GAN Driven Semi-distant Supervision for Relation Extraction

Pengshuai Li[1], **Xinsong Zhang**[1], **Weijia Jia**[2,1*] and **Hai Zhao**[1*]
[1]Dept. of CSE, Shanghai Jiao Tong University, Shanghai, China
[2]State Key Lab of IoT for Smart City, CIS, University of Macau, Macao, SAR China
{pengshuai.li,xszhang0320}@sjtu.edu.cn
{jia-wj,zhaohai}@cs.sjtu.edu.cn

Abstract

Distant supervision has been widely used in relation extraction tasks without hand-labeled datasets recently. However, the automatically constructed datasets comprise numbers of wrongly labeled negative instances due to the incompleteness of knowledge bases, which is neglected by current distant supervised methods resulting in seriously misleading in both training and testing processes. To address this issue, we propose a novel semi-distant supervision approach for relation extraction by constructing a small accurate dataset and properly leveraging numerous instances without relation labels. In our approach, we construct accurate instances by both knowledge base and entity descriptions determined to avoid wrong negative labeling and further utilize unlabeled instances sufficiently using generative adversarial network (GAN) framework. Experimental results on real-world datasets show that our approach can achieve significant improvements in distant supervised relation extraction over strong baselines.

1 Introduction

Relation extraction aims to identify relations for a pair of entities in a sentence to construct relation triples like [*Steve Jobs, Founder, Apple*]. It has been well studied by supervised approaches with hand-labeled data. However, supervised methods are limited to costly hand-labeled training sets and hard to be extended to large-scale relations. To break the bottleneck of hand-labeled training set, distant supervision (Mintz et al., 2009) automatically construct datasets with knowledge bases. It assumes that if two entities have a known relation in a knowledge base, all sentences that mention these two entities will probably express the same relation and can be called positive instances. At the same time, it treats sentences as negative instances whose entity pairs do not have a known relation in knowledge bases. Due to the strong assumption, instances are likely to be mislabeled. To alleviate the wrong labeling problem, distant supervised methods have been implemented with multi-instance learning and neural networks (Riedel et al., 2010; Hoffmann et al., 2011; Zeng et al., 2015; Lin et al., 2016, 2017). However, previous works focus on positive instances and few methods have addressed the issue of false negative instances.

The false negative instances, which contain true relations, are misclassified sentences in the negative set due to the incomplete nature of knowledge bases. For example, over 70% of people included in Freebase have no known place of birth (Dong et al., 2014). As shown in Figure 1, *S1* presents the relation *place of birth*, while it is labeled as a negative instance. The other three sentences are mislabeled in the same way. The missing relation triples in knowledge bases yield numbers of false negative instances in the automatically labeled dataset. These instances will not only mislead the training method to an unreliable convergence but also make the measurement criteria inaccurate in the testing process. Table 1 compares the precision of automatic and manual evaluation methods for top N predictions by the previous relation extractor (Lin et al., 2016) on the NYT dataset. From the table, we can see that manual evaluation is more precise than automatic evaluation by over 19.8%.

* Corresponding authors: Weijia Jia, Hai Zhao, {jia-wj, zhaohai}@cs.sjtu.edu.cn. This work is supported by National China 973 Project No. 2015CB352401; Chinese National Research Fund (NSFC) Key Project No. 61532013 and No. 61872239. 0007/2018/A1, DCT-MoST Joint-project No. 025/2015/AMJ,FDCT,SAR Macau, China, and University of Macau Grant Nos: MYRG2018-00237-RTO, CPG2019-00004-FST and SRG2018-00111-FST, National Key Research and Development Program of China (No. 2017YFB0304100), Key Projects of National Natural Science Foundation of China (U1836222 and 61733011), and Key Project of National Society Science Foundation of China (No. 15-ZDA041).

Proceedings of NAACL-HLT 2019, pages 3026–3035
Minneapolis, Minnesota, June 2 - June 7, 2019.

ID	Instances	Dataset Label	Predicted Label
S1	[James Hillier] was born in [Brantford], Ontario.	NA	PB
S2	Dr. Fortner will be interred in [Bedford], [Indiana] with his parents.	NA	LC
S3	"This is an expression of what has been going on", archbishop [Phillip Aspinall] of [Australia] said at a news briefing here.	NA	PN
S4	What Dr. Sims did is called user-driven innovation by [Eric Von Hippel], a professor at the [Massachusetts Institute of Technology]'s Sloan School of management.	NA	PC

PB: /person/place of birth LC: /location/contains PN: /person/nationality
PC: /person/company NA: non-relation

Figure 1: Illustration of the false negative instances in relation extraction by distant supervision. Instances are selected from a widely used dataset NYT (Riedel et al., 2010).

