Article

Improving Distant Supervised Relation Extraction with Noise Detection Strategy

Xiaoyan Meng ^{1,2,3,4}, Tonghai Jiang ^{1,3,*}, Xi Zhou ^{1,2,3}, Bo Ma ^{1,2,3}, Yi Wang ^{1,2,3} and Fan Zhao ^{1,2,3}

- ¹ The Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences, Urumqi 830011, China; mxy@xjau.edu.cn (X.M.); zhoxi@ms.xjb.ac.cn (X.Z.); mabo@ms.xjb.ac.cn (B.M.); wangyi@ms.xjb.ac.cn (Y.W.); zhaofan@ms.xjb.ac.cn (F.Z.)
- ² University of Chinese Academy of Sciences, Beijing 100049, China
- ³ Xinjiang Laboratory of Minority Speech and Language Information Processing, Chinese Academy of Sciences, Urumqi 830011, China
- ⁴ College of Computer and Information Engineering, Xinjiang Agricultural University, Urumqi 830052, China
- * Correspondence: jth@ms.xjb.ac.cn; Tel.: +86-099-1383-7795

Abstract: Distant supervised relation extraction (DSRE) is widely used to extract novel relational facts from plain text, so as to improve the knowledge graph. However, distant supervision inevitably suffers from the noisy labeling problem that will severely damage the performance of relation extraction. Currently, most DSRE methods are mainly focused on reducing the weights of noisy sentences, ignoring the bag-level noise where all sentences in a bag are wrongly labeled. In this paper, we present a novel noise detection-based relation extraction approach (NDRE) to automatically detect noisy labels with entity information and dynamically correct them, which can alleviate both instance-level and bag-level noisy problems. By this means, we can extend the dataset from the Web tables without introducing more noise. In this approach, to embed the semantics of sentences from corpus and web tables, we firstly propose a powerful sentence coder that employs an internal multi-head self-attention mechanism between the piecewise max-pooling convolutional neural network. Second, we adopt a noise detection strategy, which is expected to dynamically detect and correct the original noisy label according to the similarity between sentence representation and entity-aware embeddings. Then, we aggregate the information from corpus and web tables to make the final relation prediction. Experimental results on a public benchmark dataset demonstrate that our proposed approach achieves significant improvements over the state-of-the-art baselines and can effectively reduce the noisy labeling problem.

Citation: Meng, X.; Jiang, T.; Zhou, X.; Ma, B.; Wang, Y.; Zhao, F. Improving Distant Supervised Relation Extraction with Noise Detection Strategy. *Appl. Sci.* **2021**, *11*, 2046. <https://doi.org/10.3390/app11052046>

Academic Editor: Keun Ho Ryu

Received: 3 February 2021

Accepted: 23 February 2021

Published: 25 February 2021

Keywords: relation extraction; distant supervision; noise detection

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Knowledge graphs (KGs) play a crucial role in natural language processing (NLP). KGs such as Freebase [1] and DBpedia [2] have shown their strong knowledge organization capability and are used as data resources in many NLP tasks including semantic search, intelligent question answering and text generation, among others. These KGs are mostly composed of relational facts, which are in the form of triplets such as <Warren Buffett, born_in, Omaha>. However, as knowledge is constantly increasing and updating, the existing KGs are far from complete. To fill this gap, relation extraction (RE), which aims to identify the relation r between a given pair of entities < e_1 , e_2 > from unstructured text, is thus an essential task in NLP.

As manually labeled data is insufficient for traditional supervised RE systems, distant supervision (DS) [3] is proposed to automatically construct large-scale labeled training data by aligning entities in text corpus and corresponding KGs. The assumption of DS is that if there is a relational fact < e_1 , r , e_2 > in the knowledge graph, all sentences mentioning

$\langle e1, e2 \rangle$ will express the relation r . Otherwise, if there is no relation between $\langle e1, e2 \rangle$ in the knowledge graph, the sentence mentioning them will be labeled as “not a relation” (NA).

DS has been widely applied in relation extraction and has achieved good results. At present, many studies have applied deep learning in DS and improved distant supervision relation extraction (DSRE) by automatically learning text features [4–6]. In addition, some scholars have found that the addition of additional knowledge can further improve DSRE, thus gaining great attention [7–10]. Among them, Deng et al. [10] proposed a hierarchical framework to fuse information from DS and web tables that share entity relationship facts, which greatly improves the efficiency of DSRE.

However, the above methods follow the strong assumption of DS, and as this strong assumption does not always hold true, it might result in a wrong labeling problem. At the same time, adding additional knowledge can improve relationship extraction, but using two-hop DS data will even introduce more new noise because the strong assumption is not changed.

On the issue of DS noise mitigation, a multi-instance learning (MIL) framework [11–13] is introduced into DSRE to relax the strong assumption to the at-least-one principle. In the MIL framework, all instances mentioning $\langle e1, e2 \rangle$ constitute an instance-bag which shares a common label r . The at-least-one principle states that at least one instance in the bag can imply the relation r , and the bag is treated as the samples for training RE models instead of instances. However, above feature-based methods rely heavily on the accurate features derived by NLP tools, which might suffer from an error propagation problem.

