SENT: Sentence-level Distant Relation Extraction via Negative Training

Ruotian Ma[1], Tao Gui[2*], Linyang Li[1], Qi Zhang[1*], Xuanjing Huang[1] and Yaqian Zhou[1]
[1]School of Computer Science, Fudan University, Shanghai, China
[2]Institute of Modern Languages and Linguistics, Fudan University, Shanghai, China
{rtma19,tgui16,linyangli19,qz,xjhuang,yqzhou}@fudan.edu.cn

Abstract

Distant supervision for relation extraction provides uniform bag labels for each sentence inside the bag, while accurate sentence labels are important for downstream applications that need the exact relation type. Directly using bag labels for sentence-level training will introduce much noise, thus severely degrading performance. In this work, we propose the use of negative training (NT), in which a model is trained using complementary labels regarding that "the instance does not belong to these complementary labels". Since the probability of selecting a true label as a complementary label is low, NT provides less noisy information. Furthermore, the model trained with NT is able to separate the noisy data from the training data. Based on NT, we propose a sentence-level framework, SENT, for distant relation extraction. SENT not only filters the noisy data to construct a cleaner dataset, but also performs a relabeling process to transform the noisy data into useful training data, thus further benefiting the model's performance. Experimental results show the significant improvement of the proposed method over previous methods on sentence-level evaluation and de-noise effect.

1 Introduction

Relation extraction (RE), which aims to extract the relation between entity pairs from unstructured text, is a fundamental task in natural language processing. The extracted relation facts can benefit various downstream applications, e.g., knowledge graph completion (Bordes et al., 2013; Wang et al., 2014), information extraction (Wu and Weld, 2010) and question answering (Yao and Van Durme, 2014; Fader et al., 2014).

A significant challenge for relation extraction is the lack of large-scale labeled data. Thus, distant supervision (Mintz et al., 2009) is proposed to gather training data through automatic alignment between a database and plain text. Such annotation paradigm results in an inevitable noise problem, which is alleviated by previous studies using multi-instance learning (MIL). In MIL, the training and testing processes are performed at the bag level, where a bag contains noisy sentences mentioning the same entity pair but possibly not describing the same relation. Studies using MIL can be broadly classified into two categories: 1) the soft de-noise methods that leverage soft weights to differentiate the influence of each sentence (Lin et al., 2016; Han et al., 2018c; Li et al., 2020; Hu et al., 2019a; Ye and Ling, 2019; Yuan et al., 2019a,b); 2) the hard de-noise methods that remove noisy sentences from the bag (Zeng et al., 2015; Qin et al., 2018; Han et al., 2018a; Shang, 2019).

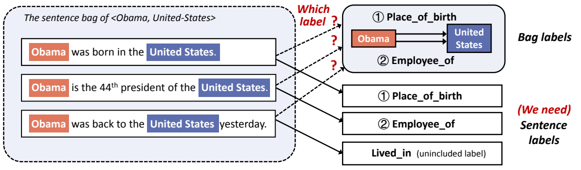

The sentence bag of <Obama, United-States>
Obama was born in the United States.
Obama is the 44th president of the United States.
Obama was back to the United States yesterday.
Which label ?
?
?
① Place_of_birth
Obama
United States
② Employee_of
Bag labels
① Place_of_birth
② Employee_of
Lived_in (unincluded label)
(We need) Sentence labels

Figure 1: Two types of noise exist in bag-level labels: 1) Multi-label noise: the exact label ("place_of_birth" or "employee_of") for each sentence is unclear; 2) Wrong-label noise: the third sentence inside the bag actually expresses "live_in" which is not included in the bag labels.

However, these bag-level approaches fail to map each sentence inside bags with explicit sentence labels. This problem limits the application of RE in some downstream tasks that require sentence-level relation type, e.g., Yao and Van Durme (2014) and Xu et al. (2016) use sentence-level relation extraction to identify the relation between the answer and the entity in the question. Therefore, several studies (Jia et al. (2019); Feng et al. (2018)) have made efforts on sentence-level (or instance-level)

[*] Corresponding authors.

distant RE, empirically verifying the deficiency of bag-level methods on sentence-level evaluation. However, the instance selection approaches of these methods depend on rewards(Feng et al., 2018) or frequent patterns(Jia et al., 2019) determined by bag-level labels, which contain much noise. For one thing, one bag might be assigned to multiple bag labels, leading to difficulties in one-to-one mapping between sentences and labels. As shown in Fig.1, we have no access to the exact relation between "place_of_birth" and "employee_of" for the sentence "Obama was born in the United States.". For another, the sentences inside a bag might not express the bag relations. In Fig.1, the sentence "Obama was back to the United States yesterday" actually express the relation "live_in", which is not included in the bag labels.

In this work, we propose the use of negative training (NT) (Kim et al., 2019) for distant RE. Different from positive training (PT), NT trains a model by selecting the complementary labels of the given label, regarding that "the input sentence does not belong to this complementary label". Since the probability of selecting a true label as a complementary label is low, NT decreases the risk of providing noisy information and prevents the model from overfitting the noisy data. Moreover, the model trained with NT is able to separate the noisy data from the training data (a histogram in Fig.3 shows the separated data distribution during NT). Based on NT, we propose SENT, a sentence-level framework for distant RE. During SENT training, the noisy instances are not only filtered with a noise-filtering strategy, but also transformed into useful training data with a re-labeling method. We further design an iterative training algorithm to take full advantage of these data-refining processes, which significantly boost performance. Our codes are publicly available at *Github*[1].

To summarize the contribution of this work:

- We propose the use of negative training for sentence-level distant RE, which greatly protects the model from noisy information.
- We present a sentence-level framework, SENT, which includes a noise-filtering and a re-labeling strategy for re-fining distant data.
- The proposed method achieves significant improvement over previous methods in terms of both RE performance and de-noise effect.