The huge bias mainly comes from false negative instances in the testing set, which severely limits the upper bound of accuracy for relation extraction. Therefore, handling false negative instances is a pivotal issue to improve the performance of distant supervised relation extraction.

Evaluations	P@100	P@200	P@300
Automatic	76.2	73.1	67.4
Manual	96.0(+19.8)	95.5(+22.4)	91.0(+23.6)

Table 1: The Precision at top N predictions (%) of the model Lin et al. (2016) upon automatic and manual evaluations on the NYT Dataset

To alleviate the effect of false negative instances, there are two possible ways. One is improving the accuracy of the automatically labeled dataset, and the other is properly leveraging unlabeled instances which cannot be labeled as positive or negative. The former way is to construct an accurate dataset by filtering credible negative instances but limited by high annotation cost and the resulting dataset size. The latter way is to train relation extraction models with abundant unlabeled instances but restricted by the prerequisite of an accurate dataset used as ground truth. Therefore, we propose a novel semi-distant supervised approach by integrating both ways to decrease the influence of false negative instances for better relation extraction.

In our approach, we additionally use entity descriptions together with a knowledge base to construct an accurate dataset. Supervised by the dataset as ground truth, to effectively exploit numbers of unlabeled instances, we train our relation extractor using a generative adversarial network (GAN) framework. In detail, We propose a three-player min-max game to generate proper relation representations for unlabeled instances in an adversarial way which minimizes the difference between labeled and unlabeled data and maximizes the probability of distinguishing from each other at the same time. The experiments demonstrate that our approach is effective and outperforms the state-of-the-art work. In summary, we make the following major contributions:

- We propose a novel semi-distant supervision method for relation extraction to alleviate the influence of false negative instances.
- To the best of our knowledge, we are the first to generate valid relation representations for sentences by an adversarial algorithm. Numbers of unlabeled instances are used to improve the performance of relation extraction. Moreover, our generative adversarial training strategy is proved effective on an additional sentiment classification with sixteen real-world datasets.
- We construct a new accurate dataset for relation extraction extended from the NYT dataset. Our approach increases the area of the Precision-Recall curve from 0.39 to 0.56 over the baselines.

2 Related Work

To extend relation extraction to large-scale datasets, distant supervision (Mintz et al., 2009) automatically labeled training sets with knowledge bases such as Freebase. Although this method is working well for large-scale relation extraction, it is trapped in the wrong labeling problem for positive instances. To deal with this problem, multi-instance learning was combined with

distant supervision (Riedel et al., 2010; Hoffmann et al., 2011). Inspired by the pioneering work, a series of later studies were conducted to further improve distant supervised relation extraction with methods such as multi-instance multi-label learning (Surdeanu et al., 2012), graph model for label generation (Takamatsu et al., 2012), partial supervision (Angeli et al., 2014), matrix completion with low-rank criterion (Fan et al., 2014) and modeling the neighbor consistency with Markov logic (Han and Sun, 2016).

However, the performance of the methods mentioned above strongly depends on the quality of human-designed features. With the development of neural models, relation features with semantic meaning can be accurately, simply and automatically extracted. Zeng et al. (2015) proposed the first neural relation extraction with distant supervision. Mnih et al. (2014), Lin et al. (2016), Zhang et al. (2018), Han et al. (2018) and Du et al. (2018) showed that attention model could improve the accuracy of neural relation extraction. Another similar work (Ji et al., 2017) assigned better attention weights with extra data like entity descriptions. DSGAN (Qin et al., 2018a), a GAN-based method, was also used to recognize true positive instances from noisy datasets. To further alleviate the effect of wrong labeling problem, soft-label training algorithm (Liu et al., 2017b), reinforcement learning methods (Feng et al., 2018; Qin et al., 2018b) and additional side information (Vashishth et al., 2018; Wang et al., 2018) have been used. Most recently, a few methods focused on the pre-training embeddings for word tokens and relations including adversarial training (Wu et al., 2017), transfer learning (Liu et al., 2018) and relation decoder (Su et al., 2018).

All the above methods mainly pay attention to positive instances. Whereas, few studies work on the quality of negative instances, which is exactly the focus of this paper. We effectively construct a reliable dataset with both entity descriptions and a knowledge base, and thus propose a novel semi-distant supervised method to extract relations precisely.

