

Denoising Distant Supervision for Relation Extraction via Instance-Level Adversarial Training

Xu Han¹, Zhiyuan Liu¹, Maosong Sun^{1,2}

¹Department of Computer Science and Technology,
State Key Lab on Intelligent Technology and Systems,
National Lab for Information Science and Technology, Tsinghua University, Beijing, China
²Beijing Advanced Innovation Center for Imaging Technology,
Capital Normal University, Beijing, China

Abstract

Existing neural relation extraction (NRE) models rely on distant supervision and suffer from wrong labeling problems. In this paper, we propose a novel adversarial training mechanism over instances for relation extraction to alleviate the noise issue. As compared with previous denoising methods, our proposed method can better discriminate those informative instances from noisy ones. Our method is also efficient and flexible to be applied to various NRE architectures. As shown in the experiments on a large-scale benchmark dataset in relation extraction, our denoising method can effectively filter out noisy instances and achieve significant improvements as compared with the state-of-the-art models.

1 Introduction

Relation extraction (RE) aims to extract relational facts from plain text via categorizing semantic relations between entities contained in text. For example, we can extract the fact (*Mark Twain*, *PlaceOfBirth*, *Florida*) from the sentence “*Mark Twain* was born in *Florida*”. Many efforts have been devoted to RE, either early works based on handcrafted features (Zelenko et al., 2003; Mooney and Bunescu, 2006) or recent works based on neural networks (Zeng et al., 2014; Santos et al., 2015). These models all follow a supervised learning approach, which is effective, but the requirement to high-quality annotated data is a major bottleneck in practice.

It is time-consuming and human-intensive to manually annotate large-scale training data. Hence, Mintz et al. (2009) propose *distant supervision* to automatically generate training sen-

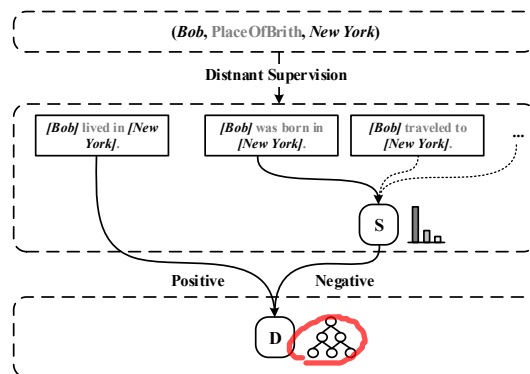


Figure 1: An example of distant supervision and adversarial denoising relation extraction.

tences via aligning KGs and text. As shown in Figure 1, distant supervision assumes that if there is a relation between two entities in a KG, all sentences that contain the two entities will be labeled with that relation. Distant supervision is an effective approach to automatically obtain training data, but it inevitably suffers from wrong labeling problems.

To address the wrong labeling problem, Riedel et al. (2010) propose multi-instance learning (MIL), and Zeng et al. (2015) extend the idea of MIL to neural models. Lin et al. (2016) further propose a neural attention scheme over multiple instances to reduce the weights of noisy instances. These methods achieve significant improvements in RE, however, still far from satisfactory. The reason is that most denoising methods simply calculate soft weights for each sentence in an unsupervised manner, which can only make a coarse-grained distinction between informative and noisy instances. Moreover, these methods cannot well cope with those entity pairs with insufficient sentences.

In order to better discriminate informative and noisy instances, inspired by the idea of adversar-

ial learning (Goodfellow et al., 2014a), we apply adversarial training over instances to enhance RE performance. The idea of adversarial training was explored in relation extraction by generating adversarial examples with a perturbation added to sentence embeddings (Wu et al., 2017), which do not necessarily correspond to real-world sentences. On contrary, we generate adversarial examples by sampling from existing training data, which may better locate real-world noise.

★ Our method contains two modules: a **discriminator** and a **sampler**, and the method will split the distantly supervised data into two parts, the confident part and the unconfident part. The **discriminator** is applied to judge which sentences are more likely to be annotated correctly, with the confident data as positive instances and the unconfident data as negative instances. The **sampler** module is used to select the most confusing sentences from unconfident data to cheat the discriminator as much as possible. Moreover, during several training epochs, we also dynamically select most informative and confident instances from the unconfident set to the confident set, so as to enrich the training instances for the discriminator.