2 Related Work

2.1 Distant Supervision for RE

Supervised relation extraction (RE) has been constrained by the lack of large-scale labeled data. Therefore, distant supervision (DS) is introduced by Mintz et al. (2009), which employs existing knowledge bases (KBs) as source of supervision instead of annotated text. Riedel et al. (2010) relaxes the DS assumption to the express-at-least-once assumption. As a result, multi-instance learning is introduced (Riedel et al. (2010); Hoffmann et al. (2011); Surdeanu et al. (2012)) for this task, where the training and evaluating process are performed in **bag-level**, with potential noisy sentences existing in each bag. Most following studies in distant RE adopt this paradigm, aiming to decrease the impact of noisy sentences in each bag. These studies include the attention-based methods to attend to useful information (Lin et al. (2016); Han et al. (2018c); Li et al. (2020); Hu et al. (2019a); Ye and Ling (2019); Yuan et al. (2019a); Zhu et al. (2019); Yuan et al. (2019b); Wu et al. (2017)), the selection strategies such as RL or adversarial training to remove noisy sentences from the bag (Zeng et al. (2015); Shang (2019); Qin et al. (2018); Han et al. (2018a)) and the incorporation with extra information such as KGs, multi-lingual corpora or other information (Ji et al. (2017); Lei et al. (2018); Vashishth et al. (2018); Han et al. (2018b); Zhang et al. (2019); Qu et al. (2019); Verga et al. (2016); Lin et al. (2017); Wang et al. (2018); Deng and Sun (2019); Beltagy et al. (2019)). Other approaches include soft-label strategy for denoising (Liu et al. (2017)), leveraging pre-trained LM (Alt et al. (2019)), pattern-based method (Zheng et al. (2019)), structured learning method (Bai and Ritter (2019)) and so forth (Luo et al. (2017); Chen et al. (2019)).

In this work, we focus on **sentence-level** relation extraction. Several previous studies also perform Distant RE on sentence-level. Feng et al. (2018) proposes a reinforcement learning framework for sentence selecting, where the reward is given by the classification scores on bag labels. Jia et al. (2019) builds an initial training set and further select confident instances based on selected patterns. The difference between the proposed work and previous works is that we do not rely on bag-level labels for sentence selecting. Furthermore, we leverage NT to dynamically separate the noisy data from

[1] https://github.com/rtmaww/SENT

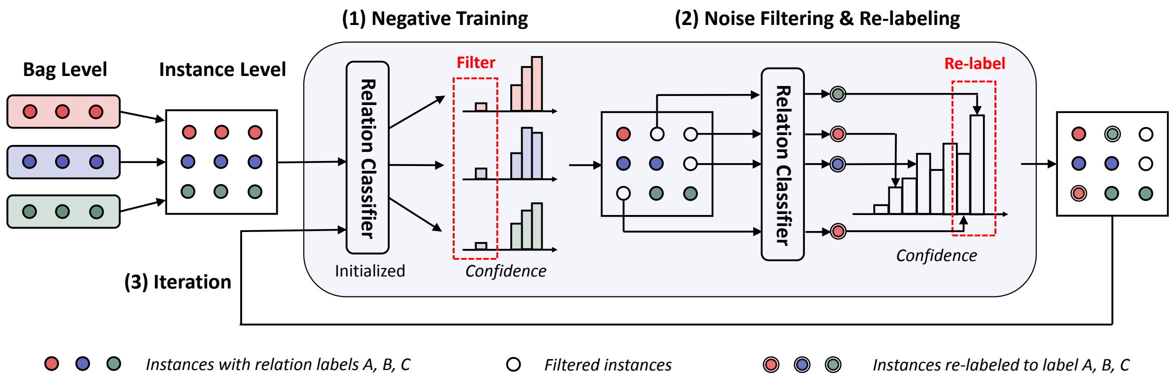


Figure 2: An overview of the proposed framework, SENT, for sentence-level distant RE. Three steps are included: (1) Negative training for separating the noisy data from the training data; (2) Noise-filtering and re-labeling; (3) Iterative training to further boost the performance.

the training data, thus can make use of diversified clean data.

2.2 Learning with Noisy Data

Learning with noisy data is a widely discussed problem in deep learning, especially in the field of computer vision. Existing approaches include robust learning methods such as leveraging a robust loss function or regularization method(Lyu and Tsang, 2020; Zhang and Sabuncu, 2018; Hu et al., 2019b; Kim et al., 2019), re-weighting the loss of potential noisy samples (Ren et al., 2018; Jiang et al., 2018), modeling the corruption probability with a transition matrix (Goldberger and Ben-Reuven, 2016; Xia et al.) and so on. Another line of research tries to recognize or even correct the noisy instances from the training data(Malach and Shalev-Shwartz, 2017; Yu et al., 2019; Arazo et al., 2019; Li et al., 2019).

In this paper, we focus on the noisy label problem in distant RE. We first leverage a robust negative loss (Kim et al., 2019) for model training. Then, we develop a new iterative training algorithm for noise selection and correction.

3 Methodology

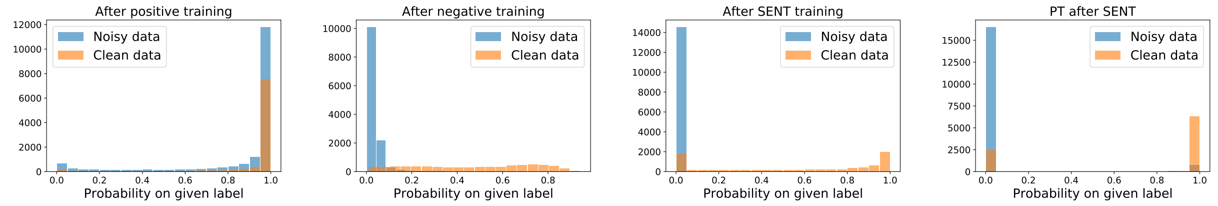
In order to achieve sentence-level relation classification using bag-level labels in distant RE, we propose a framework, SENT, which contains three main steps (as shown in Fig.2): (1) Separating the noisy data from the training data with negative training (Sec.3.1); (2) Filtering the noisy data as well as re-labeling a part of confident instances (Sec.3.2); (3) Leveraging an effective training algorithm based on (1) and (2) to further boost the performance (Sec.3.3).