3 Method

In the distant supervised relation extraction paradigm, all sentences labeled by a relation triple constitute a bag, and each sentence is called an instance. The relation triple is described as [*head, relation, tail*], where $head$ and $tail$ are both entities. We extract relation features from labeled training bags and then predict relations for unseen bags in the test set. This section presents our method about constructing an accurate dataset, the sentence encoder for relation representation and the semi-supervised way for relation extraction.

3.1 Dataset Construction

To reduce false negative instances, we construct a new reliable dataset extended from a widely used dataset NYT (Riedel et al., 2010) with entity descriptions. Entity descriptions are crawled from Wikipedia with entity name matching[1]. We assume that if an entity is relevant to another entity, its name is possibly mentioned in the description of the other entity. For example, the entity *Apple Inc.* is mentioned in the description of *Steve Jobs*. To verify the assumption, we count the number of all the accurate positive instances whose entity descriptions mention the other entity name in the NYT corpora. There are 163,108 positive sentences in total, in which 161,392 ones contain entity pairs that related to each other in their descriptions at least once. In other words, over 98.9% instances in positive set fitting our assumption indicates that most entity pairs in positive instances contain each other in their descriptions. Therefore, a former negative instance has a big chance to be credible negative if any of its entities is not mentioned in the description of the other one. Excluding instances that contain entity pairs related to each other in their descriptions, we can obtain more confident negative instances. Finally, we filter credible positive and negative instances from the dataset, and the other instances are unlabeled ones that cannot be labeled as positive or negative.

3.2 Sentence Encoder

3.2.1 Input Embedding

We pre-train input embeddings of word tokens including word and position embeddings. Word embeddings are distributed representations that map each word to a vector $word \in \mathcal{R}^w$, where the parameter w indicates the dimension of the vector. The vectors are trained in advance by $word2vec$ in the setting of Skip-gram (Mikolov et al., 2013). In the task of relation extraction, the relative positions of input tokens are important information.

[1]One entity name may refer to multiple entities which have their own pages. In our work, all the matched pages are collected together to obtain its description.

Position embeddings are defined as the combination of the relative distances from the current word to $head$ and $tail$. For instance, the relative distances from [*co-founder*] to [*Steve Jobs*] and [*Apple*] are respectively 3 and -6 in the sentence *Steve Jobs was the co-founder and the CEO of the Apple*. We encode distances to vectors $position \in \mathcal{R}^p$, where p is the dimension. The position embeddings are initialized randomly and updated in the training process. Finally, word embeddings and position embeddings are concatenated together to feed the neural model. We denote all the words in an instance as an initial vector sequence $b^* = \{x_1, \cdots, x_i, \cdots, x_q\}$, where $x_i \in \mathcal{R}^{w+p}$ and q is the number of words in the instance b^*.

3.2.2 Convolutional Encoder

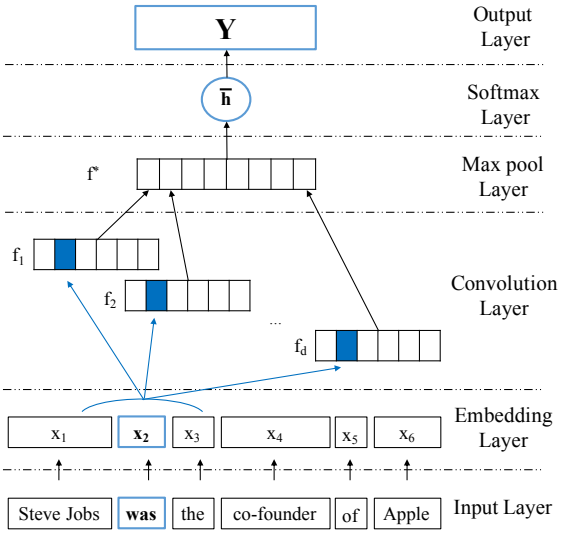


Figure 2: The architecture of our sentence encoder illustrating the procedure for handling one instance and predicting the relation between [Apple] and [Steve Jobs]. All forms of f are hidden states, $\bar{h}$ is relation representation of the sentence and Y represents relation labels.

Convolutional Neural Network (CNN) is a widely used structure for sentence encoder as shown in Figure 2. With the input embeddings, the convolutional layer extracts local features with a sliding window of length k over the input tokens. In the figure, we extract local features from 3 ($k = 3$) adjacent word tokens with dot production between convolutional kernels and input embeddings. The convolutional kernels are weight vectors represented by $W \in \mathcal{R}^{d \times k(w+p)}$ and the number of kernels is d. In summary, the convolutional operation follows the equation,

$$f_{ij} = W_i \cdot [x_{j-1}; x_j; x_{j+1}], \tag{1}$$

where $[x; y]$ denotes the vertical concatenation of x and y. f_{ij} presents j-th value of the i-th filter, where i and j are in range $[1, d]$ and $[1, q]$ respectively. Out-of-range input values such as x_0 and x_{q+1} are taken to be zero. A max-pooling operation selects the most important features of each f_i with $f_i^* = max(f_{ij})$, where $f^* \in \mathcal{R}^d$. Furthermore, PCNN (Zeng et al., 2015) improves the max-pooling operation with a piecewise method whose outputs of convolutional filters are divided into three segments by $head$ and $tail$ entities. Therefore, the max pooling procedure is performed in three segments separately.

