Adversarial Discriminative Denoising for Distant Supervision Relation Extraction

Bing Liu[1], **Huan Gao**[1], **Guilin Qi**[1] *, **Shangfu Duan**[1], **Tianxing Wu**[1], **Meng Wang**[1]
[1]School of Computer Science and Engineering, Southeast University, Nanjing, China
{liubing_cs, hg, gqi, sf_duan, wutianxing}@seu.edu.cn
wangmengsd@outlook.com

Abstract

Distant supervision has been widely used to generate labeled data automatically for relation extraction by aligning knowledge base with text. However, it introduces much noise, which can severely impact the performance of relation extraction. Recent studies have attempted to remove the noise explicitly from the generated data but suffer from (1) the lack of effective way of introducing explicit supervision to the denoising process and (2) the difficulty of optimization caused by the sampling action in denoising result evaluation. To solve these issues, we propose an adversarial discriminative denoising framework, which provides an effective way of introducing human supervision and exploiting it along with the potential information underlying the noisy data in a unified framework. Besides, we employ a continuous approximation of sampling action to guarantee the holistic denoising framework differentiable. Experimental results show that very little human supervision is sufficient for our approach to outperform the state-of-the-art methods significantly.

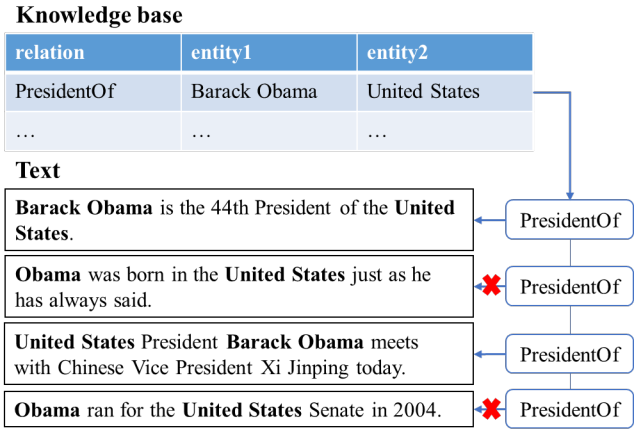


Figure 1: Examples labeled by distant supervision.

Introduction

Relation extraction (RE) is a crucial task in natural language processing (NLP), which aims at predicting the semantic relations expressed in the text (Zelenko, Aone, and Richardella 2003; Bunescu and Mooney 2005; Zhou et al. 2005). It has a wide range of applications such as knowledge base (KB) completion and question answering. Although supervised methods have achieved excellent performance in RE (Zeng et al. 2014; dos Santos, Xiang, and Zhou 2015; Wang et al. 2016), they suffer from the lack of costly human annotations.

Distant supervision (DS) for RE can generate training data automatically by aligning KB with text (Mintz et al. 2009). As illustrated in Figure 1, if the triplet *<Barack Obama, PresidentOf, United States>* exists in KB, all sentences containing *Barack Obama* and *United States* will be selected as training instances of relation *PresidentOf*. However, the second and fourth sentences in Figure 1 do not mention relation *PresidentOf*. The entities may co-occur in a sentence just because they are related to the same topic. The percentage of such false positives (FP) may be quite high. For example, (Riedel, Yao, and McCallum 2010) reported up to 31% of FP when matching Freebase with New York Times (NYT) articles. This noise will hinder the performance of DS based RE models trained on such noisy data.

Previous works have focused on building DS models with noise adaptability. (Riedel, Yao, and McCallum 2010; Hoffmann et al. 2011; Zeng et al. 2015) formulated DS as multi-instance learning problem based on *expressed-at-least-once*[1] assumption. (Lin et al. 2016; Ji et al. 2017) modeled DS in bag[2] level and exploited neural network with sentence-level attention. These methods can alleviate the side effect of the noise to a certain extent. However, they all assume that there is no case where all the sentences containing the same entities are FP, which is actually a common phenomenon in DS. Explicit noise reduction is a solution to this problem (Qin, Xu, and Wang 2018b). To enable explicit noise reduction, a few challenges need to be addressed.

The first challenge comes from the lack of an effective way of introducing explicit supervision to the denoising process. This makes it difficult to distinguish between the true positive instances and the noise. (Takamatsu, Sato, and Nak-

*Corresponding author: Guilin Qi, gqi@seu.edu.cn

[1]If two entities participate in a relation, at least one sentence that mentions these two entities might express that relation.

[2]All sentences containing the same entity pair make up a bag.

agawa 2012) removed noisy sentences using artificial syntactic patterns, which is very time-consuming and unable to scale. Recent works mainly focused on exploiting the potentially useful information underlying the original data, such as using reinforcement learning (RL) (Feng et al. 2018; Zeng et al. 2018; Qin, Xu, and Wang 2018b) and generative adversarial net (GAN) (Qin, Xu, and Wang 2018a). These methods are in an unsupervised manner, which can only make a coarse-grained distinction between the true positive instances and the noise. (Luo et al. 2017) proposed to characterize the noise with dynamic transition matrix, which is also unsupervised but can introduce prior knowledge of the data quality to guide the training process. However, this prior knowledge can hardly be obtained. Thus, an effective way of introducing explicit supervision remains to be studied.

