Active Relation Discovery: Towards General and Label-aware OpenRE

Anonymous ACL submission

Abstract

Open Relation Extraction (OpenRE) aims to discover and label novel relations from open domains. Previous methods mainly suffer from two problems: (1) Insufficient capacity to discriminate between known and novel relations. When extending conventional test settings to a more *general setting* where test data might also come from seen classes, existing OpenRE approaches have a significant performance decline. (2) *Secondary labeling* must be performed before practical application. Existing methods cannot label human-readable and meaningful types for novel relations, which is urgently required by the downstream tasks. To address these issues, we propose the **A**ctive **R**elation **D**iscovery (**ARD**) framework, which utilizes relational outlier detection for discriminating known and novel relations and involves active learning for labeling novel relations. Extensive experiments[1] on three real-world datasets show that ARD significantly outperforms state-of-the-art methods on both conventional and our proposed general OpenRE settings.

1 Introduction

Open Relation Extraction (OpenRE) aims at discovering and extracting potential novel relations from open-domain corpora. Novel relations are cropping up at a rate of tens of thousands per year (Shi and Weninger, 2018), while most of the rapidly emerging relations are still unlabeled and under-explored, mixed with pre-defined relations. These relations cannot be well handled by supervised RE methods due to the fixed predefined relation schema.

Some recent preliminaries have noticed the challenge of learning emerging relations and explored methods for OpenRE. Previous works can be divided into two main paradigms: pattern-based and clustering-based methods. Specifically, pattern-based method (Angeli et al., 2015; Cui et al., 2018) utilize statistical or neural approaches to heuristically extract relation patterns, then clustering-based methods (Elsahar et al., 2017; Wu et al., 2019) are proposed to aggregate instances representing the same novel relation.

However, previous works mainly have two shortcomings in real scenarios:

(1) The widely used traditional setting can't comprehensively reflect what OpenRE in the real world entails. The traditional setup for OpenRE is that models are judged based on their ability to discriminate among unseen classes, assuming the absence of known relation during the test phase. While the ability to learn novel relations is, by all means, a trait that any OpenRE model should possess, it is merely one side. The other important, yet so far unstudied, trait is the ability to discriminate between the known and unseen relations. The relation distribution in the real world is intricate, mixed with known and unseen relations. Therefore, it's unrealistic to assume that we will never encounter known relations during the test stage.

In the light of above facts, we evaluate existing OpenRE model on a *General OpenRE* setting: test data might also come from known relations. Empirical experiments in Table 1 show that the state-of-the-art OpenRE model (Wu and Weld, 2010; Hu et al., 2020; Zhang et al., 2021) performs poorly under this setting.

(2) The results produced by current OpenRE models require *secondary labeling* before they can be practically applied. In other words, for a certain novel relation, the model cannot assign it a surface name with a specific meaning. As the foundation of a series of downstream tasks, labels with actual meaning are urgently desired. However, due to the absence of human knowledge, both pattern-based and clustering-based methods lack the ability to name novel relation types as human-readable and meaningful. Pattern-based methods rely heavily on the surface phrase, yet relations between entities

[1] The source code will be available for reproducibility.

are often not directly represented by the span in the sentence. Clustering-based methods merely cluster instances that express the same relations, but do not provide concrete representation of the novel relations. Both methods require manual re-labeling of the novel relations found. This gap between model and practice hinders model application in real-world scenarios.

To address above mentioned issues, we propose the **Active Relation Discovery (ARD)** framework shown in Figure 1. Targeted improvements are made in two aspects: (1) To avoid the model being confused by the set of mixed known and novel relations, we developed a relational outlier detection algorithm that separates known and novel relations by treating novel relations as outliers, performing stably under the General OpenRE setting.

(2) To assign meaningful labels to novel relations, the incorporation of human knowledge is inevitable. To minimize the labor cost, we propose an active learning algorithm. Specifically, we introduce the *representative instance*, which denotes an instance can offer rich information of unknown relations. Only a handful of representative instances requires manual labeling, and then the model can automatically label the novel relations in a supervised manner.

In summary, our contributions are in three folds:

(1) We reveal two major shortcomings of current OpenRE approaches, and introduce a new setting called *General OpenRE*, which can realistically measure the capabilities of the model.

(2) We propose ARD, a practical framework that not only adapts to the General OpenRE utilizing relational outlier detection, but also exploits active learning to assign more meaningful and human-readable labels to novel relations.

(3) We conduct extensive experiments on both conventional and General OpenRE settings to show that our framework can achieve significant improvements in three real-world datasets. Detailed analysis demonstrates the effectiveness of each part of ARD.

2 Related Work

Open Relation Extraction. Whereas supervised RE (Liu et al., 2013; Zhang and Wang, 2015) relies heavily on manual annotation and the inherent inadequacy of predefined relation schema, OpenRE gains increasing attention. The method of OpenRE can be broadly divided into two categories: pattern-based and clustering-based. Pattern-based approaches extract relation patterns from textual corpora (Banko et al., 2007; Fader et al., 2011; Stanovsky and Dagan, 2016). These methods apply heuristic algorithms to describe relations between marked entities with relation patterns consisting of several key phrases in texts. Due to the ambiguity of relations obtained by the pattern-based methods, the focus of research in recent years has been primarily on clustering-based methods.

Clustering-based method (Shinyama and Sekine, 2006; Elsahar et al., 2017; Wu et al., 2019) cluster instances in the feature space into novel relation types. Wu et al. (2019) enhances unsupervised clustering-based methods by introducing Siamese Network to measure instance similarity.

As described in Section 1, there are two main problems with the current OpenRE: (1) They focus only on the discrimination of novel relations, supposing that test sets only have novel relations. (2) The model output is not directly usable by downstream tasks. In response, we propose a General OpenRE setup and incorporate outlier detection and active learning into OpenRE.

Active Learning in Relation Extraction. The key idea behind active learning (Settles, 2009) is that the learning algorithm is allowed to ask for true meaningful labels of some selected unlabelled instances. Various criterion (Zhang et al., 2012; Fu and Grishman, 2013; Qian et al., 2014) have been proposed to choose these instances on traditional supervised RE tasks. To our best knowledge, we firstly integrate active learning into OpenRE, enabling meaningful tags of the novel relation type with the addition of human knowledge.

Generalized Zero-Shot Learning(GZSL). The motivation for the General OpenRE setting is similar to that of the GZSL. Traditionally, ZSL approaches (Romera-Paredes and Torr, 2015; Zhang and Saligrama, 2015) assume that only the unseen classes are present in the test set. (Chao et al., 2016) first challenged this implausible setting and proposed the GZSL setting: test data might also come from seen classes. GZSL approaches (Rahman et al., 2018; Huang et al., 2019) focus on mitigating the strong bias caused by known classes and preventing novel classes from being categorized as one of the seen classes. While in our General OpenRE setting, we concentrate more on the distinction between known and novel classes.

