Open Set Relation Extraction via Unknown-Aware Training

Jun Zhao[1]*, **Xin Zhao**[1]*, **Wenyu Zhan**[1], **Qi Zhang**[1]†, **Tao Gui**[2]†
Zhongyu Wei[3], **Yunwen Chen**[4], **Xiang Gao**[4] **Xuanjing Huang**[1]
[1]School of Computer Science, Fudan University
[2]Institute of Modern Languages and Linguistics, Fudan University
[3]School of Data Science, Fudan University
[4]DataGrand Information Technology (Shanghai) Co., Ltd.
{zhaoj19,qz,tgui}@fudan.edu.cn,{zhaoxin21,wyzhan21}@m.fudan.edu.cn

Abstract

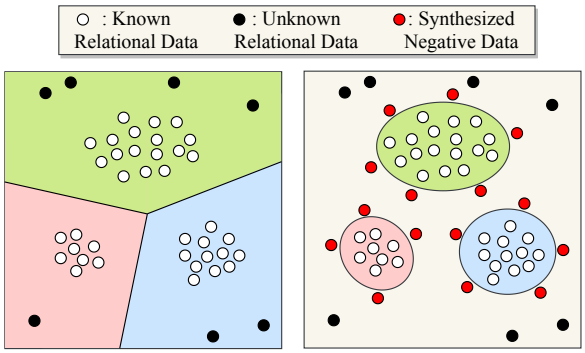
The existing supervised relation extraction methods have achieved impressive performance in a closed-set setting, where the relations during both training and testing remain the same. In a more realistic open-set setting, unknown relations may appear in the test set. Due to the lack of supervision signals from unknown relations, a well-performing closed-set relation extractor can still confidently misclassify them into known relations. In this paper, we propose an unknown-aware training method, regularizing the model by dynamically synthesizing negative instances. To facilitate a compact decision boundary, "difficult" negative instances are necessary. Inspired by text adversarial attacks, we adaptively apply small but critical perturbations to original training instances and thus synthesizing negative instances that are more likely to be mistaken by the model as known relations. Experimental results show that this method achieves SOTA unknown relation detection without compromising the classification of known relations.

1 Introduction

Relation extraction (RE) is an important basic task in the field of natural language processing, aiming to extract the relation between entity pairs from unstructured text. The extracted relation facts have a great practical interest to various downstream applications, such as dialog system (Madotto et al., 2018), knowledge graph (Lin et al., 2015), web search (Xiong et al., 2017), among others.

Many efforts have been devoted to improving the quality of extracted relation facts (Han et al., 2020). Conventional supervised relation extraction is oriented to **known** relations with pre-specified schema. Hence, the paradigm follows a *closed-set setting*, meaning that during both training and testing the relations remain the same. Nowadays, neural RE methods have achieved remarkable success within this setting (Wang et al., 2016; Wu and He, 2019); and in contrast, open relation extraction (OpenRE) is focused on discovering constantly emerging **unknown** relations. Common practices include directly tagging the relational phrases that link entity pairs (Zhan and Zhao, 2020), and clustering instances with the same relation (Hu et al., 2020; Zhao et al., 2021). However, relation extraction in real applications follows an *open-set setting*, meaning that both known and unknown relations are **mixed** within testing data.* This requires that a model can not only distinguish among the known relations, but also filter the instances that express unknown relations. The ability to filter these instances is also called none-of-the-above (NOTA) detection (Gao et al., 2019).

Unfortunately, a well-performing closed-set model can still confidently make arbitrarily wrong predictions when exposed to unknown test data (Nguyen et al., 2015; Recht et al., 2019). As



(a) Supervised decision boundary (b) Expected decision boundary

Figure 1: The decision boundary optimized only on the known relations cannot cope with an open set setting, in which the input may come from the relations unobserved in training. We target at regularizing the decision boundary by synthesizing difficult negative instances.

*Equal Contributions.
†Corresponding authors.

*Some sentences even express no specific relations.

shown in fig. 1 (a), the decision boundary is optimized only on the known relational data (white points), leading to a three-way partition of the whole space. Consequently, the unknown relational data (black points), especially those far from the decision boundary, will be confidently classified into one of the known relations. By contrast, a more compact decision boundary (as shown in fig. 1 (b)) is desirable for NOTA detection. However, the compact decision boundary requires "difficult" negative data (red points in fig. 1 (b)) to be used, so strong supervision signals can be provided. It is important to note that synthesizing such negative data is a non-trivial task.

In this work, we propose an unknown-aware training method, which simultaneously optimizes known relation classification and NOTA detection. To effectively regularize the classification, we iteratively generate negative instances and optimize a NOTA detection score. During the testing phase, instances with low scores are considered as NOTA and filtered out. The key of the method is to synthesize "difficult" negative instances. Inspired by text adversarial attacks, we achieve the goal by substituting a small number of critical tokens in original training instances. This would erase the original relational semantics and the model is not aware of it. By using gradient-based token attribution and linguistic rules, key tokens that express the target relation are found. Then, the tokens are substituted by misleading normal tokens that would cause the greatest increase of NOTA detection score, thus misleading negative instances, which are more likely to be mistaken by the model as known relations, are synthesized. Human evaluation shows that almost all the synthesized negative instances do not express any known relations. Experimental results show that the proposed method learns more compact decision boundary and achieve state-of-the-art NOTA detection performance. Our codes are publicly available at Github.[†]

The contributions are threefold: (1) we propose a new unknown-aware training method for more realistic open-set relation extraction. The method achieves state-of-the-art NOTA detection, without compromising the classification of known relations; (2) the negative instances are more challenging to the model, when compared to the mainstream synthesis method [‡] (e.g., generative adversarial network (GAN)-based method); (3) the comprehensive evaluation and analysis facilitate future research on the pressing but underexplored task.

2 Related Works

Open-set Classification: The open-set setting considers knowledge acquired during training phase to be incomplete, thereby new unknown classes can be encountered during testing. The pioneering explorations in (Scheirer et al., 2013) formalize the open-set classification task, and have inspired a number of subsequent works, which roughly fall into one of the following two groups.

The first group explores model regularization using unknown data. Larson et al. (2019) manually collect unknown data to train a $(n+1)$-way classifier with one additional class, where $(n+1)^{th}$ class represents the unknown class. Instead of manually collecting unknown data, Zheng et al. (2020) generate feature vectors of unknown data using a generative adversarial network (Goodfellow et al., 2014). Zhan et al. (2021) use MixUp technique (Thulasidasan et al., 2019a) to synthesize known data into unknown data.