Then, we summarize f^* to $\bar{h}$ by a non-linear function such as the hyperbolic tangent. The final feature vector $\bar{h}$ is fed into output layer after the softmax method $\hat{p} = softmax(W_r \bar{h} + b_r)$, where $W_r \in \mathcal{R}^{z \times d}$ and $b_r \in \mathcal{R}^z$ are variables, $\hat{p} \in \mathcal{R}^z$ is the estimated probability for each class and z is the number of relations. A cost function for one instance is the negative log-likelihood of the relations,

$$J_{truth}(\hat{p}, y, \theta) = -\frac{1}{z} \sum_{j=1}^{z} y_j \log \hat{p}_j, \tag{2}$$

where $y \in \mathcal{R}^z$ is the one-hot represented ground truth and θ presents all the parameters.

3.3 Semi-distant Supervision

The architecture of our semi-distant supervision is shown in Figure 3. To sufficiently utilize the reconstructing dataset including accurately labeled instances and unlabeled ones, we propose a generative adversarial training strategy, which transforms unlabeled instances (x_{ul}) to labeled data (x_l) space by generating valid relation representations (x_{gen}) and making the distribution of labeled instances $p(x_l)$ equal to that of generative data $p(x_{gen})$ in relation space[2]. Inspired by Goodfellow et al. (2014), we further devise a three-player min-max game to generate valid data distribution $p(x_{gen})$ with sentence encoder, generative and discriminative modules. The generative module minimizes the difference of $p(x_l)$ and $p(x_{gen})$, and the discriminative module maximizes the probability of distinguishing from each other at the same

[2] $p(x_l)$ and $p(x_{gen})$ represents the data distribution of labeled and generative instances.

Algorithm 1 GAN driven Semi-Distant Supervision algorithm

Require: discriminator D, generator G, sentence encoder S, s_i, s_j and s_k are hyper-parameters to indicate iterator number of each module

Initialize the parameters of D, G, S with random weights θ_d, θ_g and θ_s
for numbers of training iterations **do**
  for s_i steps **do**
    Sample mini-batch of n samples from accurate instances set presented as x
    Sample mini-batch of m samples from unlabeled instances set presented as c
    Fix G and S, update D by ascending its stochastic gradient:
      $\nabla_{\theta_d}[\frac{1}{n}\sum_{u=1}^{n}\log D(x_u) + \frac{1}{m}\sum_{v=1}^{m}\log(1 - D(G(c_v)))]$
  end for
  for s_j steps **do**
    Sample mini-batch of m samples from unlabeled instances set presented as c
    Fix D and S, update G by descending its stochastic gradient:
      $\nabla_{\theta_g}\frac{1}{m}\sum_{v=1}^{m}\log(1 - D(G(c_v)))$
  end for
  for s_k steps **do**
    Sample mini-batch of n samples from accurate instances set presented as x
    Sample mini-batch of m samples from unlabeled instances set presented as c
    Fix D and G, update S by descending its stochastic gradient:
      $\nabla_{\theta_s}[-\frac{1}{nz}\sum_{u=1}^{n}\sum_{j=1}^{z} y_{l_j}\log\hat{p}(y_j|x_u) - \frac{1}{mz}\sum_{v=1}^{m}\sum_{j=1}^{z} y_{g_j}\log\hat{p}(y_j|G(c_v))]$
  end for
end for

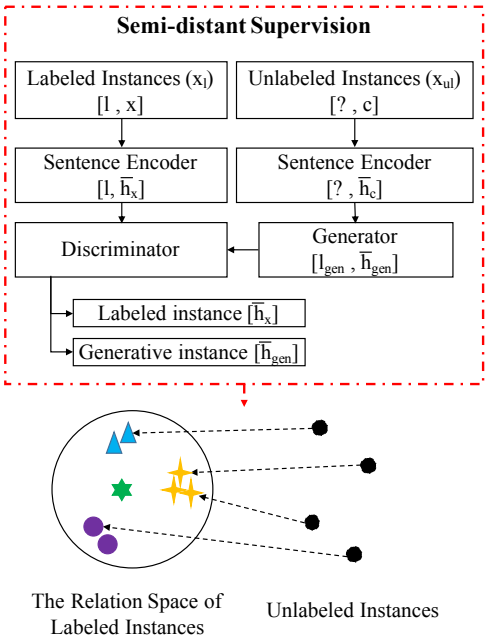


Figure 3: The architecture of GAN driven semi-distant supervision for relation extraction. $\bar{h}_x$ and $\bar{h}_c$ are relation representations of labeled and unlabeled instances respectively. $\bar{h}_{gen}$ is generated relation representation by the generator. l and l_{gen} represent accurate label and generated label respectively. The symbols in relation space represent labeled instances.