The discriminator and the sampler are trained adversarially. As shown in Figure 1, during the training process, the actions of the sampler will admonish the discriminator to focus on improving those most confusing instances. Since noisy instances are ineffective to decrease the loss functions of both sampler and discriminator, the noise will be gradually filtered out during the adversarial training. Finally, the sampler can effectively distinguish those informative instances from the unconfident data, and the discriminator can well categorize relations between entities in text. As compared with the aforementioned MIL denoising methods, our method achieves more efficient noise detection in finer granularity.

We conduct experiments on a real-world dataset derived from New York Times (NYT) corpus and Freebase. Experimental results demonstrate that our adversarial denoising method effectively reduces noise and significantly outperforms other baseline methods.

2 Related Works

2.1 Relation Extraction

Relation extraction is an important task in NLP, which aims to extract relational facts from text

corpora. Many efforts are devoted to RE, especially in supervised RE, such as early kernel-based models (Zelenko et al., 2003; GuoDong et al., 2005; Mooney and Bunescu, 2006). Mintz et al. (2009) align plain text with KGs and propose a distantly supervised RE model, by assuming all sentences that mention two entities can describe their relations in KGs.

However, distant supervision inevitably accompanies with the wrong labeling problem. Riedel et al. (2010) and Hoffmann et al. (2011) apply the multi-instance learning (MIL) mechanism for RE, which considers the reliability of each instance and combines multiple sentences containing the same entity pair together to alleviate the noise problem.

In recent years, neural models (Zhang and Wang, 2015; Zeng et al., 2017; Miwa and Bansal, 2016) have been widely used in RE. These neural models are capable of accurately capturing textual relations without explicit linguistic analysis. Based on these neural architectures and the MIL mechanism, Lin et al. (2016) propose a sentence-level attention to reduce the influence of incorrectly labeled sentences. To summarize, these MIL models generally make soft weight adjustment for informative and noisy instances. Some works further adopt external information to improve denoising performance: Ji et al. (2017) incorporate external entity descriptions to enhance attention representations; Liu et al. (2017) manually set label confidences to denoise entity-pair level noises.

More sophisticated mechanisms, such as reinforcement learning (Feng et al., 2018; Zeng et al., 2018), have recently also been adapted to select positive sentences from noisy data. However, these complex mechanisms usually require much time to fine-tune and the convergence is not yet well guaranteed in practice. In this paper, we propose a novel fine-grained denoising method for RE via adversarial training. The method is simple and effective to be applied in various neural architectures and to scale up to large-scale data.

2.2 Adversarial Training

Szegedy et al. (2013) propose to generate adversarial examples by adding noise in the form of small perturbations to the original data. These noise examples are often indistinguishable for humans but lead to models' wrong predictions.

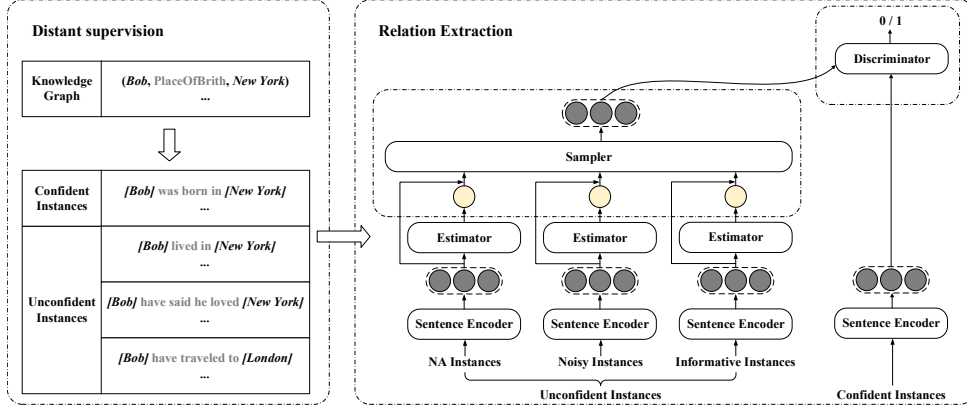


Figure 2: The overall framework of the instance-level adversarial training model for relation extraction. The discriminator module is used to judge whether an instance is labeled correctly, and the instance will be considered coming from the confident set if the prediction is yes. The sampler module is used to select the most confusing instances from unconfident data to cheat the discriminator.

Goodfellow et al. (2014b) analyze adversarial examples and propose adversarial training for image classification tasks. Afterwards, Goodfellow et al. (2014a) propose a mature adversarial training framework and use the framework to train generative models.