The second group approaches this problem by discriminative representation learning, which facilitates open-set classification by widening the margin between known and unknown classes. MSP (Hendrycks et al., 2017) is a maximum posterior probability-based baseline and ODIN (Liang et al., 2018) enlarges the difference between known and unknown classes by adding temperature scaling and perturbations to MSP. More recently, different optimization objectives such as large margin loss (Lin and Xu, 2019) and gaussian mixture loss (Yan et al., 2020) are adopted to learn more discriminative representations. Shu et al. (2017); Xu et al. (2020); Zhang et al. (2021) also impose gaussian assumption to data distribution to facilitate distinct unknown data.

Open-set Relation Extraction: Open-set RE is a pressing but underexplored task. Most of the existing RE methods manually collect NOTA data and adopt a $(n+1)$ way classifier to deal with NOTA relations (Zhang et al., 2018; Zhu et al., 2019; Ma et al., 2021). However, the collected NOTA data with manual bias cannot cover all NOTA relations and thus these methods cannot effectively deal with open-set RE (Gao et al., 2019).

[†]https://github.com/XinZhao0211/OpenSetRE.

[‡]A quantitative analysis will be provided in Sec. 5.2.

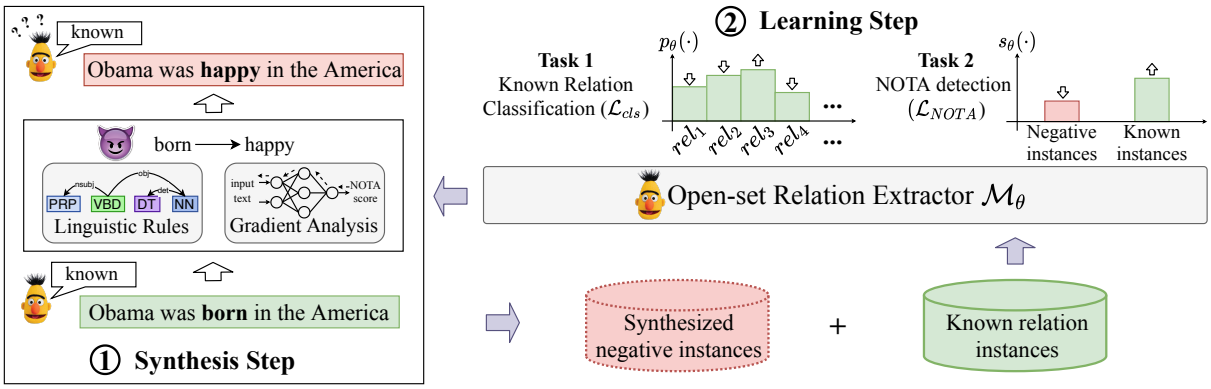


Figure 2: Overview of the proposed unknown-aware training method. The training loop consists of two iteration steps: the synthesis step consists in adaptively synthesizing "difficult" instances according to the states of the model; while in the learning step, an optimization of the dual objectives of both known relation classification and NOTA relation detection is performed, based on the known and synthesized instances.

Our method avoids the bias and the expensive cost of manually collecting NOTA data by automatically synthesizing negative data. Compared with general open-set classification methods, our method takes relational linguistic rules into consideration and outperforms them by a large margin.

3 Approach

We start by formulating the *open-set* relation extraction task. Let $\mathcal{K} = \{r_1, ..., r_n\}$ denote the set of known relations and NOTA indicates that the instance does not express any relation in $\mathcal{K}$. Given a training set $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{N}$ with N positive samples, consisting of relation instance x_i with a pre-specified entity pair [§] and relation $y_i \in \mathcal{K}$, we aim to learn a open-set relation extractor $\mathcal{M} = \{p_\theta(y|x), s_\theta(x)\}$, where θ denote the model parameters. $p_\theta(y|x)$ is the classification probability on the known relations (The NOTA label is excluded from $p_\theta(y|x)$). NOTA detection score $s_\theta(x)$ is used to distinguish between known relations and NOTA. x is classified as NOTA if $s_\theta(x)$ is less than the threshold α. Conversely, x is classified into a known relation $\hat{y} = \arg\max_y p_\theta(y|x)$.

3.1 Method Overview

We approach the problem by an unknown-aware training method, which dynamically synthesizes "difficult" negative instances and optimizes the dual objectives of both known relation classification and NOTA detection. As shown in fig. 2, the training loop consists of two iteration steps:

① **Synthesis Step**: This step aims to synthesize "difficult" negative instances for model regularization. We draw inspiration from text adversarial attacks to achieve the goal. Specifically, $\mathcal{B} = \{(x_i, y_i)\}_{i=1}^{B}$ represents a training batch sampled from $\mathcal{D}_{\text{train}}$. For each $(x, y) \in \mathcal{B}$, we synthesize a negative instance by substituting the key relational tokens of x with misleading tokens. First, both the attribution method and relational linguistic rules are used to find key tokens expressing the target relation y. Second, the misleading token w_i^{mis} is searched for each key token w_i, along the direction of the gradient $\nabla_{\boldsymbol{w}_i} s_\theta(x)$. By substituting w_i with w_i^{mis}, it is expected for $s_\theta(x)$ to experience its greatest increase, so it is difficult for the model to correctly detect the derived negative instance x' as NOTA.

② **Learning Step**: This step aims to optimize the open-set relation extractor $\mathcal{M} = \{p_\theta(y|x), s_\theta(x)\}$. Based on the training batch $\mathcal{B}$ from $\mathcal{D}_{\text{train}}$, we optimize $p_\theta(y|x)$ to accurately classify known relations. To effectively detect NOTA instances, we further synthesize negative batch $\mathcal{B}' = \{(x_i', \text{NOTA})\}_{i=1}^{B}$ and optimize the model to widen the gap of $s_\theta(x)$ between $x \in \mathcal{B}$ and $x' \in \mathcal{B}'$. Consequently, instances with low $s_\theta(x)$ scores are filtered out before being fed into $p_\theta(y|x)$.

Next, we elaborate on the model structure of $\mathcal{M}$ (sec. 3.2) and the technical details of the synthesis step (sec. 3.3) and the learning step (sec. 3.4).