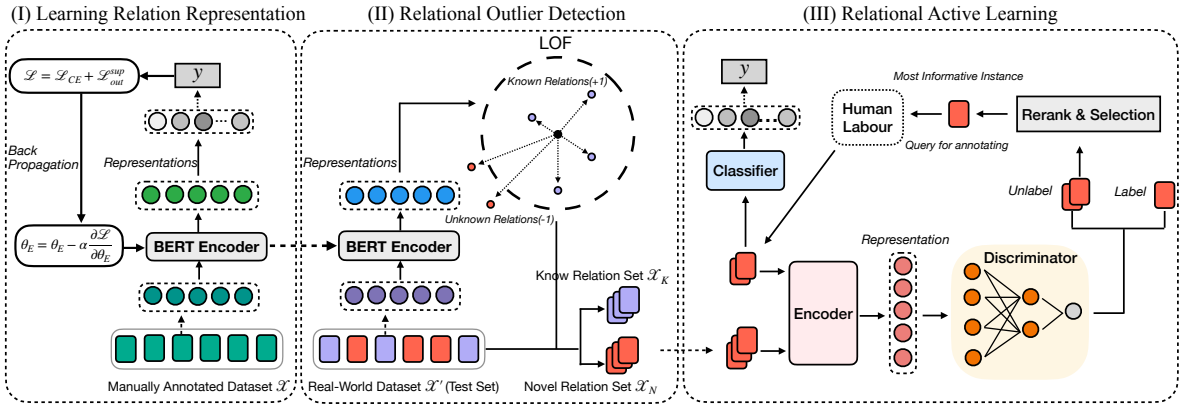


Figure 1: An illustration of our proposed Active Relation Discovery (ARD) framework.

3 Task Formulation

General OpenRE formulates the task slightly differently from traditional OpenRE setting. The original train set is a large-scale manually annotated corpus $\mathcal{X} = \{x_j^{r_i} | r_i \in \mathcal{R}_K\}$, where relations in $\mathcal{R}_K$ are pre-defined as "known relations". Obviously, we assume that there exists a relation set $\mathcal{R}_N$ that contains "novel relations" in another corpus without annotations. In the real-world scenario, we need to process the dataset whose instances express relations both in $\mathcal{R}_K$ and $\mathcal{R}_N$, distinguish known and novel relations, then label each instance.

Under this fact, we first consider the *novel relation discovery*, in which we solely focus on the mining of unseen relations. At this stage, We pre-train the model on $\mathcal{X}$ and obtain a trained encoder E. Then for a concrete dataset (test set) $\mathcal{X}' = \{x_j^{r_i}, x_j^{\prime r_i'} | r_i \in \mathcal{R}_K, r_i' \in \mathcal{R}_N\}$. The model will unsupervisedly divide $\mathcal{X}'$ into a "known relation set" $\mathcal{X}_K$ and a "novel relation set" $\mathcal{X}_N$.

$\mathcal{X}_K$ can be easily labeled for sufficient information obtained from $\mathcal{X}$. Secondly, we focus on the *annotation of novel relations* $\mathcal{X}_N$. In this phase, we integrate the intuition of active learning by utilizing limited labor to facilitate the novel relation annotation performance. Our model queries a small set of informative samples in $\mathcal{X}_N$ for manual labeling and then trains a classifier to annotate novel relations.

4 Methodology

4.1 Overview

The overview of the method is illustrated in Figure 1. We will detailedly introduce our work into three components: (1) **Relation representation**, in which we extend to transform semantic relations into low-dimension dense representations. (2) **Relational Outlier Detection**, where the model automatically detects a novel relation set from real-world datasets and feeds them into the active learning stage. (3) **Relational Active Learning**, where the model selects the most informative instances to train a powerful classifier for novel relation.

4.2 Relation Representation

Given a dataset $X = \{x_1, ..., x_n\}$, an instance x is a word (token) sequence $\{w_1, w_2, ..., w_n\}$ with two marked entities e_h and e_t. We use triplets of relation facts (e_h, r, e_t) to denote that there is a relation r between the marked entity pair. And x^r indicates an instance that expresses the relation r. Specifically, we define four special markers $\langle e_h \rangle$, $\langle /e_h \rangle$, $\langle e_t \rangle$, and $\langle /e_t \rangle$ to locate the head entity and the tail entity. We denote the indices of $\langle e_h \rangle$ and $\langle e_t \rangle$ as START(h) and START(t). An instance is represented as:

$$x = ..., \langle e_h \rangle, w_{\text{START}(h)+1}, ..., w_{\text{END}(h)}, \langle /e_h \rangle, ..., \langle e_t \rangle, w_{\text{START}(t)+1}, ..., w_{\text{END}(t)}, \langle /e_t \rangle, ... \quad (1)$$

We use pre-trained language model (i.e. BERT (Devlin et al., 2019)) to encode each token w_t to the corresponding representation $\boldsymbol{h}_t \in \mathbb{R}^d$, where d is denotes the dimension of representation vectors.

For an instance $x_i \in S$, we use the concatenation of representations of two start positions ($w_{\text{START}(h)}$ and $w_{\text{END}(h)}$) as the representation of the relation:

$$\boldsymbol{h}_r(x_i) = [\boldsymbol{h}_{\text{START}(h)}, \boldsymbol{h}_{\text{START}(t)}], \quad (2)$$

These extra tokens play a similar role like position embeddings in conventional RE tasks (Zeng et al., 2015). The relation representation $\boldsymbol{h}_r(x_i)$ will be utilized to predict the relation type r.

As mentioned previously, $\mathcal{X}$ are used to fine-tune the pre-trained language model. Notably, along

with the traditional cross-entropy loss, we integrate a supervised contrastive loss $\mathcal{L}_{\text{out}}^{\text{sup}}$ described in Khosla et al. (2020):

$$\mathcal{L}_{\text{out}}^{sup} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_p / \tau\right)}{\sum_{a \in A(i)} \exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_a / \tau\right)}, \tag{3}$$

Here, $P(i) \equiv \{p \in A(i) : \tilde{\boldsymbol{y}}_p = \tilde{\boldsymbol{y}}_i\}$ is the set of indices of all positives in the mini-batch distinct from i. Contrastive loss allows for tighter clustering of intra-class instances and a more dispersed distribution of inter-class instances. The essence behind the employment of contrastive loss is to gain relation representations that are more friendly to outlier detection and active learning. The performance of our relation representation on supervised RE can also be found in Appendix F.

4.3 Relational Outlier Detection

After pre-training, $\mathbb{E}_\theta$ could encode an instance x into a dense vector $\boldsymbol{h}_r(x)$ as the relation representation. In the feature space, due to the similarity of the semantics, representations that express the same relation tend to densely gather (forming n separate clusters) and ones that express different relations tend to disperse. Figure 3 illustrated the distribution of different representations. Since the instances express unseen relations have not been pre-trained, in other words, the model has not seen the semantics, the instances are not projected near any clusters. We utilize this property to design local outlier factor (LOF) to reflect the local density of instances in the feature space.