Another challenge relates to the evaluation of the denoising result. The existing approaches (Feng et al. 2018; Qin, Xu, and Wang 2018b; 2018a) performed the evaluation by sampling from the noisy data according to the noise recognizer and then assessing the resulting subset. However, the sampling action can lead to non-differentiability, which hinders the use of the evaluators that back-propagate gradients to guide the optimization of the noise recognizer in a holistic manner. These works tried to address the non-differentiability through policy learning, which tended to suffer from high variance.

To solve the above challenges, we propose an adversarial discriminative denoising framework, which can not only acquire the denoising ability by exploiting the beneficial information underlying DS-generated data but also further get boosted via introducing very little human annotations efficiently. To guarantee the model differentiable, we employ a continuous approximation of sampling action when evaluating the denoising result, which helps fast convergence. Compared with state-of-the-art methods, our approach can achieve significant improvement just using very little human supervision and has better applicability when the noise proportion is large.

We demonstrate the advantages of our denoising framework on two datasets: (1) the widely used DS dataset NYT (Riedel, Yao, and McCallum 2010) where our approach achieves state-of-the-art performance and (2) a customized noisy dataset based on the SemEval-2010 Task 8 dataset, which is used to give a further interpretation of our approach. Our contributions can be summarized as follows:

- We propose an adversarial discriminative denoising framework, which provides an highly-efficient way of introducing human supervision for removing the noise.
- We employ a continuous approximation of sampling action to guarantee the holistic model differentiable.
- We conduct the experiments on two datasets, and the results show our approach outperforms the existing methods significantly using very little human supervision and have better applicability when the noise proportion is large.

Related Work

DS Models with Noise Adaptability

(Mintz et al. 2009) applied distant supervision to generate labeled data automatically for RE by assuming that all the sentences containing the same entity pair express their relation in KB. However, this initial work ignored the introduced noise, including false positives (Riedel, Yao, and McCallum 2010; Takamatsu, Sato, and Nakagawa 2012; Lin et al. 2016) and false negatives (Ritter et al. 2013; Xu et al. 2013; Min et al. 2013), which impedes the performance of RE models severely. In this paper, we mainly care about false positive noise. To alleviate the effect of the noise, (Riedel, Yao, and McCallum 2010) proposed *expressed-at-least-once* assumption and formulated DS as multi-instance learning problem. Following this assumption, researchers advanced this paradigm using probabilistic graphical models (Hoffmann et al. 2011; Surdeanu et al. 2012), and neural network methods (Zeng et al. 2015). As attention methods raised concern, (Lin et al. 2016; Ji et al. 2017) exploited neural network with sentence-level attention to build RE models with noise adaptability. These studies can alleviate the effect of the noise to a certain extent. However, they did not reduce the noise explicitly and suffer from the case where all sentences containing the same entities are false positive, which is a common phenomenon in DS. In this paper, we aim to remove the noise explicitly from DS-generated data.

Explicit Noise Reduction

Cleaning the noisy data is difficult because there always lacks explicit supervision to remove the noise. (Takamatsu, Sato, and Nakagawa 2012) removed noisy sentences using artificial syntactic patterns, which is time-consuming and thus unable to scale. To avoid human participation, recent works mainly focused on removing the noise by mining the potentially useful information underlying the noisy data, using RL (Feng et al. 2018; Zeng et al. 2018; Qin, Xu, and Wang 2018b) and GAN (Qin, Xu, and Wang 2018a). These methods are all unsupervised and cannot introduce human supervision. As a result, they can only make a coarse-grained distinction between the true positive instances and the noise. (Luo et al. 2017) proposed a flexible method named dynamic transition matrix, which can not only work in unsupervised paradigm but also exploiting prior knowledge of data quality to guide the training process if it is provided. Unfortunately, it is tough to obtain such prior knowledge. To mitigate the lack of explicit supervision, we provide an effective way of introducing sufficient human supervision just using very few human annotations.

Adversarial Learning

Adversarial learning has gained great success in computer vision (Goodfellow et al. 2014) and NLP (Ganin et al. 2016). These works use discriminators as metrics to provide a different way for generator or predictor learning. Our work draws on such idea of adversarial learning. Besides, researchers have shown backpropagation is highly effective for adversarial learning in continuous image generation and representation learning (Chen et al. 2016; Larsen

et al. 2016). However, the evaluation of the denoising result always involves sampling action, which will cause non-differentiability to the optimization process (Qin, Xu, and Wang 2018a; 2018b). (Hu et al. 2017) applied continuous approximation of text samples to avoid non-differentiability when solving text generation problem. Inspired by this work, we address this problem by employing a continuous approximation to sampling action.

Methodology

In this section, we present an adversarial discriminative denoising framework, which can efficiently filter out FP noise from the DS-generated positive data.