Specifically, we denote the input data in this task as $\mathbf{S}^* = \{(s_1, y_1^*), \dots, (s_N, y_N^*)\}$, where $y_i^* \in \mathbb{R} = \{1, \dots, C\}$ is the bag-level label of the i^{th} input sentence s_i. Obviously, this is a noisy dataset drawn from a noisy distribution $\mathbf{D}^*$ because these bag-level labels y^* come from the distant label of each entity bag. For each s_i containing a pair of entities $< e_1, e_2 >$, y_i^* is one of the relation facts[2] that $< e_1, e_2 >$ participates in in the database. Such annotation method indicates that y_i^* is a potential noisy label for s_i. Here, we denote $\mathbf{D}$ as the real data distribution without noise, and the clean dataset drawn from $\mathbf{D}$ as $\mathbf{S} = \{(s_1, y_1), \dots, (s_N, y_N)\}$. The ambition of this work is to find the best estimated parameters θ of the real mapping $f : x \rightarrow y, (x, y) \in \mathbf{D}$ based on the noisy data $\mathbf{S}^*$. We design three steps for achieving this goal: (1) Recognizing the set of noisy data $\mathbf{S}_n^*$ from $\mathbf{S}^*$ using negative training, where $\mathbf{S}_n^* = \{(s_i, y_i^*) \mid y_i^* \neq y_i\}$. (2) Refining $\mathbf{S}^*$ by noise-filtering and re-labeling, e.g., $\mathbf{S}_{refined}^* = (\mathbf{S}^* \setminus \mathbf{S}_n^*) \cup \mathbf{S}_{n,relabeled}^*$, where $\mathbf{S}_{n,relabeled}^* = \{(s_i, y_i) \mid (s_i, y_i^*) \in \mathbf{S}_n^*\}$. (3) Iteratively perform (1) and (2) so the refined dataset $\mathbf{S}_{refined}^*$ approaches the real dataset $\mathbf{S}$.

3.1 Negative Training on Distant Data

In order to perform robust training on the noisy distant data, we propose the use of negative Training (NT), which trains based on the concept that "the input sentence does not belong to this complementary label". We find that NT not only

[2] Here, we randomly choose one of the multiple bag labels for injective relation classification. See details in Sec.4.2.



(a) Positive training (b) Negative training (c) After SENT (d) PT after SENT

Figure 3: Data distribution when training with PT and SENT. (a) During PT, the confidence of the clean and noisy data increase simultaneously; (b) During NT, the confidence of the noisy data is much lower than that of the clean data; (c) After training with the SENT method, the clean and noisy data are further separated; (d) PT after SENT helps improve the convergence of the clean data.

provides less noisy information, but also separates the noisy and clean data during training.

3.1.1 Positive Training

Positive training (PT) trains the model towards predicting the given label, based on the concept that "the input sentence belongs to this label". Here, given any input s with a label $y^* \in \mathbb{R} = \{1, 2, \ldots, C\}$, $\mathbf{y} \in \{0, 1\}^C$ is the C-dimension one-hot vector of y^*. We denote $\mathbf{p} = f(s)$ as the probability vector of a sentence given by a relation classifier $f(\cdot)$. With the cross entropy loss function, the loss defined in typical positive training is:

$$\mathcal{L}_{PT}(f, y^*) = -\sum_{k=1}^{C} y_k \log p_k \tag{1}$$

where p_k denotes the probability of the k^{th} label. Optimizing on Eq.1 meets the requirement of PL, as the probability of the given label approaches 1 with the loss decreasing.

3.1.2 Negative Training

In negative training (NT), for each input s with a label $y^* \in \mathbb{R}$, we generate a complementary label $\overline{y^*}$ by randomly sampling from the label space except y^*, e.g., $\overline{y^*} \in \mathbb{R} \backslash \{y^*\}$. With the cross entropy loss function, we define the loss in negative training as:

$$\mathcal{L}_{NT}(f, y^*) = -\sum_{k=1}^{C} \overline{y_k} \log(1 - p_k) \tag{2}$$

Different from PT, Eq.2 aims to reduce the probability value of the complementary label, as $p_k \rightarrow 0$ with the loss decreasing.

To further illustrate the effect of NT, we train the classifier with PT and NT respectively on a constructed TACRED dataset with 30% noise (details shown in Sec.4.1). A histogram[3] of the training data after PT and NT is shown in Figs. 3(a),(b), which reveals that, when training with PT, the confidence of clean data and noisy data increase with no difference, resulting in the model to overfit noisy training data. On the contrary, when training with NT, the confidence of noisy data is much lower than that of clean data. This result confirms that the model trained with NT suffers less from overfitting noisy data with less noisy information provided. Moreover, as the confidence value of clean data and noisy data separate from each other, we are able to filter noisy data with a certain threshold. Fig.4 shows the details of the data-filtering effect. After the first iteration of NT, a modest threshold contributes to 97% precision noise-filtering with about 50% recall, which further verifies the effectiveness of NT on noisy data training.

3.2 Noise Filtering and Re-labeling

In Section 3.1, we have illustrated the effectiveness of NT on training with noisy data, as well as the capability to recognize noisy instances. While filtering noisy data is important for training on distant data, these filtered data contain useful information that can boost performance if properly re-labeled. In this section, we describe the proposed noise-filtering and label-recovering strategy for refining distant data based on NT.