time. Sentence encoder is proposed as the third player, which extracts relation features from all the instances and produces a pre-trained relation representation for unlabeled instances. With the sentence encoder, we can control relation features contained in the generated representations.

Therefore, the discriminative module D will try to distinguish labeled data from generative data, while the generative module G makes $p(x_{gen}) \approx p(x_l)$. In addition, the sentence encoder S extracts relation features with all the training instances p_{all}. The training procedure is a three-player min-max game as the following equation,

$$\begin{aligned}\min_{S,G}\max_{D} V(S,D,G) = E_{x\sim p_{all}}[\log S(x)] \\ + E_{x\sim p_{x_l}}[\log D(x)] \\ + E_{c\sim p_{x_{gen}}}[\log(1 - D(G(c)))],\end{aligned} \tag{3}$$

In generative adversarial training, the discriminative module is trained by maximizing the gap between labeled data and generative data with the following equation,

$$J_D(x, c, \theta_d) = \log D(x) + \log(1 - D(G(c))), \tag{4}$$

where x and c are instances from accurately labeled set (x_l) and unlabeled set (x_{ul}) respectively.

θ_d presents parameters for the discriminator. $D(x)$ and $D(G(c))$ are defined as follows,

$$D(x) = \sigma(W_d \bar{h}_x), \tag{5}$$

$$D(G(c)) = \sigma(W_d(\bar{h}_c + W_g)), \tag{6}$$

where W_d and W_g are variables for discriminative and generative modules respectively. σ is the sigmoid function. The generative module is trained to make the generated relation representations more similar to real ones by the following loss function, where θ_g presents parameters.

$$J_G(c, \theta_g) = \log(1 - D(G(c))) \tag{7}$$

Finally, we train our sentence encoder S by optimizing the following loss function,

$$\begin{aligned} J_S(x, c, \theta_s) = &-\frac{1}{z}\sum_{j=1}^{z} y_{l_j} \log \hat{p}(y_j|x) \\ &-\frac{1}{z}\sum_{j=1}^{z} y_{g_j} \log \hat{p}(y_j|G(c)), \end{aligned} \tag{8}$$

where y_g means a one-hot vector which labels the most possible relation for unlabeled instances generated by the sentence encoder. θ_s represents parameters of S. The complete training procedure for generative adversarial training is shown as Algorithm 1.

4 Experiments

The experiments are proposed to answer the following three questions, 1) Is the proposed semi-distant supervision method effective for the task of relation extraction? 2) Is the constructed dataset credible enough? 3) Is the generative adversarial training helpful to relation extraction and other semi-supervised tasks?

4.1 Experimental Settings

4.1.1 Dataset

We conduct experiments on a widely used dataset NYT (Riedel et al., 2010) and its new version Accurate-NYT (A-NYT). A-NYT is a credible dataset filtered by our data construction module. We follow the previous work (Lin et al., 2016) to partition training and testing sets for NYT and A-NYT. Besides, we apply sixteen real-world datasets[3] (Liu et al., 2017a) to further verify the effectiveness of our generative adversarial training strategy on the task of sentiment classification. The dataset details are shown in Table 2.

[3]The datasets are Amazon product reviews and movie reviews.

Dataset	Positive	Negative	Unlabeled	Classes
NYT	163,108	579,428	-	53
A-NYT	163,108	240,453	338,975	53
Books	1,000	1,000	2,000	2
Electronics	1,000	998	2,000	2
DVD	1,000	1,000	2,000	2
Kitchen	1,000	1,000	2,000	2
Apparel	1,000	1,000	2,000	2
Camera	999	998	2,000	2
Health	1,000	1,000	2,000	2
Music	1,000	1,000	2,000	2
Toys	1,000	1,000	2,000	2
Video	1,000	1,000	2,000	2
Baby	1,000	900	2,000	2
Magazine	1,000	970	2,000	2
Software	1,000	915	475	2
Sports	1,000	1,000	2,000	2
IMDB	994	1,006	2,000	2
MR	986	1,014	2,000	2

Table 2: Statistics of the datasets

4.1.2 Evaluation Metrics and Baselines

On the dataset NYT and A-NYT, we evaluate our method in the classical held-out evaluation. It evaluates our models by comparing relation facts discovered from the test sentences with those in Freebase. Specifically, we report both the aggregate Precision-Recall (PR) curves and Precision at top N predictions (P@N) in our experiments. For the other datasets, we compute the precision of all the predictions.

