applied sciences

MDPI

Article

EANT: Distant Supervision for Relation Extraction with Entity Attributes via Negative Training

Xuxin Chen and Xinli Huang *

School of Computer Science and Technology, East China Normal University, Shanghai 200062, China
* Correspondence: xlhuang@cs.ecnu.edu.cn

Abstract: Distant supervision for relation extraction (DSRE) automatically acquires large-scale annotated data by aligning the corpus with the knowledge base, which dramatically reduces the cost of manual annotation. However, this technique is plagued by noisy data, which seriously affects the model's performance. In this paper, we introduce negative training to filter them out. Specifically, we train the model with the complementary label based on the idea that "the sentence does not express the target relation". The trained model can discriminate the noisy data from the training set. In addition, we believe that additional entity attributes (such as description, alias, and types) can provide more information for sentence representation. On this basis, we propose a DSRE model with entity attributes via negative training called EANT. While filtering noisy sentences, EANT also relabels some false negative sentences and converts them into useful training data. Our experimental results on the widely used New York Times dataset show that EANT can significantly improve the relation extraction performance over the state-of-the-art baselines.

Keywords: distant supervision; relation extraction; negative training; entity attributes

check for updates

Citation: Chen, X.; Huang, X. EANT: Distant Supervision for Relation Extraction with Entity Attributes via Negative Training. *Appl. Sci.* **2022**, *12*, 8821. https://doi.org/10.3390/app12178821

Academic Editor: Keun Ho Ryu

Received: 8 August 2022
Accepted: 30 August 2022
Published: 2 September 2022

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

Copyright: © 2022 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) describe real-world concepts and their relations in graphs and play an essential role in natural language processing tasks, such as recommendation systems [1,2], search engines [3], and Q&A systems [4,5]. Currently, open-source available KGs such as FreeBase [6], DBPedia [7], and Wikidata [8] have been widely used. Although they already contain hundreds of millions of relation facts, they still cannot cover the nearly infinite number of facts in the real world. To improve the completeness of KGs, relation extraction (RE) is proposed to automatically extract the relation between two given entities from unstructured text.

Most existing supervised RE methods require large amounts of annotated data, which is often expensive and time-consuming. To address this problem, Mintz et al. [9] proposed a distant supervision (DS) approach to generate training data automatically. It heuristically aligns the corpus with the KGs and assumes that if two entities have a relation in the KGs, then sentences that contain both entities may express the relation. However, it is clear that this assumption is too strong and inevitably introduces noise problems, such as false positives (FP) and false negatives (FN). The former is because the sentence does not actually express the target relation but is incorrectly marked as positive; the latter is due to the sentence actually expressing the target relation but being marked as negative because of the missing corresponding facts in the KGs. These two types of noise seriously affect the RE performance of the model.

The current noise reduction methods can be broadly classified into four categories: (1) feature-based methods that manually construct suitable features and filter noise using machine learning; (2) representation-enhancing methods that improve the contribution of valid sentences in the dataset using attention mechanisms [10], extra information [11], pre-trained models [12], etc.; (3) noise-filtering methods that filter out noisy sentences using

reinforcement learning [13,14], adversarial learning [15], pattern matching [16,17], etc.; (4) relabeling methods that assign correct relation labels to sentences using soft labeling [18], re-labeling [19], noise modeling [20], etc.

However, none of these approaches take full advantage of the rich information contained in entity attributes (e.g., description, alias, and types). Such information can provide richer background knowledge of entities for sentence representation, and incorporating them can improve the performance of the model. As shown in Figure 1, given the sentence "Larry Page and Sergey Brin founded Google in 1998 on the back of PageRank algorithm.", the task of RE is to discover the relation "*Founder_of*" between the entity "*Larry Page*" and "*Google*". From this example, we can clearly see the role of entity attribute information: the description of "*Larry Page*" mentions that "Larry Page is the founding CEO of Google", which undoubtedly provides valid evidence for the relation "*Founder_of*". Additionally, the alias of "*Google*" as "Google Inc" can provide more information of "Google is a company" to the entity "*Google*". By extracting these features, we can effectively determine similar relations. Finally, types information can reduce the probability that sentences are classified as impossible relations. For example, if "*Larry Page*" is an instance of "person" and the tail entity "*Google*" is an instance of "company", then it definitely not be an instance of the relation "*Place_of_birth*".

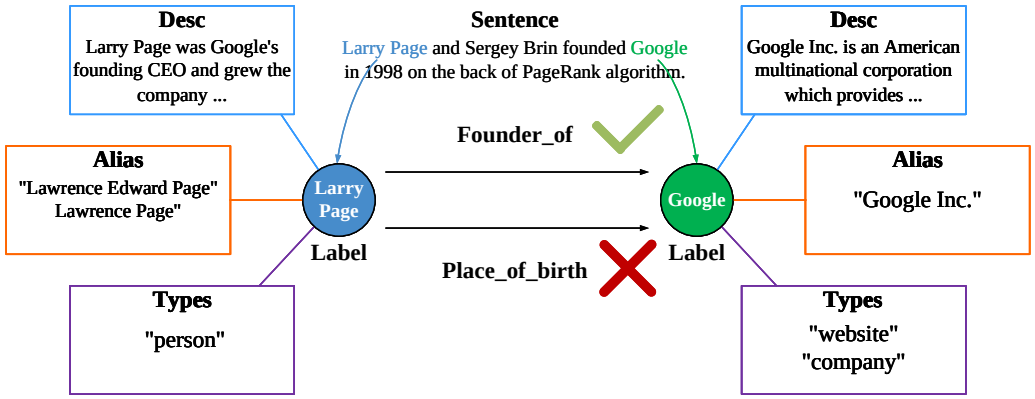


Figure 1. An example of entity attributes in the relation extraction task.