Problem Definition

Given a DS-generated dataset, which can be partitioned as positive data $\mathcal{D}^p = \{x_i^p\}_{i=1}^{|\mathcal{D}^p|}$ and negative data $\mathcal{D}^n = \{x_i^n\}_{i=1}^{|\mathcal{D}^n|}$ with respect to a specific relation r, explicit noise reduction aims to filter out the false positive instances from $\mathcal{D}^p$ with the help of a handful of human annotations $\mathcal{D}^l = \{x_i^l\}_{i=1}^{|\mathcal{D}^l|}$ ($|\mathcal{D}^l| \ll |\mathcal{D}^p|$) of relation r.

We formulate a positive instance predictor P, which outputs the probability $P(x)$ that an instance x mentions relation r. It functions as the noise recognizer. Its training procedure is provided with data composed by instances from $\mathcal{D}^p$ labeled 1 and instances from $\mathcal{D}^n$ labeled 0.

Also, we formulate a data source discriminator D, which outputs the probability $D(x)$ that an instance x is from $\mathcal{D}^l$ other than $\mathcal{D}^p$. It functions as the metrics of the distinguishability between $\mathcal{D}^l$ and the weighted $\mathcal{D}^p$, in which each instance x is assigned a weight $P(x)$. The weights determine their distinguishability. The training procedure of D is provided with data composed by instances from $\mathcal{D}^l$ labeled 1 and the weighted instances from $\mathcal{D}^p$ labeled 0.

Overview

The overview of our approach is shown in Figure 2. The adversarial learning process is carried out on handful manually labeled data $\mathcal{D}^l$ and a mass of DS-generated data, partitioned as positive data $\mathcal{D}^p$ and negative data $\mathcal{D}^n$, concerning a specific relation r. The framework consists of two core modules: (1) a positive instance predictor P, whose primary task is to predict if an instance x_p is expressing relation r, and (2) a data source discriminator D aiming to detect if an instance x_d is from $\mathcal{D}^l$ or the weighted $\mathcal{D}^p$. P not only tries to satisfy the constraints from DS but also attempts to assign $\mathcal{D}^p$ with proper weights to make it indistinguishable from $\mathcal{D}^l$. D functions as a critic of the distinguishability between $\mathcal{D}^l$ and the weighted $\mathcal{D}^p$ and will adjust its ability once the weights on $\mathcal{D}^p$ change. Under this adversarial setting, P can not only obtain its denoising ability through the supervision from DS but also get boosted in the competition with D. In order to make the whole architecture differentiable, we employ a continuous approximation of sampling action when evaluating the denoising result of P and get the weighted $\mathcal{D}^p$. Owing to this strategy, D can guide the optimization of P in a holistic manner, and the whole framework can be optimized via the standard backpropagation algorithm.

Adversarial Discriminative Denoising Framework

We now describe our framework in detail, by presenting the predictor, continuous approximation of sampling action and the discriminator, respectively.

Predictor. To obtain a valid P, we train it using two sources of guidance: (1) the DS-generated labeled data (this supervision information is actually from KB) and (2) the feedback from D which represents the distinguishability between $\mathcal{D}^l$ and the weighted $\mathcal{D}^p$.

Although the DS-generated data contains much noise, it can provide beneficial information due to the correctly labeled instances. To take advantage of this information, we provide P with $\mathcal{D}^p$ labeled 1 and $\mathcal{D}^n$ labeled 0 and try to minimize the cross-entropy classification loss defined as:

$$L'_p = -[\mathbb{E}_{x\sim\mathcal{D}^p}[\log P(x)] + \mathbb{E}_{x\sim\mathcal{D}^n}[\log(1-P(x))]] \quad (1)$$

Another goal of P is to reduce the distinguishability between $\mathcal{D}^l$ and the weighted $\mathcal{D}^p$. Thus, we treat this distinguishability as another part of the loss of P. Here, we use D to measure this distinguishability, and worse distinguishability will bring about more substantial loss L_d to D. Considering this negative correlation, we use $-L_d$ as the metric of the distinguishability. Thus, the objective of P can be formulated as minimizing the following loss function:

$$L_p = (1-\lambda)L'_p + \lambda(-L_d) \quad (2)$$

where λ is used to tune the trade-off between these two loss items. The loss item $-L_d$ seeks to provide P with the information of what the true positive instances are like, as well as avoid P being misled by L'_p. When λ is set as 0, P will work in unsupervised paradigm. Therefore, human participation is not a must but can be introduced if needed in our approach.

Continuous Approximation of Sampling Action. When evaluating the rationality of the predicting result of P with human participation, a conventional idea is sampling on $\mathcal{D}^p$ following the likelihoods from P and then measuring the similarity between the resulting subset and the manually annotated data $\mathcal{D}^l$. However, the sampling action will lead to non-differentiability and hinder the model be optimized in a holistic manner.

To solve this issue, we approximate the sampling action by assigning each instance x in $\mathcal{D}^p$ with $P(x)$ as the weight and let the weights play a role in measuring the similarity between $\mathcal{D}^p$ and $\mathcal{D}^l$ (as shown in Equation 3). In this continuous approximation setting, the instances with higher weights have more effect on the measurement, which is equivalent to more frequent participation in sampling setting. Therefore, this similarity is controlled by the weights.

