Abstains from Prediction: Towards Robust Relation Extraction in Real World

Jun Zhao[1*], **Yongxin Zhang**[1*], **Nuo Xu**[1], **Tao Gui**[1]†, **Qi Zhang**[1]†, **Yunwen Chen**[2], **Xiang Gao**[2]
[1] School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing,
Fudan University, Shanghai, China
[2] DataGrand Information Technology (Shanghai) Co., Ltd., Shanghai, China
{zhaoj19,yongxinzhang20,tgui,qz}@fudan.edu.cn
xun22@m.fudan.edu.cn
{chenyunwen,gaoxiang}@datagrand.com

Abstract

Supervised learning is a classic paradigm of relation extraction (RE). However, a well-performing model can still confidently make arbitrarily wrong predictions when exposed to samples of unseen relations. In this work, we propose a relation extraction method with rejection option to improve robustness to unseen relations. To enable the classifier to reject unseen relations, we introduce contrastive learning techniques and carefully design a set of class-preserving transformations to improve the discriminability between known and unseen relations. Based on the learned representation, inputs of unseen relations are assigned a low confidence score and rejected. Off-the-shelf open relation extraction (OpenRE) methods can be adopted to discover the potential relations in these rejected inputs. In addition, we find that the rejection can be further improved via readily available distantly supervised data. Experiments on two public datasets prove the effectiveness of our method capturing discriminative representations for unseen relation rejection.

1 Introduction

Relation extraction aims to predict the relation between entities based on their context. The extracted relational facts play a vital role in various natural language processing applications, such as knowledge base enrichment (Distiawan et al., 2019), web search (Xiong et al., 2017), and question answering (Honovich et al., 2021).

To improve the quality of extracted relational facts and benefit downstream tasks, many efforts have been devoted to this task. *Supervised relation extraction* is a representative paradigm built upon the closed world assumption (Gallaire and Minker, 1978). Benefiting from artfully designed network architectures (Miwa and Bansal, 2016; Huang and Wang, 2017; Zhang et al., 2018) and valuable knowledge in pretrained language model (Du et al., 2018; Verga et al., 2018; Wu and He, 2019; Baldini Soares et al., 2019), models effectively capture semantic-rich representations and achieves superior results. However, conventional supervised relation extraction suffer from the lack of large-scale labeled data. To tackle this issue, *distantly supervised relation extraction* has attracted much attention. The existing works mainly focus on how to alleviate the noise generated in the automatic annotation. Common approaches include selecting informative instances (Lin et al., 2016), incorporating extra information (Zhang et al., 2019), and designing sophisticated training (Ma et al., 2021).

Although a supervised relation classifier achieves excellent performance on known relations, real-world inputs are often mixed with samples of unseen relations. A well-performing model can still confidently make arbitrarily wrong predictions when dealing with these unseen relations (Nguyen et al., 2014; Recht et al., 2019). The unrobustness is rooted in the *Shortcut* feature (Geirhos et al., 2020) of neural networks. Models optimized by a supervised objective does not actively learn features beyond the bare minimum necessary to discriminate between known relations. As shown in Figure 1, if there is only president relation in the training data between Obama and the United States, the model tends

* Equal contribution.
† Corresponding authors.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 798-810, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

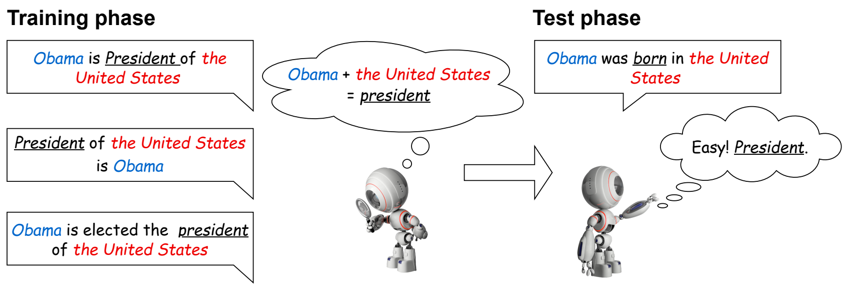
Training phase
Obama is President of the United States
President of the United States is Obama
Obama is elected the president of the United States
Obama + the United States = president
Test phase
Obama was born in the United States
Easy! President.

Figure 1: Neural models tend to use the simplest way to meet the supervised objective (*Shortcut* phenomenon (Geirhos et al., 2020)), which would lead to negative predictions on unseen relations. Hence, for the unseen relations, we hope neural models can reject prediction through embracing sufficient features.

Model/ Dataset	SpanBERT	Roberta	CP
Ori(F_1-score)	0.919	0.928	0.936
Mix(ΔF$_1$-score)	0.317↓	0.310↓	0.310↓

Table 1: Supervised RE models' performance when encountering new relations. These models are from previous papers(Joshi et al., 2019; Liu et al., 2019; Peng et al., 2020). Ori: all relations in the test set are present in the training set. Mix: 50% of the relations in the test set do not appear in the training set.

to predict the president relation when it encounters them again. However, entities are not equivalent to relation definitions. Models severely biased to the extraction of overly simplistic features can easily fail to generalize to discriminate between known and unseen relations. As shown in Table 1, when the unseen relations appears in the test set, the supervised RE models' F_1-score drops by at least 30 points.