Moreover, since the DSRE dataset is automatically generated from aligned KGs, previous methods often lack explicit supervised information to capture the noise. Therefore, we propose to introduce negative training (NT) [21] to train the noise selector to achieve the separation of noisy data. Different from positive training (PT), negative training is based on the idea that "the sentence **does not** express the target relation", and the complementary labels are selected to train the model. A complementary label indicates a label other than the original label. As shown in Figure 1, the correct label for the sentence should be "*Founder_of*". However, under the strong assumption of DS, the sentence may be marked as other labels, resulting in incorrect information provided to PT. In negative training, the sentence is assigned a complementary label other than "*Founder_of*", such as "*Place_of_birth*", to train the model that the sentence is not "*Place_of_birth*". Since the probability of selecting a real label as a complementary label is low, it is assumed that the label provides the correct information to NT in this case. In other words, the relation expressed in the sentence may not be "*Founder_of*", but it is certainly not "*Place_of_birth*". The resulting model is able to separate the noisy data from the training data.

In this paper, we propose a DSRE model with **E**ntity **A**ttributes via **N**egative **T**raining, called **EANT**. Furthermore, we argue that some of the noisy data simply lack correct labels but still contain useful information. If reliable labels can be assigned to these sentences, they will be transformed into useful training data, which will help improve the model's performance. Therefore, we relabel the noisy sentences according to their confidence levels.

The contributions of this paper can be summarized as follows:

- We incorporate additional entity attribute information to provide more information for sentence representation. To reflect the importance of different attributes, we use knowledge graph embedding to assign weights to them.
- We introduce negative training into the DSRE task and train the model with the complementary label. By negative training, the model is able to widen the confidence gap between clean and noisy data to achieve noise filtering.
- The experimental results on the New York Times dataset show that the proposed approach can significantly improve the relation extraction performance over the state-of-the-art baselines.

2. Related Work

Mintz et al. [9] proposed distant supervision to address the lack of large-scale annotated data in relation extraction supervision methods. The method aligns the KGs with the corpus and labels sentences containing entity pairs from the KGs as the relation in the KGs. Although this approach generates annotated data quickly without manual effort, the generated data contains much noise, which degrades the performance of the RE model. To solve the noise problem, Riedel et al. [22] proposed the at-least-once assumption, where sentences with the same entity pairs were formed into a bag. It is the bag that expresses the relation rather than an individual sentence. Meanwhile, they designed lexical features, syntactic features, named entity labeling features, and predicted relations using machine learning methods, improving the model's results.

Although these feature-based approaches are well designed, this is not enough. The NLP tools they used had inevitable errors, which led to the propagation or accumulation of errors. With the successful application of deep learning methods, scholars have started to extract features using deep neural networks.

These approaches can be broadly classified into the following three categories: the first category of approaches makes full use of the non-noisy sentences in the dataset to extract rich information to generate more accurate representations for the sentences. Lin et al. [10] used sentence-level selective attention to assign different weights to each sentence in the dataset to distinguish their impact. Ji et al. [11] used entity description information as supplementary background knowledge to strengthen the learning of entity representation. Jat et al. [23] argued that only a few words in a sentence were relevant to the expressed relation and assigning higher weights to these words can help to improve the RE performance.

The second category of approaches attempts to filter out noisy sentences from the original dataset to eliminate their influence. Feng et al. [13] and Qin et al. [14] trained an instance selector using reinforcement learning with a policy function to decide whether to keep an instance or not. At the same time, the performance of the relation classifier was readjusted as a reward or penalty for reinforcement learning. Qin et al. [15] chose generative adversarial networks to train generators and discriminators that achieved similar noise filtering effects. Jia et al. [16] extracted the expression pattern of the relation from the sentences and removed the sentences that did not match pattern as noise.

The third category of approaches is trained with noisy data and aims to generate relation labels for instances that better match the true expression of the sentence. Liu et al. [18] proposed a soft-label noise reduction method that dynamically assigned a new label to a sentence during the training period based on the semantic information of the sentence. Shang et al. [19] used unsupervised deep clustering to generate reliable labels for noisy sentences. Luo et al. [20] described the noise patterns by dynamically generating a transition matrix, which indicated the possibility that the DS-labeled relation was confused for each instance.

In this paper, we combine the features of the above three approaches and incorporate the attributes of entities as additional information into the model to obtain a more accurate sentence representation. Meanwhile, we use the NT method to train a noise filter to filter

noisy sentences from the dataset. After filtering out the noisy sentences, we assign possible labels according to their confidence levels to utilize the information in them.