Discriminator. D aims to detect if an instance is from $\mathcal{D}^l$ or the weighted $\mathcal{D}^p$. To improve D's discriminating ability, we optimize it by minimizing the loss function defined as:

$$\begin{aligned} L_d = &- \mathbb{E}_{x\sim\mathcal{D}^l}[\log D(x)] \\ &- \frac{1}{\sum_{x\sim\mathcal{D}^p} P(x)} \sum_{x\sim\mathcal{D}^p} P(x)\log(1-D(x)) \end{aligned} \quad (3)$$

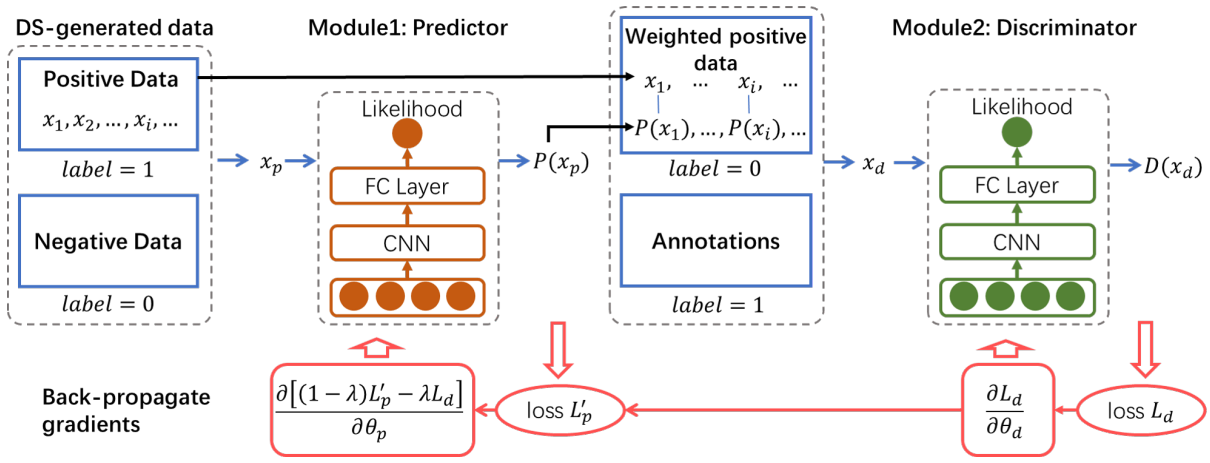


Figure 2: Overview of the adversarial discriminative denoising framework.

Essentially, D is the metrics of the weights, and P adjusts itself according to the feedback about the weights. In $\mathcal{D}^p$, the true positive instances are more difficult to be correctly recognized by D than the false positive ones. In order to puzzle D and cause more losses to it, P will assign higher weights on the true positive ones while lower weights on the noisy ones. As the adversary of P, D have to pay more attention to the instances with high weights so as to avoid major losses. This will drive P to avert mistaking noisy data as the correct ones.

Architecture of Predictor and Discriminator

In the adversarial discriminative denoising architecture, it is straightforward to generalize P and D to any derivable classification models, which might be more appropriate for the data at hand. In this paper, we use a convolutional neural network (CNN) based model similar to the one proposed by (Zeng et al. 2014) to both P and D. Considering their task complexity is comparative, we employ the same model structure to them for simplicity. In this model, we extract sentence-level features via a convolutional layer and then concatenate it with the lexical-level features to get the final sentence representation. Finally, we feed the sentence representation into a fully-connected (FC) layer and obtain a probability. For P, the probability represents the likelihood that an instance mentions relation r. For D, the probability represents the likelihood that a sentence is from $\mathcal{D}^l$ other than $\mathcal{D}^p$. The parameters of P and D are denoted as θ_p and θ_d respectively.

Optimization

In order to demonstrate the convergence of our adversarial discriminative architecture, we first note that the loss functions of P and D can be embedded into the following value function with which P and D play a two-player min-max game:

$$\min_P \max_D V(\theta_p, \theta_d) = (1-\lambda)L'_p - \lambda L_d \tag{4}$$

where we are seeking the parameters (θ_p, θ_d) that deliver a saddle point given by

$$\begin{aligned} \hat{\theta_p} &= \arg\min_{\theta_p} V(\theta_p, \theta_q) \\ \hat{\theta_d} &= \arg\max_{\theta_d} V(\theta_p, \theta_q) \end{aligned} \tag{5}$$

Thus, the optimization problem involves a minimization with respect to θ_p, as well as a maximization with respect to θ_d. During training, P and D are competing against each other, in an adversarial way, over the value function above.

For simplicity in implementation, we rewrite the loss function of P as Equation (6) since the first item on the right side of Equation (3) is irrelevant with θ_p:

$$\begin{aligned} L_p =& (1-\lambda)L'_p + \\ & \frac{\lambda}{\sum_{x \sim \mathcal{D}^p} P(x)} \sum_{x \sim \mathcal{D}^p} P(x) \log(1 - D(x)) \end{aligned} \tag{6}$$

The update rules of θ_p and θ_d are as follows:

$$\begin{aligned} \theta_p &\leftarrow \theta_p - \mu \frac{\partial L_p}{\partial \theta_p} \\ \theta_d &\leftarrow \theta_d - \mu \frac{\partial L_d}{\partial \theta_d} \end{aligned} \tag{7}$$

where μ is the learning rate. Algorithm 1 provides the complete pseudo-code of the learning procedure.

