Distant Supervision for Relation Extraction with Hierarchical Attention-Based Networks

Anonymous ACL submission

Abstract

Distant supervision employs external knowledge bases to automatically label corpora. The labeled sentences in a corpus are usually packaged and trained for relation extraction using a multi-instance learning paradigm. The automated distant supervision inevitably introduces label noises. Previous studies that used sentence-level attention mechanisms to de-noise neither considered correlation among sentences in a bag nor correlation among bags. This paper proposes hierarchical attention-based networks that can de-noise at both sentence and bag levels. In the calculation of bag representation, we provide weights to sentence representations using sentence-level attention that considers correlations among sentences in each bag. Then, we employ bag-level attention to merge the similar bags by considering their correlations and to provide properer weights in the calculation of bag group representation. Experimental results on the New York Times datasets show that the proposed method outperforms the state-of-the-art ones.

1 Introduction

Relation extraction (RE) is a task that predicts attributes and relations for entities in sentences, which forms a foundation of many NLP applications such as structured search, sentiment analysis, and question answering. Conventional RE methods, such as (Zelenko et al., 2003; Culotta and Sorensen, 2004; Bunescu and Mooney, 2005), usually adopt fully supervised learning paradigm. To achieve good performance, these methods require a large well-labeled training corpus. However, labeling large corpora involves great economic and time costs, which prevents the real-world usage of these traditional supervised methods.

The emergence of distant supervision (Mintz et al., 2009) greatly promotes the corpus annotation, where free text in the corpus is aligned with an external knowledge base to generate labels automatically. Distant supervision holds an assumption that if two entities exhibit a relation, all sentences with these entities will express the same relation. Obviously, the assumption will result in quite a few wrong labels if distant supervision is directly applied to each sentence. To push distant supervision into real usage, researchers resort to multi-instance learning (Riedel et al., 2010; Hoffmann et al., 2011; Surdeanu et al., 2012), which learns relations at the bag level. That is, the sentences with the same entity pair are packaged, indicating that at least one sentence in the bag expresses the relation (defined by the entity pair). Consequently, the label of the bag is that relation. These methods aim to learn bag representations and relations associated with bags so that they are robust to whether or not a single sentence in the bag fully expresses the relation.

Bag	Sentence	Weight	Truth
b_1	s_1. The pope also issued emotional appeals to President **Saddam Hussein** of **Iraq** and ...	high	yes
	s_2. ... the removal of **Saddam Hussein** and the birth of democracy in ... **Iraq** worth the effort.	high	yes
	s_3. Nobody who experienced **Iraq** under the tyranny of **Saddam Hussein** could imagine...	low	yes
b_2	s_4. ... served as inspiration for **Hedi Slimane** 's Dior Homme show in **Paris**.	high	no

Table 1: The weight distribution of sentences in the bags expressing relation *place_of_birth*, calculated by existing attention models. "Yes" and "No" stand for whether or not each sentence actually expresses this relation. Thus, s_3 and s_4 have wrong weight assignment.

In order to generate bag representation more effectively, many recent studies (Lin et al., 2016; Ji et al., 2017; Liu et al., 2017; Alt et al., 2019; Yuan et al., 2019) used attention mechanisms to assign different weights to the sentences in a bag, indicating their probabilities of accurately expressing the relation assigned to the bag. The weights are usually calculated by considering every single sentence and the vector-represented relation. However, these methods with sentence-level attention mechanisms still have some defects.

First, they completely ignore associations be-

tween sentences in a bag, which may lead to incorrect weight assignment in bag representation calculation. For the example, in Table 1, if we use the method in (Lin et al., 2016), which only considers the sentence itself and the relation *place_of_birth*, to calculate the weights for sentences s_1, s_2, and s_3 in bag b_1, we will obtain a low weight for s_3, indicating that s_3 does not express the relation. However, if we consider the correlation between s_3 and s_1 (and s_2), we may obtain a different result, i.e., s_3 does express the relation, which is exactly the ground truth in the example. Second, they may have a wrong judgment of the relation for an entire bag when the bag has insufficient representation. Inappropriate weights can adversely affect the method (Yuan et al., 2019) that trains RE models at the bag-group level. In Table 1, since bag b_2 only has one sentence, it may be assigned a low weight by traditional methods. If we can augment its representation by adding more information from similar sentences, we may obtain a correct result.