3.2 Open-set Relation Extractor

Instance Encoder and Classifier: Given an input instance $x = \{w_1, .., w_n\}$ with four reserved

[§]We assume that the entity recognition has already been done and an instance expresses at most one relation between the entity pair.

special tokens $[E_1], [\backslash E_1], [E_2], [\backslash E_2]$ marking the beginning and end of the head and tail entities, the instance encoder aims to encode the relational semantics into a fixed-length representation $\boldsymbol{h} = \boldsymbol{enc}(x) \in \mathbb{R}^d$. We adopt BERT (Devlin et al., 2018), a common practice, as the implementation of the encoder. We follow Baldini Soares et al. (2019) to concatenate the hidden states of special tokens $[E_1]$ and $[E_2]$ as the representation of the input instance.

$$\boldsymbol{w}_1, .., \boldsymbol{w}_n = \text{BERT}(w_1, .., w_n) \quad (1)$$

$$\boldsymbol{h} = \boldsymbol{w}_{[E_1]} \oplus \boldsymbol{w}_{[E_2]}, \quad (2)$$

where $\boldsymbol{w}_i, \boldsymbol{w}_{[E_1]}, \boldsymbol{w}_{[E_2]}$ denotes the hidden states of token $w_i, [E_1], [E_2]$, respectively. $\oplus$ denotes the concatenation operator. The classification probability on known relations $p_\theta(\cdot|x)$ can be derived through a linear head $\boldsymbol{\eta}(\cdot)$:

$$\boldsymbol{\eta}(\boldsymbol{h}) = W_{\text{cls}}\boldsymbol{h} + b \quad (3)$$

$$p_\theta(\cdot|x) = \text{Softmax}(\boldsymbol{\eta}(\boldsymbol{h})), \quad (4)$$

where $W_{\text{cls}} \in \mathbb{R}^{n \times d}$ is the weight matrix transforming the relation representation to the logits on n known relations and b is the bias.

NOTA Detection Score: The goal of distinguishing between known and NOTA relations requires the modeling of the data density. However, directly estimating $\log p(x)$ can be computationally intractable because it requires sampling from the entire input space. Inspired by Liu et al. (2020) in the image understanding task, the free energy function $E(\boldsymbol{h})$ is theoretically proportional to the probability density of training data. Considering that it can be easily derived from the linear head $\boldsymbol{\eta}(\cdot)$ without additional calculation, the negative free energy function is used to compute the NOTA detection score as follows:

$$s_\theta(x) = -E(\boldsymbol{h}) = \log \sum_{j=1}^{n} e^{\boldsymbol{\eta}(\boldsymbol{h})_j}, \quad (5)$$

where $\boldsymbol{\eta}(\boldsymbol{h})_j$ denotes the j^{th} logit value of $\boldsymbol{\eta}(\boldsymbol{h})$. The detection score has shown to be effective in out-of-distribution detection (Liu et al., 2020). Based on the classification probability $p_\theta(\cdot|x)$ and NOTA detection score $s_\theta(x)$, the open-set relation extractor $\mathcal{M}$ works in the following way:

$$\hat{y} = \begin{cases} \arg\max_y p_\theta(y|x) & \mathcal{S}(x) > \alpha \\ \texttt{NOTA} & \mathcal{S}(x) \leq \alpha, \end{cases} \quad (6)$$

where α is the detection threshold.

3.3 Iterative Negative Instances Synthesis

"Difficult" negative instances are the key to effective model regularization. $x = \{w_1, .., w_n\}$ is a training instance with a label y. To synthesize negative instance x', we perturb each key token w_i, which expresses the relation y, with a misleading token w_i^{mis}. The substitutions are expected to erase original relational semantics without the model being aware of it. Based on the attribution technique and relational linguistic rules, a score $I(w_i, x, y)$ is developed to measure the contribution of a token $w_i \in x$ to relation y as follows:

$$I(w_i, x, y) = a(w_i, x) \cdot t(w_i, y) \cdot dp(w_i, x), \quad (7)$$

where $a(w_i, x)$ denotes an attribution score reweighted by two linguistic scores $t(w_i, y)$, $dp(w_i, x)$. We rank all tokens according to $I(w_i, x, y)$ in descending order and take the first ϵ percent of tokens as key tokens to perform substitutions. Next, we elaborate on (1) how to calculate the attribution score $a(w_i, x)$ and linguistic scores $t(w_i, y)$, $dp(w_i, x)$; (2) how to select misleading tokens for substitution.

Gradient-based Token Attribution: Ideally, when the key tokens are removed, instance x will no longer express the original known relation y, and the NOTA detection score $s_\theta(x)$ would drop accordingly. Therefore, the contribution of a token w_i to relational semantics can be measured by a counterfactual:

$$c(w_i, x) = s_\theta(x) - s_\theta(x_{-w_i}), \quad (8)$$

where x_{-w_i} is the instance after removing w_i. However, to calculate the contribution of each token in instance x, n forward passes are needed, which is highly inefficient. Fortunately, a first-order approximation of contribution $c(w_i, x)$ can be obtained by calculating the dot product of word embedding $\boldsymbol{w}_i$ and the gradient of $s_\theta(x)$ with respect to $\boldsymbol{w}_i$, that is $\nabla_{\boldsymbol{w}_i} s_\theta(x) \cdot \boldsymbol{w}_i$ (Feng et al., 2018). The contribution of n tokens can thus be computed with a single forward-backward pass. Finally, a normalized attribution score is used, in order to represent the contribution of each token:

$$a(w_i, x) = \frac{|\nabla_{\boldsymbol{w}_i} s_\theta(x) \cdot \boldsymbol{w}_i|}{\sum_{j=1}^{n} |\nabla_{\boldsymbol{w}_j} s_\theta(x) \cdot \boldsymbol{w}_j|}. \quad (9)$$

Linguistic Rule-based Token Reweighting: As a supplement to the attribution method, linguistic

rules that describe the pattern of relational phrases can provide valuable prior knowledge for the measure of tokens' contribution. Specifically, the following two rules are used. Rule 1: *If a token w_i significantly contributes to relation y, it should appear more frequently in the instances of y, and rarely in the instances of other relations*. By following this rule, `tf-idf` statistic (Salton and Buckley, 1987) $t(w_i, y)$[¶] is used to reflect the contribution of token w_i to relation y (Appendix A.1 contains additional details about the statistic). Rule 2: *Tokens that are part of the dependency path between the entity pair usually express the relation between the entity pair, while shorter dependency paths are more likely to represent the relation* (ElSahar et al., 2018). Following the rule, stanza[‖] is used to parse the instance and the dependency score as calculated as follows:

$$dp(w_i, x) = \begin{cases} |x|/|\mathcal{T}| & w_i \in \mathcal{T} \\ 1, & otherwise, \end{cases} \tag{10}$$

where $\mathcal{T}$ denotes the set of tokens in the dependency path between the entity pair. $|x|$, $|\mathcal{T}|$ denote the number of tokens in instance x and set $\mathcal{T}$, respectively. Eq. 10 indicates that the tokens in $\mathcal{T}$ are given a higher weight, and the shorter the path, the higher the weight.