Formally, given any two representations $\boldsymbol{h}_r(x_i), \boldsymbol{h}_r(x_j)$ of instances x_i, x_j, we denote $d(\boldsymbol{h}_r(x_i), \boldsymbol{h}_r(x_j))$ as the Euclidean distance between them. Then, we define k-th distance, denoted as $d_k(\boldsymbol{h}_r(x_i))$, to represent the distance from $\boldsymbol{h}_r(x_i)$ to the k-th nearest neighbour. The reachability distance between $\boldsymbol{h}_r(x_i)$ and $\boldsymbol{h}_r(x_j)$ is represented as follows:

$$\begin{aligned} \text{rd}_k(\boldsymbol{h}_r(x_i), \boldsymbol{h}_r(x_j)) = \max\{&d_k(\boldsymbol{h}_r(x_j)), \\ &d(\boldsymbol{h}_r(x_i), \boldsymbol{h}_r(x_j))\}, \end{aligned} \tag{4}$$

We then compute the density to measure the average distance of reach-ability distance:

$$\text{den}_k(h_r(x_i)) = 1 / \frac{\sum_{h_r(x_j) \in N_k(h_r(x_i))} \text{rd}_k(h_r(x_i), h_r(x_j))}{|N_k(h_r(x_i))|}, \tag{5}$$

where $N_k(\boldsymbol{h}_r(x_i))$ denotes all the points whthin in k-th distance of $\boldsymbol{h}_r(x_i)$.

The computation of local outlier factor is:

$$\text{LOF}_k(\boldsymbol{h}_r(x_i)) = \frac{\sum_{\boldsymbol{h}_r(x_j) \in N_k(\boldsymbol{h}_r(x_i))} \frac{\text{den}_k(\boldsymbol{h}_r(x_j))}{\text{den}_k(\boldsymbol{h}_r(x_i))}}{|N_k(\boldsymbol{h}_r(x_i))|}, \tag{6}$$

where the larger LOF is, the more likely $\boldsymbol{h}_r(x_i)$ is an outlier point, i.e, an instance that expresses a novel relation. Our model could unsupervisedly detect the instances with novel relations.

4.4 Relational Active Learning

To this end, the model could divide the real-world dataset into a "known relation set" $\mathcal{X}_K$ and an "novel relation set" $\mathcal{X}_N$. In view of the fact that $\mathcal{X}_K$ can be conveniently and precisely annotated, we focus on labeling meaningful types for discovered instances in $\mathcal{X}_N$ in this subsection.

To retrieve human-readable labels and avoid subsequent secondary labeling, we need to incorporate human knowledge into the relation learning phase through active learning. Our primary goal is to find a small part of instances with the most information and artificially label them. Then we use the labeled data to train a classifier in a supervised manner. The problem of how to find instances with most information essentially is the problem of how to find the instances that are most likely to express "novel relations". Inspired by this, we propose the following Relation Active Learning module:

In the beginning, we randomly label a small part of data in $\mathcal{X}_N$. The labeled dataset is denoted as $\mathcal{X}_L$ and the rest of the unlabeled data is denoted as $\mathcal{X}_U$. We assume that all the instances x are i.d.d according to a latent distribution $P(x)$. Correspondingly, their labels are distributed by the conditional distribution $P(y|x)$.

Neural Encoder We adopt a neural encoder to learn the distribution of $\mathcal{X}_L$ and $\mathcal{X}_U$ in the latent feature space. Our framework is independent of the choice of neural encoders, in this case, we adopt BERT (Devlin et al., 2019) as the encoder. The goal of the neural encoder is to encode $\mathcal{X}_L$ and $\mathcal{X}_U$ into the same feature space and try to fool a discriminator to correctly predict if the instance is "representative". The loss function of the encoder is:

$$\begin{aligned} \mathcal{L}_e = &- \mathbb{E}_{x \sim P_{\mathcal{X}_L}}[\log(D_\psi(E_\theta(x)))] \\ &- \mathbb{E}_{x \sim P_{\mathcal{X}_U}}[\log(1 - D_\psi(E_\theta(x)))], \end{aligned} \tag{7}$$

Discriminator A binary classifier (or a discriminator): $\mathcal{X} \to \{-1, 1\}$ is adopted to select the most

informative samples. We utilize adversarial training to leverage the information of both $\mathcal{X}_L$ and $\mathcal{X}_U$. The discriminator is adversarially trained to accurately distinguish if the instance expresses a novel relation. The loss function is a flipped version of the encoder:

$$\begin{aligned}\mathcal{L}_d = &- \mathbb{E}_{x\sim P_{\mathcal{X}_L}}[\log(1 - D_\psi(E_\theta(x)))] \\ &- \mathbb{E}_{x\sim P_{\mathcal{X}_U}}[\log(D_\psi(E_\theta(x)))],\end{aligned} \tag{8}$$

Naturally, we could jointly optimize the two objective functions by allocate two parameters: $\mathcal{L} = \lambda\mathcal{L}_e + \lambda'\mathcal{L}_d$.

Active learning At each training step, we select k instances with the highest confidence of the discriminator as the most informative instances. Then the instances will be manually annotated and then used to train the classifier. At this point, the discussion of manual annotations needs to be further developed. Considering the explosive growth of the number of relations, an annotating process that supports online and continual learning of novel relations needs to be designed. Thus, we propose a practical and easy-to-implement annotation procedure. At the start, for each selected instance x_i, the annotator only needs to judge if x_i has the same relation class as any instances of $\mathcal{X}_L$. x_i will be indexed as a novel relation if it doesn't share the same relation with instances in $\mathcal{X}_L$, or labeled as one known relation. After the procedure, the standard of relations would be easy to design than before the active learning begins. This manner effectively ensures the ability to continual learning and online learning of our framework, expediently fitting the real situation. Subsequently, $\mathcal{X}_L$ will be fed into a classifier, which is a one-layer MLP with an output layer, optimized by cross-entropy objective function, denoted as $\mathcal{L}_c$ and parameterized by γ:

$$\mathcal{L}_c = \sum_{i\in|\mathcal{X}_L|} -\log p(y_L^{(i)}|x_L^i, \gamma). \tag{9}$$

5 Experiments

In this section, we verify the performance of the model on three large-scale OpenRE datasets and their variants, and at the same time, a series of auxiliary experiments are carried out to prove the effectiveness of the model. Finally, we give a detailed analysis of the efficacy of our ARD framework.

Algorithm 1 Training for Active Learner, λ, λ', k are hyper-parameters.

Input: Labeled data $(\mathcal{X}_L, Y_L)$, unlabeled data $\mathcal{X}_U$, initialized encoder model with θ, discriminator model with ψ, classifier with γ
while not converge **do**
  Sample mini-batches (x_L, y_L) from $(\mathcal{X}_L, Y_L)$
  Sample mini-batches (x_U) from $(\mathcal{X}_U)$
  Compute $\mathcal{L}_e$ by Eq. 7
  Update θ w.r.t $\mathcal{L}_e$
  Compute $\mathcal{L}_d$ by Eq. 8
  Update ψ w.r.t $\mathcal{L}_d$
  Select k most informative instances $\{x_1, ..., x_k\}$ by the output of d
  for $i \leftarrow 1$ to k do **do**
    if x_i has the same relation as $x_j^r \in \mathcal{X}_L$ **then**
      Label x_i with r and append x_i to $\mathcal{X}_L$
    else
      Label x_i with a new index and append x_i to $\mathcal{X}_L$
    end if
  end for
  Update γ w.r.t $\mathcal{L}_c$
end while

5.1 Baseline

To demonstrate the effectiveness of our ARD models, we compare our models with three state-of-the-art models: (1) **RSN-CV** (Wu et al., 2019) employs conditional entropy and virtual adversarial learning to train Siamese Network to measure instance similarity. (2) **SelfORE** (Hu et al., 2020) utilizes self-training to iteratively learn relation representations and clusters. (3) **OHRE** (Zhang et al., 2021) integrate hierarchy information into relation representations for better novel relation extraction.