To address the above issues, this paper proposes a novel hierarchical attention-based network for distant supervision RE, which considers the correlation among sentences and that among bags. Experimental results demonstrate the proposed model can obtain better representations for bags and bag groups, deriving a better RE performance. The contributions of the paper are three-fold:

- We propose a novel framework for distant supervision RE, which uses a hierarchical attention mechanism to generate better representations for bags and bag groups. The framework can easily incorporate different pre-trained sentence encoders.
- The proposed hierarchical attention runs at both sentence and bag levels, which uses a similarity-based principle to calculate the correlation among sentences and that among bags. The correlations derive proper weights for sentences and bags when generating bag and bag group representations, respectively.
- We conducted comprehensive experiments to show how the hierarchical attention works and the advantages of the proposed model against the state-of-the-art models.

2 Related Work

Relation extraction serves as a basic function for many NLP applications. Traditional supervised RE methods, such as (Zelenko et al., 2003; Culotta and Sorensen, 2004; Bunescu and Mooney, 2005), require a great deal of annotated data for model training, which is time-consuming and labor-intensive. Mintz et al. (2009) first proposed distant supervision for RE, which can automatically label corpora by aligning free text with external knowledge bases. Besides RE, distant supervision was also used in sentiment analysis (Go et al., 2009), part-of-speech tagging (Plank and Agić, 2018), and named entity recognition (Lee et al., 2016). However, it is accompanied by the wrong label problem. To alleviate the negative impact of mislabeled sentences, some studies (Riedel et al., 2010; Hoffmann et al., 2011; Surdeanu et al., 2012) employed multi-instance learning to package sentences with the same entity pair and learn bag representations.

Thereafter, representation learning-based methods made great progress. Zeng et al. (2015) combined multi-instance learning, convolutional neural networks, and segmented maximum pooling to build a mislabeled-sentence robust RE model. Miwa and Bansal (2016) added the sequence and structural information of dependency trees to the neural networks. Zhou et al. (2016) used an attention mechanism of bidirectional long-term short-term memory networks to capture the most important semantic information in the sentence. Vashishth et al. (2018) used additional side information from knowledge bases and employed graph convolution networks to encode syntactic information from text to improves performance of RE.

Recently, with the advantages of attention mechanism being known, it began to be used to build distant supervision RE models. In (Lin et al., 2016), a sentence-level attention was designed to score all sentences in a bag so as to evaluate their contributions to the bag representation. This scheme was widely used in (Ji et al., 2017; Qin et al., 2018b; Christou and Tsoumakas, 2021). Liu et al. (2017) proposed a soft-label method to reduce the influence of mislabeled sentences. Unlike our proposed method, all the above methods ignored the correlation between instances, resulting in the loss of supervision information. Yuan et al. (2019) proposed a selective attention for bag representation and a cross-bag attention for bag-group representation. Unlike our proposed method, their method cannot solve poor bag representation problem (i.e., the bag has inadequate sentences) for few-sentence bags in the calculation of bag-group representation.