In this work, we propose a robust relation extraction method in real world settings. By integrating rejection option, the classifier can effectively detect whether inputs express unseen relations instead of making arbitrary bad predictions. Specifically, we introduce contrastive training techniques to achieve this goal. A set of carefully designed class-preserving transformations are used to learn sufficient features, which can enhance the discriminability between known and unknown relation representations. The classifier built on the learned representation is confidence-calibrated. Thereby samples of unseen relations are assigned a low confidence score and rejected. Off-the-shelf OpenRE methods can be used to discover potential relations in these samples. In addition, we find the rejection can be further improved via the readily available distantly-supervised data. Experimental results show the effectiveness of our method capturing discriminative representations for unseen relation rejection.

To summarize, the main contributions of our work are as follows: (1) We propose a relation extraction method with rejection option, which is still robust when exposed to unseen relations. (2) We design a set of class-preserving transformations to learn sufficient features to discriminate known and novel relations. In addition, we propose to use readily available distantly-supervised data to enhance the discriminability. (3) Extensive experiments on two academic datasets prove the effectiveness of our method capturing discriminative representations for unseen relation rejection.

2 Related Work

2.1 Relation Extraction

Relation extraction has advanced for more than a couple of decades. Supervised/Distantly supervised relation extraction is oriented at predefined relational types. Researchers have explored different network architectures (Zhang et al., 2018), training strategies (Ma et al., 2021) and external information (Zhang et al., 2019). Superior results have been achieved. Open relation extraction is oriented at emerging unknown relation. Well-designed extraction forms (e.g. sequence labelling (Fader et al., 2011), clustering (Zhao et al., 2021)) are used to deal with relations without pre-specified schemas. Different from them, we consider a more general scenario, in which known and unknown relations are mixed in the input. We effectively separate them by a rejection option, which enables us to use the optimal paradigm to deal with the corresponding relations.

2.2 Classification with Rejection Option

Most existing classification methods are based on the closed world assumption. However, inputs are often mixed with samples of unknown classes in real-world applications. The approaches used to handle it roughly fall into one of two groups. The first group calculates the confidence score based on the classifier output. The score can be used to measure whether an input belongs to unknown classes. Maximum softmax probability (MSP) (Hendrycks and Gimpel, 2017) is a represetative method and Liang et al. (2018) further improve MSP by introducing temperature scaling. Furthermore, Shu et al. (2017) build a multi-class classifier with a 1-vs-rest final layer of sigmoids to reduce the open space risk. The second group considers classification with rejection option as an outlier detection problem. Off-the-shelf outlier detection algorithms (Breunig et al., 2000; Schölkopf et al., 2001; Liu et al., 2008) are leveraged. Different optimization objectives such as large margin loss (Lin and Xu, 2019), gaussian mixture loss (Yan et al., 2020) are adopted to learn more discriminative representations to facilitate anomaly detection. Recently, Zhang et al. (2021) propose to learn the adaptive decision boundary (ADB) that serves as the basis for judging outliers.

3 Approach

In this paper, we propose a robust relation extraction method in real world settings. By integrating rejection option, the classifier can effectively detect whether inputs express unseen relations instead of making arbitrary bad predictions. Off-the-shell OpenRE methods can be used to discover potential relations in these rejected samples.

The problem setting in this work is formally stated as follows. Let $\mathcal{K} = \{\mathcal{R}_1, ..., \mathcal{R}_k\}$ be a set of known relations and $\mathcal{U} = \{\mathcal{R}_{k+1}, ..., \mathcal{R}_n\}$ be a set of unseen relations where $\mathcal{K} \cap \mathcal{U} = \emptyset$. Let $\mathcal{X}$ be an input space. Given the training data $\mathcal{D}^{\ell} = \{(x_i^{\ell}, y_i^{\ell})\}_{i=1,...,N}$ where $x_i^{\ell} \in \mathcal{X}$, $y_i^{\ell} \in \mathcal{K}$, we target constructing a mapping rule $f : \mathcal{X} \rightarrow \{\mathcal{R}_1, ..., \mathcal{R}_k, \mathcal{R}^*\}$ where $\mathcal{R}^*$ denotes rejection option. Let $\mathcal{D}^u = \{(x_i^u, y_i^u)\}_{i=1,...,M}$ be the testing dataset where $y_i^u \in \mathcal{K} \cup \mathcal{U}$. An desirable mapping rule f should meet the following objective as much as possible:

$$f(x) = \begin{cases} y_i^u & y_i^u \in \mathcal{K} \\ \mathcal{R}^* & y_i^u \in \mathcal{U}. \end{cases}$$

3.1 Method Overview

We approach the problem by introducing contrastive learning techniques. As illustrated in Figure 2, the proposed method comprises four major components: relation representation encoder $\boldsymbol{g}(\cdot)$, confidence-calibrated classifier $\boldsymbol{\eta}(\cdot)$, class-preserving transformations $\mathcal{T}$, and the OpenRE module.