5.2 Datasets and Setting

Datasets Three datasets and their variants are used to evalutate our model: FewRel (Han et al., 2018), New York Times Freebase(NYT+FB) (Marcheggiani and Titov, 2016) and FewRel2.0 (Gao et al., 2019), the first two of which have been widely used in previous RE works (Simon et al., 2019; Hu et al., 2020; Zhang et al., 2021). We follow the division of the datasets from previous works. More details about the dataset can be found in the Appendix A.

To verify the cross-domain capability of the model comprehensively, we also use FewRel2.0

dataset whose training and test sets are from completely different domains. As an advanced version of FewRel, FewRel2.0 incorporates knowledge transferring.

Datasets Processing As described above, in the original OpenRE setting, there are no overlapping relations in the training and test sets. The relations in the test set are all novel relations. To measure the performance of the model in our proposed *General OpenRE* setting, we resample the original dataset and gain two variants: *noisy* and *imbalanced*. In the test sets of the two variants, there are known relations with different distributions exist. In other words, the original dataset corresponds to the conventional setting and the noisy and imbalanced variants to the general setting.

To obtain the noisy variant, we randomly select 40% samples from original training sets. Given that the number of samples for each novel relation is identical in FewRel and FewRel2.0, we further construct the imbalanced variant to explore the performance of the model in the presence of class imbalance. Specifically, we build on the noisy variant by randomly discarding a portion of the samples with different probabilities for each relation class in the test set, yielding class imbalance in test set.

5.3 Evaluation Settings

Following previous works, we apply instance-level evaluation metrics to evaluate the model, covering B^3 (Bagga and Baldwin, 1998), V-measure (Rosenberg and Hirschberg, 2007) and Adjusted Rand Index(ARI) (Hubert and Arabie, 1985).

For quantitative validation, we divide $\mathcal{X}_N$ into $\mathcal{X}_N^{train}$ and $\mathcal{X}_N^{test}$, which account for 40% and 60% respectively. The active learning module selects the instance with the most information in $\mathcal{X}_N^{train}$ and trains the relation classifier. In the test phase, we merge $\mathcal{X}_K$ and $\mathcal{X}_N^{test}$, report metric scores on it. As the baselines are semi-supervised, $\mathcal{X}_N^{train}$ is also applied to the training of the baseline models to ensure a fair comparison.

For FewRel and NYT+FB, the seminal set size for Active Learning module is 32. The sample size k is 32 and we sample a total of 8 epochs. In other words, a whole of 288 samples is manually labeled. As for FewRel2.0, we choose a smaller sample size: $k = 8$ and keep semianl set size as 32. Finally, 96 informative samples are annotated.

5.4 Main Experiment

Table 1 shows the quantitative evaluation results on three datasets and their variants, from which we observe that: (1) Our ARD model outperforms state-of-the-art models by a large margin. Specifically, B^3, V-measure and ARI increased by 15.2, 16.7, and 15.9 respectively compared to OHRE on FewRel. Compared with other semi-supervised methods, the gap is even larger, rises of over 25 are achieved by ARD. This proves that ARD can efficiently discover and learn representations of novel relations at a fraction of the labor cost. (2) An universal and consistent decline in performance of baseline models from the original datasets to noisy variants and then to unbalanced variants. This demonstrates that the *General OpenRE* setting is more challenging and more practical for the real scenario. The $F1$ score for RSN-CV drops dramatically from the original data to the noisy variant by 19.5. In contrast, the ARD model outperforms on both the noisy and imbalanced variants than on the original dataset, even with a $F1$ score boosting by 7.2 on FewRel. This indicates the relation discovery procedure and relational active learning is robust in different scenarios. (3) The state-of-the-art models perform poorly on FewRel2.0. This is entirely to be expected, as the instances in the test set are from non-generic and low-resource domains such as biomedicine. ARD, on the other hand, still shows strong stability, confirming the cross-domain capability of the model. Further, to substantiate the applicability of our framework, we perform deploy ARD to a real medical dataset, as detailed in Appendix D.

5.5 Analysis on Active Learning

The Efficiency of Active Learing Table 2 shows the results of our active learning approach compared to various active learning baseline models including DBAL (Gal et al., 2017), CoreSet (Sener and Savarese, 2018), SRAAL (Zhang et al., 2020). It can be observed that in each iteration, our model outperforms the other models, indicating that our method can consistently sample informative samples. Time efficiency analysis and case study can be found in the Appendix G and Appendix C.

The Impact of Different Encoder and Scope of Query Figure 2 shows the experimental results on noisy FewRel with different encoders and query ranges. The "query ranges" represents the ratio of $\mathcal{X}_N^{train}$ to $\mathcal{X}_N$, We also explore the impact of β-

Data -set	Model	B^3			V-measure			ARI
		F1	Prec.	Rec.	V	Hom.	Comp.	
FR (Ori)	RSN-CV	57.3	51.5	64.5	69.4	66.3	72.8	44.9
	SelfORE	57.5	57.4	57.6	63.1	63.3	63.0	53.8
	OHRE	58.4	48.6	73.2	68.4	62.3	75.7	48.2
	Ours	**73.6**	**70.7**	**76.8**	**85.1**	**84.9**	**85.3**	**64.1**
NYF (Ori)	RSN-CV	45.1	35.0	63.6	61.1	54.0	70.3	31.7
	SelfORE	54.3	52.0	56.9	71.0	70.2	71.8	54.6
	OHRE	36.1	23.8	75.0	52.8	41.2	73.4	24.5
	Ours	**51.4**	**45.0**	**60.0**	**72.3**	**75.0**	**69.8**	**45.1**
FR2.0 (Ori)	RSN-CV	26.7	17.2	59.4	48.1	38.3	64.8	10.0
	SelfORE	35.3	24.6	62.5	60.0	51.9	71.3	27.0
	OHRE	21.3	14.9	36.9	12.9	10.4	16.8	6.6
	Ours	**48.8**	**43.2**	**56.1**	**65.1**	**60.8**	**70.1**	**34.4**
FR (Noi)	RSN-CV	37.8	25.8	70.7	61.8	51.5	77.3	24.4
	SelfORE	50.8	48.2	53.7	55.4	55.7	55.1	42.0
	OHRE	28.3	17.4	75.9	56.8	44.8	77.7	21.3
	Ours	**80.8**	**75.7**	**86.7**	**90.2**	**89.0**	**91.4**	**71.3**
NYF (Noi)	RSN-CV	40.9	29.5	66.7	58.7	49.7	71.5	28.8
	SelfORE	46.2	41.8	51.5	65.8	64.0	67.7	44.5
	OHRE	32.6	21.8	64.7	44.4	34.5	62.3	24.3
	Ours	**71.3**	**60.8**	**86.2**	**72.9**	**70.5**	**75.5**	**51.0**
FR2.0 (Noi)	RSN-CV	21.4	27.1	17.6	25.0	26.3	23.9	7.3
	SelfORE	31.7	23.3	49.8	53.5	46.8	62.4	25.4
	OHRE	20.7	12.6	57.9	46.8	36.8	64.3	11.2
	Ours	**55.0**	**52.8**	**57.4**	**69.3**	**65.1**	**74.0**	**38.5**
FR (Imb)	RSN-CV	35.9	23.5	75.4	61.4	50.0	79.5	22.6
	SelfORE	46.4	42.8	50.6	51.8	52.1	51.5	36.9
	OHRE	23.9	14.3	73.1	52.4	40.5	74.3	18.3
	Ours	**76.5**	**74.7**	**78.4**	**86.5**	**86.8**	**86.2**	**67.8**
FR2.0 (Imb)	RSN-CV	20.5	14.8	33.5	26.5	21.6	34.4	7.9
	SelfORE	30.9	22.0	52.0	52.8	45.4	63.2	23.8
	OHRE	20.0	12.0	59.7	43.6	33.4	62.7	10.5
	Ours	**52.4**	**50.1**	**54.9**	**67.4**	**63.2**	**72.2**	**36.4**

Table 1: Main results on three original datasets and their variants. Ori, Noi, and Imb stand for original, noisy and imbalanced respectively. Ori corresponds to the conventional setting. Noi, and Imb refer to the general setting. Results are the average of 3 experiments with different random seeds.