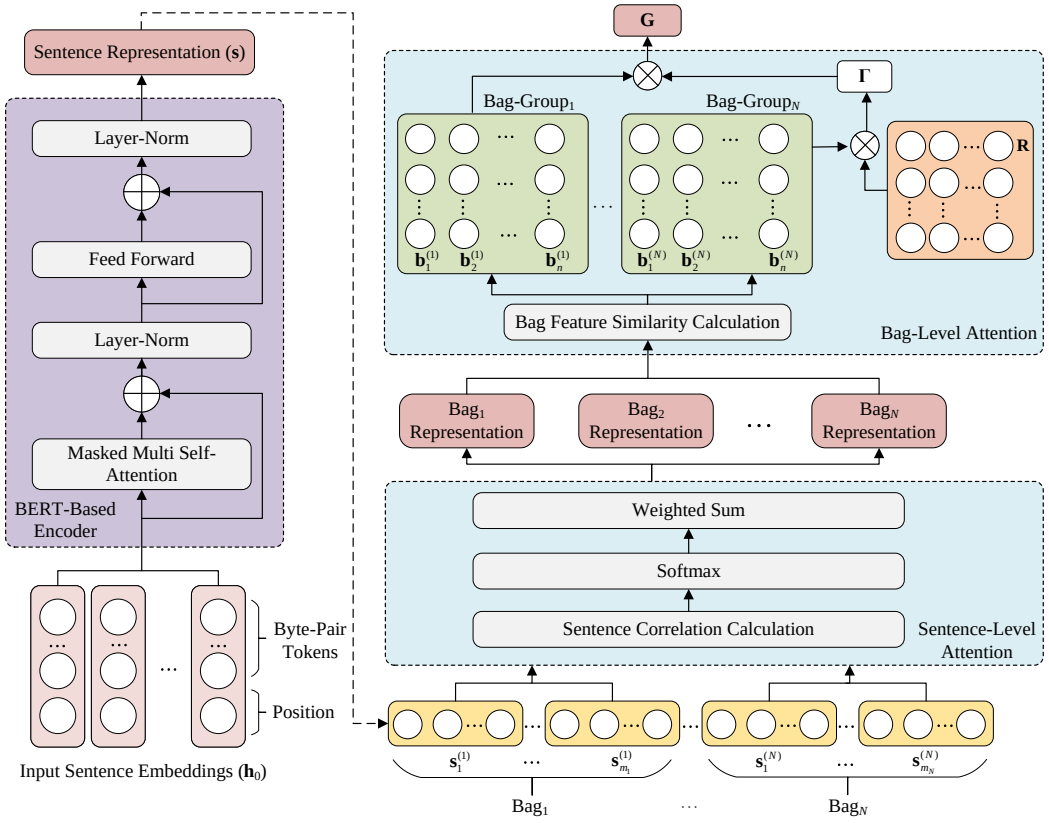

Figure 1: The framework of the proposed hierarchical attention-based model.

3 The Proposed Method

In this section, we first present the problem statement and the proposed framework. Then, we go to the details of our solution.

3.1 Problem Statement and Framework

Problem Statement This study still follows the multi-instance learning paradigm. A set of sentences with the same entity pair $< h, t >$ forms a bag. A training set contains N bags and each bag associates a relation $\mathbf{r}$. A relation is represented as a H-dimensional random vector. There are totally K relations in the training set, forming a relation embedding matrix $\mathbf{R} = [\mathbf{r}_1, \cdots, \mathbf{r}_K]$. We group n similar bags together to form a bag group $\mathbf{g}$. Our goal is to learn bag-group representations $\mathbf{G}$ from input sentences (with labels obtained by distant supervision). The learned model can predict relations for unlabeled bags with entity pairs.

Framework The framework of our proposed hierarchical attention-based model is shown in Figure 1. Our model mainly has three components:

- *Sentence Encoder* accepts the basic input sentence embeddings (consisting of byte-pair tokens and positions) to generate more effective feature representations through some pre-trained language models, such as BERT (Devlin et al., 2018) or PCNN (Zeng et al., 2015). In Figure 1, the Sentence Encoder is implemented as a BERT-based one, which performs the best in our experiments.

- *Sentence-Level Attention* aims to learn representations for bags, where each bag consists of a certain number of sentences with the same entity pair associated with some relation. It considers the correlation among sentences in a bag and assigns them proper weights in the calculation of bag representation.

- *Bag-Level Attention* aims to learn representations for bag groups, where each bag group consists of a certain number of similar bags with the same relation. In the calculation of bag-group representation, it assigns a weight to each bag in a bag group by evaluating the degree of the bag matching the label (i.e., a relation vector) of the bag group.

3.2 Sentence Encoder

Following the usual practice of distant supervision RE, like (Christou and Tsoumakas, 2021), we need

to construct a low-dimensional distributed representation for a sentence by concatenating the relation (conveyed by an entity pair in the sentence) and the sentence embedding (obtained from a pre-trained language model such as BERT).

Input Embeddings The input embedding $\mathbf{h}_0$ for a sentence that will be fed into a pre-trained language model is created by summing over the positional and byte-pair embeddings for each token in the structured input.