Our overview starts from the first two components. There is no doubt that an encoder and classifier are the basic components of a supervised relation extractor. However, the supervised training objective does not encourage the model to learn features beyond the bare minimum necessary to discriminate between known relations. Consequently, the classifier can misclassify unseen relations to known relations with high confidence.

In order to calibrate the confidence of the classifier, we introduce contrastive learning techniques. Given training batch $\mathcal{B}$, an augmented batch $\widetilde{\mathcal{B}}$ is obtained by applying random transformation $t \in \mathcal{T}$ to mask partial features. Then the supervised contrastive learning objective max/minimize the representation agreement according to whether their relations are the same. By doing this, the model is forced to find more features to discriminate between relations and the classifier can be calibrated. Based on the confidence-calibrated classifier, unknown relations are rejected if the maximum softmax probability of the classifier does not exceed a preset threshold θ.

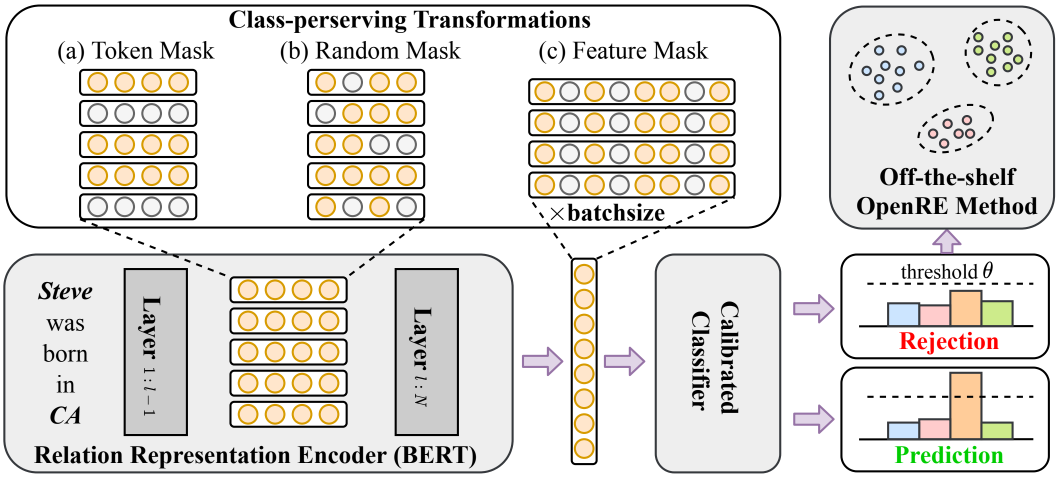

Figure 2: An overview of the proposed method. Three steps are included: (1) Contrastive training techniques and a set of class-preserving transformations are utilized to learn sufficient features. (2) The classifier extract known relations and rejects samples of unseen relations according to these features. (3) Off-the-shelf OpenRE method (SelfORE) is incorporated to discovery unseen relations in these rejected samples.

In order to discriminate unknown relations rather than just detect their existence, we further integrate the off-the-shelf OpenRE method into our framework. The samples rejected by the classifier are sent to the OpenRE module to detect potential unknown relations.

3.2 Relation Representation Encoder

Given a relation instance $x_i^\ell = (\boldsymbol{w}_i, h_i, t_i) \in \mathcal{D}^\ell$ where $\boldsymbol{w}_i = \{w_1, w_2, ..., w_n\}$ is the input sentence and $h_i = (s^h, e^h)$, $t_i = (s^t, e^t)$ mark the position of head and tail entities, relation representation encoder $\boldsymbol{g}(\cdot)$ aims to encode contextual relational information to a fixed-length representation $\boldsymbol{r}_i = \boldsymbol{g}(x_i) \in \mathbb{R}^d$. We opt for simplicity and adopt the commonly used BERT (Devlin et al., 2018) to obtain $\boldsymbol{r_i}$ while various other choices of the network architecture are also allowed without any constraints. Formally, the process of obtaining $\boldsymbol{r_i}$ is:

$$\boldsymbol{h}_1, ..., \boldsymbol{h}_n = \mathrm{BERT}(w_1, ..., w_n) \tag{1}$$

$$\boldsymbol{h}_{ent} = \mathrm{MAXPOOL}(\boldsymbol{h}_s, ..., \boldsymbol{h}_e) \tag{2}$$

$$\boldsymbol{r}_i = \langle \boldsymbol{h}_{head} | \boldsymbol{h}_{tail} \rangle \,, \tag{3}$$

where $\boldsymbol{h}_1, ..., \boldsymbol{h}_n$ is the result of the input sentence after BERT encoding, subscript s and e represent the start and end positions of the entity, $\boldsymbol{h}_{ent}$ represents the result of the maximum pooling of the entity, $\boldsymbol{h}_{ent}$ can be divided into head entity $\boldsymbol{h}_{head}$ and tail entity $\boldsymbol{h}_{tail}$, and $\langle\cdot|\cdot\rangle$ is the concatenation operator.