VAE (Higgins et al., 2016) and BERT as encoders. As well, we report the results when using a random selection of 30% and the full amount of $\mathcal{X}_N^{train}$ for training. From the results we observe that: (1) Generally, the model performance is proportional to the size of $\mathcal{X}_N^{train}$. However, the results are improved marginally as the number of samples increase. But the model still yields better performance when the query range is 40%. (2) The comparisons between the VAE and the BERT encoder are in line with intuition. Although VAE is intuitive and can be more easily trained, BERT still shows superiority in empirical results. (3) The performance of the model trained with 288 samples (approximately

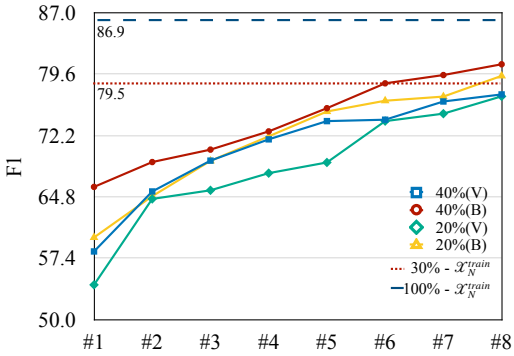


Figure 2: F1-measure on noisy FewRel, (V) denotes the β-VAE and (B) denotes the BERT encoder.

Dataset	Model	Epoch					
		#1	#2	#3	#4	#5	#6
FewRel	DBAL	58.8	64.8	71.6	70.4	74.1	76.9
	CoreSet	60.1	61.8	66.1	68.4	70.9	75.4
	SRAAL	61.9	64.7	65.7	69.8	73.7	73.9
	Ours	**66.0**	**69.0**	**70.5**	**72.7**	**75.5**	**78.5**
NYY+FB	DBAL	47.4	48.6	51.4	53.3	54.9	55.5
	CoreSet	45.4	49.5	52.0	55.3	56.8	59.2
	SRAAL	50.2	51.9	54.0	55.6	56.2	56.9
	Ours	**56.8**	**62.5**	**66.6**	**68.3**	**69.3**	**69.9**
FewRel2.0	DBAL	46.9	50.5	51.4	51.7	52.2	53.7
	CoreSet	44.0	45.5	50.3	51.9	53.0	54.2
	SRAAL	45.0	49.7	51.8	52.0	52.8	53.9
	Ours	**48.8**	**51.2**	**52.4**	**53.2**	**53.5**	**54.5**

Table 2: F1-measure for various active learning methods on noisy FewRel dataset.

8% of $\mathcal{X}_N^{train}$) is similar to that of a random selection of 30%. When trained with the full amount of $\mathcal{X}_N^{train}$, the $F1$ is 6.1% higher than ARD while costing 12 times as much in human effort. The results demonstrate the effectiveness of ARD.

The Impact of Different Sampling Strategies In order to prove the effectiveness of the active learning method, we conduct a further ablation experiment. As mentioned above, our sampling strategy is to select the k instances with the highest confidence for manual labeling. In the ablation experiment, we test two other sampling strategies: selecting the k instances with the lowest confidence; randomly selecting k instances. The comparison results are shown in Table 3.

It can be seen that after being trained by instances with the highest confidence, the model achieves the most improvement. In contrast, instances with the lowest confidence contribute very little to improving the performance of the model. Even with the continuous increase of training data, the improvement is extremely little. The results

Epoch	Lowest	Random	Highest
#1	57.1	58.7	66.0
#2	57.1	60.4	69.0
#3	57.8	65.6	70.5
#4	57.6	67.2	72.7
#5	57.6	67.7	75.5
#6	58.1	67.7	78.5

Table 3: Comparisons of F1-measure between different sampling strategies on noisy FewRel dataset.

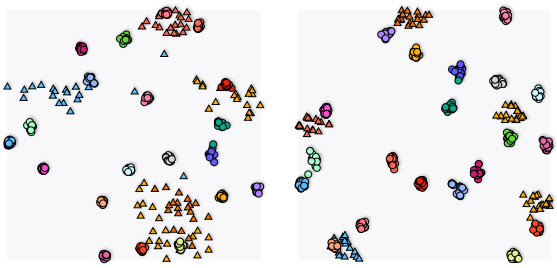
Dataset	Model	Epoch				
		#1	#2	#3	#4	#5
FewRel	ARD	66.0	69.0	70.5	72.7	75.5
	w/o LOF	62.8	64.4	68.5	70.5	73.4
NYT+FB	ARD	56.8	62.5	66.6	68.3	69.3
	w/o LOF	47.7	53.8	57.2	60.3	64.4
FewRel2.0	ARD	48.8	51.2	52.4	53.2	53.5
	w/o LOF	42.9	44.2	45.6	46.4	48.2

Table 4: Ablation experiment over novel relation discovery module on noisy FewRel dataset.

prove that the active learning model does select the most informative instances.

5.6 Analysis on Relational Outlier Detection

ARD employs novel relation discovery module to distinguish between known and novel relations, preserving the active learning module to more efficiently select informative novel relations without being distracted by known relations. To demonstrate the effectiveness and significance of the novel relation discovery, we perform ablation experiments over LOF algorithm on three noisy variants. Table 4 shows the experimental results, and we note that: (1) Despite the robust learning ability of active learning on novel relations, the model performances show different degrees of degradation after the removal of the LOF algorithm. (2) Average of decline of $F1$ scores in each epoch on the FewRel, NYT+FB, and FewRel2.0 datasets is 2.82, 8.02, 6.36 respectively, with the most severe drop on NYT+FB. The phenomenon is intuitive, as the NYT+FB dataset contains the most known relations; the more noise (known relations) there is, the more confused the active learning module becomes about the novel relations. The results demonstrate the novel relation discovery module plays a key role as "noise reduction". The impact of different outlier detection algorithms and qualitative analysis can be seen in Appendix E



(a) Train using only traditional cross-entropy loss. (b) Plus contrastive loss.

Figure 3: t-sne visualization of relation representation. The known and novel relations are distinguished by circular and triangular symbols respectively.