3.2.1 Filtering Noisy Data

As discussed before, it is intuitive to construct a filtering strategy based on a certain threshold after NT. However, in distant RE, the long-tail problem cannot be neglected. During training, the

[3] When drawing the histogram, we omitted the large amount of "NA"-class data (80% of the training data) for a clearer representation of the positive-class data.

degree of convergence is disparate among different classes. Simply setting a uniform threshold might harm the data distribution with instances of long-tail relations largely filtered out. Therefore, we leverage a dynamic threshold for filtering noisy data. Suppose the probability of class c of the i^{th} instance is $p_c^i \in (0, p_c^h)$, where p_c^h is the maximum probability value in class c. Based on empirical experience, we assume the probability values follow a distribution where the noisy data are largely distributed in low-value areas and the clean data are generally distributed in middle- or high-value areas. Therefore, the filtering threshold of class c is set to:

$$Th_c = Th \cdot p_c^h, p_c^h = \max_{i=1}^{N}\{p_c^i\} \tag{3}$$

where Th is a global threshold. In this way, the noise-filtering threshold not only relies on the degree of convergence in each class, but also dynamically changes during the training phase, thus making it more suitable for noise-filtering on long-tail data.

3.2.2 Re-labeling Useful Data

After noise-filtering, the noisy instances are regarded as unlabeled data, which also contain useful information for training. Here, we design a simple strategy for re-labeling these unlabeled data. Given the set of filtered data $D_u = \{s_1, \ldots, s_m\}$, we use the classifier trained in this iteration to predict the probability vectors $\{\mathbf{p}^1, \ldots, \mathbf{p}^m\}$. Then, we re-label these instances by:

$$\hat{y}_i = \arg\max_k\{p_k^i\}, if \max_k\{p_k^i\} > Th_{relabel} \tag{4}$$

where p_k^i is the probability of the i^{th} instance in class k, and $Th_{relabel}$ is the re-label threshold.

3.3 Iterative Training Algorithm

Although effective, simply performing a pipeline of NT, noise-filtering and re-labeling fail to take full advantage of each part, thus the model performance can be further boosted through iterative training.

As shown in Fig.2, for each iteration, we first train the classifier on the noisy data using NT: for each instance, we randomly sample K complementary labels and calculate the loss on these labels with Eq.(2). After M-epochs negative training, the noise-filtering and re-labeling processes are carried out for updating the training data. Next, we perform a new iteration of training on the newly-refined data. Here, we re-initialize the classifier in every iteration for two reasons: First, re-initialization ensures that in each iteration, the new classifier is trained on a dataset with higher quality. Second, re-initialization introduces randomness, thus contributing to more robust data-filtering. Finally, we stop the iteration after observing the best result on the dev set. We then perform a round of noise-filtering and re-labeling with the best model in the last iteration to obtain the final refined data.

Fig.3(c) shows the data distribution after certain iterations of SENT. As seen, the noise and clean data are separated by a large margin. Most noisy data are successfully filtered out, with an acceptable number of clean data mistaken. However, we can see that the model trained with NT still lacks convergence (with low-confidence predictions). Therefore, we train the classifier on the iteratively-refined data with PT for better convergence. As shown in Fig.3(d), the model predictions on most of the clean data are in high confidence after PT training.

4 Experiments

The experiments in this work are divided into two parts, respectively conducted on two datasets: the NYT-10 dataset (Riedel et al., 2010) and the TACRED dataset (Zhang et al., 2017).

The first part is the effectiveness study on sentence-level evaluation for distant RE. Different from bag-level evaluation, a sentence-level evaluation compute Precision (Prec.), Recall (Rec.) and F1 metric directly on all of the individual instances in the dataset. In this part, we adopt the NYT-10 data set for sentence-level training, following the setting of Jia et al. (2019), who publishes a manually labeled sentence-level test set. [4] Besides, they also publish a test set for evaluating noise-filtering ability. Details of the adopted dataset are shown in Table 1.

We construct the second part of experiments (Sec.4.4) to better understand SENT's behaviors. Since no labeled training data are available in the distant supervision setting, we construct a noisy dataset with 30% noise from a labeled dataset, TACRED (Zhang et al., 2017) [5]. We regard this constructed dataset as noisy-TACRED. The reason

[4] https://github.com/PaddlePaddle/Research/tree/master/NLP/ACL2019-ARNOR

[5] https://github.com/yuhaozhang/tacred-relation

Datasets		**NYT-10**	**noisy-TACRED**
#Label num.		24	41
Train	#Instances	371461	68124
	#Positive	110518	26575
	#Noise	Unknown	20586
Dev	#Instances	2379	22631
	#Positive	337	5436
Test	#Instances	2164	15509
	#Positive	323	3325

Table 1: Statistics of datasets[6]. "Positive" means positive instances that are not labeled as "NA". Note that the positive instances of noisy-TACRED include false-positive noise and the noise number in NYT-10 is unknown due to the inaccurate annotations.

we choose this dataset is that 80% instances in the training data are "no_relation". This "NA" rate is similar to the NYT data which contains 70% "NA" relation type, thus analysis on this dataset is more credible.

When constructing noisy-TACRED, the noisy instances are uniformly selected with 30% noise ratio. Then, each noisy label is created by sampling a label from a complementary class with a weight of class frequency (in order to maintain the data distribution). Note that the original dataset consists of 80% "no_relation" data, which means 80% of the noisy instances are "false-positive" instances, corresponding to the large amount of "false-positive" noise in NYT-10. Details of the noisy-TACRED are also shown in Table 1.

4.1 Baselines

We compare our SENT method with several strong baselines in distant RE. These compared methods can be categorized as: bag-level denoising methods, sentence-level denoising methods, sentence-level non-denoising methods.