Misleading Token Selection: Negative instances are synthesized by substituting key tokens with misleading tokens. Note that we have obtained the gradient of $s_\theta(x)$ with respect to each token w_i in the attribution step. Based on the gradient vectors, a misleading token is selected from vocabulary $\mathcal{V}$ for each key token w_i as follows:

$$w_i^{\text{mis}} = \arg\max_{w_j \in \mathcal{V}} \nabla_{\boldsymbol{w}_i} s_\theta(x) \cdot \boldsymbol{w}_j. \tag{11}$$

Substituting w_i with w_i^{mis} is expected to cause the greatest increase in $s_\theta(x)$, so the synthesized negative instance is misleading to the model. To avoid that w_i^{mis} is also a key token of a known relation, the top 100 tokens with the highest `tf-idf` statistic of each relation are removed from the vocabulary $\mathcal{V}$, when performing the substitution. Human evaluation results show that almost all the synthesized negative instances do not express any known relation. In addition, we provide two real substitution cases in tab. 7.

[¶]The statistic is based on the whole training set and does not change with a specific instance x.

[‖]https://stanfordnlp.github.io/stanza/depparse.html

3.4 Unknown-Aware Training Objective

In this section, we introduce the unknown-aware training objective for open-set relation extraction. Based on the synthesized negative samples, an optimization of the dual objectives of both known relation classification and NOTA relation detection is performed. Specifically, at the m^{th} training step, A batch of training data $\mathcal{B}_m = \{(x_i, y_i)\}_{i=1}^B$ is sampled from $\mathcal{D}_{\text{train}}$. Cross entropy loss is used for the optimization of known relation classification:

$$\mathcal{L}_{cls} = \frac{1}{B}\sum_{i=1}^{B}(-\log p_\theta(y_i|x_i)), \tag{12}$$

where $p_\theta(\cdot|x_i)$ is the classification probability on the known relations (eq. 4). For each instance x in $\mathcal{B}_m$, we synthesize a negative sample x' as described in sec. 3.3, and finally obtain a batch of negative samples $\mathcal{B}'_m = \{(x'_i, \texttt{NOTA})\}_{i=1}^B$. To learn a compact decision boundary for NOTA detection, we use the binary sigmoid loss to enlarge the gap of detection scores $s_\theta(\cdot)$ between known and synthesized instances as follows:

$$\begin{aligned} \mathcal{L}_{\text{NOTA}} = &-\frac{1}{B}\sum_{i=1}^{B}\log\sigma(s_\theta(x_i)) \\ &-\frac{1}{B}\sum_{i=1}^{B}\log(1-\sigma(s_\theta(x'_i))) \end{aligned} \tag{13}$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the `sigmoid` function. The overall optimization objective is as follows:

$$\mathcal{L} = \mathcal{L}_{cls} + \beta \cdot \mathcal{L}_{\text{NOTA}}, \tag{14}$$

where β is a hyper-parameter to balance the two loss term.

4 Experimental Setup

4.1 Datasets

FewRel (Han et al., 2018). FewRel is a human-annotated dataset, which contains 80 types of relations, each with 700 instances. We take the top 40 relations as known relations. The middle 20 relations are taken as unknown relations for validation. And the remaining 20 relations are unknown relations for testing. Our training set contains 22,400 instances from the 40 known relations. Both the validation and test set consist of 5,600 instances, of which 50% are from unknown relations. **Note that the unknown relations in the test set and the validation set do not overlap.**

Method	FewRel			TACRED		
	ACC↑	AUROC↑	FPR95↓	ACC↑	AUROC↑	FPR95↓
MSP (Hendrycks et al., 2017)	$63.69_{1.71}$	$83.60_{2.12}$	$62.93_{4.05}$	$71.83_{1.99}$	$89.24_{0.32}$	$43.20_{4.15}$
DOC (Shu et al., 2017)	$63.96_{1.00}$	$84.46_{0.97}$	$59.38_{1.92}$	$70.08_{0.59}$	$89.40_{0.25}$	$42.83_{1.66}$
ODIN (Liang et al., 2018)	$66.78_{1.57}$	$84.47_{2.16}$	$55.98_{3.03}$	$72.37_{2.32}$	$89.42_{0.30}$	$40.83_{3.09}$
MixUp (Thulasidasan et al., 2019b)	$66.30_{0.45}$	$84.95_{1.38}$	$57.44_{0.37}$	$72.85_{1.60}$	$89.80_{0.59}$	$40.30_{3.77}$
Energy (Liu et al., 2020)	$71.54_{1.05}$	$85.53_{1.84}$	$46.88_{1.50}$	$75.15_{0.14}$	$90.34_{0.12}$	$35.30_{2.86}$
Convex (Zhan et al., 2021)	$71.19_{1.51}$	$86.23_{0.81}$	$46.00_{2.67}$	$71.55_{1.17}$	$90.16_{0.58}$	$37.40_{3.28}$
SCL (Zeng et al., 2021)	$65.52_{1.48}$	$86.71_{1.23}$	$58.04_{3.24}$	$72.70_{2.17}$	$90.22_{0.67}$	$35.80_{3.67}$
Ours	$\mathbf{74.00}_{0.56}$	$\mathbf{88.73}_{0.67}$	$\mathbf{41.17}_{1.37}$	$\mathbf{76.97}_{1.81}$	$\mathbf{91.02}_{0.59}$	$\mathbf{30.27}_{2.29}$

Table 1: Main results of open-set relation extraction. The subscript represents the corresponding standard deviation (e.g., $74.00_{0.56}$ indicates 74.00 ± 0.56). The results of **ACC** on n known relations are provided in tab.6.