To alleviate the noise problem, many RE studies based on the MIL framework employ neural networks with a selective attention mechanism to assign weights to different instances within the bag [14–18], and all achieve good results. These selective attention methods still assign a certain weight to the noisy instances (false positive instances); especially when a bag composed of single instance is wrongly labeled, as illustrated in Table 1, selective attention will not work on such a bag-level noise problem. According to the statistics [19], nearly 28% of one-instance bags in the NYT dataset are incorrectly labeled, which seriously hurts the performance of RE.

Table 1. Examples of instances of wrong labeling by distant supervision. There are two bags of instances B1 and B2 in the table, which are labeled with “play” and “famous_in”, respectively. The bold words in instances are entities. The first instance in B1 is correctly labeled, and the second is wrongly labeled due to the strong assumption of distant supervision. In this situation, the selective attention mechanism might assign 0.9 and 0.1 weights to the two instances, respectively. Meanwhile, only the one instance in B2 is mislabeled due to the incomplete knowledge graph and its weight will be set to 1 in selective attention. Therefore, there is instance-level noise in B1 and bag-level noise in B2, respectively.

Bag	Instance	Correct	Weight	Noise
B1 (play)	Kristen Stewart is set to play Princess Diana in an upcoming movie.	True	0.9	Instance-level
	Kristen Stewart can be mysterious and fragile and ultimately strong as well, which is similar to Princess Diana .	False	0.1	
B2 (famous_in)	Will Smith was born in Philadelphia , Pennsylvania on September 25, 1968.	False	1	Bag-level

In this paper, we propose a novel noise detection-based relation extraction model (NDRE), which can automatically distinguish the true positive and false positive cases in the training process by evaluating the correlation between sentences and tags, so as to alleviate the noisy labeling problem at both the instance and bag levels and avoid adding new noise labels while integrating two-hop DS data, thus further improving the performance of DSRE.

Specifically, in the proposed framework: 1) To learn a comprehensive sentence representation for each sentence in corpus and web tables (high-quality relational tables

extracted from <http://websail-fe.cs.northwestern.edu/TabEL/#content-code> (access date: 19 April 2013) [10], we firstly combine the multi-head self-attention mechanism and piecewise convolution neural network (PCNN) for the sentence encoder; 2) In this stage, we use a noise detection strategy to address the issue of noisy labeling. To evaluate the correlation of sentences and labels, we calculate the similarity between entity-aware embeddings and each sentence representation. According to the similarity score, we can judge whether the sentence can express the labeled relation; that is, whether the sentence is true positive or false positive. Then, the label of the detected false positive sentence will be dynamically corrected to NA during training; 3) To fuse information from corpus and web tables, we utilize a bag aggregation method to balance their impact on the predictive relation. The experimental results on a real-world DS dataset prove that our model performs better than the state-of-the-art baselines.

The remainder of the article consists of the following: In Section 2, we introduce the related work of relation extraction. The detailed methodology of the NDRE is illustrated in Section 3. Section 4 shows the results of the experiments and the analysis of results in detail. Finally, Section 5 presents our conclusion and discussion of future work.

2. Related Work

Relation extraction is an important research task in the NLP area. Many works regard RE as a supervised classification task and have achieved good results. One of the main defects of these conventional supervised relation extraction methods is the lack of manually annotated training data, which is expensive and time-consuming. To deal with this issue, distant supervision was proposed by Mintz et al. [3] to generate a large amount of training data automatically by aligning text to corresponding KGs.

At present, many DSRE works are combined with a neural network model to achieve good results. Socher et al. [4] and Zeng et al. [5] employed a recursive neural network (RNN) and a convolutional neural network (CNN) to obtain text representation, respectively. Zeng et al. [6] attempted to propose a piecewise convolution neural network for RE. They considered the relative position information of each word and entities and adopted piecewise pooling on entity position to retain more fine-grained information in the sentence. In addition, the addition of additional knowledge associated with text can further improve DSRE and become one of the new research directions. Ji et al. [7] proposed a sentence-level attention model that made full use of extracted entity descriptions to provide more background knowledge. Vashishth et al. [8] employed graph convolution networks to capture syntactic information from text and utilized available side information such as entity type and relation alias to improve RE. Beltagy et al. [9] combined distant supervision with a directly supervised data and used it to improve weights of relevant sentences. Deng et al. [10] proposed a hierarchical framework to fuse information from DS and web tables that share relational facts about entities to further improve RE.

The above methods have achieved good results, but because of the strong assumption of DS, the data generated inevitably contains sentences that are mistaken for their strong assumption. Moreover, adding two-hop DS data may bring new noise. Therefore, how to fully join the existing entity data and additional knowledge and get rid of the noise problem is an urgent problem for RE.

The multi-instance learning (MIL) framework is an important method of noise reduction in DSRE. Riedel et al. [11] introduced a multi-instance single-label learning framework to RE and assumed that the positive instance must exist and the instance with the highest confidence can express the label. Hoffmann et al. [12] and Surdeanu et al. [13] adopted multi-instance multi-label learning to model multiple relations between entities.