3. Methodology

In this section, we will introduce our proposed DSRE model, EANT, in detail. The structure of the model is shown in Figure 2. It contains three modules: a sentence encoder, an entity attribute encoder and a negative training process. The sentence encoder is used to convert sentences into corresponding distributed representations. The entity attribute encoder learns entity representations incorporating entity attribute information. A noise filter is trained by negative training, and it is decided whether the filtered noise sentence needs to be relabeled according to its confidence.

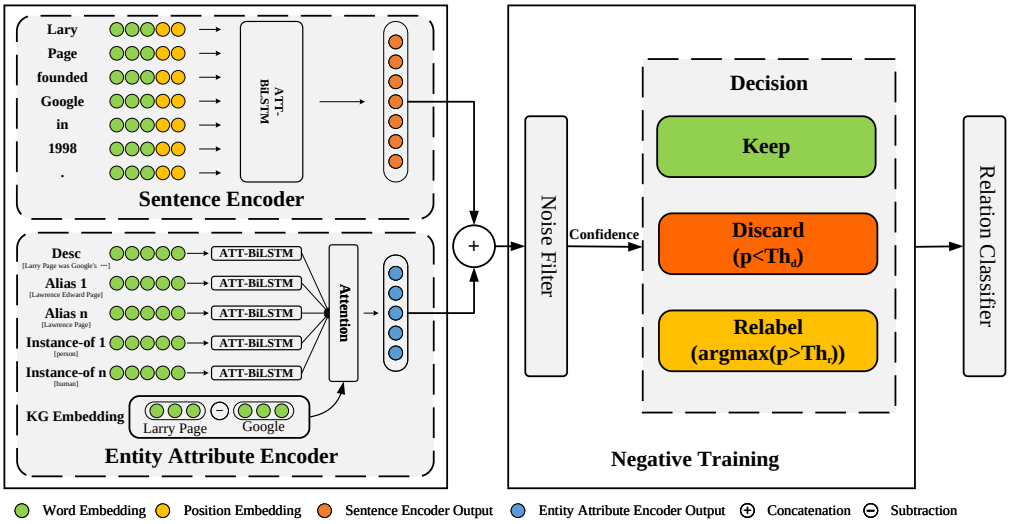


Figure 2. Illustration of our model EANT. In the figure, p denotes the confidence level of the sentence. Th_d represents the discard threshold. Th_r represents the relabel threshold.

3.1. Problem Definition

Before we formally start to introduce the model, we first give the problem definition. We denote the dataset obtained by DS as $\mathcal{D}^* = \{\mathcal{S}^*, \mathcal{R}\}$, where $\mathcal{S}^* = \{< s_1, r_1^* >, \ldots, < s_n, r_n^* >\}$ is the set of sentences s_i expressing the relation r_i^*. s_i is a sentence associated with two entities $< e_1, e_2 >$, and $r_i^* \in \mathcal{R}$ is the noisy relation label of sentence s_i generated by DS. Following the at-least-once assumption, sentences with the same entity pair are packaged into a bag $\mathcal{B} = \left\{\{s_1, s_2, \ldots, s_m\}, r_i^*\right\}$, indicating that this bag of sentences expresses the target relation r_i^*. The goal is to select sentences from $\mathcal{S}^*$ that correctly express the labeled relations to form a clean dataset $\mathcal{D} = \{\mathcal{S}, \mathcal{R}\}$, where $\mathcal{S} = \mathcal{S}^* - \mathcal{S}_{noise}$. We then train the relation classifier on the dataset $\mathcal{D}$ to predict the relation of each bag.

3.2. Sentence Encoder

For the input sentence $s = \{w_1, w_2, \ldots, w_n\}$, we first transform it into a low-dimensional vector consisting of word embeddings and position embeddings. The word embeddings map each token w_i in the sentence to a k-dimensional vector of real values to form distributed representations of the words. This distributed representation of words has been shown to greatly capture syntactic and semantic information as well as similarity properties [24]. It can be obtained by looking up the embedded representation matrix, e.g., *Word2vec* (https://code.google.com/p/word2vec/, accessed on 1 September 2022) and *GloVe* [25]. The position embeddings are a p-dimensional vector proposed by Zeng et al. [26] to represent the relative distance of each word to the head and tail entity and are widely used in RE tasks. These two embeddings are then concatenated together to form the initial representation of the sentence $\mathbf{s} \in \mathbb{R}^{n \times (d_k + 2d_p)}$.

BiLSTM has been proved to extract complete, sequential information of every word in a sentence. Therefore, we use BiLSTM to learn the semantic features of sentences. The hidden state of the BiLSTM is denoted as:

$$\mathbf{H} = (\mathbf{h_1}, \mathbf{h_2}, \ldots, \mathbf{h_N})^T, \tag{1}$$

where $\mathbf{h_i} = [\overrightarrow{h_i} : \overleftarrow{h_i}]$ is a concatenation of the forward hidden state $\overrightarrow{h_i}$ and the backward hidden state $\overleftarrow{h_i}$ for the i-th word. N is the sentence length. T is the transpose operation.

We then input $\mathbf{H}$ to the attention layer to distinguish the importance of different words. The sentence representation $\mathbf{s}$ obtained by the sentence encoder is denoted as:

$$\begin{aligned} \mathbf{M} &= \tanh(\mathbf{H}), \\ \mathbf{a} &= \mathrm{softmax}\left(\mathbf{w}^T\mathbf{M}\right), \\ \mathbf{s} &= \mathbf{H}\mathbf{a}^T, \end{aligned} \tag{2}$$

where $\mathbf{w}^T$ is a trained parameter vector.