3.3 Confidence-calibrated Classifier

In order to alleviate overconfidence to unseen relations, we introduce contrastive learning techniques to calibrate classifier. A well-calibrated classifier should not only accurately classify known relations, but also give low confidence to unseen relations, that is, $\max_y p(y|x)$.

Given a training batch $\mathcal{B} = (x_i^\ell, y_i^\ell)_{i=1}^B$, we obtain a augmented batch $\widetilde{\mathcal{B}} = (\widetilde{x}_i^\ell, y_i^\ell)_{i=1}^B$ by applying random transformation $t \in \mathcal{T}$ on $\mathcal{B}$. For brevity, the superscript ℓ is omitted in the subsequent elaboration of this section. For each labeled sample $(\widetilde{x}_i, y_i)$, $\widetilde{\mathcal{B}}$ can be divided into two subsets $\widetilde{\mathcal{B}}_{y_i}$ and $\widetilde{\mathcal{B}}_{-y_i}$. $\widetilde{\mathcal{B}}_{y_i}$ denotes a set that contains samples of relation y_i and $\widetilde{\mathcal{B}}_{-y_i}$ contains the rest. The supervised contrastive

Proceedings of the 21st China National Conference on Computational Linguistics, pages 798-810, Nanchang, China, October 14 - 16, 2022.

learning objective is defined as follows:

$$\mathcal{L}_{cts}^{sup}(\mathcal{B},\mathcal{T}) = \frac{1}{2B}\sum_{j=1}^{2B}\mathcal{L}_{cts}(\widetilde{x}_i, \widetilde{\mathcal{B}}_{y_i}\backslash\{\widetilde{x}_i\}, \widetilde{\mathcal{B}}_{-y_i}) \tag{4}$$

$$\mathcal{L}_{cts}(x,\mathcal{D}^+,\mathcal{D}^-) = -\frac{1}{|\mathcal{D}^+|}log\frac{\sum_{x'\in\mathcal{D}^+} q(x,x')}{\sum_{x'\in\mathcal{D}^+\cup\mathcal{D}^+} q(x,x')} \tag{5}$$

$$q(x,x') = \exp(sim(\boldsymbol{z}(x),\boldsymbol{z}(x'))/\tau), \tag{6}$$

where $|\mathcal{D}|$ denotes the number of samples in $\mathcal{D}$, $sim(x,x')$ denotes the cosine similarity between x and x' and τ denotes a temperature coefficient. Following Chen et al. (2020), we use a additional projection layer $\boldsymbol{t}$ to obtain the contrastive feature $\boldsymbol{z}(x) = \boldsymbol{t}(\boldsymbol{g}(x))$.

Benifiting from contrastive training, the encoder $\boldsymbol{g}(\cdot)$ learns rich features to discriminate between known and novel relations. Accordingly, we train a confidence-calibrated classifier $\boldsymbol{\eta}(\cdot)$ upon $\boldsymbol{g}(\cdot)$ as follows:

$$\mathcal{L} = \mathbb{E}_{(x,y)\sim\mathcal{D}^\ell}[\mathcal{L}_{ce}(\boldsymbol{\eta}(\boldsymbol{g}(x_i)), y)], \tag{7}$$

where $\mathcal{L}_{ce}$ is the cross entropy loss. In addition, we can easily obtain a large number of training data $\mathcal{D}^{dist}$ through distant supervision. None of the y_i^{dist} in $\mathcal{D}^{dist}$ are known relation, that is, $\{y_i^{dist}\}\cap\{y_j^\ell\} = \emptyset$. These data are only used as negative examples, so the noise in the data will not be a problem. We force the classifier output distribution of negative examples to approximate the uniform distribution by optimizing the cross-entropy between them. Using $\mathcal{D}^{dist}$, we optimize model by following objective instead of equation 7.

$$\mathcal{L}^{dist} = \mathcal{L} + \lambda\mathbb{E}_{x\sim\mathcal{D}^{dist}}[\mathcal{L}_{ce}(\boldsymbol{\eta}(\boldsymbol{g}(x)), y_{uni})], \tag{8}$$

where $\mathcal{L}$ refers to the optimization objective of equation 7. λ is the hyperparamters that balances the known relation data and distantly supervised data. We can achieve good results simply by setting λ to 1 without adjustment. y_{uni} represents a uniform distribution.

Based on the confidence-calibrated classifier, we specify the rejection rule $f(\cdot)$ as follows:

$$f(x_i) = \begin{cases} y & max_y p(y|x_i) > \theta \\ \mathcal{R}^* & Otherwise, \end{cases} \tag{9}$$

where θ is a threshold hyperparameters, the posterior probability $p(y|x_i)$ is the output of classifier $\boldsymbol{\eta}$ and $\mathcal{R}^*$ denotes the rejection option.