We adopt the following baselines for distant supervised relation extraction.

Zeng et al. (2015) extracted relation features with piecewise convolutional neural network (PCNN).

Lin et al. (2016) integrated PCNN with selective attention mechanism (PCNN+ATT).

Wu et al. (2017) added adversarial noise at the level of the word embeddings (PCNN+ATT+AT).

Liu et al. (2017b) relabeled the training instances dynamically by the relation extractor (PCNN+ATT+SL).

Qin et al. (2018a) designed a GAN to recognize true positive samples (PCNN+DSGAN).

Liu et al. (2018) shortened the training instances with the parser tree and pre-trained word embeddings with transfer learning, which is the latest state-of-the-art work.

Self-Training (ST) is a semi-supervised method that can be integrated with PCNN+ATT for unlabeled data, which generates relation types for unlabeled instances with the model itself.

4.1.3 Parameters

In our experiments, we use the $word2vec$ in the setting of Skip-gram to train the word embeddings on NYT set. To train our model efficiently, we iterate by randomly selecting a batch from the training set until convergence and apply sentence-level attention mechanism following the previous work (Lin et al., 2016). The parameter n and m are batch sizes for accurate and unlabeled datasets respectively. We update the gradient with adaptive moment estimation (Kingma and Ba, 2015). Furthermore, $L2$ regularization and dropout (Srivastava et al., 2014) are adopted to avoid overfitting. Finally, we use a grid search and cross-validation to determine the optional parameters as shown in Table 3. The hyper-parameters s_i, s_j and s_k are training steps for different modules of generative adversarial training. Since the other parameters have little effect on the results, we follow the settings as the previous work (Lin et al., 2016).

batch size (n, m)	50
s_i, s_j, s_k	2, 1, 2
filter number d	230
kernel size k	3
word dimension w	50
position dimension p	10
learning rate	0.001
dropout probability	0.5
L2 regularization strength	0.0001

Table 3: Parameter settings

4.2 Overall Performance of Semi-Distant Supervision

The overall performance of our method compared with baselines for distant supervised relation extraction is shown in Table 4. We can see that our semi-distant supervised method achieves much better results than the baselines on all metrics. The huge improvement comes from both the accurate dataset and the effective training strategy which leverages unlabeled instances properly.

4.3 Effect of Dataset A-NYT

In this section, we apply two previous methods Zeng et al. (2015) and Lin et al. (2016) on NYT and A-NYT. PR curves for NYT are reported in their papers, while PR curves for A-NYT come from our implementations of the two baselines. NYT and A-NYT share the same positive instances, while A-NYT set has less and credible negative instances. As shown in Figure 4(a), methods on A-NYT always obtain better performance. The huge gap between PR curves is caused by false negative instances in NYT, which are not used for training and testing in A-NYT. To prove that results on A-NYT are according to the actual situation, we do manual evaluations at P@Ns. As shown in Table 5, the huge bias caused by false negative instances on NYT is dramatically alleviated on the dataset A-NYT.

P@N	100	200	300	Mean	PR
Zeng et al. (2015)	72.3	69.7	64.1	68.7	0.33
Lin et al. (2016)	76.2	73.1	67.4	72.2	0.35
Wu et al. (2017)	81.0	74.5	71.7	75.7	0.34
Liu et al. (2017b)	87.0	84.5	77.0	82.8	0.34
Qin et al. (2018a)	78.0	75.5	72.3	75.3	0.35
Liu et al. (2018)	87.0	83.0	78.0	82.7	0.39
Our Method	**96.0**	**93.5**	**93.0**	**94.2**	**0.56**

Table 4: Overall performance at P@Ns(%) and PR curve areas

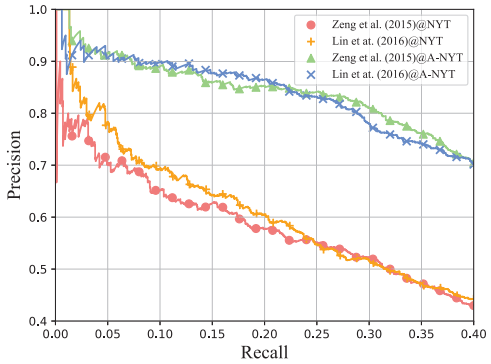
Evaluations	P@100	P@200	P@300
Automatic@NYT	76.2	73.1	67.4
Manual@NYT	96.0(+19.8)	95.5(+22.4)	91.0(+23.6)
Automatic@A-NYT	93.0	89.5	88.0
Manual@A-NYT	96.0(+3.0)	92.5(+3.0)	90.7(+2.7)