3.3. Entity Attributes Encoder

Entity attributes can provide rich background knowledge for an entity, which can help improve the performance of RE. We extract entity descriptions, entity aliases, and entity types as shown in Figure 1 from the public database of Freebase [6] and Wikidata [8]. We first query the word embeddings of these attributes and input them to an attention-based BiLSTM encoder to obtain a vector representation $\mathbf{s_a}$ of each attribute, as $\mathbf{s}$ in Equation (2).

Furthermore, we note that not all attribute information is beneficial for RE, which may contain irrelevant information and introduce some additional noisy information. Therefore, we propose to use knowledge graph embedding TransE [27] to discriminate useful attributes. TransE treats relations as a translation from head entity e_h to tail entity e_t, i.e., $e_h + r \approx e_t$. Therefore, the relation can be approximated in the following way: $\mathbf{r} = \mathbf{e_t} - \mathbf{e_h}$. We calculate the weight of each attribute by multiplying $\mathbf{s_a}$ with $\mathbf{r}$. The final representation of the entity attributes information is denoted as follows:

$$\begin{aligned} \alpha_i &= \frac{\exp(\mathbf{s_a} \cdot \mathbf{r})}{\sum_{i=1} \exp(\mathbf{s_a} \cdot \mathbf{r})}, \\ \mathbf{d}_e &= \sum_{i=1}^{l} \alpha_i \mathbf{s_a}, \end{aligned} \tag{3}$$

where α_i is the attention weight of attribute $\mathbf{s_a}$. The entity attribute embedding $\mathbf{d}_e$ is calculated by the weighted average of all attributes.

Finally, we concatenate the sentence encoder output with the entity attribute encoder output to obtain the new sentence representation. It can be expressed as:

$$\overline{\mathbf{s}} = [\mathbf{s} : \mathbf{d_h} : \mathbf{d_t}]. \tag{4}$$

To calculate the confidence level for each relation, we feed the sentence representation into a softmax classifier after a linear transformation, as follows:

$$\mathrm{p}(r|s, \theta) = \mathrm{Softmax}(\mathbf{M}_s\overline{\mathbf{s}} + \mathbf{b}_s), \tag{5}$$

where $\mathbf{M}_s$ is the transformation matrix and $\mathbf{b}_s$ is the bias.

3.4. Negative Training

Our goal is to enable the model to be trained robustly on noisy datasets, so we need to distinguish between noisy data and clean data as much as possible. In other words, we aim to widen the gap between the confidence of clean data and noisy data. In positive

training, which is based on the idea that "the sentence **does** express the target relation", the cross-entropy loss function is often defined as follows:

$$\mathcal{L}_{PT} = -\sum_{i=1}^{|\mathcal{R}|} r_i \log \mathrm{p}(r_i|s,\theta), \tag{6}$$

where $|\mathcal{R}|$ denotes the number of relations in relation set $\mathcal{R}$, and r_i is an $|\mathcal{R}|$-dimensional one-hot vector. As the loss decreases, the probability of a sentence to a given label will be close to 1.

In contrast, the negative training is based on the idea that "the sentence **does not** express the target relation", and the complementary label is chosen to train the model. Specifically, for an input s with label $r_i \in \mathcal{R}$, we randomly sample a complementary label r^* from $\mathcal{R}$, but not r_i. Since we want the input to be far away from this complementary label, that is, to make its probability far away from 1. Therefore, we propose the loss function as follows:

$$\mathcal{L}_{NT} = -\sum_{i=1}^{|\mathcal{R}|} r^* \log(1 - \mathrm{p}(r^*|s,\theta)). \tag{7}$$

This allows the probability value of the complementary labels to be optimized to zero, so that the probability value of other classes increases, achieving the purpose of NL.

3.5. Noise Filtering and Relabeling

After negative training, the model is able to distinguish between noisy data and clean data. We can simply set thresholds to filter out the noise in them. However due to the serious long-tail problem in the DSRE task, it is necessary to set a corresponding threshold for each relation to prevent the tail relation from being filtered out. Here, we propose a simple method for setting filtering thresholds:

$$Th_d = Th \cdot \max_{i=1}^{N} \left\{ p_r^i \right\}, \tag{8}$$

where Th is a global threshold, and p_r^i denotes the probability value of an instance with relation r. N is the number of all instances with relation r.

As mentioned above, some of these filtered sentences may simply lack the correct labels but still contain useful information on their own. When these sentences are assigned reliable labels, they will become useful training data. Similarly, we devise a simple strategy to relabel these sentences with the following equation:

$$r_i^{'} = \arg\max_{r} \left\{ p_r^i > Th_r \right\}, \tag{9}$$

where Th_r is the relabel threshold. Note that we only relabel those false-negative sentences. Finally, we perform positive training on the clean dataset to obtain the final relation classifier.