PCNN+SelATT (Lin et al., 2016): A *bag-level* RE model which leverages an attention mechanism to reduce noise effect.

PCNN+RA_BAG_ATT (Ye and Ling, 2019) short for PCNN+ATT_RA+BAG_ATT, a *bag-level* model containing both intra-bag and inter-bag attentions to alleviate noise.

CNN+RL$_1$ (Qin et al., 2018): A RL-based *bag-level* method. Different from **CNN+RL$_2$**, they redistribute the filtered data into the negative examples.

CNN+RL$_2$ (Feng et al., 2018): A *sentence-level* RE model. It jointly train a instance selector and a CNN classifier using reinforcement learning (RL).

ARNOR (Jia et al., 2019): A *sentence-level* RE model which selects confident instances based on the attention score on the selected patterns. It is the state-of-the-art method in sentence level.

CNN (Zeng et al., 2014), **PCNN** (Zeng et al., 2015) and **BiLSTM** (Zhang et al., 2015) are typical architectures used in RE.

BiLSTM+ATT (Zhang et al., 2017) leverages an attention mechanism based on BiLSTM to capture useful information.

BiLSTM+BERT (Devlin et al., 2019): Based on BiLSTM, it utilizes the pre-trained BERT representations as word embedding.

4.2 Implementation Details

As SENT is a model-agnostic framework, we implement the classification model with two typical architectures: BiLSTM and BiLSTM+BERT. Since BiLSTM is also the base model of ARNOR, we can compare these two methods more fairly. During SENT training, we use the 50-dimension glove vectors as word embedding. While for PT after SENT, we randomly initialize the 50-dimension word embedding as the same in ARNOR. In both training phases, we use 50-dimension randomly-initialized position and entity type embedding. We train a single-layer BiLSTM with hidden size 256 using the adam optimizer at a learning rate of 5e-4. When implemented with BiLSTM+BERT, the setting is the same as those with BiLSTM except that we use a 768-dimension fixed BERT representation as word embedding (we use the "bert-base-uncased" pre-trained model). We tune the hyperparameters on the development set via a grid search. Specifically, when training on the NYT dataset, we train the model for 10 epochs in each iteration, with the global data-filtering threshold $Th = 0.25$, the re-labeling threshold $Th_{relabel} = 0.7$ and negative samples number $K = 10$. When training on the noisy-TACRED, we train for 50 epochs in each iteration, with $Th = 0.15$, $Th_{relabel} = 0.85$ and $K = 50$.

To deal with the multi-label problem, we utilize a simple method by randomly selecting one of the bag labels for each sentence. Such random selection turns the multi-label noise into the wrong-label noise, which is easier to handle. According to Surdeanu et al. (2012), there are 31% wrong-label noise and 7.5% multi-label noise in NYT-10, and incorrect selection may result in 4% extra wrong-

[6] Statistics of NYT-10 are quoted from (Jia et al., 2019).

Method	Dev			Test		
	Prec.	Rec.	F1	Prec.	Rec.	F1
CNN(Zeng et al., 2014)	38.32	65.22	48.28	35.75	64.54	46.01
PCNN(Zeng et al., 2015)	36.09	63.66	46.07	36.06	64.86	46.35
BiLSTM(Zhang et al., 2015)	36.71	66.46	47.29	35.52	67.41	46.53
BiLSTM+ATT(Zhang et al., 2017)	37.59	64.91	47.61	34.93	65.18	45.48
BERT(Devlin et al., 2019)	34.78	65.17	45.35	36.19	70.44	47.81
BiLSTM+BERT(Devlin et al., 2019)	36.09	73.17	48.34	33.23	72.70	45.61
PCNN+SelATT(Lin et al., 2016)	46.01	30.43	36.64	45.41	30.03	36.15
PCNN+RA_BAG_ATT(Ye and Ling, 2019)	49.84	46.90	48.33	56.76	50.60	53.50
CNN+RL$_1$ (Qin et al., 2018)	37.71	52.66	43.95	39.41	61.61	48.07
CNN+RL$_2$ (Feng et al., 2018)	40.00	59.17	47.73	40.23	63.78	49.34
ARNOR(Jia et al., 2019)	62.45	58.51	60.36	65.23	56.79	60.90
SENT (BiLSTM)	66.71$_{\pm 0.30}$	57.27$_{\pm 0.30}$	61.63$_{\pm 0.29}$	71.22$_{\pm 0.58}$	59.75$_{\pm 0.62}$	64.99$_{\pm 0.34}$
SENT (BiLSTM+BERT)	**69.94**$_{\pm 0.51}$	**63.11**$_{\pm 0.61}$	**66.35**$_{\pm 0.11}$	**76.34**$_{\pm 0.56}$	**63.66**$_{\pm 0.17}$	**69.42**$_{\pm 0.13}$

Table 2: Main results on sentence-level evaluation. Compared baselines include normal RE model (the first part of the table) and models for distant RE (the second part of the table). We ran the model three times to get the average results.

label noise, which can be filtered out through NT identically with wrong-label instances.

4.3 Sentence-Level Evaluation

Table 2 shows the results of SENT and other baselines on sentence-level evaluation, where the results of SENT are obtained by PT after SENT. We can observe that: 1) Bag-level methods fail to perform well on sentence-level evaluation, indicating that it is difficult for these bag-level approaches to benefit downstream tasks with exact sentence labels. This result is consistent with the results in Feng et al. (2018). 2) When performing sentence-level training on the noisy distant data, all baseline models show poor results, including the preeminent pre-trained language model BERT. These results indicate the negative impact of directly using bag-level labels for sentence-level training regardless of noise. 3) The proposed SENT method achieves a significant improvement over previous sentence-level de-noising methods. When implemented with BiLSTM, the model obtains a 4.09% higher F1 score than ARNOR. Moreover, when implemented with BiLSTM+BERT, the F1 score is further improved by 8.52%. 4) The SENT method achieves much higher precision than the previous de-noising methods when maintaining comparable or higher recall, indicating the effectiveness of the noise-filtering and re-labeling approaches.