TACRED (Zhang et al., 2017). TACRED is a large-scale relation extraction dataset, which contains 41 relations and a `no_relation` label indicating no defined relation exists. Similar to FewRel, we take the top 21 relations as known relations. The middle 10 relations are taken as unknown relations for validation. The remaining 10 relations and `no_relation` are unknown relations for testing. We randomly sample 9,784 instances of known relations to form the training set. Both the validation and test set consist of 2,000 instances, of which 50% are from unknown relations. Unknown relations in the validation set and the test set still do not overlap.

For the specific composition of relations in each dataset, please refer to Appendix A.4.

4.2 Compared Methods

To evaluate the effectiveness of the proposed method, we compare our method with mainstream open-set classification methods, which can be roughly grouped into the following categories: **MSP** (Hendrycks et al., 2017), **DOC** (Shu et al., 2017), **ODIN** (Liang et al., 2018), **Energy** (Liu et al., 2020), and **SCL** (Zeng et al., 2021) detect unknown data through a carefully designed score function or learning a more discriminative representation. No synthesized negative instances are used in these methods. **MixUp** (Thulasidasan et al., 2019b), and **Convex** (Zhan et al., 2021) use synthesized negative instances to regularize the model. Please refer to the appendix A.3 for a brief introduction to these methods.

We do not compare **BERT-PAIR** (Gao et al., 2019) because it is only applicable to the few-shot setting. We use **DOC** (Shu et al., 2017) with a BERT encoder as an alternative method for it.

4.3 Metrics

Following previous works (Liu et al., 2020; Zeng et al., 2021), we treat all unknown instances as one `NOTA` class and adopt three widely used metrics for evaluation. (1) **FPR95**: The false positive rate of `NOTA` instances when the true positive rate of known instances is at 95%. The smaller the value, the better. (2) **AUROC**: the area under the receiver operating characteristic curve. It is a threshold-free metric that measures how well the detection score ranks the instances of known and `NOTA` relations. (3) **ACC**: The classification accuracy on n known relations and one `NOTA` relation, measuring the overall performance of open-set RE.

4.4 Implementation Details

We use the AdamW as the optimizer, with a learning rate of $2e-5$ and batch size of 16 for both datasets. Major hyperparameters are selected with grid search according to the model performance on a validation set. The detection threshold is set to the value at which the true positive rate of known instances is at 95%. The regularization weight β is 0.05 selected from $\{0.01, 0.05, 0.1, 0.15, 0.5\}$. See the appendix A.2 for the processing of sub-tokens. The dependency parsing is performed with stanza 1.4.2. All experiments are conducted with Python 3.8.5 and PyTorch 1.7.0, using a GeForce GTX 2080Ti with 12GB memory.

5 Results and Analysis

5.1 Main Results

In this section, we evaluate the proposed method by comparing it with several competitive open-set classification methods. The results are reported in tab. 1, from which we can observe that our method

Method	FewRel				TACRED			
	ACC↑	AUROC↑	FPR95↓	Δs_θ ↓	ACC↑	AUROC↑	FPR95↓	Δs_θ ↓
Baseline	71.54	85.53	46.88	–	75.15	90.34	35.30	–
Gaussian	71.81	86.67	46.81	4.35	74.73	90.16	35.47	4.48
Gaussian†	<u>72.93</u>	86.66	<u>42.69</u>	**0.02**	75.17	90.38	34.73	**0.03**
MixUp	72.86	86.17	43.90	2.34	75.95	89.35	<u>33.20</u>	1.90
Real	71.75	86.52	46.08	3.55	<u>76.10</u>	89.92	33.67	3.91
GAN	72.11	<u>86.77</u>	45.69	4.01	76.06	<u>90.46</u>	34.30	4.10
Ours	**74.00**	**88.73**	**41.17**	<u>1.73</u>	**76.97**	**91.02**	**30.27**	<u>1.36</u>

Table 2: The unknown-aware training with various negative instance synthesis methods. The numbers in **bold** and <u>underlined</u> indicate the best and second-best results, respectively. To quantify the difficulty of negative instances, we calculate the average difference Δs_θ between the NOTA detection score of known and negative instances. Obviously, the smaller the difference, the more difficult it is for the model to distinguish the two types of instances.

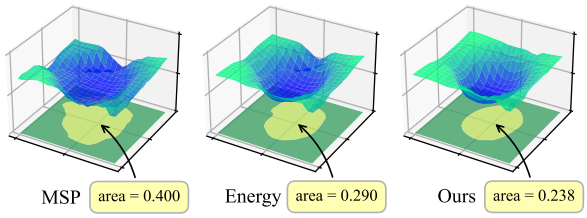


Figure 3: Decision boundary visualization. Energy can be seen as a degenerate version of our method when removing unknown-aware training. The vertical axis represents the difference between the detection threshold α and the NOTA score $s_\theta(x)$, normalized to the range of $[-1, 1]$. When an instance falls within the yellow region below zero, the model classifies it as a known relation. Conversely, when a sentence falls within the green region above zero, the model identifies it as NOTA.

achieves state-of-the-art NOTA detection (reflected by FPR95 and AUROC) without compromising the classification of known relations (reflected by ACC). In some baseline methods (e.g., MSP, ODIN, Energy, SCL), only instances of known relations are used for training. Compared with them, we explicitly synthesize the negative instances to complete the missing supervision signals, and the improvement in NOTA detection shows the effectiveness of the unknown-aware training. To intuitively show the changes of the decision boundary, we use the method of Yu et al. (2019) to visualize the decision boundary of the model in the input space. As can be seen from fig. 3, a more compact decision boundary is learned with the help of unknown-aware training. Although methods such as MixUp, and Convex also synthesized negative instances, our method is still superior to them. This may be due to the fact that our negative instances are more difficult and thus beneficial for an effective model regularization (we provide more results in sec. 5.2 to support the claim).

5.2 Negative Instance Synthesis Analysis

In this section, the unknown-aware training objective is combined with the various negative instance synthesis methods to fairly compare the performance of these synthesis methods. The results are shown in tab. 2. `Baseline` means no negative instances are used. `Gaussian` takes Gaussian noise as negative instances and `Gaussian`† adds the noise to known instances. `MixUp` synthesizes negative instances by convexly combining pairs of known instances. `Real` means using real NOTA instances**. `GAN` synthesizes negative instances by Generative Adversarial Network (Ryu et al., 2018).