Adversarial training has also been explored in NLP. Miyato et al. (2016) propose adversarial training for text classification by adding perturbations to word embeddings. The idea of perturbation addition has further been applied in other NLP tasks including language models (Xie et al., 2017) and relation extraction (Wu et al., 2017). Different from (Wu et al., 2017) that generates pseudo adversarial examples by adding perturbations to instance embeddings, we perform adversarial training by sampling adversarial examples from real-world noisy data. The adversarial examples in our method can better correspond to the real-world scenario for RE. Hence our method is more favorable to solve the wrong labeling problem in distant supervision, which will be shown in experiments.

3 Methodology

In this section, we introduce the details of our instance-level adversarial training model for denoising RE. For this model, we split the entire training data into two parts, the set of those confident instances \mathcal{I}_c and the set of those unconfident instances \mathcal{I}_u . A sentence encoder is adopted to represent sentence semantics with embeddings. The adversarial training framework consists of a sampler and a discriminator, corresponding to the noise filter and the relation classifier respectively.

3.1 The Framework

As shown in Figure 2, the overall framework of our instance-level adversarial training model includes a discriminator D and a sampler S , in which S samples adversarial examples from the unconfident set \mathcal{I}_u , and D learns to judge whether a given instance is from \mathcal{I}_c or \mathcal{I}_u .

We assume that each instance $s \in \mathcal{I}_c$ exposes implicit semantics of its labeled relation r_s . In contrast, those instances $s \in \mathcal{I}_u$ are not trusted to be labeled correctly during the adversarial training. Hence, we implement D as a function $D(s, r_s)$ to judge whether a given instance s exposes implicit semantics of its labeled relation r_s : if yes, the instance comes from \mathcal{I}_c ; while if no, the instance comes from \mathcal{I}_u .

The training process is a min-max game and can be formalized as follows,

$$\phi = \min_{p_u} \max_D (E_{s \sim p_c} [\log(D(s, r_s))] + E_{s \sim p_u} [\log(1 - D(s, r_s))]) \quad (1)$$

where p_c is the confident data distribution, and the sampler S samples adversarial examples from the unconfident data according to the probability distribution p_u .

After sufficient training, S tends to sample those informative instances in \mathcal{I}_u rather than those noisy instances, and D becomes a relation classifier of good robustness to noisy data. We will give the detailed introduction to the sampler in Section 3.2 and the discriminator in Section 3.3.

3.2 Sampler

The sampler module aims to select the most confusing sentences from the unconfident set \mathcal{I}_u to the discriminator as much as possible by optimizing the probability distribution p_u . Hence, we need to calculate the *confusing score* for each instance in the unconfident set \mathcal{I}_u .

Given an instance s , we can use neural sentence encoders to represent its semantic information as an embedding \mathbf{y} . The details of neural encoders will be introduced in Section 3.4. Here, we can simply calculate the confusing score according to the sentence embedding \mathbf{y} as follows,

$$C(s) = \mathbf{W} \cdot \mathbf{y}, \quad (2)$$

where \mathbf{W} is a separating hyperplane. We further define $P_u(s)$ as the *confusing probability* over \mathcal{I}_u ,

$$P_u(s) = \frac{\exp(C(s))}{\sum_{s \in \mathcal{I}_u} \exp(C(s))}. \quad (3)$$

In the unconfident set, we regard those instances with high $D(s, r_s)$ scores as the confusing instances, because they will fool the discriminator D to make wrong decision. An optimized sampler will assign larger confusing score to those most confusing instances. Hence, we formalize the loss function to optimize the sampler module as follows:

$$\mathcal{L}_S = - \sum_{s \in \mathcal{I}_u} P_u(s) \log(D(s, r_s)). \quad (4)$$

When optimizing the sampler, we regard the component $P_u(s)$ as parameters for updating.

Note that, when an instance is labeled as $r_s = \text{NA}$, it indicates the relation of this instance is not available, either unsure or having no relation. Since these instances are always wrongly predicted into other relations, in order to let the discriminator restrain this tendency, we specifically define $D(s, \text{NA})$ as the average score of the instance over all feasible relations:

$$D(s, \text{NA}) = \frac{1}{|\mathcal{R}| - 1} \sum_{r \in \mathcal{R}, r \neq \text{NA}} D(s, r), \quad (5)$$

where \mathcal{R} indicates the set of relations.