In recent years, to alleviate the noise problem, many RE studies based on the MIL framework have employed neural networks with a selective attention mechanism to assign weights to different instances within the bag. Lin et al. [14] proposed a CNN-based model with a sentence-level attention mechanism to assign more weights to effective instances and reduce the weights of noisy instances. Liu et al. [15] introduced a soft-label method to

exploit valid information from correctly labeled entity pairs to help noisy instances. Jat et al. [16] proposed a word attention model based on the Bidirectional Gated Recurrent Unit (BiGRU) and an entity-centric attention model to identify key words in sentences. Xiao et al. [17] proposed a hybrid attention-based Transformer block to get word-level features and constitute the bag representation. He et al. [18] presented a reinforcement-learning-based framework to choose positive instances and make full use of unlabeled instances.

Although the above methods have achieved good results, they still use noisy sentences in training and do not consider the problem of bag-level noise labeling, so do not fundamentally get rid of the noise problems.

In order to alleviate instance-level and bag-level mislabeling problems, we propose the use of the NDRE to detect noisy instances with entity information and dynamically correct the wrong labels to NA, which can reduce the influence of noisy instances and bags, which in turn can effectively fuse corpus and two-hop DS data and avoid introducing new noise tags, thus greatly improving the performance of RE.

3. Proposed Method

In this section, we describe our proposed method, the NDRE, in detail. The overall framework of the NDRE consists of three main components as shown in Figure 1.

Figure 1. The architecture of our proposed method (the noise detection-based relation extraction model, NDRE) used for distant supervised relation extraction. MLP: multi-layer perceptron.

Then, we describe our proposed method, the NDRE, in detail: (1) A sentence encoder which includes two parts: the input representation and encoding layer. The input representation takes the combination of word embedding and position embedding as a vector representation of each sentence in the package. The encoding layer consists of three modules: convolution, multi-head self-attention and piecewise max-pooling, which are used to extract semantic features implicit in sentence representation; (2) A noise-detection strategy

that can automatically detect noisy labels during the training process. We introduce the embedding of entities to evaluate whether the sentences in the package can express the target relation. Sentences that cannot express the target relation will be regarded as noise, and we remove the noise by modifying their labels; (3) A bag aggregation method, which first obtains the bag-level representation of the corpus and the web tables through selective attention, and then combines these two bags in a balanced manner to obtain the final sentence-bag representation.

In the following, we first introduce the task definition and notation, and then provide a detailed formalization of the NDRE.

3.1. Task Definition and Notation

Following the multi-instance learning paradigm, we are given a bag of sentences (instances) $\mathbb{S}_{h,t} = \mathbb{S}_{h,t}^1 \cup \mathbb{S}_{h,t}^2 = \{s_1, s_2, \dots\}$ with a pair of target entities (h, t) , where $\mathbb{S}_{h,t}^1$ is composed of all sentences mentioning (h, t) in corpus and $\mathbb{S}_{h,t}^2$ contains some sentences mentioning the anchors of (h, t) discovered in the web. The anchors are defined as two entities in web tables co-occurring with (h, t) , which are denoted as $\{(h_1, t_1), (h_2, t_2), \dots\}$. The goal of relation extraction is to predict the relation r from a predefined relation set $\mathbb{R} = \{r_1, r_2, \dots, r_l\}$, where l is the number of distinctive relation categories. If no relation exists, the label of the sentence will be assigned NA.

3.2. Sentence Encoder

3.2.1. Input Representation

Given a sentence $s \in \mathbb{S}_{h,t}$ consisting of m words $s = \{w_1, w_2, \dots, w_m\}$, each word w_i is embedded into a low-dimensional dense-vector $\mathbf{v}_i \in \mathbb{R}^{d^w}$, where d^w is the dimension of word embedding. Following previous works [6,14], our initial word embedding is pre-trained by the skip-gram method [20]. To define the position feature of the target entities and words, position embedding as proposed by Zeng [5] is applied in our work.

Position embedding describes the relative distances between the current word w_i to the two target entities, which are further mapped into two low-dimensional vectors, $\mathbf{p}_i^{e1}$ and $\mathbf{p}_i^{e2}$, respectively, where $\mathbf{p}_i^{e1}, \mathbf{p}_i^{e2} \in \mathbb{R}^{d^p}$. As illustrated in Figure 2, in the sentence “Donald Trump became the first billionaire president of the United States in January 2017.”, the relative distances of the word (*president*) to the head entity (*Donald Trump*) and tail entity (*the United States*) are 5 and -2 .

Figure 2. An example of position embedding.

We concatenate the word embedding and position embedding as the final word representation $x_i = [\mathbf{v}_i; \mathbf{p}_i^{e1}; \mathbf{p}_i^{e2}]$. In this way, the input sentence representation of s can be denoted as a vector sequence $S = \{s = \{x_1, x_2, \dots, x_m\}\}$, where $x_i \in \mathbb{R}^d (d = d^w + 2 \times d^p)$.

3.2.2. Encoding Layer

The piecewise convolutional neural network (PCNN) [6] has been utilized by a number of previous RE works and its powerful capability of capturing local features and positional information of sentences has been proven. However, PCNN usually performs badly when processing long sentences due to its limitation of not being able to obtain the global dependencies of sentences. Therefore, a multi-head self-attention mechanism is employed in our framework to obtain long-distance dependency information from multiple perspectives.