4. Experiments

4.1. Dataset and Evaluation Metrics

We evaluate the performance of the proposed EANT model on a widely used distant supervised dataset, New York Times (NYT). It was generated by aligning Freebase with the New York Times corpus. The dataset contains a total of 52 relations and an NA label, which indicates no relation between entities. The detailed statistics of this dataset can be found in Table 1.

Table 1. Statistics of NYT dataset.

NYT	Train	Test
#Sentences	522,611	172,448
#Entity pairs	281,270	96,678
#Triplets	18,252	1950
#Relations	53	53

In this work, we perform a hold-out evaluation [10] to demonstrate the effectiveness of the model. The hold-out approach compares the relations predicted by the model with those in Freebase. Specifically, we plot the precision-recall (PR) curve to show the trade-off between precision and recall and report the precision at Top N (P@N) metric to consider the accuracy values at different cutoffs. The area enclosed by the PR curve and the two axes intuitively represents the performance of the model, i.e., the larger the area of the PR curve, the better the performance of the model. P@N is the accuracy performance of the top N predictions with the highest prediction scores after sorting the results in descending order. Therefore, a higher value of P@N indicates better performance of the model.

4.2. Baselines

We compare with several typical baselines for DSRE tasks, as follows:

- **Mintz**: Mintz et al. [9] proposed a multi-class logistic classifier optimized using L-BFGS with Gaussian regression model for DSRE.
- **MultiR**: Hoffmann et al. [28] proposed an undirected graphical model based on factor graphs and used the SampleRank framework for multi-instance learning.
- **MIMLRE**: Surdeanu et al. [29] jointly modeled multi-instance and multi-label learning with a graphical model.
- **PCNN**: Zeng et al. [30] proposed to introduce a piecewise max-pooling mechanism based on traditional convolutional neural networks (CNN) to enhance the representation of sentences.
- **PCNN+ATT**: Lin et al. [10] proposed a PCNN-based model and distinguished the importance of different sentences using sentence-level attention.
- **PCNN+ATT+Soft**: Liu et al. [18] proposed a PCNN-based model which dynamically generated a soft label to replace the hard label as the gold label for each bag during the training process.
- **BGWA**: Jat et al. [23] proposed a word-level attention model based on bi-directional gated recurrent unit (Bi-GRU) to identify key phrases in the sentence.
- **RESIDE**: Vashishth et al. [31] proposed a graph convolution network (GCN)-based model which employed relevant side information and syntactic information for DSRE.
- **A2DSRE**: Shi et al. [32] proposed a DSRE framework based on adaptive dependency-path and additional KG supervision.

4.3. Experimental Settings

In the experiment, we adjusted other hyperparameters on the training data using the cross-validation method to minimize the loss and determine the hyperparameters of the model via a grid search. After adjustment, we chose the following settings: we used a 50-dimensional glove vector (https://github.com/stanfordnlp/GloVe, accessed on 1 September 2022) as the word embedding for initialization. The number of BiLSTM network layers was set to 1, and the size of the hidden vector was set to 128. We use the stochastic gradient descent (SGD) algorithm for optimization, and the learning rate was set to 0.05. Table 2 shows all the parameters used in our experiments.

Table 2. Parameter settings.

Parameter	Value
Word dimension, d_w	50
Position dimension, d_p	5
Hidden layer dimension, d_h	128
Learning rate, α	0.05
Batch size, b	128
Global filter threshold, Th	0.25
Relabel threshold, Th_r	0.7
Dropout rate, p_{drop}	0.5

4.4. Results

As shown in Figure 3, we plot the P-R curve comparison results of our model against several baselines on the NYT dataset. From the figure, we can observe that (1) the neural models outperform human feature-based methods, which illustrates the limitations of manually designed features and the advanced nature of neural network methods; (2) compared with PCNN and PCNN+ATT, the PCNN+ATT+Soft model achieves better performance, which indicates that assigning suitable labels to noisy sentences can make full use of the information contained in them; (3) RESIDE and A2DSRE achieve sub-optimal results, which indicates that utilizing additional information in the knowledge graph can effectively improve the performance of model relation extraction; (4) our approach achieves the best performance compared to the other baseline models. This demonstrates the effectiveness of incorporating entity attribute information and negative training. The entity attribute information can provide richer information for entity representation, and the negative training can effectively filter out the noisy data in the dataset and improve the quality of the training data.

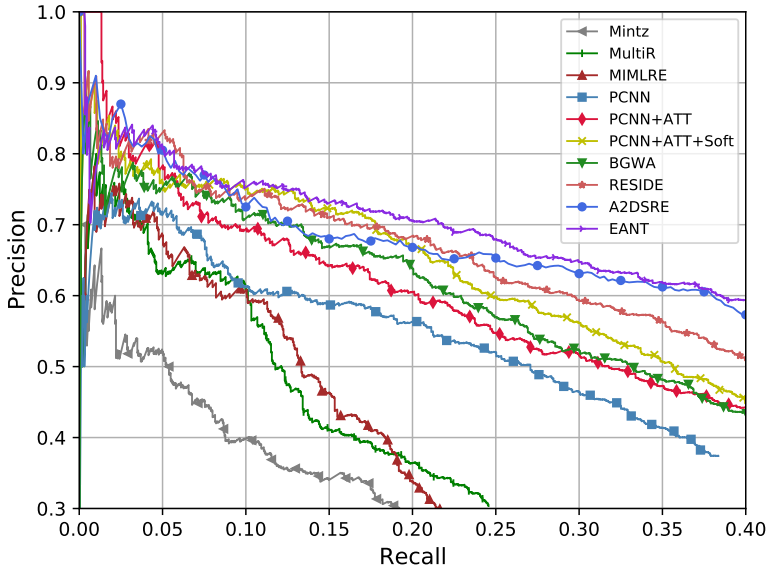


Figure 3. Precision-recall curves of EANT and baselines.