5.7 Visualization of Relation Representations

In order to intuitively demonstrate the distribution of novel relations relative to known relations and, on the other hand, to illustrate the benefits of introducing contrastive loss, we visualize the relation representation $\boldsymbol{h}_r(x)$ after dimension reduction using t-SNE (Maaten and Hinton, 2008).

As illustrated in Figure 3, instances of the same known relation type are densely clustered with a high local density, while instances of novel relations distribute dispersedly. This fact strongly supports the premise of the LOF algorithm. Also, comparing subfigures 3(a) and 3(b), we observe that contrastive loss firmly constrains the distribution of intra-class instances. In pre-experiments on FewRel, the introduction of contrastive loss boosts the accuracy in distinguishing known and novel relations from 79.3% to 83.9%.

6 Conclusion and Future Work

The paper proposes Active Relation Discovery (ARD), which aims at accurately discovering and meaningfully annotating new semantic relations under the *General OpenRE* setting. By introducing outlier detection and active learning, ARD solves two problems: (1) *Sufficient capabilities to distinguish between known and novel relations*, with robust performance under General OpenRE settings. (2) *Avoiding Secondary labeling of downstream tasks*. Extensive experiments are conducted to demonstrate the effectiveness of ARD.

As a pioneering work in OpenRE, several directions can be further explored: (1) Better methods to discriminate and annotate novel relations in *General OpenRE* setting. (2) Combination with bootstrapping methods to partially replace active learning. (3) Combination with lifelong learning to continuously incorporate novel relations.

References

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of ACL-IJCNLP*, pages 344–354.

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566. Citeseer.

Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *Proceedings of IJCAI*, pages 2670–2676.

Marcia Afonso Barros, André Lamúrias, Diana Sousa, Pedro Ruas, and Francisco M Couto. 2020. Covid-19: A semantic-based pipeline for recommending biomedical entities. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. 2016. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *European conference on computer vision*, pages 52–68. Springer.

Francisco M Couto and Andre Lamurias. 2018. Mer: a shell script and annotation server for minimal named entity recognition and linking. *Journal of cheminformatics*, 10(1):1–10.

Lei Cui, Furu Wei, and Ming Zhou. 2018. Neural open information extraction. In *Proceedings of ACL*, pages 407–413.

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.

Hady Elsahar, Elena Demidova, Simon Gottschalk, Christophe Gravier, and Frederique Laforest. 2017. Unsupervised open relation extraction. In *Proceedings of ESWC*, pages 12–16.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1535–1545. Association for Computational Linguistics.

Lisheng Fu and Ralph Grishman. 2013. An efficient active learning framework for new relation types. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 692–698.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192. PMLR.

Tianyu Gao, Xu Han, Hao Zhu, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2019. FewRel 2.0: Towards more challenging few-shot relation classification. 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 6250–6255, Hong Kong, China. 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. *arXiv preprint arXiv:1810.10147*.

María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. 2013. The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics*, 46(5):914–920.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2016. beta-vae: Learning basic visual concepts with a constrained variational framework.

Lixiang Hong, Jinjian Lin, Shuya Li, Fangping Wan, Hui Yang, Tao Jiang, Dan Zhao, and Jianyang Zeng. 2020. A novel machine learning framework for automated biomedical relation extraction from large-scale literature repositories. *Nature Machine Intelligence*, pages 1–9.

Xuming Hu, Lijie Wen, Yusong Xu, Chenwei Zhang, and S Yu Philip. 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.

He Huang, Changhu Wang, Philip S Yu, and Chang-Dong Wang. 2019. Generative dual adversarial network for generalized zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 801–810.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of classification*, 2(1):193–218.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33.

Sangrak Lim, Kyubum Lee, and Jaewoo Kang. 2018. Drug drug interaction extraction from the literature using a recursive neural network. *PloS one*, 13(1):e0190926.

ChunYang Liu, WenBo Sun, WenHan Chao, and Wanxiang Che. 2013. Convolution neural network for relation extraction. In *International Conference on Advanced Data Mining and Applications*, pages 231–242. Springer.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE.

Shengyu Liu, Buzhou Tang, Qingcai Chen, and Xiaolong Wang. 2016. Drug-drug interaction extraction via convolutional neural networks. *Computational and mathematical methods in medicine*, 2016.

Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

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.

Longhua Qian, Haotian Hui, Ya'nan Hu, Guodong Zhou, and Qiaoming Zhu. 2014. Bilingual active learning for relation classification via pseudo parallel corpora. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 582–592.

Chanqin Quan, Lei Hua, Xiao Sun, and Wenjun Bai. 2016. Multichannel convolutional neural network for biological relation extraction. *BioMed research international*, 2016.

Shafin Rahman, Salman Khan, and Fatih Porikli. 2018. A unified approach for conventional zero-shot, generalized zero-shot, and few-shot learning. *IEEE Transactions on Image Processing*, 27(11):5652–5667.

Bernardino Romera-Paredes and Philip Torr. 2015. An embarrassingly simple approach to zero-shot learning. In *International conference on machine learning*, pages 2152–2161. PMLR.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 410–420.

Sunil Kumar Sahu and Ashish Anand. 2018. Drug-drug interaction extraction from biomedical texts using long short-term memory network. *Journal of biomedical informatics*, 86:15–24.

Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. 2001. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471.

Ozan Sener and Silvio Savarese. 2018. Active learning for convolutional neural networks: Acore-set approach. *stat*, 1050:21.

Burr Settles. 2009. Active learning literature survey.

Baoxu Shi and Tim Weninger. 2018. Open-world knowledge graph completion. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Yusuke Shinyama and Satoshi Sekine. 2006. Preemptive information extraction using unrestricted relation discovery. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 304–311. Association for Computational Linguistics.

Etienne 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.

Gabriel Stanovsky and Ido Dagan. 2016. Creating a large benchmark for open information extraction. In *Proceedings of EMNLP*, pages 2300–2305.

Fei Wu and Daniel S Weld. 2010. Open information extraction using wikipedia. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 118–127. 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.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of EMNLP*, pages 1753–1762.

Beichen Zhang, Liang Li, Shijie Yang, Shuhui Wang, Zheng-Jun Zha, and Qingming Huang. 2020. State-relabeling adversarial active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8756–8765.

Dongxu Zhang and Dong Wang. 2015. Relation classification via recurrent neural network. *arXiv preprint arXiv:1508.01006*.

Hong-Tao Zhang, Min-Lie Huang, and Xiao-Yan Zhu. 2012. A unified active learning framework for biomedical relation extraction. *Journal of Computer Science and Technology*, 27(6):1302–1313.

Kai Zhang, Yuan Yao, Ruobing Xie, Xu Han, Zhiyuan Liu, Fen Lin, Leyu Lin, and Maosong Sun. 2021. Open hierarchical relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5682–5693.

Ziming Zhang and Venkatesh Saligrama. 2015. Zero-shot learning via semantic similarity embedding. In *Proceedings of the IEEE international conference on computer vision*, pages 4166–4174.

Zhehuan Zhao, Zhihao Yang, Ling Luo, Hongfei Lin, and Jian Wang. 2016. Drug drug interaction extraction from biomedical literature using syntax convolutional neural network. *Bioinformatics*, 32(22):3444–3453.

Deyu Zhou, Lei Miao, and Yulan He. 2018. Position-aware deep multi-task learning for drug–drug interaction extraction. *Artificial intelligence in medicine*, 87:1–8.

