppy was inside a sealed garbage bag lying in vomit and near death.	Content-Container	Content-Container	**9.65**	**7.83**	**7.00**
			Entity-Origin	16.97	13.69	8.93
	Group2: The puppy was born in a barn where Layla made a soft, bed out of hay in an empty horse stall.	Entity-Origin	**Content-Container**	**13.49**	**9.29**	10.14
			Entity-Origin	15.49	11.78	**9.37**

FewRel	Sentence	Label	Cluster Center	Euclidean Distance		
				ZS-BERT	MTB-BERT	RCL
Similar Relations	The Doctor tries to restore the universe with the help of River and the alternative universe versions of his companions Amy Pond (Karen Gillan) and Rory Williams (Arthur Darvill).	part_of	**member_of**	**16.55**	**26.13**	28.76
			part_of	17.90	26.42	**26.06**
	Later in the game, she joins Snake in rescuing Dr Marv, but dies when Jaeger (as Gray Fox in Metal Gear D) destroys the bridge she is on.	part_of	**member_of**	**16.10**	**25.63**	31.34
			part_of	16.85	26.10	**27.19**
Similar Entities	***Group1:*** In May 2015, Vienna hosted the Eurovision Song Contest following Austria's victory in the 2014 contest.	follows	follows	**13.57**	**7.52**	**11.41**
			part_of	16.11	26.01	26.01
	Group1: Thus, the song was succeeded as Romanian representative at the 2002 Contest by Monica Anghel & Marcel Pavel with "Tell Me Why".	part_of	**follows**	**14.89**	**25.46**	28.22
			part_of	18.11	26.54	**27.78**
	Group2: On 1 September 1939, the Second World War began with the German Invasion of Poland, and two days later the United Kingdom declared war on Germany.	part_of	part_of	**15.16**	**22.98**	**21.81**
			follows	18.64	30.77	28.90
	Group2: During the War of 1812, Rolette, like many other French-Canadian Fur Traders in the Old Northwest, was an active supporter of the British Empire against the United States.	follows	**part_of**	**15.15**	**24.63**	25.00
			follows	18.13	25.28	**23.54**

Figure 6: Case study of similar relations and similar entities on two datasets. "Euclidean Distance" is the euclidean distance between the relation representation of input sentence and the cluster center of the relation. The target entities of input sentence are marked in orange.