Following previous work [10], we also evaluated our method on the P@N metric, which is the precision of the top N sentences ranked by prediction confidence. Here, we report P@100, P@200, P@300 and their averages for each model, and the comparison results are shown in Table 3. The best value is shown in bold, while the second best value is underlined. From the table, we can observe that our model achieves the best performance on most of the metrics. Specially, compared with the latest baseline model (i.e., A2DSRE), EANT achieves a 2.34% improvement on the P@Mean metric. From this, we can conclude

that our model can extract richer sentence representations from sentences and effectively reduce the interference of noisy data.

Table 3. P@N of EANT and baselines.

Method	P@100	P@200	P@300	P@Mean
Mintz [9]	54.0	50.5	45.3	49.9
MultiR [28]	75.0	65.0	62.0	67.3
MIMLRE [29]	70.0	64.5	60.3	64.9
PCNN [30]	72.3	69.7	64.1	68.7
PCNN+ATT [10]	76.2	73.1	67.4	72.2
PCNN+ATT+Soft [18]	87.0	**84.5**	77.0	82.8
BGWA [23]	82.0	75.0	72.0	76.3
RESIDE [31]	84.0	78.5	75.6	79.4
A2DSRE [32]	87.0	79.0	77.6	81.2
EANT(ours)	**87.6**	83.2	**78.6**	**83.1**

4.5. Ablation Study

The proposed EANT model mainly consists of the sentence encoder, entity attribute encoder and the negative training process. In order to verify the effectiveness of different modules of our model, we designed two variants of the model to perform the ablation experiments.

Figure 4 depicts the PR curves of the different variants of the model. **EANT-woEA** is the method without the entity attribute encoder, which only uses the sentence encoder to obtain the sentence representation. **EANT-woNT** is the method without negative training, which obtains the final relation classifier by positive training only. From the figure, we can observe that removing any one module will result in a loss of model performance. This illustrates the positive effect of incorporating entity attribute information and negative training on model performance improvement.

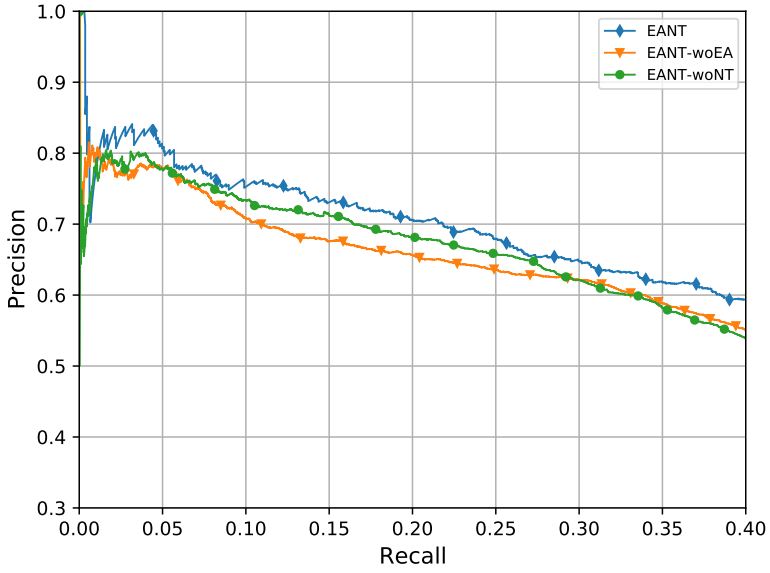


Figure 4. Comparison of EANT variant models.

4.6. Case Study

As discussed in Section 3, EANT can filter noise and re-label some of them by negative training. Table 4 shows some examples of the model-refined DSRE dataset. The first column indicates the sentence in the dataset, the second column denotes the original label, and the third column shows the decision made by the model for that sentence. The first sentence is

labeled as relation *"/people/person/profession"* in the original dataset, but we can see that it does not express this relation. The model calculates the confidence value of the sentence and identifies that it is below the threshold Th_d of the relation *"/people/person/profession"*, so the sentence is refined as *"NA"*. Similarly, the fourth sentence's confidence value is higher than the threshold Th_r of the relation *"/business/company/locations"*, so EANT relabeled such false-negative instance with the highest confidence relation label. From these examples, we can see the effectiveness of the model in distinguishing noisy data.

Table 4. Some examples of the model-refined NYT dataset.