As shown in Figure 3, given an input representation S , we firstly apply a convolution kernel with a sliding window over S , where the window's size is set to k . In order to keep

the original size of the input, we add $\frac{k-1}{2}$ padding tokens on both sides of the sentence boundaries. The hidden output of the convolution layer can be expressed as:

$$\mathbf{H}^C = \{\mathbf{h}_1^c, \mathbf{h}_2^c, \dots, \mathbf{h}_m^c\}, \mathbf{h}_i^c = \mathbf{W}_c \mathbf{x}_{i-\frac{k-1}{2}:i+\frac{k-1}{2}} + \mathbf{b}_c \quad (1)$$

where $\mathbf{W}_c \in \mathbb{R}^{d^c \times (k \times d)}$, $\mathbf{b}_c \in \mathbb{R}^{d^c}$, and $\mathbf{H}^C \in \mathbb{R}^{m \times d^c}$.

Figure 3. Infrastructure of our sentence encoder. Relu: Rectified Linear Unit.

Next, to exploit the global dependency of a sentence, we adopt the multi-head self-attention mechanism of Transformer [21], which has successfully achieved promising results in most NLP tasks. The basic idea of a self-attention mechanism is to find the interaction between each word and the whole sentence. The importance or the weight of each word in the sentence can be calculated by the degree of its interaction and is utilized to adjust the representation of the sentence. In this way, the global information of the whole sentence will be contained in the new representation.

Formally, a self-attention mechanism is defined as follows:

$$Attention(\mathbf{a}, \mathbf{a}, \mathbf{a}) = softmax(\frac{\mathbf{a}\mathbf{a}^T}{\sqrt{d^a}})\mathbf{a} \quad (2)$$

where the *Softmax* ($\cdot$) function is used to calculate the weights that are applied to the words, T is a transpose operation, $\mathbf{a}\mathbf{a}^T$ is the dot product and d^a is the dimension of $\mathbf{a}$, and the purpose of scaling through $\sqrt{d^a}$ is to avoid the dot product result being too large.

Further, we adopt a multi-head operation on the self-attention mechanism to extract the features about the different levels of a sentence. We get the sentence representation $\mathbf{H}^a$, as represented in the following equation:

$$head_j = Attention(\mathbf{W}_j^Q \mathbf{H}^c, \mathbf{W}_j^K \mathbf{H}^c, \mathbf{W}_j^V \mathbf{H}^c) (j = 1, 2, \dots, A) \quad (3)$$

$$\mathbf{H}^a = \mathbf{W}^a [\mathbf{head}_1; \mathbf{head}_2; \dots; \mathbf{head}_A] \quad (4)$$

where $\mathbf{W}_j^Q, \mathbf{W}_j^K, \mathbf{W}_j^V \in \mathbb{R}^{d^h \times d^c}$ and $\mathbf{W}^a \in \mathbb{R}^{d^h \times d^c}$ are learnable parameters; d^h is the dimension of the head; and $[\cdot]$ is a concatenate operation. Notice that A is the number of heads and $d^c = d^h \times A$. The parameters of different heads are independent of each other.

Then, piecewise max-pooling is applied over $\mathbf{H}^a = \{\mathbf{H}_1^a, \mathbf{H}_2^a, \dots, \mathbf{H}_m^a\}$ to capture the sentence-level structural information, which is formulated as

$$\mathbf{s} = relu([\mathbf{Pool}(\mathbf{H}_{1:h}^a); \mathbf{Pool}(\mathbf{H}_{h+1:t}^a); \mathbf{Pool}(\mathbf{H}_{t+1:m}^a)]) \quad (5)$$

where $\mathbf{H}^a$ is segmented according to the position of (h, t), $\text{Pool}(\cdot)$ is a max-pooling operation and $\text{relu}(\cdot)$ is a non-linear activation function. As a result, $\mathbf{s} = \mathbb{R}^{m \times (3 \times d^c)}$ concatenated by three pooling results is the final sentence vector representation.

3.3. Noise Detection Strategy

As mentioned before, DSRE suffers from the wrong labeling problem whether the data comes from a corpus or web table. Especially when there is the only one sentence in the bag, if the sentence is labeled incorrectly, this noise problem will greatly damage the performance of the model. Unlike previous works that just relied on an attention module to alleviate noise, we propose a noise-detection strategy to detect and revise the noise label in the training process. Specifically, since RE aims to predict the relation between pairs of entities, we combine the embeddings of entities into our model. We employ the similarity between entity embeddings and sentence representation to estimate whether the current sentence can imply a relation between the target entity pair. If the current positive relation label does not exist in the sentence, the label will be modified to NA and the sentence is treated as invalid in the training of current relations. In this way, the noisy labeling problem of distant supervision will be largely eliminated.

Motivated by the translation-based knowledge graph embedding approaches [22–24], we utilize the difference vector between the target entity pair as an additional feature. Specifically, we use $\mathbf{r}_{ht} = \mathbf{t} - \mathbf{h}$ to reflect some information of relation between entities, where $\mathbf{h}, \mathbf{t} \in \mathbb{R}^{d^w}$ are pre-trained word embeddings.