Cleaning Noisy Dataset with Predictor

After the adversarial learning process, we obtain a valid P concerning relation r. Then, we apply P to $\mathcal{D}^p$ and filter out the instances whose score is below a certain threshold thr. The cleansed positive data will be used as the positive data of relation r in the training stage of RE models.

Algorithm 1 Minibatch stochastic gradient descent training of the adversarial discriminative denoising framework

Input: Positive data $\mathcal{D}^p$ and negative data $\mathcal{D}^n$ generated by DS; Manually labeled data $\mathcal{D}^l$.
Output: Parameters θ_p of P and parameters θ_d of D.

for number of training iterations **do**
  for k_1 steps **do**
    Sample minibatch of m manually labeled positive samples $\{x_1^l, x_2^l, ..., x_m^l\}$ from $\mathcal{D}^l$.
    Sample minibatch of m positive samples $\{x_1^p, x_2^p, ..., x_m^p\}$ from $\mathcal{D}^p$.
    if it is the first iteration **then**
      Assign 1 as the confidence score s_i of each sentence x_i^p in $\mathcal{D}^l$.
    else
      Assign $P(x_i^p)$ as the confidence score s_i of each sentence x_i^p in $\mathcal{D}^l$.
    Update P by descending its stochastic gradient:

$$\nabla_{\theta_d} - \left[\frac{1}{m}\sum_{i=1}^{m}\log(D(x_i^l)) + \frac{1}{\sum_{i=1}^{m} s_i}\sum_{i=1}^{m} s_i \log(1 - D(x_i^p))\right]$$

  for k_2 steps **do**
    Sample minibatch of m positive samples $\{x_1^p, x_2^p, ..., x_m^p\}$ from $\mathcal{D}^p$.
    Sample minibatch of m negative samples $\{x_1^n, x_2^n, ..., x_m^n\}$ from $\mathcal{D}^n$.
    Update D by descending its stochastic gradient:

$$\nabla_{\theta_p} - \left[(1-\lambda)\left(\frac{1}{m}\sum_{i=1}^{m}\log P(x_i^p) + \frac{1}{m}\sum_{i=1}^{m}\log(1 - P(x_i^n))\right) - \lambda\left(\frac{1}{\sum_{i=1}^{m} P(x_i^p)}\sum_{i=1}^{m} P(x_i^p)\log(1 - D(x_i^p))\right)\right]$$

Experiments

Our experiments were intended to (1) show the effectiveness of our approach and (2) illustrate the effect of the adversarial setting.

Datasets and Evaluation Metrics

We carried out the experiments on the widely used NYT dataset[3] (Riedel, Yao, and McCallum 2010) and a customized dataset (naming CusDT).

NYT. This dataset was generated by aligning Freebase with NYT corpus. The aligned sentences from the years 2005-2006 are used for training, and the aligned sentences from 2007 are used for test. There are 53 relations including a special relation NA which indicates there is no relation between the two entities.

When cleaning the positive training data $\mathcal{D}^p$ of each relation r except NA, very few annotations $\mathcal{D}^l$ were manually picked up from it, and all the training data of other relations were treated as $\mathcal{D}^n$. In total, we annotated 1,500 instances, which is less than $\leq 1\%$ of the original positive data.

CusDT. We constructed this dataset based on SemEval 2010 task 8 [4], which consists of 10,717 annotated examples of 9 relations and an artificial relation *Other* similar with relation NA in NYT. Firstly, we chose one relation except *Other* as the target relation. Then, we partitioned the corpus into three parts: (1) *target*, consisting of instances of the target relation; (2) *other*, consisting of instances of the other relations except *Other*; and (3) *na*, consisting of instances of relation *Other*. Finally, we built $\mathcal{D}^l$, $\mathcal{D}^p$ and $\mathcal{D}^n$ respectively. $\mathcal{D}^l$ is composed of N^l instances from *target*. $\mathcal{D}^p$ is composed of N_t^p true positive instances from *target* and N_f^p false positive instances from *other*. $\mathcal{D}^n$ is composed of N_{na}^n instances from *na*. To support experimental analysis, we constructed four groups of CusDT datasets, and each group had 9 datasets corresponding to the 9 target relations. Each group of datasets had an annotations scale $R_l = N^l/(N_t^p + N_f^p)$ of 2% and a noise proportion $R_n = N_f^p/(N_t^p + N_f^p)$ ranging from 0.3 to 0.75 by step 0.15.