3.4 Class-preserving Transformations

Transformations is the core component of contrastive learning. Our intuition in designing transformation is that feature masks at different views force the model to find more features to discriminate between known relations. These new features can play a vital role in recognizing unseen relations. Why do the above methods work? As shown in Figure 1, due to the *shortcut* phenomenon, the model is more inclined to remember the relations between entities and it would make mistakes when predicting new relations between the same entity pair. Intuitively through the mask mechanism, the model could mask out some features that belong to Obama and the United States, and then it will have to find more other features to distinguish *the president of* from other relations. Therefore it will not learn the *Shortcut* bias of *Obama + the United States = the president of*. In this work, we design three class-preserving transformations to mask partial features as follows.

Token Mask. Token mask works in the process of sentence encoding. In this transformation, we randomly mask a certain proportion of tokens to generate a new view of relation representation.

Random Mask. Random mask also works in the process of sentence encoding. Instead of completely masking representation of selected tokens, each dimension of the representation of each word is considered independently in this transformation.

```
Algorithm 1: Robust Relation Extraction
  Input: known relation dataset $\mathcal{D}^{\ell}$, distantly
  supervised dataset $\mathcal{D}^{dist}$ (optional), testing dataset $\mathcal{D}^{u}$, transformation set $\mathcal{T}$, model parameters
  $\Theta$, $\Phi$ for encoder and classifier, OpenRE module $\mathcal{O}$ and learning rate $\alpha$.
Training Phase
repeat
    sample a training batch $\mathcal{B}$ from $\mathcal{D}^{\ell}$;
    obtain transformed batch $\widetilde{\mathcal{B}} = t(\mathcal{B}), t \sim \mathcal{T}$;
    enrich representation by contrastive training (equ 4): $\Theta = \Theta - \alpha \nabla_{\Theta} \mathcal{L}_{cts}^{sup}$;
    sample a distant batch $\mathcal{B}^{dist}$ from $\mathcal{D}^{dist}$;
    optimize classifier by supervised training (equ 7 or 8):
    $\{\Theta, \Phi\} = \{\Theta, \Phi\} - \alpha \nabla_{\{\Theta,\Phi\}} \mathcal{L}^{dist}$;
until convergence;
Testing Phase
Filter the unseen relations subset $\mathcal{D}^{rej}$ from $\mathcal{D}^{u}$ by the rejection rule $f$ (equ 9);
Output predictions $\{y_i^u\}$ for the rest samples of known relations;
Run the OpenRE module $\mathcal{O}$ to obtain potential relations in $\mathcal{D}^{rej}$;
```

Feature Mask. Feature mask works after sentence encoding. Given a relation instance $x_i^{\ell} \in \mathcal{D}^{\ell}$, we first obtain its relation representation $\boldsymbol{r}_i = \boldsymbol{g}(x_i)$. Then we randomly mask a certain proportion of feature dimensions of $\boldsymbol{r}_i$ to generate a new view.

It is certain that a more complicated and diverse transformations will bring additional improvement. This will be one of our future work.

3.5 OpenRE Module

We introduce the OpenRE module for the integrity of the framework, although it is not our main concerns. Based on the rejection rules f described in section 3.3, we can classify samples of known relations while rejecting unseen relations. In this section, we take a step forward. By integrating the off-the-shelf OpenRE method, we try to discover the potential unseen relations in the rejected samples instead of only detecting their existence. We adopt SelfORE (Hu et al., 2020), a clustering-based OpenRE method, as the building block of our OpenRE module. Various other methods can also be used as the alternative to SelfORE without any constraints. More details about OpenRE methods can be found in the related papers. Overall, the method proposed in this paper is detailed in algorithm 1.

4 Experimental Setup

In this section, we describe the datasets for training and evaluating the proposed method. We also detail the baseline models for comparison. Finally, we clarify the implementation details.

4.1 Datasets

We conduct our experiments on two well-known relation extraction datasets. In addition, a distantly supervised dataset are used in a auxiliary way.

FewRel. Few-Shot Relation Classification Dataset (Han et al., 2018). FewRel is a human-annotated dataset containing 80 types of relations, each with 700 instances. We use the top 40 relations as known and the middle 20 relations as unseen. Since the relations of FewRel dataset is exactly the same as that of FewRel-Distance, we hold out the last 20 relations for the use of distant supervision. The training set contains 25600 randomly selected samples of known relations. In order to evaluate the rejection performance to the unseen relations, the test/validation set contains 3200/1600 samples composed of known and unseen relations.

TACRED. The TAC Relation Extraction Dataset (Zhang et al., 2017). TACRED is a human-annotated large-scale relation extraction dataset that covers 41 relation types. Similar to the setting of FewRel, we use the top 31 relations as known and the rest 10 relations as unseen. The training set consists of 18113 randomly selected samples of known relations. The size of validation set and test set are 900 and 1800 respectively, including known and unseen relations. It should be noted that 50% of the unseen relation samples in the validation set and test is `no_relation`.

FewRel-distant. FewRel-distant contains the distantly-supervised data obtained by the authors of FewRel before human annotation. We use this dataset as the distantly supervised data in our experiments.