Correlation between effectiveness and difficulty. (1) `Gaussian` with the largest Δs_θ performs even worse than `Baseline` in TACRED, suggesting that overly simple negative instances are almost ineffective for model regularization. (2) Our method synthesizes the second difficult negative instances (reflected by Δs_θ) and achieves the best performance (reflected by ACC, AUROC, FPR95), which shows that the difficult negative instances are very beneficial for effective model regularization. (3) The difficulty of negative instances of competitive methods (e.g., `MixUp`, `Real`, `GAN`) is lower than that of `Ours`, which indicates that it is non-trivial to achieve our difficulty level. (4) Although `Gaussian`† synthesizes

**We use the data from SemEval-2010 (Hendrickx et al., 2010). The overlap relations are manually removed.

Dataset	NOTA	Known-Original	Known-Other	Controversial
FewRel	92	2	1	5
TACRED	90	3	0	7

Table 3: Human evaluation of our negative instances. More than 90% of the negative instances do not express any known relations.

the most difficult negative instances, our method still significantly outperforms `Gaussian`†. One possible reason is that overly difficult instances may express the semantics of known relations. This leads to the following research question.

Do our synthetic negative instances really not express any known relations? We conduct human evaluation to answer this question. Specifically, we randomly select 100 synthesized negative instances on each dataset and asked human judges whether these instances express known or `NOTA` relations. The evaluation is completed by three independent human judges. We recruit 3 graduates in computer science and English majors from top universities. All of them passed a test batch. Each graduate is paid $8 per hour. The results are shown in tab. 3, from which we can observe that: (1) More than 90% of the negative instances do not express any known relations (`NOTA`). (2) Very few instances remain in the original known relations (`Known-Original`) or are transferred to another known relation (`Known-Other`). (3) There are also some instances that are `Controversial`. Some volunteers believe that the instances express known relations, while others believe that the instances are `NOTA`. In general, our synthesis method achieves satisfactory results, but there is still potential for further improvement.

5.3 Ablation Study

To study the contribution of each component in our method, we conduct ablation experiments on the two datasets and show the results in tab. 4. First, the attribution score measures the impact of a token on NOTA detection of the model. The dependency score and `tf-idf` statistic reflect the matching degree between a token and the relational linguistic rules. When the three scores are removed, there may be some key relational phrases that can not be correctly identified and the performance decline accordingly. It is worth mentioning that the model parameters change dynamically with the training process, thus

Method	ACC↑	AUROC↑	FPR95↓
w/o attribution score	73.81	88.34	41.32
w/o dependency score	73.89	88.55	41.88
w/o tfidf statistic	73.92	87.64	42.42
w/o iterative synthesis	72.61	86.90	44.71
w/o misleading tokens	71.87	86.99	46.35
Ours	**74.00**	**88.73**	**41.17**
w/o attribution score	75.47	90.71	35.10
w/o dependency score	76.73	90.93	30.57
w/o tfidf statistic	76.68	90.46	34.43
w/o iterative synthesis	76.75	90.57	32.77
w/o misleading tokens	75.80	90.41	33.53
Ours	**76.97**	**91.02**	**30.27**

Table 4: Abalation study of our method. The upper (resp. lower) part lists the results on FewRel (resp. TACRED).

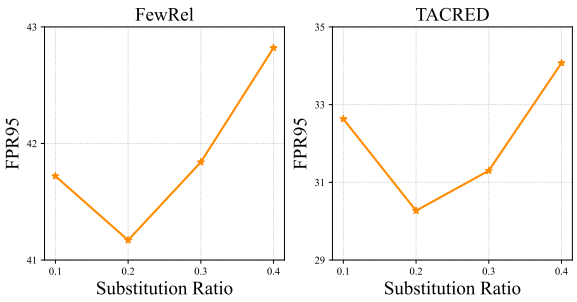


Figure 4: FPR95 with different substitution ratio.

iteratively synthesizing negative instances is crucial for effective regularization. When the practice is removed, the static negative instances can not reflect the latest state of the model, and thus the performance degrades significantly. Finally, we remove misleading token selection by substituting the identified key tokens with a special token `[MASK]` and the performance is seriously hurt, which indicates that misleading tokens play an important role in synthesizing difficult instances.

5.4 Hyper-parameter Analysis

We synthesize negative instances by substituting ϵ percent of key tokens with misleading tokens. In this section, we conduct experiments to study the influence of substitution ratio ϵ on NOTA detection. From fig. 4 we obtain the following observations. When the substitution ratio gradually increases from 0, the performance of NOTA detection is also improved (Note that the smaller the value of FPR95, the better). This means that an overly small substitution ratio is not sufficient to remove all relational phrases. The residual relational tokens are detrimental to model regularization. When the substitution ratio exceeds a certain threshold (i.e.,

0.2), a continued increase in the substitution ratio will lead to a decline in detection performance. One possible reason is that too high a substitution ratio can severely damage the original sentence structure, resulting in negative instances that differ too much from the real NOTA instances.

6 Conclusions

In this work, we propose an unknown-aware training method for open-set relation extraction, which is a pressing but underexplored task. We dynamically synthesize negative instances by the attribution technique and relational linguistic rules to complete the missing supervision signals. The negative instances are more difficult than that of other competitive methods and achieve effective model regularization. Experimental results show that our method achieves state-of-the-art NOTA detection without compromising the classification of known relations. We hope our method and analysis can inspire future research on this task.

Limitations

We synthesize negative instances by substituting relational phrases with misleading tokens. However, the relational semantics in some instances may be expressed implicitly. That is, there are no key tokens that directly correspond to the target relation. Therefore, we cannot synthesize negative instances based on these instances. Additionally, we consider substitution ratio ϵ as a fixed hyperparameter. It may be a better choice to dynamically determine ϵ based on the input instance. We leave these limitations as our future work.

Acknowledgements

The authors wish to thank the anonymous reviewers for their helpful comments. This work was partially funded by National Natural Science Foundation of China (No.62206057,62076069,61976056), Shanghai Rising-Star Program (23QA1400200), Program of Shanghai Academic Research Leader under grant 22XD1401100, and Natural Science Foundation of Shanghai (23ZR1403500).