Noise Reduction	Prec.	Rec.	F1
CNN+RL$_2$	40.58	**96.31**	57.10
ARNOR	76.37	68.13	72.02
SENT (BiLSTM)	80.00	88.46	84.02
SENT (BiLSTM+BERT)	**84.33**	85.67	**84.99**

Table 3: The noise-filtering effect evaluated on a noise-annotated test set of NYT-10.

4.3.1 Noise-Filtering Effect on Distant Data

In order to prove the effectiveness of SENT in de-noising distant data, we conduct a noise-filtering experiment following ARNOR. We use a test set published by ARNOR, which consists of 200 randomly selected sentences with an "is_noise" annotation. We perform a noise-filtering process as described in Sec.3.2.1, and calculate the de-noise accuracy. As seen in Table 3, the SENT method achieves remarkable improvement over ARNOR in F1 score by 12%. While improving in precision, SENT achieves 20% gain over ARNOR in recall. As ARNOR initializes the training data with a small part of frequent patterns, these patterns might limit the model from generalizing to various correct data. Different from ARNOR, SENT leverages negative training to automatically learn the correct patterns, showing better ability in diversity and generalization.

	Method	Prec.	Rec.	F1
Clean Data	BiLSTM+ATT	67.7	63.2	65.4
	BiLSTM	61.4	61.7	61.5
Noisy Data	BiLSTM+ATT	32.8	43.8	37.5
	BiLSTM	37.8	45.5	41.3
	SENT (BiLSTM)	**66.0**	**52.9**	**58.7**

Table 4: Model performance on clean and noisy-TACRED. When trained on noisy data, the performance of base models degrades dramatically while SENT achieves comparable results with the models trained on clean data.

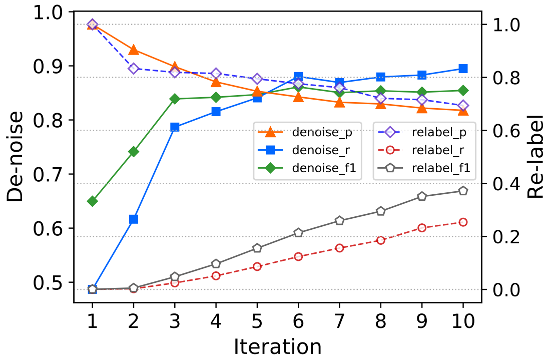


Figure 4: Data-refining details on noisy-TACRED.

4.4 Analysing SENT on "Labeled Noise"

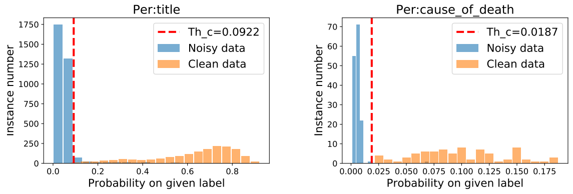
In this section, we analyze the effectiveness of the data-refining process with a self-constructed noisy data set: noisy-TACRED (details in Table 1).

4.4.1 Performance on Noisy-TACRED

Table 4 shows the results of training on TACRED and noisy-TACRED. As seen, the baseline model degrades dramatically on the noisy data, with the LSTM dropping by 20.2%. However, after training with SENT, the BiLSTM model can achieve comparable results with the model that trained on the clean data. Note that the de-noising method is quite helpful in promoting the precision score, yet the recall is still lower than that on clean data.

4.4.2 Effects of Data-Refining

We also evaluate the noise-filtering and label-recovering ability on the noisy-TACRED training set, as shown in Fig.4. We can observe that: 1) SENT achieves about 85% F1 score on the noisy-TACRED data. This result is consistent with the noise-filtering results obtained on the NYT dataset (with 200 sampled instances), validating the de-noising ability of SENT on different datasets. 2) As the training iteration progressed, the precision of noise-filtering decreases with the recall promoting. More noise-filtering contributes to a cleaner dataset, while it might bring more false-noise mistakes. Therefore, we stop the iteration when the model reaches the best score on the development set. 3) As for label-recovering, SENT can achieve about 70% precision with about 25% recall. Here, the threshold setting is also a trade-off that we prefer to adopt a modest value for more accurate re-labeling.



(a) Head relation.

(b) Long-tail relation.

Figure 5: Data distribution of a head relation (per:title) and a long-tail relation (per:cause_of_death) during NT. The dynamically designed thresholds benefits filtering.

4.4.3 Effects of Dynamically Filtering

As described in Sec.3.2, we design a dynamic filtering threshold for long-tail data. The effect of this strategy is shown in Fig.5. As seen, the degree of convergence of the long-tail relation "per:cause_of_death" is much lower than that of the head relation. Simply setting a uniform threshold would harm the data distribution with instances of "per:cause_of_death" largely filtered. While with a dynamically determined threshold, both data from the head and the long-tail relations are appropriately filtered.