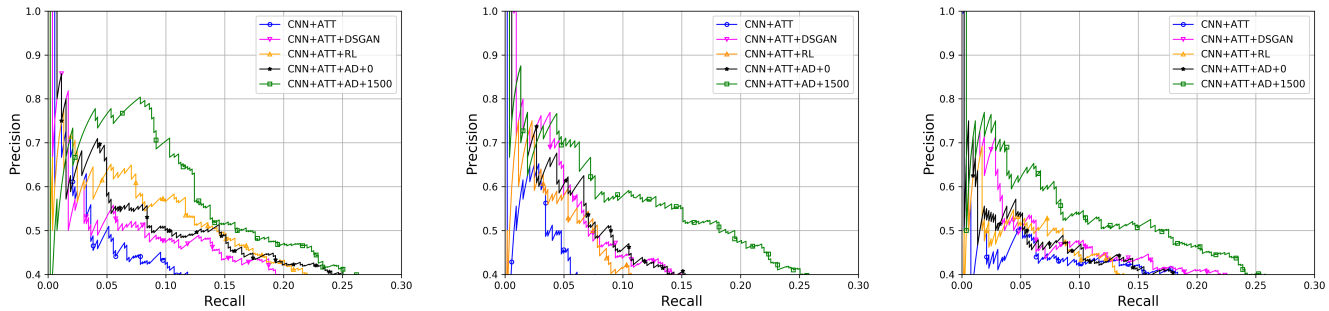
Evaluation Metrics. Following previous works, we evaluated the denoising result on the NYT dataset. Due to the absence of true labels in the NYT dataset, we could not evaluate the denoising result accurately. Considering the ultimate purpose of removing noise is to enhance the relation extractor, we trained the relation extractor on the cleansed dataset and compared its performance with that on the original dataset (Qin, Xu, and Wang 2018a; 2018b).

We evaluated the performance of the relation extractor via heldout evaluation, which is the most widely used method (Zeng et al. 2015; Qin, Xu, and Wang 2018a). Heldout evaluation compares the relation facts discovered from the test articles with those in Freebase to provide an **approximate** measure of precision without time-consumed human evaluation. Besides, we observed that the test data have a large amount of overlapped triplets with the training data. To make the evaluation result more reasonable, we removed the aligned instances by these triplets from the test data.

Heldout evaluation cannot measure the denoising perfor-

[3] http://iesl.cs.umass.edu/riedel/ecml/

[4] http://docs.google.com/View?docid=dfvxd49s_36c28v9pmw



(a) 31% noise proportion (b) 50% noise proportion (c) 70% noise proportion

Figure 3: Aggregate PR curves of CNN-ATT based model.

mance accurately because whether a triplet exists in KB does not strictly means the aligned text mentions the corresponding relation (Riedel, Yao, and McCallum 2010). Although heldout evaluation is useful, it can not support the analysis relying on accurate evaluation, including analyzing the effect of the adversarial setting in our approach. Thus, we performed the accurate analysis on CusDT. Since CusDT has the true labels of the data, we can evaluate the denoising performance accurately with F1-score.

Parameter Settings

For the parameters of P and D, we set the window size s^w of the convolution layer as 3, the kernel number d^k as 100 and the dimension of position embedding d^p as 5. As for word embeddings, we directly used the word embedding file released by (Lin et al. 2016)[5], whose dimension d^w is 50. For the parameters of the adversarial discriminative architecture, we set the trade-off factor λ as 0.2 and the learning rate μ as 1e-3. We employed a dropout rate of 0.5 during the training procedure.

Performance on DS

In this section, we aim to illustrate the denoising performance of our approach and compare it with the state-of-the-art methods. Specifically, we chose the CNN based model with attention mechanism (CNN+ATT) proposed by (Lin et al. 2016) as the DS relation extractor. Then, we chose two state-of-the-art unsupervised denoising methods based on RL (Qin, Xu, and Wang 2018b) (CNN+ATT+RL) and GAN (Qin, Xu, and Wang 2018a) (CNN+ATT+DSGAN) as the comparison. Besides, we trained our approach in two ways, including using none annotation (CNN+ATT+AD+0) and 1500 annotations (CNN+ATT+AD+1500).

To adequately compare these methods and explore their applicabilities when the noise proportion ranges, we carried out three groups of experiments on the datasets with different noise proportions. (Riedel, Yao, and McCallum 2010) has reported 31% of noise in the original NYT dataset. To increase the amount of noise, we reallocated a certain amount of instances of NA into the positive data as noise. As a result, we got two additional datasets with 50% and 70% noise respectively.

The three charts in Figure 3 show the results of these experiments. It is noticeable that the PR curves seem worse than that in the existing works because we have removed the test data aligned by the overlapped triplets. We carry out the analysis as follows.

Degree of Human Supervision. In these experiments, we only annotated 1,500 instances, which is less than 1%, even 0.5% in some cases, of the original dataset to clean. However, the three charts show CNN+ATT+AD+1500 is much better than the baseline CNN+ATT. Thus, very little human supervision is sufficient for our approach to enhance the performance of the relation extractor significantly.

Effectiveness. (1) CNN+ATT+AD+0 always performs better than the baseline CNN+ATT in the three charts. It shows our approach can take effect without human supervision by exploiting the potential beneficial information underlying the original noisy data. (2) CNN+ATT+AD+1500 achieves excellent improvements than CNN+ATT, especially in (b) and (c) when the noise proportion is greater than 50%. Thus, our approach can achieve better denoising performance by introducing human supervision high-efficiently.