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. Association for Computational Linguistics.

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.

Hady ElSahar, Elena Demidova, Simon Gottschalk, Christophe Gravier, and Frédérique Laforest. 2018. Unsupervised open relation extraction. *CoRR*, abs/1801.07174.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium. Association for Computational Linguistics.

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.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Xu Han, Tianyu Gao, Yankai Lin, Hao Peng, Yaoliang Yang, Chaojun Xiao, Zhiyuan Liu, Peng Li, Jie Zhou, and Maosong Sun. 2020. More data, more relations, more context and more openness: A review and outlook for relation extraction. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 745–758, Suzhou, 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. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

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

Uppsala, Sweden. Association for Computational Linguistics.

Dan Hendrycks, Kevin Gimpel, 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.

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

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. 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 1311–1316, Hong Kong, China. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2181–2187. AAAI Press.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. *neural information processing systems*.

Ruotian Ma, Tao Gui, Linyang Li, Qi Zhang, Xuanjing Huang, and Yaqian Zhou. 2021. SENT: Sentence-level distant relation extraction via negative training. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6201–6213, Online. Association for Computational Linguistics.

Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Mem2Seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1468–1478, Melbourne, Australia. Association for Computational Linguistics.

Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436.

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.

Seonghan Ryu, Sangjun Koo, Hwanjo Yu, and Gary Geunbae Lee. 2018. Out-of-domain detection based on generative adversarial network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 714–718, Brussels, Belgium. Association for Computational Linguistics.

Gerard Salton and Chris Buckley. 1987. Term weighting approaches in automatic text retrieval. Technical report, USA.

Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. 2013. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772.

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. Association for Computational Linguistics.

Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. 2019a. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. 2019b. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems*, 32.

Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan Liu. 2016. Relation classification via multi-level attention CNNs. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1298–1307, Berlin, Germany. Association for Computational Linguistics.

Shanchan Wu and Yifan He. 2019. Enriching pre-trained language model with entity information for relation classification. *CoRR*, abs/1905.08284.

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.

Hong Xu, Keqing He, Yuanmeng Yan, Sihong Liu, Zijun Liu, and Weiran Xu. 2020. A deep generative distance-based classifier for out-of-domain detection with mahalanobis space. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1452–1460, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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. Association for Computational Linguistics.

Fuxun Yu, Zhuwei Qin, Chenchen Liu, Liang Zhao, Yanzhi Wang, and Xiang Chen. 2019. Interpreting and evaluating neural network robustness. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4199–4205. International Joint Conferences on Artificial Intelligence Organization.

Zhiyuan Zeng, Keqing He, Yuanmeng Yan, Zijun Liu, Yanan Wu, Hong Xu, Huixing Jiang, and Weiran Xu. 2021. Modeling discriminative representations for out-of-domain detection with supervised contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 870–878, Online. Association for Computational Linguistics.

Junlang Zhan and Hai Zhao. 2020. Span model for open information extraction on accurate corpus. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9523–9530.

Li-Ming Zhan, Haowen Liang, Bo Liu, Lu Fan, Xiao-Ming Wu, and Albert Y.S. Lam. 2021. Out-of-scope intent detection with self-supervision and discriminative training. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3521–3532, Online. 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.

Yuhao Zhang, Peng Qi, and Christopher D. Manning. 2018. Graph convolution over pruned dependency trees improves relation extraction. *CoRR*, abs/1809.10185.

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. Association for Computational Linguistics.

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

Yinhe Zheng, Guanyi Chen, and Minlie Huang. 2020. Out-of-domain detection for natural language understanding in dialog systems. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 28:1198–1209.

Hao Zhu, Yankai Lin, Zhiyuan Liu, Jie Fu, Tat-Seng Chua, and Maosong Sun. 2019. Graph neural networks with generated parameters for relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1331–1339, Florence, Italy. Association for Computational Linguistics.

A Appendix

A.1 Tf-idf statistic

We consider a token w_i to contribute significantly to a known relation $y \in \mathcal{K}$ if it occurs frequently in the instances of relation y and rarely in the instances of other relations. `Tf-idf` statistic (Salton and Buckley, 1987) can well characterize this property. Specifically, `Tf-idf` consists of term frequency and inverse document frequency. The term frequency $tf(w_i, y)$ describes how often a token w_i appears in the instances of relation y:

$$tf(w_i, y) = \frac{n(w_i, y)}{\sum_{w_j \in \mathcal{V}} n(w_j, y)}, \qquad (15)$$

where $n(w_i, y)$ denotes the number of times the token w_i appears in the instances of relation y. Obviously, some tokens (e.g., the stop words) have high tf values in different relational instances. However, they do not contribute to the relational semantics. The inverse document frequency describes whether the token w_i appears only in the instances of specific relations:

$$idf(w_i) = log \frac{|\mathcal{K}|}{|\{y : n(w_i, y) \neq 0\}|}, \qquad (16)$$

where $|\mathcal{K}|$ denotes total number of known relations and $|\{y : n(w_i, y) \neq 0\}|$ denotes the number of known relations that token w_i appears in their instances. Finally, we calculate $t(w_i, y)$ as follows:

$$t(w_i, y) = tf(w_i, y) \times idf(w_i). \quad (17)$$

The `tf-idf` statistic $t(w_i, y)$ measures the contribution of token w_i to the relation semantics of y. We calculate and store the statistics based on the entire training set $\mathcal{D}_{\text{train}}$ before the training loop start. During the training, the statistic of each token in the vocabulary is fixed.

A.2 How to Deal With Sub-tokens?

BERT adopts BPE encoding to construct vocabularies. While most tokens are still single tokens, rare tokens are tokenized into sub-tokens. In this section, we introduce how to deal with sub-tokens when performing the substitution. First, the `tf-idf` statistics and the dependency scores are calculated at the token level and require no additional process. If a token consists of n sub-tokens, we calculate its attribution score by summing the scores of all its sub-tokens. In addition, the misleading token of this token is only selected from the tokens that also have n sub-tokens according to $\arg\max_{w_j \in \mathcal{V}_n} \sum_{k=1}^{n} \nabla_{\boldsymbol{w}_{i,k}} s_\theta(x) \cdot \boldsymbol{w}_{j,k}$. $\mathcal{V}_n$ denotes a vocabulary, in which all tokens consist of n sub-tokens. $\boldsymbol{w}_{i,k}$ denotes the embedding of the k^{th} sub-token of the token w_i.