Sentence	Sentence Label	Refined Label
Critic's notebook correction: march 18, 2005, friday a critic's notebook article on tuesday about gertrude stein misstated the title of an opera she wrote with virgil thomson.	*/people/person /profession*	NA
China should throw its influence as a global trading giant behind efforts to revive talks aimed at lowering trade barriers, susan c. schwab, the united states trade representative, said in Beijing.	*/location/location /contains*	NA
It is the museum of contemporary art San diego, which has sites in La Jolla and San diego, not the San diego museum.	*/location/neighborhood /neighborhood_of*	NA
In the decades he has spent in executive suites, Michael Armstrong was identified with some of the best-known totems of American corporate : I.B.M., hughes, at&t and Comcast.	NA	*/business/company /locations*
But the full extent of the death toll began to be known only on thursday after a reporter from china central television, or cctv, filed a story from Pingshi, Hunan.	NA	*/location/location /contains*

5. Conclusions

In this paper, we propose a new distant supervised relation extraction method, EANT, to alleviate the noise interference problem. EANT enhances the sentence representation by incorporating entity attribute information and filters out the noisy sentences in the dataset by negative training. The negative training selects complementary labels to train the model instead of the original labels. The obtained model can widen the gap between the confidence of clean and noisy data to achieve the noise filtering effect. Subsequently, for some potentially valid noisy sentences, we assign possible labels to them according to their confidence so we can make full use of them. The experimental results on the New York Times dataset show that our approach effectively improves the performance of the model for the relation extraction task.

In the future, we plan to explore pre-trained models to enhance sentence representation and capture useful semantic and syntactic properties of text through deep language models. This information is then used to adjust the entity attribute filtering strategy to reduce the impact of noisy entity attribute information.

Author Contributions: Conceptualization, X.C.; Data curation, X.C.; Formal analysis, X.C.; Funding acquisition, X.H.; Investigation, X.C.; Methodology, X.C.; Project administration, X.H.; Resources, X.C.; Software, X.C.; Supervision, X.H.; Validation, X.C.; Visualization, X.C.; Writing—original draft, X.C.; Writing—review & editing, X.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Key Research and Development Plan of China under Grant 2019YFB2012803, in part by the Key Project of Shanghai Science and Technology Innovation Action Plan under Grant 19DZ1100400 and Grant 18511103302, in part by the Key Program of Shanghai Artificial Intelligence Innovation Development Plan under Grant 2018-RGZN-02060, and in part by the Key Project of the "Intelligence plus" Advanced Research Fund of East China Normal University.

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

References

1. Cao, Y.; Wang, X.; He, X.; Hu, Z.; Chua, T.S. Unifying Knowledge Graph Learning and Recommendation: Towards a Better Understanding of User Preferences. In Proceedings of the World Wide Web Conference (WWW '19), San Francisco, CA, USA, 13–17 May 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 151–161. [CrossRef]
2. Wang, X.; He, X.; Cao, Y.; Liu, M.; Chua, T.S. KGAT: Knowledge Graph Attention Network for Recommendation. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining (KDD '19), Anchorage, AK, USA, 4–8 August 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 950–958. [CrossRef]
3. Fensel, D.; Şimşek, U.; Angele, K.; Huaman, E.; Kärle, E.; Panasiuk, O.; Toma, I.; Umbrich, J.; Wahler, A. Why We Need Knowledge Graphs: Applications. In *Knowledge Graphs: Methodology, Tools and Selected Use Cases*; Springer International Publishing: Cham, Switzerland, 2020; pp. 95–112. [CrossRef]
4. Cui, W.; Xiao, Y.; Wang, H.; Song, Y.; Hwang, S.w.; Wang, W. KBQA: Learning Question Answering over QA Corpora and Knowledge Bases. *Proc. VLDB Endow.* **2017**, *10*, 565–576. [CrossRef]
5. Huang, X.; Zhang, J.; Li, D.; Li, P. Knowledge Graph Embedding Based Question Answering. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19), Melbourne, Australia, 11–15 February 2019; Association for Computing Machinery: New York, NY, USA, 2019; pp. 105–113. [CrossRef]
6. 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 (SIGMOD '08), Vancouver, BC, Canada, 10–12 June 2008; Association for Computing Machinery: New York, NY, USA, 2008; pp. 1247–1250. [CrossRef]
7. Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; Ives, Z. DBpedia: A Nucleus for a Web of Open Data. In Proceedings of the Semantic Web, Busan, Korea, 11–15 November 2007; Aberer, K., Choi, K.S., Noy, N., Allemang, D., Lee, K.I., Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., et al., Eds.; Springer: Berlin/Heidelberg, Germany, 2007; pp. 722–735.
8. Vrandečić, D.; Krötzsch, M. Wikidata: A free collaborative knowledgebase. *Commun. ACM* **2014**, *57*, 78–85. [CrossRef]
9. Mintz, M.; Bills, S.; Snow, R.; Jurafsky, D. 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, Singapore, 2–7 August 2009; Association for Computational Linguistics: Suntec, Singapore, 2009; p. 1003. [CrossRef]
10. 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 (Volume 1: Long Papers), Berlin, Germany, 7–12 August 2016; Association for Computational Linguistics: Berlin, Germany, 2016; pp. 2124–2133. [CrossRef]
11. Ji, G.; Liu, K.; He, S.; Zhao, J. Distant Supervision for Relation Extraction with Sentence-Level Attention and Entity Descriptions. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, AAAI'17, San Francisco, CA, USA, 4–9 February 2017; AAAI Press: Palo Alto, CA, USA, 2017; pp. 3060–3066.
12. Alt, C.; Hübner, M.; Hennig, L. 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, Florence, Italy, 28 July–2 August 2019; Association for Computational Linguistics: Florence, Italy, 2019; pp. 1388–1398. [CrossRef]
13. Feng, J.; Huang, M.; Zhao, L.; Yang, Y.; Zhu, X. Reinforcement learning for relation classification from noisy data. In Proceedings of the 32nd AAAI Conference on Artificial Intelligence, AAAI 2018, New Orleans, LO, USA, 2–7 February 2018; McIlraith, S.A., Weinberger, K.Q., Eds.; AAAI Press: Palo Alto, CA, USA, 2018, pp. 5779–5786.
14. Qin, P.; Xu, W.; Wang, W.Y. Robust Distant Supervision Relation Extraction via Deep Reinforcement Learning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; Association for Computational Linguistics: Melbourne, Australia, 2018; pp. 2137–2147. [CrossRef]
15. Qin, P.; Xu, W.; Wang, W.Y. DSGAN: Generative Adversarial Training for Distant Supervision Relation Extraction. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; Association for Computational Linguistics: Melbourne, Australia, 2018; pp. 496–505. [CrossRef]
16. Jia, W.; Dai, D.; Xiao, X.; Wu, H. ARNOR: Attention Regularization based Noise Reduction for Distant Supervision Relation Classification. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; Association for Computational Linguistics: Florence, Italy, 2019; pp. 1399–1408. [CrossRef]