Table 5: P@Ns(%) of Lin et al. (2016) upon automatic and manual evaluations

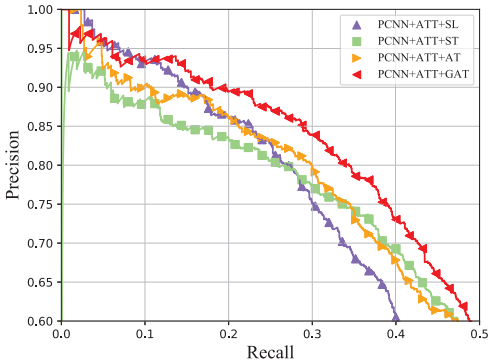
4.4 Effect of Generative Adversarial Training for Relation Extraction

To further demonstrate the effectiveness of our training strategies, we compare Generative Adversarial Training (GAT) with other baselines on the partially labeled dataset A-NYT as shown in Figure 4(b). The figure gives the following insights, 1) PCNN+ATT+ST and PCNN+ATT+AT do not work well, which is caused by the low quality of unlabeled instances. 2) PCNN+ATT+SL works as well as our models at low recall rate because of its excellent ability to extract notable features. Unfortunately, it falls far behind all the baselines at high recall rate, which means it tends to converge to a local optimum. 3) Our model achieves solid PR curves at all range of recall rate.

Meanwhile, we propose a detailed comparison of baselines with P@Ns and PR curve areas as shown in Table 6. From the table, we can see that our training strategy achieves much better result than the other baselines, which indicates that abundant unlabeled instances are helpful to extract relations only if used appropriately. Going



(a) PR curves of baselines on NYT and A-NYT [4]



(b) PR curves of our model and baselines on A-NYT with all the labeled and unlabeled data

Figure 4: PR curves for the comparisons (Better view in color)

deeper in the table, PCNN+ATT+SL works well at top predictions but obtains the worst PR curve. Our semi-distant supervised model with adversarial generations is useful for leveraging unlabeled instances properly.

P@N	100	200	300	Mean	PR
Zeng et al. (2015)‡	91.0	88.5	87.6	89.0	0.513
Lin et al. (2016)‡	93.0	89.5	88.0	88.2	0.513
Qin et al. (2018a)‡	90.0	91.0	88.3	89.8	0.524
Liu et al. (2018)‡	93.0	93.5	91.3	92.6	0.503
PCNN+ATT+SL	**96.0**	93.0	90.6	93.2	0.466
PCNN+ATT+ST	92.0	88.0	85.3	88.4	0.519
PCNN+ATT+AT	95.0	92.0	88.6	91.9	0.526
PCNN+ATT+GAT	**96.0**	**93.5**	**93.0**	**94.2**	**0.558**

Table 6: P@Ns(%) and PR curve areas on A-NYT. Methods with ‡ do not use unlabeled data. PCNN+ATT+AT is a semi-supervised extension of original method (Wu et al. (2017)) to make the unlabeled instances consistent with their predictions.

4.5 Effect of Generative Adversarial Training for Sentiment Classification

To verify the expandability of generative adversarial training, we conduct additional experiments on the task of sentiment classification. We implement our model and three baselines based on the Long Short-Term Memory (LSTM) network[5]. The results are shown in Table 7, from which we see that, 1) Self-training obtains poor results compared with the basic LSTM model, which means they fail to utilize unlabeled data correctly. 2) Adversarial training improves the performance on only three of the datasets and performs poorly on others, which means they possibly rely on the quality of unlabeled data. 3) LSTM+GAT achieves better results than the baselines on most of the datasets because of generating high-quality representations for unlabeled sentences.