4.2 Baselines and Evaluation Metrics

MSP (Hendrycks and Gimpel, 2017). MSP assumes that correctly classified examples tend to have greater maximum softmax probabilities than examples of unseen classes. Thereby the maximum softmax probabilities are used as confidence score for unseen classes detection.

MSP-TC (Liang et al., 2018). MSP-TC uses maximum softmax probabilities with temperature scaling and small perturbations to enhance the separability between known and unseen classes, allowing for more effective detection.

DOC (Shu et al., 2017). DOC builds n 1-vs-rest sigmoid classifiers for n known classes respectively. The maximum probability of these binary classifiers is considered as the confidence score for unseen classes detection.

LMCL (Lin and Xu, 2019). Large margin cosine loss (LMCL) aims to learn a discriminative deep representations. It forces the model to not only classify correctly but also maximize inter-class variance and minimize intra-class variance. Based on the learned representations, local outlier factor (LOF) is used to detect unseen classes.

ADB (Zhang et al., 2021). Labeled known classes samples are first used for representation learning. Then the learned representations are utilized to learn the adaptive spherical decision boundaries for each known classes. Samples outside the hypersphere will be rejected for recognition.

Evaluation Metrics. We follow previous work (Zhang et al., 2021; Lin and Xu, 2019) and take all the unseen relations as one rejected class. The accuracy and macro F1 metrics are used as the scoring function to evaluate the unseen relation detection.

4.3 Implementation Details

We use the Adam (Kingma and Ba, 2015) as the optimizer, with a learning rate of $1e-4$ and batch size of 100 for all datasets. If the results don't improve on the validation set for 10 epochs, we stop the training to avoid overfitting. All experiments are conducted using a NVIDIA GeForce RTX 3090 with 24GB memory.

5 Results and Analysis

In this section, we present the experimental results of our method on FewRel and TACRED datasets to demonstrate the effectiveness of our method.

5.1 Main Results

Our experiments in this section focus on the following three related questions.

Can the proposed method effectively detect unseen relations? To answer this question, we consider all the known relations as one predicted class and the rest unseen relations as one rejected class. Table 2 reports model performances on FewRel, TACRED datasets, which shows that the proposed method achieves state-of-the-art results on unseen relation detection. Benefiting from the contrastive training objectives and the carefully designed transformations, the *Shortcut* phenomenon is effectively alleviated, and the model learns sufficient features to discriminate between known and unseen relations. Therefore, the proposed method consistently outperforms the compared baselines by a large margin in different mixing-ratio settings.

Does the detection of unseen relations impair the extraction of known relations? Integrating the rejection option can make the classifier more robust in real applications. However, we do not want the unseen relations detection impair known relations classification, which is the basic function of the

Dataset	Method	25%		50%		75%	
		Accuracy	F_1-score	Accuracy	F_1-score	Accuracy	F_1-score
FewRel	MSP (Hendrycks and Gimpel, 2017)	0.805	0.781	0.786	0.786	0.797	0.774
	MSP-TC (Liang et al., 2018)	0.802	0.772	0.769	0.769	0.786	0.768
	DOC (Shu et al., 2017)	0.794	0.768	0.781	0.781	0.784	0.761
	LMCL (Lin and Xu, 2019)	0.810	0.785	0.740	0.740	0.835	0.777
	ADB (Zhang et al., 2021)	0.801	0.800	0.837	0.799	0.837	0.784
	Ours	**0.888**	**0.852**	**0.844**	**0.824**	**0.838**	**0.827**
TACRED	MSP (Hendrycks and Gimpel, 2017)	0.758	0.691	0.698	0.688	0.734	0.650
	MSP-TC (Liang et al., 2018)	0.789	0.687	0.674	0.670	0.765	0.671
	DOC (Shu et al., 2017)	0.793	0.687	0.707	0.678	0.775	0.681
	LMCL (Lin and Xu, 2019)	0.737	0.705	0.667	0.684	0.785	0.654
	ADB (Zhang et al., 2021)	0.772	0.714	0.711	0.710	0.767	0.699
	Ours	**0.827**	**0.758**	**0.723**	**0.742**	**0.788**	**0.715**

Table 2: Main results of unseen relation detection with different known class proportions (25%, 50% and 75%) on two relation extraction datasets. Compared with the best results of all baselines, our method improves F_1-score by an average of 2.6%, 3.5% on FewRel and TACRED dataset, respectively.

Dataset	Method	25%	50%	75%
FewRel	MSP	0.730	0.769	0.814
	MSP-TC	0.675	0.771	0.764
	DOC	0.737	0.780	0.805
	LMCL	0.765	0.767	0.809
	ADB	0.778	0.770	0.810
	Ours	**0.827**	**0.793**	**0.828**
TACRED	MSP	0.610	0.619	0.668
	MSP-TC	0.378	0.438	0.639
	DOC	0.628	0.627	0.686
	LMCL	0.616	0.615	0.687
	ADB	0.625	**0.640**	0.665
	Ours	**0.637**	0.633	**0.688**

Table 3: Macro F_1-score of known relation classification with different proportion of known relations.

classifier. From table 3 we can observe that the proposed model not only effectively detect unseen relations, but also accurately classify known relations. This demonstrate that the designed transformation will not affect the original relational semantics, so the rich features obtained by comparative learning remain discriminability for the known relations.