3.3 Discriminator

Given an instance s and its embedding \mathbf{y} , the discriminator is responsible for judging whether its labeled relation r_s is correct. We implement the

discriminator based on the semantic relatedness between \mathbf{r}_s and \mathbf{y} ,

$$D(s, r_s) = \sigma(\mathbf{r}_s \cdot \mathbf{y}), \quad (6)$$

where $\sigma(\cdot)$ is the sigmoid function.

An optimized discriminator will assign high scores to those instances in \mathcal{I}_c and low scores to those instances in \mathcal{I}_u . Hence, we formalize the loss function to optimize the discriminator module as follows:

$$\mathcal{L}_D = - \sum_{s \in \mathcal{I}_c} \frac{1}{|\mathcal{I}_c|} \log(D(s, r_s)) - \sum_{s \in \mathcal{I}_u} P_u(s) \log(1 - D(s, r_s)). \quad (7)$$

When optimizing the discriminator, we regard the component $D(s, r_s)$ as parameters for updating. Note that, the objective functions of the sampler in Eq. 4 and the discriminator in Eq. 7 are adversarial to each other.

In practice, the data set is usually too large to be frequently traversed due to intractable large amounts of computation. For convenience of training efficiency, we can simply sample subsets to approximate the probability distribution. Hence, we formalize a new loss function for optimization:

$$\tilde{\mathcal{L}}_D = - \sum_{s \in \hat{\mathcal{I}}_c} \frac{1}{|\hat{\mathcal{I}}_c|} \log(D(s, r_s)) - \sum_{s \in \hat{\mathcal{I}}_u} Q_u(s) \log(1 - D(s, r_s)), \quad (8)$$

where $\hat{\mathcal{I}}_c$ and $\hat{\mathcal{I}}_u$ are subsets sampled from \mathcal{I}_c and \mathcal{I}_u respectively, and $Q_u(s)$ is the corresponding approximation to $P_u(s)$ in Eq. 3:

$$Q_u(s) = \frac{\exp(C(s)^\alpha)}{\sum_{s \in \hat{\mathcal{I}}_u} \exp(C(s)^\alpha)}. \quad (9)$$

Note that α is a hyper-parameter that controls the sharpness of the confusing probability distribution. For consistency, we also approximate \mathcal{L}_S in Eq. 4 as:

$$\tilde{\mathcal{L}}_S = - \sum_{s \in \hat{\mathcal{I}}_u} Q_u(s) \log(D(s, r_s)). \quad (10)$$

$\tilde{\mathcal{L}}_S$ and $\tilde{\mathcal{L}}_D$ are used to optimize our adversarial training model.

3.4 Instance Encoder

Given an instance s containing two entities, we apply several neural network architectures to encode the sentence into continuous low-dimensional embeddings \mathbf{y} , which are expected to capture the implicit semantics of the labeled relation between two entities.

3.4.1 Input Layer

The input layer aims to map discrete language symbols (i.e., words) into continuous input embeddings. Given an instance s containing n words $\{w_1, \dots, w_n\}$, we use Skip-Gram (Mikolov et al., 2013) to embed all words into k_w -dimensional space $\{\mathbf{w}_1, \dots, \mathbf{w}_n\}$. For each word w_i , we also embed its relative distances to the two entities into two k_p -dimensional vectors, and then concatenate them as an unified position embedding \mathbf{p}_i (Zeng et al., 2014). We finally get the k_i -dimensional input embeddings for the following encoding layer,

$$\begin{aligned} \mathbf{s} &= \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \\ &= \{[\mathbf{w}_1; \mathbf{p}_1], \dots, [\mathbf{w}_n; \mathbf{p}_n]\}. \end{aligned} \quad (11)$$

3.4.2 Encoding Layer

In the encoding layer, we select four typical architectures including CNN (Zeng et al., 2014), PCNN (Zeng et al., 2015), RNN (Zhang and Wang, 2015) and BiRNN (Zhang and Wang, 2015) to further encode input embeddings of the instance into sentence embeddings.