A.3 Compared Methods

To validate the effectiveness of the proposed method, we compare our method with mainstream open-set classification methods.

MSP (Hendrycks et al., 2017). MSP assumes that correctly classified instances tend to have greater maximum softmax probability than samples of unknown classes. Therefore, the maximum softmax probability is used as the detection score.

DOC (Shu et al., 2017). DOC builds a 1-vs-rest layer containing m binary sigmoid classifiers for m known classes. The maximum probability of m binary classifiers is used as the detection score.

ODIN (Liang et al., 2018). Based on MSP, ODIN uses temperature scaling and small perturbations to separate the softmax score distributions between samples of known and unknown classes.

MixUp (Thulasidasan et al., 2019b). MixUp trains the model on convexly combined pairs of instances, which is effective to calibrate the softmax scores.

Energy (Liu et al., 2020). Instead of maximum softmax probability, this method uses the free energy $E(x) = -\log \sum_{k=1}^{K} e^{f_k(x)}$ as the detection score of the unknown data.

Convex (Zhan et al., 2021). The method learns a more discriminative representation by generating synthetic outliers using inlier features.

SCL (Zeng et al., 2021). SCL proposes a supervised contrastive learning objective, learning a more discriminative representation for unknown data detection.

A.4 Relations comprising the datasets

In this subsection, we present the known relations contained in the training set, the unknown relations included in the validation set, and the unknown relations present in the test set, as shown in Table 5.

Relations in FewRel: **Training Set**: P241, P22, P460, P4552, P140, P39, P118, P674, P361, P1408, P410, P931, P1344, P1303, P1877, P407, P105, P3450, P991, P800, P40, P551, P750, P106, P364, P706, P127, P150, P131, P159, P264, P102, P974, P84, P155, P31, P740, P26, P177, P206 **Validation Set**: P135, P403, P1001, P59, P25, P412, P413, P136, P178, P1346, P921, P123, P17, P1435, P306, P641, P101, P495, P466, P58 **Testing Set**: P57, P6, P2094, P1923, P463, P1411, P710, P176, P355, P400, P449, P276, P156, P137, P27, P527, P175, P3373, P937, P86
Relations in FewRel: **Training Set**: per:stateorprovince_of_death, org:shareholders, org:alternate_names, per:country_of_birth, org:city_of_headquarters, per:age, per:cities_of_residence, per:children, org:members, org:founded per:title, org:website, per:alternate_names, org:country_of_headquarters, per:stateorprovinces_of_residence, per:cause_of_death, per:charges org:political_religious_affiliation, org:parents, org:dissolved, per:spouse, **Validation Set**: org:subsidiaries, per:city_of_birth, per:date_of_death, per:stateorprovince_of_birth, per:employee_of, org:member_of, per:origin, per:date_of_birth, per:countries_of_residence, org:founded_by **Testing Set**: org:stateorprovince_of_headquarters, per:country_of_death, per:religion, per:city_of_death, org:number_of_employees_members, per:parents, per:schools_attended, per:siblings, per:other_family, org:top_members_employees, no_relation

Table 5: Relations comprising each dataset.

A.5 Additional Results

Classification Accuracy: One of our key claims is that the proposed method achieves state-of-the-art SOTA detection without compromising the classification of known relations. In this section, we provide an additional **ACC** metric, in which only the instances of n known relations are used to calculate the classification accuracy. The metric exactly indicates whether `NOTA` detection impairs the classification of known relations. From tab. 6 we can observe that our method is comparable to the existing method, which supports the key claim at the beginning of the paragraph.

Two Real Substitution Cases: To intuitively show the effectiveness of the proposed synthesis method,

Method	FewRel	TACRED
MSP (Hendrycks et al., 2017)	$93.13_{0.41}$	$94.77_{0.98}$
DOC (Shu et al., 2017)	$93.25_{0.17}$	$93.70_{0.16}$
ODIN (Liang et al., 2018)	$93.11_{0.38}$	$94.88_{0.57}$
MixUp (Thulasidasan et al., 2019b)	$93.19_{0.41}$	$94.37_{1.28}$
Energy (Liu et al., 2020)	$93.36_{0.18}$	$94.97_{0.54}$
Convex (Zhan et al., 2021)	$91.97_{0.96}$	$93.10_{0.21}$
SCL (Zeng et al., 2021)	$93.45_{0.08}$	$95.20_{0.50}$
Ours	$\mathbf{93.50}_{0.37}$	$\mathbf{95.53}_{0.17}$

Table 6: The results of **ACC** on n known relations. The subscript represents the corresponding standard deviation (e.g., $93.50_{0.37}$ indicates 93.50 ± 0.37).

we conduct a case study based on the "Instrument" relation from FewRel and the "Spouse" relation from TACRED. The tokens with top-10 `tf-idf` statistics and a substitution case of each relation are shown in tab. 7, from which we can observe that: (1) the tokens with high `tf-idf` statistics have a strong semantic association with the target relation (such as Instrument-bass, Spouse-wife). (2) By substituting only two critical tokens in original training instances, the target relation is completely erased.

Relation: Instrument (musical instrument that a person plays) **Tokens with top 10 tf-idf statistics**: bass, saxophone, guitar, player, trumpet, trombone, composer, drums, organ, cello **Original Training Instance**: In 1961, McIntosh composed a **song** for [**trumpet**]$_{tail}$ legend [Howard Mcghee]$_{head}$. **Synthesized Negative Instance**: In 1961, McIntosh composed a **verse** for [**Mississippi**]$_{tail}$ legend [Howard Mcghee]$_{head}$.
Relation: Spouse (a husband or wife, considered in relation to their partner.) **Tokens with top 10 tf-idf statistics**: wife, husband, married, survived, died, grandchildren, children, heidi, sons, robert **Original Training Instance**: "[his]$_{head}$ **family** was at his bedside", his **wife**, [Barbara Washburn]$_{tail}$, said Thursday. **Synthesized Negative Instance**: "[his]$_{head}$ **friend** was at his bedside", his **captain**, [Barbara Washburn]$_{tail}$, said Thursday.

Table 7: Case study of the proposed negative samples synthesis method. The relation semantics between the given entity pair is completely erased by substituting only 2 tokens (tokens in red).