Can the model achieve superior performance under different threshold settings? We show the receiver operating characteristic (ROC) curve in Figure 3. The area under ROC curve (AUROC) summarize the performance of a classifier detecting unseen relations across different thresholds. From Figure 3 we can observe that the AUROC of the proposed method is the largest. Therefore, the proposed method has certain advantages under different threshold settings.

5.2 Ablation Study

To understand the effects of each component of the proposed model, we conduct an ablation study on it and report the results (Macro-F_1) on the two dataset in Table 4. The results show that the detection of unseen relations is degraded if any transformation is removed. It indicates that (1) These transformations force model learn sufficient features through mask mechanism from different views. The learned features are beneficial for the detection of unseen relations. (2) Since the transformations are from different views, they can be superimposed and further enhance the detection of unseen relations. In addition, we find that distantly supervised data can significantly improve the detection of unseen relations. Because there are a large number of diverse relations in the external knowledge base, we can easily construct a large number of negative samples. So this improvement can be seen as a free lunch.

5.3 Relation Representation Visualization

To intuitively show the influence of the rich features learned through contrastive training, we visualize the relational representation with t-SNE (van der Maaten and Hinton, 2008). We select five semantically

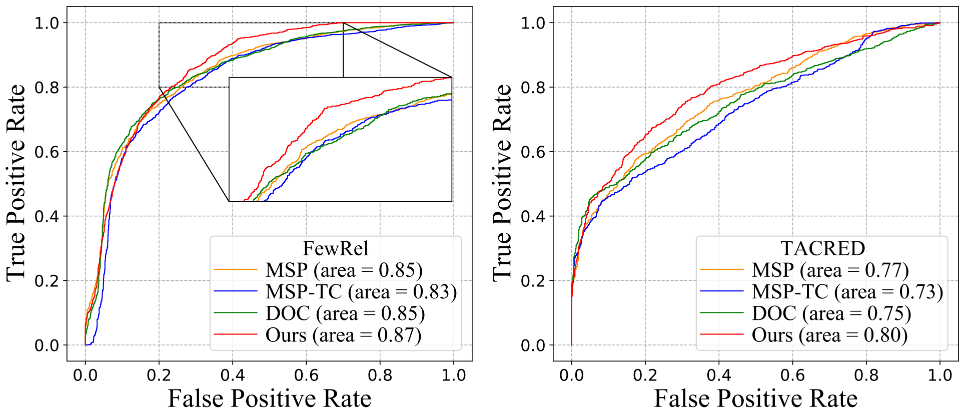

Figure 3: ROC curves on two datasets.

Dataset	Method	25%	50%	75%
FewRel	w/o Feature Mask	0.845	0.807	0.816
	w/o Random Mask	0.846	0.814	0.809
	w/o Token Mask	0.833	0.810	0.803
	w/o Distant	0.810	0.805	0.815
	Ours	**0.852**	**0.824**	**0.827**
TACRED	w/o Feature Mask	0.753	0.728	0.703
	w/o Random Mask	0.740	0.735	0.706
	w/o Token Mask	0.750	0.738	0.706
	w/o Distant	0.716	0.700	0.684
	Ours	**0.758**	**0.742**	**0.715**

Table 4: Abalation study of our method.

similar known relations from FewRel dataset, and randomly select 40 samples for each of them. 100 hard samples of unseen relations misclassified by MSP method are selected to show the superiority of our method. From the visualization results in Figure 4, we can observe that, before training (upper left), the relation representations are scattered in the semantic space. After supervised training (upper right), samples can be roughly divided by relation, but different relations are still close to each other. This is consistent with the *Shortcut* feature in neural network. We note that samples of unseen relations are mixed with known relation samples. After contrastive training (down left), model learns sufficient features to discriminate unseen relations. Therefore, samples of unseen relations are effectively separated. Finally, a best relation representation are obtained by applying both supervised and contrastive optimization (down right).

5.4 A Case Study on OpenRE

For the samples rejected by the classifier, the off-the-shelf OpenRE method can be used to discovery potential unseen relations. In this section, we provide a brief case study to show the discovered unseen relations by SelfORE (Hu et al., 2020). OpenRE module outputs the cluster assignment of these

Extracted surface-form	Golden surface-form
university	schools_attended
was found	founded
charges with	charges
died in	country_of_death
was born in	date_of_birth

Table 5: Extracted and golden surface-form relation names on TACRED.