Here, we adopt the Euclidean distance to measure the similarity of $\mathbf{r}_{ht}$ and the sentence representations. A two-layer feed-forward network is then applied to the distance, which is formally denoted as

$$sml = \mathbf{W}^{f2} \sigma(\mathbf{W}^{f1} d(\mathbf{A}\mathbf{s}, \mathbf{r}_{ht}) + \mathbf{b}^{f1}) + \mathbf{b}^{f2} \quad (6)$$

where $\mathbf{W}^{f1} \in \mathbb{R}^{d^w \times 1}$, $\mathbf{W}^{f2} \in \mathbb{R}^{1 \times d^w}$, $\mathbf{b}^{f1} \in \mathbb{R}^{d^w}$, $\mathbf{b}^{f2} \in \mathbb{R}^1$ and $\mathbf{A} \in \mathbb{R}^{d^w \times (3 \times d^c)}$ are learnable parameters; $\sigma(\cdot)$ is the sigmoid function; and $d(\cdot)$ is the Euclidean distance function.

Then, we correct the original relation label r using the similarity score sml as follows:

$$r' = \begin{cases} r, & \text{if } sml \in [0, \phi] \\ \text{NA}, & \text{otherwise} \end{cases} \quad (7)$$

where ϕ is a pre-defined threshold. If the similarity score sml is less than the threshold ϕ , we think the sentence can express the current relation r . Otherwise, r will be modified to NA.

3.4. Bag Aggregation

After we use the noise-detection strategy to eliminate the noise, we get a new bag-level label $\{r'_1, r'_2, \dots\}$ instead of just one label r for $\mathbb{S}_{h,t}$. It is remarkable that maybe not all sentences in $\mathbb{S}_{h,t}$ have the positive label r , and different positive sentences express r to different degrees.

Based on the above-mentioned reasons, we adopt frequently used selective attention over $\mathbf{S} = \{s_1, s_2, \dots\}$ to get the bag-level representation $\mathbf{B}$. Selective attention with a query vector of relation r'_i can make the noisy sentences invalid and no longer contribute to the representation of the bag. In addition, via selective attention, higher weights are assigned for the sentences expressing r more clearly. The representation of the bag is computed as follows:

$$\begin{cases} \alpha_i = \text{softmax}(\mathbf{q}_{r'_i}^T, \mathbf{s}_i) \\ \mathbf{B} = \sum_{i=1}^n \alpha_i \mathbf{s}_i \end{cases} \quad (8)$$

where $\mathbf{q}_{r'_i} \in \mathbb{R}^{(3 \times d^c)}$ is a learnable query vector of relation r'_i and n is the number of sentences in the current bag.

Due to the fact that the sentences contained by $\mathbb{S}_{h,t}^1$ and $\mathbb{S}_{h,t}^2$ are not an order of magnitude, $\mathbb{S}_{h,t}^1$ and $\mathbb{S}_{h,t}^2$ will get their own representation through Formula (8), which are denoted as $\mathbf{B}^1$ and $\mathbf{B}^2$. To balance their impact on relation prediction, they are assigned different proportions in the final sentence-bag representation $\mathbf{y}$, which is formulated as

$$\mathbf{y} = \lambda \mathbf{B}^1 + (1 + \lambda) \mathbf{B}^2 \quad (9)$$

where λ is a variable from 0 to 1 and is calculated by

$$\lambda = \sigma(\mathbf{W}^\lambda [\mathbf{B}^1; \mathbf{B}^2; \mathbf{q}_r] + \mathbf{b}^\lambda) \quad (10)$$

where $\mathbf{W}^\lambda \in \mathbb{R}^{(1 \times (9 \times d^w))}$ is the transformation matrix, $\mathbf{b}^\lambda \in \mathbb{R}^1$ is the bias, and $\mathbf{q}_r \in \mathbb{R}^{3 \times d^c}$ is a learnable query vector of bag label r .

3.5. Classification and Objective Function

DSRE can be regarded as a classification task. In order to get the predicted relation category, $\mathbf{y}$ is fed into a multi-layer perceptron (MLP) to get the final output, which corresponds to all relation categories. Then, we employ a $|I|$ -way *softmax* function over the final output to calculate the conditional probability P of each relation category:

$$P = \text{softmax}(\text{MLP}(\mathbf{y})) \quad (11)$$

Supposing that there are N bags in the training set $\{B_1, B_2, \dots, B_N\}$ with their corresponding target relation labels $\{L_1, L_2, \dots, L_N\}$, we define the objective function using cross-entropy, which is written as:

$$J(\theta) = -\frac{1}{N} \sum_{i=1}^N \log P(L_i | B_i, \theta) \quad (12)$$

Given that the parameters set the θ of the model, we use the mini-batch stochastic gradient descent (SGD) optimizer to minimize the objective function.

4. Experiments and Results

4.1. Dataset

We evaluated our model on a widely used NYT dataset developed by Riedel et al. [11]. The NYT dataset was generated automatically by aligning the entities in Freebase with the New York Times corpus. This dataset contains 53 relations, including a negative class NA, which means the relation of an entity pair is unavailable. In addition, we used the WikiTable corpus dataset aligned with the NYT. The dataset and more details on how to build it were provided by Deng et al. [10]. In this paper, NYT and WikiTable corpus are denoted as $\mathbb{S}_{h,t}^1$ and $\mathbb{S}_{h,t}^2$, respectively.