Advantages over Unsupervised Methods. (1) CNN+ATT+AD+0 always has similar performance to CNN+ATT+RL and CNN+ATT+DSGAN. In another word, our framework has a similar ability to mine the implicit information with these unsupervised methods. (2) CNN+ATT+AD+1500 outperforms CNN+ATT+RL and CNN+ATT+DSGAN over the most range of recall, especially when the noise proportion is greater than 50%. It shows our approach outperforms existing unsupervised methods significantly by introducing very little human supervision.

Applicability. Comparing the three charts, we can see that CNN+ATT+RL, CNN+ATT+DSGAN and CNN+ATT+AD+0 get decreased performance as the noise proportion increases and can hardly improve the

[5]https://github.com/thunlp/NRE

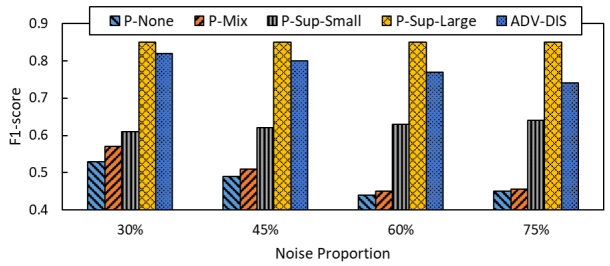


Figure 4: Denoising results of the variant methods.

baseline when the noise occupies most of the proportion (70%). This phenomenon indicates that the increasing noise proportion makes it more difficult to mine the implicit information. On the contrary, CNN+ATT+AD+1500 can still enhance the relation extractor effectively. Thus, our approach has much more wide applicability than these unsupervised methods when the noise proportion ranges.

Effect of the Adversarial Setting

Our framework consists of a positive instance predictor P and a data source discriminator D. Introducing human annotations via the competition between P and D is the key of our approach. In this section, we aim to demonstrate the effect of this adversarial setting.

We designed another four variant methods as the comparison. Commonly, these methods only exploit P trained on $\mathcal{D}^p$ and $\mathcal{D}^n$ and treat the noise reduction as a binary classification problem. The differences between them lie in $\mathcal{D}^p$. Specifically, these methods are:

- P-None, which use the original $\mathcal{D}^p$. This method does not introduce any human supervision.
- P-Mix, which mixes $\mathcal{D}^l$ and $\mathcal{D}^p$ as $\mathcal{D}^p$. This method introduces human annotations in a simple way.
- P-Sup-Small, which uses $\mathcal{D}^l$ as $\mathcal{D}^p$. This method gives up mining the implicit information in $\mathcal{D}^p$ and treats the problem in supervised paradigm.
- P-Sup-Large, which also treats the problem in supervised paradigm, but uses much more human annotations than P-Sup-Small. It will combine $\mathcal{D}^l$ with a certain amount of positive instances from $\mathcal{D}^p$ as $\mathcal{D}^p$.

We have carried out four groups of experiments on the four groups of CusDT datasets and took the average F1-score on each group of datasets as the denoising performance. Figure 4 shows the results of these variant methods. We can observe that: (1) ADV-DIS can achieve much better performance than P-None. This shows our approach can introduce human supervision effectively to improve the denoising result. (2) ADV-DIS always performs better than P-Mix, especially when the noise proportion is large. Thus, the adversarial setting is a much more effective way of introducing human supervision than the simple way used by P-Mix. (3) P-Sup-Small always performs worse than ADV-DIS. We infer the main reason that the supervised method is inferior to our approach is that it cannot use the implicit information underlying the original data. It is a great advantage for our approach to combine both the explicit supervision and the implicit information in a unified framework. (4) Although ADV-DIS performs not as well as P-Sup-Large, it can get close to P-Sup-Large using much less human annotations (less than 10%). Thus, our approach has much less dependence on annotations than the supervised method.

Table 1: Examples from the experimental results on NYT.

True positive sentences	Score1	Score2
Timothy_f._geithner, president of the **federal_reserve_bank_of_new_york**, discussed the issue in a recent speech.	0.42	0.86
Robert_shaye, the founder of **new_line_cinema**, a division of time warner that will celebrate its 40th anniversary next year ...	0.49	0.82
False positive sentences	Score1	Score2
Under **california** law , gov. **arnold_schwarzenegger** has 14 days to call a special election to fill mr. cunningham 's seat ...	0.60	0.43
Under its licensing agreement with **dreamworks**, the studio that produced the **steven_spielberg** film , abc was obligated to ...	0.54	0.31

Case Study

Table 1 shows several examples of relation */business/person/company* from the experimental results on NYT dataset. The scores represent the likelihoods these sentences are positive. *Score1* comes from CNN+ATT+AD+0, while *Score2* comes from CNN+ATT+AD+1500. We can observe that: after introducing a handful of human annotations in an adversarial mode, the scores of the true positive sentences are increased while the scores of the false positive instances are decreased. The resulting scores make it easier to separate these sentences.