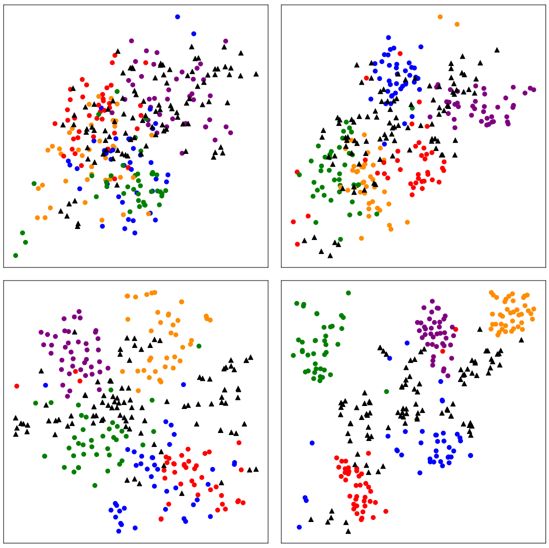

Figure 4: Visualization of the relation representation after t-SNE dimension reduction. The representations are colored with their ground-truth relation labels. Black triangles indicate unknown relations. These four from top left to bottom right sequentially illustrate the relation representation of initial state, after supervised optimization, after contrastive optimization, after both of them.

rejected samples. We extract the relation names using the frequent n-gram in each cluster and the extraction results are shown in table 5. By integrating the OpenRE module, our method complete (1) the classification of known relations, (2) the rejection of unseen relations, (3) discovery of unseen relations. Based on the above process, robust relation extraction in real applications is realized.

6 Conclusions

In this work, we introduce a relation extraction method with rejection option to improve the robustness in real-world applications. The proposed method employs contrastive training techniques and a set of carefully designed transformations to learn sufficient features. The classification of known relations and rejection of unseen relations can be done with these features. Unseen relations in the rejected samples can be discovered by incorporating off-the-shelf OpenRE methods. Experimental results show that our method outperforms SOTA methods for unseen relation rejection.

References

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy, July. Association for Computational Linguistics.

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. 2000. Lof: Identifying density-based local outliers. *SIGMOD Rec.*, 29(2):93–104, May.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Bayu Distiawan, Gerhard Weikum, Jianzhong Qi, and Rui Zhang. 2019. Neural relation extraction for knowledge base enrichment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 229–240.

Jinhua Du, Jingguang Han, Andy Way, and Dadong Wan. 2018. Multi-level structured self-attentions for distantly supervised relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2216–2225, Brussels, Belgium, October-November. Association for Computational Linguistics.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1535–1545, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.

Hervé Gallaire and Jack Minker, editors, 1978. *On Closed World Data Bases*, pages 55–76. Springer US, Boston, MA.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

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

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. *CoRR*, abs/2104.08202.

Xuming Hu, Lijie Wen, Yusong Xu, Chenwei Zhang, and Philip S. Yu. 2020. Selfore: Self-supervised relational feature learning for open relation extraction. *CoRR*, abs/2004.02438.

Yi Yao Huang and William Yang Wang. 2017. Deep residual learning for weakly-supervised relation extraction. *CoRR*, abs/1707.08866.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2019. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *arXiv e-prints*, page arXiv:1907.10529, July.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Shiyu Liang, Yixuan Li, and R. Srikant. 2018. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*.

Ting-En Lin and Hua Xu. 2019. Deep unknown intent detection with margin loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5496, Florence, Italy, July. 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.

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.

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 e-prints*, page arXiv:1907.11692, July.

Ruotian Ma, Tao Gui, Linyang Li, Qi Zhang, Yaqian Zhou, and Xuanjing Huang. 2021. SENT: sentence-level distant relation extraction via negative training. *CoRR*, abs/2106.11566.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using LSTMs on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1116, Berlin, Germany, August. Association for Computational Linguistics.

Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. 2014. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *CoRR*, abs/1412.1897.

Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2020. Learning from Context or Names? An Empirical Study on Neural Relation Extraction. *arXiv e-prints*, page arXiv:2010.01923, October.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR.

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, 07.

Lei Shu, Hu Xu, and Bing Liu. 2017. DOC: Deep open classification of text documents. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2911–2916, Copenhagen, Denmark, September. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously self-attending to all mentions for full-abstract biological relation 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 872–884, New Orleans, Louisiana, June. 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.

Chenyan Xiong, Russell Power, and Jamie Callan. 2017. Explicit semantic ranking for academic search via knowledge graph embedding. In *Proceedings of the 26th international conference on world wide web*, pages 1271–1279.

Guangfeng Yan, Lu Fan, Qimai Li, Han Liu, Xiaotong Zhang, Xiao-Ming Wu, and Albert Y.S. Lam. 2020. Unknown intent detection using Gaussian mixture model with an application to zero-shot intent classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1050–1060, Online, July. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark, September. Association for Computational Linguistics.

Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2205–2215, Brussels, Belgium, October-November. Association for Computational Linguistics.

Ningyu Zhang, Shumin Deng, Zhanlin Sun, Guanying Wang, Xi Chen, Wei Zhang, and Huajun Chen. 2019. Long-tail relation extraction via knowledge graph embeddings and graph convolution networks. 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 3016–3025, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Hanlei Zhang, Hua Xu, and Ting-En Lin. 2021. Deep open intent classification with adaptive decision boundary. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14374–14382.

Jun Zhao, Tao Gui, Qi Zhang, and Yaqian Zhou. 2021. A relation-oriented clustering method for open relation extraction.