4.2. Comparison with Baselines

Following previous studies [3,11], we conducted our experiments in a held-out evaluation instead of costly human evaluation. Held-out evaluation compares predicted relational facts in text with existing facts in Freebase. To verify the effort of our model, we adopted precision–recall (PR) curves as the evaluation metrics, which can show the trade-off between precision and recall. To further quantify PR curves, we also report precision at recall (P@Recall) and the area under the curves (AUC). P@Recall indicates the precision values at concrete recall rates, and AUC evaluates the overall performance of PR curves. We selected the following baselines to compare with our model, the NDRE:

- Mintz: A multi-class logistic regression model based on text syntactic and lexical features proposed by Mintz et al. [3];
- MultiR: A probabilistic graphical model in a multi-instance learning framework proposed by Hoffmann et al. [12];

- PCNN+ATT: A CNN-based model combined with piecewise max pooling and selective attention over sentences proposed by Lin et al. [6];
- PCNN+ATT+SL: An entity-pair level model with soft labels that exploits semantic/syntactic information from correctly labeled instances, proposed by Liu et al. [15];
- BGWA: A BiGRU-based model with word-level attention proposed by Jat et al. [16];
- RESIDE: A GCN model based on Graph Convolutional Network (GCN) with utilizing syntactic dependency, entity type, and relation alias information, proposed by Vashishth et al. [8];
- Direct+Dis: This model, by (Beltagy et al., employs directly supervised data as supervision for the attention weights of DS data [9];
- REDS2: A two-hop distant supervision model fusing data from NYT and web tables proposed by Deng et al. [10];
- PCNN+PU: A bag-level reinforcement-learning-based method utilizing relation embedding and unlabeled instances proposed by He et al. [18];

4.3. Experimental Settings

4.3.1. Word and Position Embeddings

Our model uses the word embeddings pre-trained by the Word2vec³ tool on NYT corpus as initial values. Position embeddings were initialized with the Xavier [25] initialization. The input of all models was the concatenation of word embeddings and position embeddings.

4.3.2. Hyper-Parameter Settings

To demonstrate the effectiveness of the NDRE, we set most hyper-parameters in the NDRE following previous work [10,14] and adjusted other hyper-parameters using cross-validation on the training data.

After adjustment, we chose the following settings in our model: the dimension of word embedding d^w was set to 50, the dimension of position embedding d^p was set to 5, the dimension of convolution layer d^c was set to 256, the window size of the convolution kernel k was set to 3, the number of multi-heads A was set to 8, the dropout rate was set to 0.5, and the threshold ϕ of similarity was set to 0.16 from experience. Following the approach in [10], we first pre-trained our sentence encoder and noise detection strategy with only data $\mathbb{S}_{h,t}^1$, then fine-tuned the whole model with data $\mathbb{S}_{h,t} = \mathbb{S}_{h,t}^1 \cup \mathbb{S}_{h,t}^2$. The epoch of pre-training was about 50, and the epoch of fine-tuning was about 180; the batch size of pre-training was fixed to 64, and the batch size of fine-tuning was fixed to 2500; the learning rate of pre-training was set to 0.005, and the learning rate of fine-tuning was set to 0.001.

4.4. Overall Evaluation Results

From Figure 4, the following observations can be made: (1) Among all the baselines, our NDRE achieves the best performance over the entire recall range; (2) The NDRE performs much better than PCNN+ATT, BGWA, and PCNN+ATT+SL. It indicates that our noise-detection strategy is superior to ordinary selective attention mechanism and soft labeling based on correctly labeled instances in alleviating the noisy labeling problem. (3) The NDRE substantially outperforms RESIDE and Direct+Dis, which utilize the extra information, and even achieves the higher performance over the recently promoted PCNN+PU. This demonstrates the effectiveness of our work.

Figure 4. Aggregate precision–recall curves of our model and compared baseline methods.

4.5. Ablation Results

To further verify the effectiveness of the main components in the NDRE, we conducted an extra ablation study under three testing modes:

- Single: Only bags composed of single instance were selected for relation extraction.
- Multiple: Only bags composed of more than one instance were selected for relation extraction.
- Whole: The whole dataset was used for relation extraction.

Figure 5 shows the PR curves of the compared models under three testing modes (single, multiple, whole), and the corresponding P@Recall and AUC values are listed in Table 2.

In particular, NDRE(-Label) denotes removing the noise-detection strategy, and NDRE(-MHA) denotes removing the multi-head self-attention mechanism. REDS2 is the first work to fuse NYT and web tables data, which is the dataset we use.

From these results, we can observe that: (1) The performance of the NDRE declines significantly when its different components are removed. Especially when the noise detection strategy is removed, it shows a 9.9%, 7.6%, and 9% decrease in terms of AUC under the three testing modes, respectively. The overall performance is NDRE > NDRE(-MHA) > NDRE(-Label) > REDS2, which shows that our sentence encoder can better capture semantic features and that the noise-detection strategy can greatly improve DSRE. (2) Under the “Single” testing mode, NDRE and NDRE(-MHA) had improvements of 12% and 4.8%, respectively, under AUC compared with REDS2. This demonstrates that the noise-detection strategy can effectively detect and filter out the noisy bag consisting of single instances. (3) Under the “Multiple” testing mode, NDRE and NDRE(-MHA) outperformed REDS2 and their AUC was improved by 14.7% and 9.2%, respectively. It indicates that the noise-detection strategy can availablely detect the wrongly labeled instance in the bag and then make full use of the information contained in the true positive instance. (4) The AUC of the NDRE under the “Whole” testing mode was 0.5094, which comprises a new state-of-the-art performance.