4.5 Ablation Study

To better illustrate the contribution of each component in SENT, we conduct an ablation study by removing the following components: final PT, re-labeling, dynamic threshold, re-initialization, NT. The test results are shown in Table 6. We can observe that: 1) Removing the final positive training affects little to the performance. This is because the model trained with NT already reaches high accuracy and the purpose of final PT is only to achieve more confidential predictions. 2) Removing the re-labeling process harms the performance, as the filtered instances are simply discarded regardless of the useful information for

Sentence	Bag label	Sentence label	Refined label
The plan filed on behalf of the state 's Democratic Congressional delegation , for instance , would make the 25th district , which zigzags 300 miles from southern Austin to Mexico , much shorter and Austin-based , which would help the incumbent Democrat , Lloyd Doggett .	place_lived place_of_birth	place_lived	NA
It would draw the lines in a way that imperils an incumbent Democrat , Representative Lloyd Doggett of Austin , and divides that most liberal of Texas cities and surrounding Travis County among three districts , all solidly Republican . "		place_of_birth	place_lived
A leather-and-metal chair bore a shameless resemblance to a Barcelona Chair by Ludwig Mies van der Rohe -LRB- who lived in Chicago -RRB- .	place_of_death	place_of_death	NA
The works of architects like Frank Lloyd Wright , Louis H. Sullivan , Ludwig Mies van der Rohe and Helmut Jahn define Chicago in many ways .		place_of_death	NA
Mr. Freed received a bachelor 's degree in architecture in 1953 from the Illinois Institute of Technology in Chicago , which was then under the direction of Ludwig Ludwig Mies van der Rohe .		place_of_death	NA
It 's really tough right now , " said Norman J. Ornstein , a resident scholar at the conservative American Enterprise Institute and a member of the PBS board . "	NA	NA	company
Three of the sailors were assigned to SEAL Delivery Team 1 , Pearl Harbor , Hawaii .	NA	NA	contains
An obituary on Wednesday about Philip Merrill , a Maryland publisher , misstated a journalism post he held as an undergraduate .	NA	NA	place_lived

Table 5: Examples showing the ability of SENT to refine the bag-level noisy data into correct data. Texts in red and blue denote the head and tail entity, respectively.

Components	Prec.	Rec.	F1
SENT (BiLSTM)	71.22	**59.75**	**64.99**
– Final PT	**72.48**	57.89	64.37
– Re-labeling	66.67	55.11	60.34
– Dynamic threshold	58.46	49.23	53.45
– Re-initialization	48.61	65.02	55.63
– NT	41.58	70.28	52.24

Table 6: An ablation study on NYT-10.

training. 3) Without dynamic threshold, clean instances from the tail classes are incorrectly filtered out, which severely degrades the performance. 4) Re-initialization also contributes a lot to the performance. The model trained on the original noisy data inevitably fits to the noisy distribution, while re-initialization helps wash out the overfitted parameters and eliminate the noise effects, thus contributing to better training and noise-filtering. 5) Training with PT instead of NT causes a dramatic decline in performance, especially on the precision, which verifies the effectiveness of NT to prevent the model from overfitting noisy data.

4.6 Case Study

As discussed, SENT is able to refine the distant RE dataset. In fact, there exists much noise in the NYT data that is difficult to tackle with bag-level methods. In Table 5, we show some examples. (1) The first two rows are the sentences in a multi-label bag. We randomly choose one of the bag labels for each sentence, and the model is able to correct the bad choice (by correcting the second sentence with "place_lived" and the first sentence with "NA"). (2) The following three rows show a bag with label "place_of_death", while this whole bag is actually a "NA" bag incorrectly labeled positive. (3) SENT can also recognize the positive samples in "NA". As shown in the last three rows, each sentence labeled as "NA" is actually expressing a positive label. In fact, such false-negative problem is frequently seen in the NYT data, which contains 70% negative instances that were labeled "NA" only because the entity pairs do not participate in a relation in the database. We believe the capacity to recognize these false-negative samples can significantly boost the performance.

5 Conclusion

In this paper, we present SENT, a novel sentence-level framework based on Negative Training (NT) for sentence-level training on distant RE data. NT not only prevent the model from overfitting noisy data, but also separate the noisy data from the training data. By iteratively performing noise-filtering and re-labeling based on NT, SENT helps re-fine the noisy distant data and achieves remarkable performance. Experimental results verify the improvement of SENT over previous methods on sentence-level relation extraction and noise-filtering effect.

Acknowledgements

The authors wish to thank the anonymous reviewers for their helpful comments. This work was partially funded by China National Key R&D Program (No. 2018YFB1005100), National Natural Science Foundation of China (No. 61976056, 62076069), Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0103).

References

Christoph Alt, Marc Hübner, and Leonhard Hennig. 2019. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1388–1398, Florence, Italy. Association for Computational Linguistics.

Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. 2019. Unsupervised label noise modeling and loss correction. In *International Conference on Machine Learning*, pages 312–321. PMLR.

Fan Bai and Alan Ritter. 2019. Structured Minimally Supervised Learning for Neural Relation Extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3057–3069, Minneapolis, Minnesota. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Waleed Ammar. 2019. Combining distant and direct supervision for neural relation extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1858–1867, Minneapolis, Minnesota. Association for Computational Linguistics.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 26:2787–2795.

Junfan Chen, Richong Zhang, Yongyi Mao, Hongyu Guo, and Jie Xu. 2019. Uncover the ground-truth relations in distant supervision: A neural expectation-maximization framework. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 326–336, Hong Kong, China. Association for Computational Linguistics.

Xiang Deng and Huan Sun. 2019. Leveraging 2-hop distant supervision from table entity pairs for relation extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 410–420, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2014. Open question answering over curated and extracted knowledge bases. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1156–1165.

Jun Feng, Minlie Huang, Li Zhao, Yang Yang, and Xiaoyan Zhu. 2018. Reinforcement learning for relation classification from noisy data. In *Proceedings of the aaai conference on artificial intelligence*.

Jacob Goldberger and Ehud Ben-Reuven. 2016. Training deep neural-networks using a noise adaptation layer.

Xu Han, Zhiyuan Liu, and Maosong Sun. 2018a. Denoising distant supervision for relation extraction via instance-level adversarial training. *arXiv preprint arXiv:1805.10959*.

Xu Han, Zhiyuan Liu, and Maosong Sun. 2018b. Neural knowledge acquisition via mutual attention between knowledge graph and text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Xu Han, Pengfei Yu, Zhiyuan Liu, Maosong Sun, and Peng Li. 2018c. Hierarchical relation extraction with coarse-to-fine grained attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 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: Human Language Technologies*, pages 541–550, Portland, Oregon, USA. Association for Computational Linguistics.