17. Zheng, S.; Han, X.; Lin, Y.; Yu, P.; Chen, L.; Huang, L.; Liu, Z.; Xu, W. 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, Florence, Italy, 28 July–2 August 2019; Association for Computational Linguistics: Florence, Italy, 2019; pp. 1419–1429. [CrossRef]
18. 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, 9–11 September 2017; Association for Computational Linguistics: Copenhagen, Denmark, 2017; pp. 1790–1795. [CrossRef]
19. Shang, Y.; Huang, H.Y.; Mao, X.L.; Sun, X.; Wei, W. Are Noisy Sentences Useless for Distant Supervised Relation Extraction? *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 8799–8806. [CrossRef]
20. Luo, B.; Feng, Y.; Wang, Z.; Zhu, Z.; Huang, S.; Yan, R.; Zhao, D. 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), Vancouver, BC, Canada, 30 July–4 August 2017; Association for Computational Linguistics: Vancouver, Canada, 2017; pp. 430–439. [CrossRef]
21. Kim, Y.; Yim, J.; Yun, J.; Kim, J. NLNL: Negative Learning for Noisy Labels. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; IEEE Computer Society: Los Alamitos, CA, USA, 2019; pp. 101–110. [CrossRef]
22. Riedel, S.; Yao, L.; McCallum, A. Modeling relations and their mentions without labeled text. In Proceedings of the Machine Learning and Knowledge Discovery in Databases, Athens, Greece, 5–9 September 2011; Balcázar, J.L., Bonchi, F., Gionis, A., Sebag, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 148–163. [CrossRef]
23. Jat, S.; Khandelwal, S.; Talukdar, P. Improving Distantly Supervised Relation Extraction using Word and Entity Based Attention. 2018. Available online: https://arxiv.org/abs/1804.06987 (accessed on 1 September 2022).
24. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. 2013. Available online: https://arxiv.org/abs/1301.3781 (accessed on 1 September 2022).
25. Pennington, J.; Socher, R.; Manning, C. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; Association for Computational Linguistics: Doha, Qatar, 2014; pp. 1532–1543. [CrossRef]
26. Zeng, D.; Liu, K.; Lai, S.; Zhou, G.; Zhao, J. 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; Dublin City University and Association for Computational Linguistics: Dublin, Ireland, 2014; pp. 2335–2344.
27. Bordes, A.; Usunier, N.; Garcia-Durán, A.; Weston, J.; Yakhnenko, O. Translating Embeddings for Modeling Multi-Relational Data. In Proceedings of the 26th International Conference on Neural Information Processing Systems—Volume 2, Lake Tahoe, NE, USA, 5–10 December 2013; Curran Associates Inc.: Red Hook, NY, USA, 2013; NIPS'13, pp. 2787–2795.
28. 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: Portland, OR, USA, 2011; pp. 541–550.
29. 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, 9–11 October 2010; Association for Computational Linguistics: Jeju Island, Korea, 2012; pp. 455–465.
30. Zeng, D.; Liu, K.; Chen, Y.; Zhao, 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, 17–21 September 2015; Association for Computational Linguistics: Lisbon, Portugal, 2015; pp. 1753–1762. [CrossRef]
31. Vashishth, S.; Joshi, R.; Prayaga, S.S.; Bhattacharyya, C.; Talukdar, P. RESIDE: Improving Distantly-Supervised Neural Relation Extraction using Side Information. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; Association for Computational Linguistics: Brussels, Belgium, 2018; pp. 1257–1266. [CrossRef]
32. Shi, Y.; Xiao, Y.; Quan, P.; Lei, M.; Niu, L. Distant Supervision Relation Extraction via adaptive dependency-path and additional knowledge graph supervision. *Neural Netw.* **2021**, *134*, 42–53. [CrossRef] [PubMed]