A Datasets Used in Experiments

Three datasets and their variants are used to evaluate our model: FewRel (Han et al., 2018), New York Times Freebase(NYT+FB) (Marcheggiani and Titov, 2016) and FewRel2.0 (Gao et al., 2019), the first two of which have been widely used in previous RE works (Simon et al., 2019; Hu et al., 2020; Zhang et al., 2021). We follow the division of the datasets from previous works. FewRel is one of the largest RE dataset. As in the previous work, we use the original train set of FewRel. The dataset contains 80 relation categories and 700 instances of each relation category. Among them, 64 relations are divided into the training set and the remaining 16 relations are chosen as the test set.

NYT+FB dataset aligns entities from the New York Times corpus with Freebase triplets. Following the setting in (Simon et al., 2019), we filter out sentence with non-binary relations and obtain 41,000 labeled sentences containing 262 relations. The training and test sets comprise 212 and 50 relations respectively.

To verify the cross-domain capability of the model comprehensively, we also use FewRel2.0 dataset whose training and test sets are from completely different domains. As an advanced version of FewRel, FewRel2.0 incorporates knowledge transferring. The test set of FewRel2.0 contains data of 10 relations (100 samples for each relation) in the biomedicine field, and the training set is exactly the same as FewRel. The statistics of the data set are shown in Table 5.

As described above, in the original OpenRE setting, there are no overlapping relations in the training and test sets. That is, the relations in the test set are all novel relations. For the purpose of measuring the performance of the model in our proposed *General OpenRE* setting, we resample the original dataset and gain two variants: *noisy* and *imbalanced*.

To obtain the noisy variant, we randomly select a random portion of the samples from the training set, whose size equals to 40% of the original test set. Given that the number of samples for each novel relation is identical in FewRel and FewRel2.0, we further construct the imbalanced variant to explore the performance of the model in the presence of class imbalance. Specifically, we build on the noisy variant by randomly discarding a portion of the samples with different probabilities for each novel relation in the test set, yielding class imbalance in

Dataset	Setting	Train		Test	
		#CLS	#SUM	#CLS	#SUM
FR	Ori	64	44,800	16	11,200
	Noi	64	40,320	64+16	4,480+11,200
	Imb	64	40,320	64+16	4,480+4,560
NYF	Ori	212	33,990	50	7,010
	Noi	212	30,591	212+50	3,399+7,010
FR2.0	Ori	64	44,800	10	1,000
	Noi	64	40,320	64+10	480+1,000
	Imb	64	40,320	64+10	480+720

Table 5: Statistical results for the dataset. #CLS represents the number of relation types and #SUM stands for the number of samples. In the addition equation $x + y$ in the table, x and y are the statistics for the known and novel relations separately.

test set. The discarding probabilities for different relations are shown in the Table 6.

B Implementation Details and Hyper-parameter Choices

To improve the experimental effect, we use $\text{BERT}_{\text{LARGE}}$ with 300M parameters in the relation representation module. We pre-train the BERT model on 3 epochs, and each epoch costs about 1 GTX 3090 GPU hour. For the discriminator, we constructed a 3-layer fully connected neural network. For active learning, λ and λ' are both 1. For optimization, different models use different optimizers. Specifically, BERT use AdamW (Loshchilov and Hutter, 2018) with a learning rate of 0.00002, for discriminator, we use Adam with a learning rate of 0.0005, and for task learner of active learning, SGD is utilized. For baseline models, we follow their original setting without modifying any parameters except the division of the dataset.

Dataset	Relation ID	P
FewRel	66-73	0.4
	74-77	0.7
	78-81	0.85
FewRel2.0	66-68	0.15
	69	0.2
	70	0.3
	71-72	0.35
	73	0.4
	74-75	0.45

Table 6: The discarding probabilities for different relations.

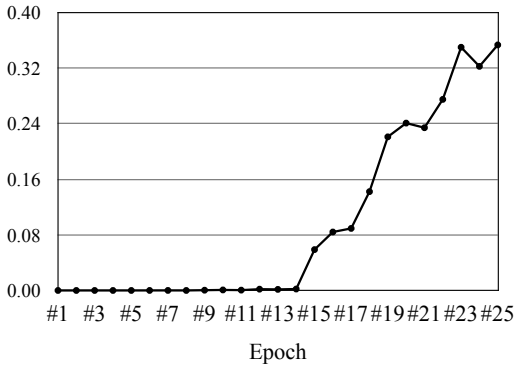


Figure 4: The confidence of the discriminator in each epoch for $\mathcal{X}_U$.

C Case Study of Active Learning

As shown in Table 7, we report 8 cases selected by discriminator in the first iteration on noisy FewRel2.0 dataset, where 64 relations are pre-trained and seen. With the highest confidence, the discriminator successfully select sentences with unseen relations.

D Practical Application on Real-world Dataset

We apply the ARD framework in real-world scenarios to verify its practicability. With the increasing number of publications about COVID-19, it is a challenge to extract personalized knowledge suitable for each researcher. Barros et al. (2020) aims to build a new semantic-based pipeline for recommending biomedical entities to scientific researchers. In this work, the researchers utilize MER (Couto and Lamurias, 2018) as NER annotation server. As a result, 9,000 articles are automatically annotated with relevant items/concepts for COVID-19. And for further relation extraction task, due to the expensive manual annotation costs, the researchers merely take initial steps towards the results, providing a small sample dataset of ten documents, with all possible relationships between the four types of entities identified by NER pipeline. Thus, we were able to establish ten different types of relations, encompassing the four ontologies (CHEBI, DO, HPO, and GO). We follow the relation types, and apply ARD framework in the results. We take sample size k of 200 and sample 25 epochs. Finally, a total of 139,479 relations between entity pairs are automatically ob-

Selected sentence	Novel relation
Ectopic overexpression of mir-497 promotes chemotherapy resistance in glioma cells by targeting *pdcd4*, a tumor suppressor that is involved in *apoptosis*.	*Biological process involves gene product*
As full-length bid is a weaker apoptogen than *tbid*,we propose that the phosphorylation of bid by jnks, followed by the accumulation of the full-length protein, delays attainment of *apoptosis*, and allows the cell to evaluate the stress and make a decision regarding the response strategy.	*Biological process involves gene product*
Pretreatment with dexamethasone 1 hour before *cyclophosphamide injection* significantly down-regulated *cyclophosphamide* induced bladder nuclear factor-u03bab dependent luminescence, ameliorated the grossly evident pathological features of acute inflammation and decreased cellular immunos-taining for nuclear factor-u03bab in the bladder.	*Ingredient of*
Trastuzumab emtansine (*t-dm1*), an antibody-drug conjugate comprising the cytotoxic agent dm1, a stable linker, and *trastuzumab*, has demonstrated substantial activity in human epidermal growth factor receptor 2 (her2), -positive metastatic breast cancer, raising interest in evaluating the feasibility and cardiac safety of t-dm1 in early-stage breast cancer (ebc).	*Ingredient of*
Here we looked for evidence of adult hippocampal *neurogenesis* using immunohistochemical techniques for the endogenous marker doublecortin (*dcx*) in 10 species of microchiropterans euthanized and perfusion fixed at specific time points following capture.	*Gene plays role in process*
Here, we explored the effects of the novel class ii-specific "histone deacety-lase inhibitors (hdacis) mc1568 and mc1575 on interleukin-8 (il-8) expres-sion and *cell proliferation* in cutaneous melanoma cell line *gr* -m and uveal melanoma cell line ocm-3 upon stimulation with phorbol 12-myristate 13-acetate (pma).	*Gene plays role in process*
Data indicate that the structurally disordered and abnormally formed ecm of *uterine fibroids* contributes to *fibroid* formation and growth.	*Classified as*
however, individuals heterozygous for both beta "e", "and", beta thalas-saemia (hbe/*beta thalassaemia*) have a severe clinical disorder which in some cases may approach that seen in *homozygous beta thalassaemia* and which is by far the commonest form of symptomatic thalassaemia in the indian subcontinent and south-east asia.	*Classified as*

Table 7: Cases selected by the confidence score of the discriminator and the novel relations, where *red* and *blue* represent the head and tail entities

.