Figure 5. Aggregate precision–recall curves of REDS2, NDRE(-Label), NDRE(-MHA), and NDRE under different testing modes: (a) single, (b) multiple, and (c) whole.

Table 2. Results of precision at recall (P@) and the area under the curves (AUC) under different testing modes.

MODE	Single				Multiple				Whole			
Model	P@0.1	P@0.2	P@0.3	AUC	P@0.1	P@0.2	P@0.3	AUC	P@0.1	P@0.2	P@0.3	AUC
REDS2	69.1	61.1	57.4	37.5	82.4	79.6	76.6	57.59	75.3	70.4	65.5	44.65
NDRE(-Label)	69.4	67.6	58.1	38.2	86.4	82.4	78.7	61.40	76.5	71.7	67.6	46.72
NDRE(-MHA)	73.5	68.0	60.3	39.3	88.6	87.5	82.4	62.90	80.9	74.9	71.3	47.89
NDRE	83.9	75.3	63.8	42.0	98.6	90.9	86.4	66.06	89.0	85.9	76.6	50.94

4.6. Case Study

To show the capabilities of the noise-detection strategy, we selected four representative instances in the training process for a case study. As shown in Table 3, each instance had a corresponding *sml* calculated in Section 3.3. If the *sml* exceeds the threshold ϕ , the label of instance will be dynamically modified to NA during training, *r* is the original relation label and *r'* is the new label modified by Noise Detection Strategy. The correct relation of the current instance is marked in blue, and the wrong relation is marked in red. From the table, we can see that: 1) Except for the first instance, none of the instances were true positive instances. The second and third instances were detected to have no relation in the instance because of their *sml*. The second instance was noisy and its label was correctly modified to NA according to the noise-detection strategy. 2) The original label of the fourth instance was *people/person/place_of_birth*, but its correct label was *people/person/place_lived*. The noise was not detected as the head entity “Peggy” had a relation *people/person/place_of_birth* with “Miami”. Thus, noise is hard to detect when there are multiple relations in an instance.

Table 3. Case study: a real example in our dataset for the noise-detection strategy. “Noise” represents whether the instance was detected as a noisy one, and “Correct” represents whether the final label of the instance was correct.

Instances	<i>sml</i> ($\phi = 0.16$)	<i>r</i>	<i>r'</i>	Noise	Correct
Buffon, one of the world 's top goalkeepers, and [Fabio_Cannavaro], [Italy]'s captain, were interviewed by the authorities.	0.04	people/person/nationality	people/person/nationality	No	Yes
The director, [Bennett_miller], had made the unflinching documentary “The cruise” in 1998, which followed the gradual descent of a near-homeless [New_York_city] tour guide.	0.68	people/person/place_lived	NA	Yes	Yes
Then he needs to be able to relate to the poor when he goes to places like [Brazil] or Morocco or [Burkina Faso].	0.83	NA	NA	No	Yes
[Peggy] was born in Miami, Florida and lived in [New_York_city] for 52 years.	0.12	people/person/place_of_birth	people/person/nationality	Yes	No

5. Conclusions

To solve the noisy labeling problem, this paper proposes a novel distant supervised method, named the NDRE, utilizing a noise-detection strategy to eliminate false positive labels and improve the quality of training. The NDRE fuses the noise-filtered information from corpus and web tables, based on a multi-head self-attention mechanism for PCNN to extract the semantic features of sentences. As a result, our model achieves better performance as compared to competitive baselines on the benchmark dataset in terms of commonly used evaluation metrics.

In the future, we plan to explore reinforcement learning for DSRE to identify noisy instances and learn the correlation between false positive and true positive instances. In addition, there is a lot of information hidden in the instances labeled as NA. We tend to make full use of this potential information to further improve relation extraction.