CNN slides a convolution kernel with the window size m over the input sequence $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ to get the k_h -dimensional hidden embeddings.

$$\mathbf{h}_i = \text{CNN}(\mathbf{x}_{i-\frac{m-1}{2}}, \dots, \mathbf{x}_{i+\frac{m-1}{2}}). \quad (12)$$

A max-pooling is then applied over these hidden embeddings to output the final instance embedding \mathbf{y} as follows,

$$[\mathbf{y}]_j = \max\{[\mathbf{h}_1]_j, \dots, [\mathbf{h}_n]_j\}. \quad (13)$$

PCNN is an extension to CNN, which also adopts a convolution kernel with the window size m to obtain hidden embeddings. Afterwards, PCNN divides the hidden embeddings into three segments $\{\mathbf{h}_1, \dots, \mathbf{h}_{e_1}\}$, $\{\mathbf{h}_{e_1+1}, \dots, \mathbf{h}_{e_2}\}$, and $\{\mathbf{h}_{e_2+1}, \dots, \mathbf{h}_n\}$, where e_1 and e_2 are entity positions. PCNN applies a piecewise max-pooling for each segment,

$$\begin{aligned} [\mathbf{y}_1]_j &= \max\{[\mathbf{h}_1]_j, \dots, [\mathbf{h}_{e_1}]_j\}, \\ [\mathbf{y}_2]_j &= \max\{[\mathbf{h}_{e_1+1}]_j, \dots, [\mathbf{h}_{e_2}]_j\}, \\ [\mathbf{y}_3]_j &= \max\{[\mathbf{h}_{e_2+1}]_j, \dots, [\mathbf{h}_n]_j\}. \end{aligned} \quad (14)$$

By concatenating all pooling results, PCNN eventually outputs a $3 \cdot k_h$ -dimensional instance embedding \mathbf{y} as follows,

$$\mathbf{y} = [\mathbf{y}_1; \mathbf{y}_2; \mathbf{y}_3]. \quad (15)$$

RNN is designed for modeling sequential data, as it keeps its hidden state changing with input embeddings at each time-step accordingly,

$$\mathbf{h}_i = \text{RNN}(\mathbf{x}_i, \mathbf{h}_{i-1}), \quad (16)$$

where $\text{RNN}(\cdot)$ is the recurrent unit and $\mathbf{h}_i \in \mathbb{R}^{k_h}$ is the hidden embedding at the time-step i . In this paper, we select gated recurrent unit (GRU) (Cho et al., 2014) as the recurrent unit. We use the hidden embedding of the last time-step as the instance embedding, i.e., $\mathbf{y} = \mathbf{h}_n$.

Bi-RNN aims to incorporate information from both sides of the sentence sequence. Bi-RNN is adopted with forward and backward directions as follows,

$$\begin{aligned} \vec{\mathbf{h}}_i &= \text{RNN}_f(\mathbf{x}_i, \vec{\mathbf{h}}_{i-1}), \\ \overleftarrow{\mathbf{h}}_i &= \text{RNN}_b(\mathbf{x}_i, \overleftarrow{\mathbf{h}}_{i+1}), \end{aligned} \quad (17)$$

where $\vec{\mathbf{h}}_i$ and $\overleftarrow{\mathbf{h}}_i$ are the hidden states at the position i of the forward and backward RNN respectively. We concatenate the hidden states from both the forward and backward RNN as the instance embedding \mathbf{y} ,

$$\mathbf{y} = [\vec{\mathbf{h}}_n; \overleftarrow{\mathbf{h}}_1]. \quad (18)$$

3.5 Initialization and Implementation Details

Here we introduce the learning and optimization details for our adversarial training model. We define the optimization function as

$$\mathcal{L} = \tilde{\mathcal{L}}_D + \lambda \tilde{\mathcal{L}}_S, \quad (19)$$

where λ is a harmonic factor. In practice, both the modules in adversarial training are optimized alternately using stochastic gradient descent (SGD). Since the framework of our model is much simpler than typical generative adversarial networks (GAN), we do not have to calibrate alternating ratio between the loss functions, and hence we can simply use a 1 : 1 ratio. It enables our model efficient for learning on large-scale data. Moreover, we can also integrate λ into the learning rate of the sampler $\tilde{\mathcal{L}}_S$, so as to avoid adjusting the hyper-parameter λ .

At the start of adversarial training, we pre-train a relation classifier on the entire training data. The relation classifier will split the entire data into a small confident data and a large unconfident data. During the adversarial training, after every few training epochs, some instances from the unconfident set that are both recommended by the sampler and recognized by the discriminator will be selected to enrich the confident set.

4 Experiments

In this section, we carry out experiments to demonstrate the effectiveness of our instance-level adversarial training method. We first introduce datasets and parameter settings. Afterwards, we compare the performance of our method with conventional neural methods and feature-based methods for RE. To further verify that our method can better discriminate those informative instances from noisy ones, we also conduct evaluations on those entity pairs with few sentences.