Relation	Count
CHEBI-CHEBI	4680
CHEBI-HP	20455
GO-HP	17254
DO-DO	14430
CHEBI-DO	2415
HP-HP	48236
HP-DO	19770
GO-CHEBI	1285
GO-DO	3615
GO-GO	7303

Table 8: Statistical results of dataset for COVID-19.

Dataset	Method		
	IF	OneClassSVM	LOF
FewRel	64.0	47.3	**83.9**
FewRel2.0	63.1	54.1	**80.3**

Table 9: F1-measure on noisy FewRel and FewRel2.0 with different outlier detection algorithms.

tained by ARD. The statistical results of the data are shown in Table 8. We also report the confidence of the discriminator in each epoch for $\mathcal{X}_U$. As can be observed from the Figure 4, the confidence is progressively increasing as the training epoch increases, which indicates that the model is becoming more confident in the classification results. In an ideal case, the confidence should converge toward 0.5.

E Impact of Different Outlier Detection Algorithms

We compare LOF with two different algorithms for the relational outlier detection, including IsolationForest (Liu et al., 2008), and OneClassSVM (Schölkopf et al., 2001). We evaluate the F1-measure of these three algorithms solely on the discovery of novel relations, the results are reported in Table 9. Our LOF algorithm outperforms by large margins, achieving 83.9% F1-measure on FewRel dataset. The principle of the IsolationForest algorithm is to cut data points and isolate data points one by one. Thus the data needs more cuts to be isolated. The main reason for the poor performance of this algorithm is a large amount of the test data. For the same type of new relations, their distribution is relatively dense, and the number of cuts will also increase. Moreover, the dimensions of relation representation are 2048, while IsolationForest has poor processing capabilities for high-dimensional features. Hence, it yields relevant poor results. OneClassSVM aims to learn a tight decision boundary from normal data and treats points outside the decision boundary as abnormal points. In the relational feature space, the distribution of known relations and novel relations are complicated. Thus the OneClassSVM is likely to learn an over-fitting decision boundary, resulting in poor performance.

F Performance of our Relation Representation on Supervised RE

To demonstrate the effectiveness of the relation representation described in the Methodology section, we conduct a series of experiments on supervised RE task. First, we conduct extensive experiments on the biomedical relation extraction benchmark DDI'13 (Herrero-Zazo et al., 2013). We make comparisons with various previous state-of-the-art methods, which fall into two groups according to the neural network architecture: convolutional neural network (CNN) based methods and recurrent neural network (RNN) based methods. For the first group, we report the results of SCNN (Zhao et al., 2016), CNN-bioWE (Liu et al., 2016) and MCCNN (Quan et al., 2016), which uses syntax word embeddings, biomedical-related embeddings and multi-channel word embeddings for feature extraction, respectively. For recurrent based networks, we report the reults of Joint AB-LSTM (Sahu and Anand, 2018), Position-aware LSTM (Zhou et al., 2018), RvNN (Lim et al., 2018) and BERE (Hong et al., 2020). Joint AB-LSTM jointly trains two bidirectional LSTM (Bi-LSTM) with different pooling mechanisms: max-pooling for one Bi-LSTM and attentive pooling for the other. Position-aware LSTM adopt position information as attention mechanism for the training of LSTM. RvNN and BERE incorporates parse-tree information to enhance the performance of prediction. Each model is trained on the training dataset to predict a relation class of five pre-defined relation types for the input sequence.

To further evaluate the performance of our representation method on large-scale distantly annotated dataset, we conduct another set of experiments on the DTI dataset. As on the DTI dataset, previous literature has shown the superiority of BERE com-

Methods	Pre.	Rec.	F1
SCNN	69.1	65.1	67.0
CNN-bioWE	75.7	64.7	69.8
MCCNN	75.9	65.2	70.2
Joint AB-LSTM	73.4	69.6	71.5
RvNN	74.4	69.3	71.7
Position-aware LSTM	75.8	70.4	73.0
BERE	76.8	71.3	73.9
Ours	**92.3**	**84.4**	**86.8**

Table 10: Results on DDI'13 dataset. The first seven rows are the results of the previous state-of-the-arts methods, and the bottom row is the performance of our method in supervised relation learning.

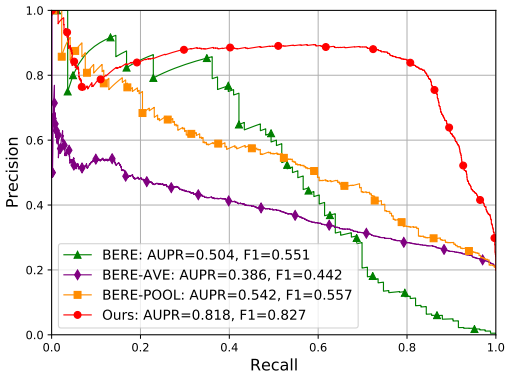


Figure 5: Precision-recall curve of BERE and our model.

pared with CNN-based and RNN-based baselines, we mainly take BERE as the baseline of our experiments. For fair comparisons, we follow the settings of BERE by using precision-recall curve, the area under the precision-recall curve and the F_1 score as the evaluation metrics. We re-run the open-source code of BERE and its two variants: BERE-AVE, BERE-POOL. BERE-AVE adopt the average pooling mechanism to aggregate the semantic information over instances in a bag. BERE-POOL uses the max-pooling strategy. The implementation details of our model on the DTI dataset are identical to the DDI'13 dataset. The precision-recall curve is shown in Figure 5, which indicates the significant performance of our representation method.

G Time Efficiency of Relational Active Learning

In practice, it is often the time spent on manual annotation that is the time-consuming bottleneck. Nevertheless, the sampling strategy for active learner should also select samples in a time-efficient manner as much as possible. We analysis the time efficiency of different active learning methods. Table 11 shows the average time for different methods to sample once on the corresponding dataset. DBAL is the most competitive baselines in terms of their achieved mean time efficiency. Our method fell marginally behind DBAL, however, our method is outperformed in accuracy by all other methods.

Dataset	Time(ms)			
	DBAL	CoreSet	SRAAL	Ours
FewRel	157.0	1145.2	409.3	465.1
NYT+FB	157.9	74.3	132.7	101.9
FewRel2.0	181.4	209.4	418.4	9.2
Average	**165.4**	476.3	320.1	192.0

Table 11: Average time token to sample once on the corresponding dataset.