Author Contributions: Conceptualization, X.Z.; methodology, X.M.; resources, F.Z.; project administration, T.J.; funding acquisition, T.J., X.Z., Y.W., and F.Z.; formal analysis, B.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the West Light Foundation of The Chinese Academy of Sciences, Grant No. 2019-XBQNXZ-A-004; the National Key Research and Development, Program of China under Grant No. 2017YFC0820704; and the Tianshan Excellent Young Scholars of Xinjiang, Grant No. 2018Q032.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; Taylor, J. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, Vancouver, BC, Canada, 10–12 June 2008; Association for Computing Machinery: New York, NY, USA; pp. 1247–1250.
2. Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; Ives, Z. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 722–735.
3. Mintz, M.; Bills, S.; Snow, R.; Jurafsky, D. Distant Supervision for Relation Extraction Without Labeled Data. In Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Singapore, 2–8 August 2009; pp. 1003–1011.
4. Socher, R.; Huval, B.; Manning, C.D.; Ng, A.Y. Semantic Compositionality Through Recursive Matrix-Vector Spaces. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Jeju Island, Korea, 12–14 July 2012; Association for Computational Linguistics: Stroudsburg, PA, USA, 2012; pp. 1201–1211.
5. Daojian, Z.; Kang, L.; Siwei, L.; Guangyou, Z.; Jun, Z. Relation classification via convolutional deep neural network. In Proceedings of the COLING 2014, the 25th international conference on computational linguistics: technical papers, Dublin, Ireland, 23–29 August 2014; 2014; pp. 2335–2344.
6. Zeng, D.; Liu, K.; Chen, Y.; Zaho, J. Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–22 September 2015; Association for Computational Linguistics: Stroudsburg, PA, USA, 2015; pp. 1753–1762.
7. Ji, G.; Liu, K.; He, S.; Zhao, J. Distant Supervision for Relation Extraction with Sentence-Level Attention and Entity Descriptions. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; pp. 3060–3066.
8. Vashishth, S.; Joshi, R.; Prayaga, S.S.; Bhattacharyya, C.; Talukdar, P. Reside: Improving distantly-supervised neural relation extraction using side information. *arXiv* **2018**, arXiv:1812.04361.
9. Beltagy, I.; Lo, K.; Ammar, W. Combining distant and direct supervision for neural relation extraction. *arXiv* **2018**, arXiv:1810.12956.
10. Deng, X.; Sun, H. Leveraging 2-hop Distant Supervision from Table Entity Pairs for Relation Extraction. *arXiv* **2019**, arXiv:1909.06007, 2019.
11. Riedel, S.; Yao, L.; McCallum, A. Modeling Relations and Their Mentions Without Labeled Text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Proceedings of the Machine Learning and Knowledge Discovery in Databases, Barcelona, Spain, 20–24 September 2010*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 148–163.
12. Hoffmann, R.; Zhang, C.; Ling, X.; Zettlemoyer, L.; Weld, D.S. 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, Portland, OR, USA, 19–24 June 2011; Association for Computational Linguistics: Stroudsburg, PA, USA, 2011; pp. 541–550.

13. Surdeanu, M.; Tibshirani, J.; Nallapati, R.; Manning, C.D. 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, Jeju Island, Korea, 13–14 July 2012; Association for Computational Linguistics: Stroudsburg, PA, USA, 2012; pp. 455–465.
14. Lin, Y.; Shen, S.; Liu, Z.; Luan, H.; Sun, M. Neural Relation Extraction with Selective Attention over Instances. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; Volume 1, pp. 2124–2133.
15. Liu, T.; Wang, K.; Chang, B.; Sui, Z. A Soft-Label Method for Noise-Tolerant Distantly Supervised Relation Extraction. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; Association for Computational Linguistics: Stroudsburg, PA, USA, 2017; pp. 1790–1795.
16. Jat, S.; Khandelwal, S.; Talukdar, P. Improving distantly supervised relation extraction using word and entity based attention. *arXiv* **2018**, arXiv:1804.06987.
17. Xiao, Y.; Jina, Y.; Cheng, R.; Hao, K. Hybrid Attention-Based Transformer Block Model for Distant Supervision Relation Extraction. *arXiv* **2020**, arXiv:2003.11518.
18. He, Z.; Chen, W.; Wang, Y.; Zhang, W.; Wang, G.; Zhang, M. Improving Neural Relation Extraction with Positive and Unlabeled Learning. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 7927–7934.
19. Li, Y.; Long, G.; Shen, T.; Zhou, T.; Yao, L.; Huo, H.; Jiang, J. Self-Attention Enhanced Selective Gate with Entity-Aware Embedding for Distantly Supervised Relation Extraction. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; 2020; volume 34, pp. 8269–8276.
20. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed Representations of Words and Phrases and Their Compositionality. In *Advances in Neural Information Processing Systems 26 (NIPS 2013), Proceedings of the Neural Information Processing Systems 2013 Conference, Lake Tahoe, NV, USA, 5–10 December 2013*; Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q., Eds.; Curran Associates Inc.: Red Hook, NY, USA, 2013; Volume 2, pp. 3111–3119.
21. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All You Need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017), Proceedings of the Neural Information Processing Systems 2017 Conference, Long Beach, CA, USA, 4–9 December 2017*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishanathan, S., Garnett, R., Eds.; Neural Information Processing Systems: San Diego, CA, USA, 2017; pp. 5998–6008.
22. Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; Yakhnenko, O. Translating Embeddings for Modeling Multi-Relational Data. In *Neural Information Processing, Proceedings of the 26th International Conference on Neural Information Processing Systems, NIPS'13, Sydney, NSW, Australia, 12–15 December 2019*; Gedeon, T., Wong, K.W., Lee, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 2787–2795.
23. Wang, Z.; Zhang, J.; Feng, J.; Chen, Z. Knowledge Graph Embedding by Translating on Hyperplanes. In Proceedings of the 28th AAAI Conference on Artificial Intelligence, Quebec, ON, Canada, 27–31 July 2014.
24. Lin, Y.; Liu, Z.; Sun, M.; Liu, Y.; Zhu, X. Learning Entity and Relation Embeddings for Knowledge Graph Completion. In Proceedings of the 29th AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015.
25. Glorot, X.; Bengio, Y. Understanding the Difficulty of Training Deep Feedforward Neural Networks. *J. Mach. Learn. Res.* **2010**, *9*, 249–256.