4.1 Datasets and Experiment Settings

4.1.1 Datasets

We conduct experiments on the benchmark dataset derived from New York Times (NYT) corpus, which is first proposed by Mintz et al. (2009) and then widely used in various distantly supervised RE works (Riedel et al., 2010; Hoffmann et al., 2011; Surdeanu et al., 2012; Zeng et al., 2015; Lin et al., 2016; Wu et al., 2017). The dataset aligns entity pairs and their relations in the KG Freebase with NYT corpus. After various essential data processing, there are 53 relation types including the NA relation in this dataset. The training data contains 522,611 sentences, 281,270 entity pairs and 18,252 relational facts. The test data contains 172,448 sentences, 96,678 entity pairs and 1,950 relational facts.

4.1.2 Parameter Settings

In our models, we select the learning rate α_d and α_s among $\{0.5, 0.1, 0.05, 0.01\}$ for training the discriminator and the sampler respectively. For other parameters, we simply follow the settings used in (Zeng et al., 2014; Lin et al., 2016; Wu et al., 2017) so that we can fairly compare the results of our adversarial denoising models with these baselines. Table 1 shows all parameters used in the experiments. During training, we select most informative and confident instances in the

unconfident set to enrich the confident set every 10 training epochs.

Discriminator Learning Rate α_d	0.1
Sampler Learning Rate α_s	0.01
Hidden Layer Dimension k_h for CNNs	230
Hidden Layer Dimension k_h for RNNs	150
Position Dimension k_p for CNNs	5
Position Dimension k_p for RNNs	3
Word Dimension k_w	50
Convolution Kernel Size m	3
Dropout Probability p	0.5

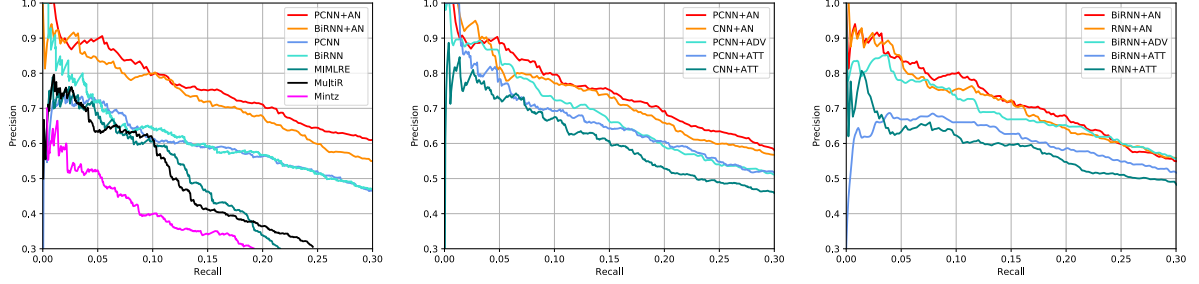
Table 1: Parameter settings.

4.2 Overall Evaluation Results

We follow Mintz et al. (2009) to conduct the held-out evaluation. We construct candidate triples by combining entity pairs in the test set with various relations and rank these triples according to their corresponding sentence representations. By regarding the triples in the KGs as correct and others as incorrect, we evaluate different models with their precision-recall results.

The evaluation results are shown in Figure 3 and Table 2. We report the results of various neural architectures including CNN, PCNN, RNN and BiRNN with various denoising methods: +ATT is the selective attention method over instances (Lin et al., 2016); +ADV is the denoising method by adding a small adversarial perturbation to instance embeddings (Wu et al., 2017); +AN is our proposed adversarial training method. We also compare our methods with feature-based models, including Mintz (Mintz et al., 2009), MultiR (Hoffmann et al., 2011) and MIML (Surdeanu et al., 2012). The results of the baseline models all come from the data reported in their papers or their open-source code. From the figure and table, we observe that:

(1) As shown in Figure 3(a), neural models significantly outperform all feature-based models over the entire range of recall. When the recall gradually grows, the performance of feature-based models drops out quickly. However, all the neural models still preserve stable and competitive precision. It demonstrates that human-designed features cannot work well in a noisy environment, and inevitable errors brought by NLP tools will further hurt the performance. In contrast, instance embeddings learned automatically by neural models



(a) Comparison of the proposed models and feature-based models. (b) Comparison of the proposed models and various CNN models. (c) Comparison of the proposed models and various RNN models.

Figure 3: Aggregate precision/recall curves of different models.