Dataset	LSTM	LSTM+ST	LSTM+AT	**LSTM+GAT**
Books	79.5	75.8	**80.5**	80.3
Elec.	80.5	77.5	**84.1**	81.5
DVD	81.7	75.8	78.6	**82.0**
Kitchen	78.0	79.3	**81.7**	81.3
Apparel	83.2	83.5	84.8	**85.2**
Camera	85.2	84.3	86.1	**86.8**
Health	84.5	84.4	81.7	**86.2**
Music	76.7	76.0	76.0	**80.3**
Toys	83.2	79.8	83.7	**84.8**
Video	81.5	79.7	80.4	**83.0**
Baby	84.7	84.3	83.0	**85.3**
Mag.	89.2	85.3	89.0	**89.5**
Soft.	84.7	84.1	83.3	**85.1**
Sports	81.7	79.8	82.3	**82.5**
IMDB	81.7	78.3	82.5	**82.8**
MR	72.7	71.8	72.3	**73.5**
Mean	81.8	80.0	81.9	**83.1**

Table 7: Precision for the sentiment classification task

5 Conclusions

In this paper, we propose a novel semi-distant supervision approach that is capable of jointly exploiting limited accurate and abundant unlabeled ones. We first construct a reliable dataset with a knowledge base and additional entity descriptions.

[4] Results on NYT are reported as their papers except Wu et al. (2017) and Qin et al. (2018a). Results of these two methods on NYT and all results on A-NYT are obtained by our implementations.

[5] Our generative adversarial training strategy is model-independent, meaning that it could be applied to other neural models.

With the dataset, the generative adversarial training strategy is proposed to deal with plenty of unlabeled instances, which generates valid relation representations. Our experiments show that the proposed approach achieves significant improvement over previous state-of-the-art baselines.

References

Gabor Angeli, Julie Tibshirani, Jean Wu, and Christopher D Manning. 2014. Combining distant and partial supervision for relation extraction. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1556–1567.

Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. 2014. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 601–610.

Jinhua Du, Jingguang Han, Andy Way, and Dadong Wan. 2018. Multi-level structured self-attentions for distantly supervised relation extraction. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2216–2225.

Miao Fan, Deli Zhao, Qiang Zhou, Zhiyuan Liu, Thomas Fang Zheng, and Edward Y Chang. 2014. Distant supervision for relation extraction with matrix completion. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 839–849.

Jun Feng, Minlie Huang, Li Zhao, Yang Yang, and Xiaoyan Zhu. 2018. Reinforcement learning for relation classification from noisy data. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, pages 5779–5786.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, pages 2672–2680.

Xianpei Han and Le Sun. 2016. Global distant supervision for relation extraction. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*, pages 2950–2956.

Xu Han, Pengfei Yu, Zhiyuan Liu, Maosong Sun, and Peng Li. 2018. Hierarchical relation extraction with coarse-to-fine grained attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2236–2245.

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 the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 541–550.

Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, pages 3060–3066.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Processings of the third International Conference for Learning Representations (ICLR)*.

Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2017. Neural relation extraction with multi-lingual attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 34–43.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2124–2133.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2017a. Adversarial multi-task learning for text classification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1–10.

Tianyi Liu, Xinsong Zhang, Wanhao Zhou, and Weijia Jia. 2018. Neural relation extraction via inner-sentence noise reduction and transfer learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2195–2204.

Tianyu Liu, Kexiang Wang, Baobao Chang, and Zhifang Sui. 2017b. A soft-label method for noise-tolerant distantly supervised relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1790–1795.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP)*, pages 1003–1011.

Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. 2014. Recurrent models of visual attention. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*, pages 2204–2212.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018a. Dsgan: Generative adversarial training for distant supervision relation extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 496–505.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018b. Robust distant supervision relation extraction via deep reinforcement learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2137–2147.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*, pages 148–163.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, pages 1929–1958.

Sen Su, Ningning Jia, Xiang Cheng, Shuguang Zhu, and Ruiping Li. 2018. Exploring encoder-decoder model for distant supervised relation extraction. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4389–4395.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 455–465.

Shingo Takamatsu, Issei Sato, and Hiroshi Nakagawa. 2012. Reducing wrong labels in distant supervision for relation extraction. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 721–729.

Shikhar Vashishth, Rishabh Joshi, Sai Suman Prayaga, Chiranjib Bhattacharyya, and Partha Talukdar. 2018. Reside: Improving distantly-supervised neural relation extraction using side information. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1257–1266.

Guanying Wang, Wen Zhang, Ruoxu Wang, Yalin Zhou, Xi Chen, Wei Zhang, Hai Zhu, and Huajun Chen. 2018. Label-free distant supervision for relation extraction via knowledge graph embedding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2246–2255.

Yi Wu, David Bamman, and Stuart Russell. 2017. Adversarial training for relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1778–1783.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1753–1762.

Xinsong Zhang, Pengshuai Li, Weijia Jia, and Hai Zhao. 2018. Multi-labeled relation extraction with attentive capsule network. *arXiv preprint arXiv:1811.04354*.