Conclusion and Future Work

In this paper, we proposed an adversarial discriminative denoising framework to remove the false positive noise explicitly from DS-generated data. This framework provides an effective way of introducing human supervision and exploiting it along with the implicit information underlying the original data in a unified framework. Specifically, this framework consists of a positive instance predictor and a data source discriminator. In the adversarial setting, the predictor first obtains its denoising ability using the original data and then gets boosted in the competition with the discriminator. The resulting predictor is valid to filter out false positive instances from DS-generated positive data. Empirically, we showed our approach can significantly improve the performance of DS model on NYT dataset and outperforms the existing noise reduction methods with a handful of human

annotations. Besides, we demonstrated the effect of the adversarial setting of our approach using customized datasets.

Actually, the presented adversarial discriminative denoising framework does not rely on the DS task. In our future work, we will generalize this framework and promote it to other noise reduction tasks.

References

Bunescu, R. C., and Mooney, R. J. 2005. Subsequence kernels for relation extraction. In *Proceedings of NIPS*, 171–178.

Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; and Abbeel, P. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of NIPS*, 2172–2180.

dos Santos, C. N.; Xiang, B.; and Zhou, B. 2015. Classifying relations by ranking with convolutional neural networks. In *Proceedings of ACL*, 626–634.

Feng, J.; Huang, M.; Zhao, L.; Yang, Y.; and Zhu, X. 2018. Reinforcement learning for relation classification from noisy data. In *Proceedings of AAAI*, 5779–5786.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. S. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research* 17:59:1–59:35.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; and Bengio, Y. 2014. Generative adversarial nets. In *Proceedings of NIPS*, 2672–2680.

Hoffmann, R.; Zhang, C.; Ling, X.; Zettlemoyer, L. S.; and Weld, D. S. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of ACL*, 541–550.

Hu, Z.; Yang, Z.; Liang, X.; Salakhutdinov, R.; and Xing, E. P. 2017. Toward controlled generation of text. In *Proceedings of ICML*, 1587–1596.

Ji, G.; Liu, K.; He, S.; and Zhao, J. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *Proceedings of AAAI*, 3060–3066.

Larsen, A. B. L.; Sønderby, S. K.; Larochelle, H.; and Winther, O. 2016. Autoencoding beyond pixels using a learned similarity metric. In *Proceedings of ICML*, 1558–1566.

Lin, Y.; Shen, S.; Liu, Z.; Luan, H.; and Sun, M. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of ACL*, 2124–2133.

Luo, B.; Feng, Y.; Wang, Z.; Zhu, Z.; Huang, S.; Yan, R.; and Zhao, D. 2017. Learning with noise: Enhance distantly supervised relation extraction with dynamic transition matrix. In *Proceedings of ACL*, 430–439.

Min, B.; Grishman, R.; Wan, L.; Wang, C.; and Gondek, D. 2013. Distant supervision for relation extraction with an incomplete knowledge base. In *Proceedings of HLT-NAACL*, 777–782.

Mintz, M.; Bills, S.; Snow, R.; and Jurafsky, D. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL*, 1003–1011.

Qin, P.; Xu, W.; and Wang, W. Y. 2018a. DSGAN: generative adversarial training for distant supervision relation extraction. In *Proceedings of ACL*, 496–505.

Qin, P.; Xu, W.; and Wang, W. Y. 2018b. Robust distant supervision relation extraction via deep reinforcement learning. In *Proceedings of ACL*, 2137–2147.

Riedel, S.; Yao, L.; and McCallum, A. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of ECML-PKDD*, 148–163.

Ritter, A.; Zettlemoyer, L.; Mausam; and Etzioni, O. 2013. Modeling missing data in distant supervision for information extraction. *Transactions of the Association for Computational Linguistics* 1:367–378.

Surdeanu, M.; Tibshirani, J.; Nallapati, R.; and Manning, C. D. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of EMNLP-CoNLL*, 455–465.

Takamatsu, S.; Sato, I.; and Nakagawa, H. 2012. Reducing wrong labels in distant supervision for relation extraction. In *Proceedings of ACL*, 721–729.

Wang, L.; Cao, Z.; de Melo, G.; and Liu, Z. 2016. Relation classification via multi-level attention cnns. In *Proceedings of ACL*, 1298–1307.

Xu, W.; Hoffmann, R.; Zhao, L.; and Grishman, R. 2013. Filling knowledge base gaps for distant supervision of relation extraction. In *Proceedings of ACL*, 665–670.

Zelenko, D.; Aone, C.; and Richardella, A. 2003. Kernel methods for relation extraction. *Journal of Machine Learning Research* 3:1083–1106.

Zeng, D.; Liu, K.; Lai, S.; Zhou, G.; and Zhao, J. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING*, 2335–2344.

Zeng, D.; Liu, K.; Chen, Y.; and Zhao, J. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of EMNLP*, 1753–1762.

Zeng, X.; He, S.; Liu, K.; and Zhao, J. 2018. Large scaled relation extraction with reinforcement learning. In *Proceedings of AAAI*, 5658–5665.

Zhou, G.; Su, J.; Zhang, J.; and Zhang, M. 2005. Exploring various knowledge in relation extraction. In *Proceedings of ACL*, 427–434.