Method		0.1	0.2	0.3	Mean
CNN+	ATT	67.5	52.8	45.8	55.4
	AN	75.3	66.3	54.3	65.3
RNN+	ATT	63.9	54.4	48.0	55.4
	AN	75.3	64.5	55.8	65.2
PCNN+	ATT	69.4	60.6	51.6	60.5
	ADV	71.7	58.9	51.1	60.6
	AN	80.3	70.2	60.3	70.3
BiRNN+	ATT	66.8	58.6	52.4	64.2
	ADV	72.8	64.6	55.3	65.2
	AN	79.1	67.3	54.1	66.8

Table 2: Precision of various models for different recall (%).

can effectively capture implicit relational semantics from noisy data for RE.

(2) Both for CNNs (CNN and PCNN) in Figure 3(b) and RNNs (RNN and BiRNN) in Figure 3(c), the models with adversarial training outperform the models with sentence-level attention. The sentence-level attention over multiple instances, which calculates soft weights for each sentence to reduce noise, only makes a coarse-grained distinction between informative and noisy instances. In contrast, the neural models trained with adversarial denoising methods generate or sample noisy adversarial examples and force the relation classifiers to overcome them. Hence, the models with adversarial training provide efficient noise reduction in finer granularity. In general, the models with our adversarial training method achieve the best results among models using adversarial training. This indicates that, as compared to generating pseudo adversarial examples by adding perturbations, our method by sampling adversarial examples from real-world instances can better discrim-

inate informative instances from noisy instances.

(3) To better compare various denoising methods, we also show evaluation results in Table 2. Since we focus more on the performance of those top-ranked results, here we show the precision scores when the recall is 0.1, 0.2, 0.3 as well as their mean. We find that complicated neural models (PCNN, BiRNN) perform better than simple neural networks (CNN, RNN) when using the same denoising methods. Both CNNs and RNNs are significantly improved by adversarial training, and our method (AN) performs consistently much better than the adversarial training baseline (ADV). The improvements brought by changing denoising methods are more significant than the improvements brought by modifying neural models. This indicates that the wrong labeling problem is the critical factor that prevents distantly supervised RE models from working effectively.

4.3 Effect of Adversarial Denoising Training

To further verify the effectiveness of our adversarial training method, we evaluate the RE performance of our method and conventional MIL denoising methods in a more challenging scenario, i.e., when entity pairs having few sentences.

For each entity pair, we randomly select one sentence, two sentences, and all sentences to construct three experimental settings respectively. We report P@100, P@200, P@300 and the mean of them in the held-out evaluation. Since PCNN is the best neural model in the above comparison, we simply use PCNN to compare our method (AN) with the recent state-of-the-art denoising method, sentence-level attention (ATT), as well as its naive versions +ONE and +AVG (Zeng et al., 2015; Lin et al., 2016). The evaluation results are shown in Table 3, and from the results we observe that:

Test Settings	One				Two				All			
P@N	100	200	300	Mean	100	200	300	Mean	100	200	300	Mean
PCNN	63.0	61.0	55.3	59.8	65.0	62.5	57.3	61.6	71.0	64.0	58.7	64.6
PCNN+ONE	73.3	64.8	56.8	65.0	70.3	67.2	63.1	66.9	72.3	69.7	64.1	68.7
PCNN+AVG	71.3	63.7	57.8	64.3	73.3	65.2	62.1	66.9	73.3	66.7	62.8	67.6
PCNN+ATT	73.3	69.2	60.8	67.8	77.2	71.6	66.1	71.6	76.2	73.1	67.4	72.2
PCNN+AN	84.0	75.0	73.0	77.3	86.0	77.0	73.7	78.9	90.0	82.0	76.3	82.8

Table 3: Top-N precision (P@N) for RE in the entity pairs with different number of instances (%).

(1) Our method achieves consistent and significant improvements as compared to the ATT method and its naive versions, especially when each entity pair only corresponds to one or two sentences. The reason is that most MIL denoising methods including ATT typically assume that at least one instance that mentions the given entity pair can express their relation, and always select at least one informative sentence for the entity pair. This assumption is not always true especially when entity pairs correspond to few sentences: it is more likely there is no instance that can express the relation of the given entity pair. In contrast, our adversarial training method is not restricted by the assumption. By conducting on instance level individually, our method keeps effective even when the instances of each entity pair are few.

(2) When taking more instances into account, all models achieve better results. PCNN+ATT and PCNN+AN achieve more improvements than those naive methods. The growth of distant supervision data brings more information for training RE models as well as more noises that may hurt performance. Our method keeps its degree of superiority to the ATT method as the data growth. This indicates that our method could provide more robust and reliable scheme to denoise distant supervision data.