Linmei Hu, Luhao Zhang, Chuan Shi, Liqiang Nie, Weili Guan, and Cheng Yang. 2019a. Improving distantly-supervised relation extraction with joint label embedding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3821–3829, Hong Kong, China. Association for Computational Linguistics.

Wei Hu, Zhiyuan Li, and Dingli Yu. 2019b. Simple and effective regularization methods for training on noisily labeled data with generalization guarantee. In *International Conference on Learning Representations*.

Guoliang Ji, Kang Liu, Shizhu He, Jun Zhao, et al. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *AAAI*, volume 3060.

Wei Jia, Dai Dai, Xinyan Xiao, and Hua Wu. 2019. ARNOR: Attention regularization based noise reduction for distant supervision relation classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1399–1408, Florence, Italy. Association for Computational Linguistics.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. 2018. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2304–2313. PMLR.

Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. 2019. Nlnl: Negative learning for noisy labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 101–110.

Kai Lei, Daoyuan Chen, Yaliang Li, Nan Du, Min Yang, Wei Fan, and Ying Shen. 2018. Cooperative denoising for distantly supervised relation extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 426–436, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Junnan Li, Richard Socher, and Steven CH Hoi. 2019. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*.

Yang Li, Guodong Long, Tao Shen, Tianyi Zhou, Lina Yao, Huan Huo, and Jing Jiang. 2020. Self-attention enhanced selective gate with entity-aware embedding for distantly supervised relation extraction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8269–8276.

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 (Volume 1: Long Papers)*, pages 34–43, Vancouver, Canada. Association for Computational Linguistics.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2124–2133.

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

Bingfeng Luo, Yansong Feng, Zheng Wang, Zhanxing Zhu, Songfang Huang, Rui Yan, and Dongyan Zhao. 2017. Learning with noise: Enhance distantly supervised relation extraction with dynamic transition matrix. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 430–439, Vancouver, Canada. Association for Computational Linguistics.

Yueming Lyu and Ivor W. Tsang. 2020. Curriculum loss: Robust learning and generalization against label corruption. In *International Conference on Learning Representations*.

Eran Malach and Shai Shalev-Shwartz. 2017. Decoupling" when to update" from" how to update". In *NIPS*.

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 ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018. Robust distant supervision relation extraction via deep reinforcement learning. *arXiv preprint arXiv:1805.09927*.

Jianfeng Qu, Wen Hua, Dantong Ouyang, Xiaofang Zhou, and Ximing Li. 2019. A fine-grained and noise-aware method for neural relation extraction. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 659–668.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. 2018. Learning to reweight examples for robust deep learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4334–4343. PMLR.

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*, pages 148–163. Springer.

Yuming Shang. 2019. Are noisy sentences useless for distant supervised relation extraction?

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*, pages 455–465, Jeju Island, Korea. Association for Computational Linguistics.

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, pages 1257–1266, Brussels, Belgium. Association for Computational Linguistics.

Patrick Verga, David Belanger, Emma Strubell, Benjamin Roth, and Andrew McCallum. 2016. Multilingual relation extraction using compositional universal schema. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 886–896, San Diego, California. Association for Computational Linguistics.

Xiaozhi Wang, Xu Han, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2018. Adversarial multi-lingual neural relation extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1156–1166, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28.

Fei Wu and Daniel S. Weld. 2010. Open information extraction using Wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 118–127, Uppsala, Sweden. Association for Computational Linguistics.

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*, pages 1778–1783, Copenhagen, Denmark. Association for Computational Linguistics.

Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning?

Kun Xu, Siva Reddy, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2016. Question answering on Freebase via relation extraction and textual evidence. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2326–2336, Berlin, Germany. Association for Computational Linguistics.

Xuchen Yao and Benjamin Van Durme. 2014. Information extraction over structured data: Question answering with Freebase. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 956–966, Baltimore, Maryland. Association for Computational Linguistics.

Zhi-Xiu Ye and Zhen-Hua Ling. 2019. Distant supervision relation extraction with intra-bag and inter-bag attentions. *arXiv preprint arXiv:1904.00143*.

Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. 2019. How does disagreement help generalization against label corruption? In *International Conference on Machine Learning*, pages 7164–7173. PMLR.

Changsen Yuan, Heyan Huang, Chong Feng, Xiao Liu, and Xiaochi Wei. 2019a. Distant supervision for relation extraction with linear attenuation simulation and non-iid relevance embedding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7418–7425.

Yujin Yuan, Liyuan Liu, Siliang Tang, Zhongfei Zhang, Yueting Zhuang, Shiliang Pu, Fei Wu, and Xiang Ren. 2019b. Cross-relation cross-bag attention for distantly-supervised relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 419–426.

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*, pages 1753–1762.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Ningyu Zhang, Shumin Deng, Zhanlin Sun, Guanying Wang, Xi Chen, Wei Zhang, and Huajun Chen. 2019. Long-tail relation extraction via knowledge graph embeddings and graph convolution networks. *arXiv preprint arXiv:1903.01306*.

Shu Zhang, Dequan Zheng, Xinchen Hu, and Ming Yang. 2015. Bidirectional long short-term memory networks for relation classification. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 73–78, Shanghai, China.

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

Zhilu Zhang and Mert Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Shun Zheng, Xu Han, Yankai Lin, Peilin Yu, Lu Chen, Ling Huang, Zhiyuan Liu, and Wei Xu. 2019.

DIAG-NRE: A neural pattern diagnosis framework for distantly supervised neural relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1419–1429, Florence, Italy. Association for Computational Linguistics.

Zhangdong Zhu, Jindian Su, and Yang Zhou. 2019. Improving distantly supervised relation classification with attention and semantic weight. *IEEE Access*, 7:91160–91168.