4.4 Case Study

Table 4 shows examples sampled by the sampler. For the frequent relation Location Contains, we use the sampler to select the positive and negative instances respectively. For each sentence, we highlight the entities in bold-face. From the table we find that: The former positive examples clearly correspond to the relation Location Contains, while those negative examples fail to reflect this relation. These examples show that our sampler is effective to dis-

Relation	Location Contains
Positive	... China's 10 most polluted cities, four, including Datong , are in Shanxi province ...
	... Manhattan's Chinatown has fought off the forces of urban decline ...
Negative	... the senior commander of U.S. forces in Baghdad , has figured out the obstacle to america 's dream for Iraq ...
	... after Japan's defeat, he said, American soldiers drove jeeps onto his family 's estate in Iwate ...

Table 4: Some examples sampled by the sampler in NYT corpus.

tinguish informative and noisy instances.

5 Conclusion

In this paper, we propose a denoising distant supervised method for RE via instance-level adversarial training. By splitting the entire data into the confident set and the unconfident set, our method trains a sampler and a discriminator adversarially. The sampler aims to select the most confusing instance from the unconfident set, and the discriminator aims to distinguish an instance which comes from either the confident set or the unconfident set. In experiments, we apply our method to various neural architectures for RE. The experimental results show that our method achieves efficient noise reduction in finer granularity and significantly outperforms the state-of-the-art baseline. Our method is also robust for those long-tail entity pairs with few instances.

In the future, we plan to explore the following directions: (1) Inspired by (Ji et al., 2017), it will be promising to adopt external knowledge, from either KBs or text, to help train more efficient samplers and discriminators for adversarial training. (2) We may also extend the instance-level adversarial training to the entity-pair level to further improve the robustness of RE models.

References

- Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *Proceedings of SSST*.
- Jun Feng, Minlie Huang, Li Zhao, Yang Yang, and Xiaoyan Zhu. 2018. Reinforcement learning for relation classification from noisy data. In *Proceedings of AAAI*.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014a. Generative adversarial nets. In *Proceedings of NIPS*.
- Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014b. Explaining and harnessing adversarial examples. In *Proceedings of ICLR*.
- Zhou GuoDong, Su Jian, Zhang Jie, and Zhang Min. 2005. Exploring various knowledge in relation extraction. In *Proceedings of ACL*.
- Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of ACL*.
- Guoliang Ji, Kang Liu, Shizhu He, Jun Zhao, et al. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *Proceedings of AAAI*.
- Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of ACL*.
- Tianyu Liu, Kexiang Wang, Baobao Chang, and Zhi-fang Sui. 2017. A soft-label method for noise-tolerant distantly supervised relation extraction. In *Proceedings of EMNLP*.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of ICLR*.
- Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL-IJCNLP*.
- Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using lstms on sequences and tree structures. In *Proceedings of ACL*.
- Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, Ken Nakae, and Shin Ishii. 2016. Distributional smoothing with virtual adversarial training. *Proceedings of ICLR*.
- Raymond J Mooney and Razvan C Bunescu. 2006. Subsequence kernels for relation extraction. In *Proceedings of NIPS*.
- Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of ECML-PKDD*.
- Cicero Nogueira dos Santos, Bing Xiang, and Bowen Zhou. 2015. Classifying relations by ranking with convolutional neural networks. *Proceedings of ACL*.
- Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of EMNLP*.
- Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.
- Yi Wu, David Bamman, and Stuart Russell. 2017. Adversarial training for relation extraction. In *Proceedings of EMNLP*.
- Ziang Xie, Sida I Wang, Jiwei Li, Daniel Lévy, Aiming Nie, Dan Jurafsky, and Andrew Y Ng. 2017. Data noising as smoothing in neural network language models. *Proceedings of ICLR*.
- Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. In *Proceedings of JMLR*.
- Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of EMNLP*.
- Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING*.
- Wenyuan Zeng, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2017. Incorporating relation paths in neural relation extraction. In *Proceedings of EMNLP*.
- Xiangrong Zeng, Shizhu He, Kang Liu, and Jun Zhao. 2018. Large scaled relation extraction with reinforcement learning. In *Proceedings of AAAI*.
- Dongxu Zhang and Dong Wang. 2015. Relation classification via recurrent neural network. *arXiv preprint arXiv:1508.01006*.