We tokenize input using byte-pair encoding (Sennrich et al., 2016) to make use of sub-word information. First we learn the most frequent character sub-strings in all words from the corpus, and then merge these frequent character sub-strings into a dictionary. Then, we add a positional embedding (Vaswani et al., 2017) to the rear of the byte-pair tokens to form the input embedding $\mathbf{h}_0$ for a sentence. The reason is that some pre-trained models (such as BERT) use the transformer encoder to learn representations of sentences through the self-attention mechanism. Usually, their self-attention mechanism does not pay attention to the position information of tokens. Adding position information into the input allows the transformer to learn a better representation.

Sentence Encoder Output The input embedding $\mathbf{h}_0$ for a sentence is further converted into a feature vector through a BERT-based model. That is, the BERT model is fine-tuned for distant supervision RE as follows: Unlike the common practice to represent a sentence by the [CLS] vector $\mathbf{h}_L$ in the last hidden layer of BERT, we need to reweight the tokens in $\mathbf{h}_L$ to obtain a better vector $\mathbf{h}'_L$ using the relation-attention mechanism in (Christou and Tsoumakas, 2021). The relation-attention uses a relation embedding $\mathbf{l}$ generated by the TransE model (Bordes et al., 2013) to adjust $\mathbf{h}_L$, emphasizing those tokens that are more relevant to the relation. Then, we concatenate $\mathbf{l}$ and $\mathbf{h}'_L$ to obtain the sentence representation as follows:

$$\tilde{\mathbf{s}} = [\mathbf{l}; \mathbf{h}'_L]. \tag{1}$$

Here, the dimensions of $\mathbf{l}$ and $\mathbf{h}'_L$ are $H/2$. Thus, the dimension of $\tilde{\mathbf{s}}$ is H, which is the same as that of a relation $\mathbf{r}$. To prevent the influence of vector length (modulus), the final sentence representation (i.e., the output of Sentence Encoder) is normalized to a unit length as follows:

$$\mathbf{s} = \tilde{\mathbf{s}}/\|\tilde{\mathbf{s}}\|_2. \tag{2}$$

3.3 Sentence-Level Attention

The Sentence-Level Attention component follows a multi-instance learning scheme that encloses a set of sentences that have the same entity pair into a bag. As Figure 1 shows, bags may have different numbers of sentences. Suppose bag i has m_i sentences denoted by $\{\mathbf{s}_1^{(i)}, \mathbf{s}_2^{(i)}, \cdots, \mathbf{s}_{m_i}^{(i)}\}$. The representation for bag i is computed as a weighted sum of all sentence vectors in it as follows:

$$\mathbf{b}_i = \sum_{j=1}^{m_i} \alpha_j^{(i)} \mathbf{s}_j^{(i)}, \tag{3}$$

where $\alpha_j^{(i)}$ is an attention weight assigned to the j-th sentence in bag $\mathbf{b}_i$. To obtain $\alpha_j^{(i)}$, we first calculate an overall similarity of a sentence against the other sentences in the bag as follows:

$$e_j^{(i)} = \sum_{j'=1,\ldots,m_i \wedge j' \neq j} \mathbf{s}_j^{(i)} {\mathbf{s}_{j'}^{(i)}}^{\mathrm{T}}. \tag{4}$$

Here, we applied the inner product similarity to a pair of sentence vectors. Since the sentence vectors are normalized, the similarity can be simplified as their inner product. (Note that all vectors in the paper are row vectors.) Then, $\alpha_j^{(i)}$ is a normalized similarity of $e_j^{(i)}$, calculated by applying the softmax function as follows:

$$\alpha_j^{(i)} = \frac{\exp\left(e_j^{(i)}\right)}{\sum_{j'=1}^{m_i} \exp\left(e_{j'}^{(i)}\right)}. \tag{5}$$

If a sentence $\mathbf{s}_j^{(i)}$ is more similar to the other sentences in the bag, it will be assigned a greater weight when computing the bag representation $\mathbf{b}_i$. Finally, the output of Sentence-Level Attention is the representation for all bags.

3.4 Bag-Level Attention

As the example in the introduction section shows, in distant supervision RE noises not only exist at the sentence level but also exist at the bag level, i.e., the bag is assigned a wrong weight, opposed to its ground truth. To address this issue, we employ an attention mechanism at the bag level. Based on the assumption that bags expressing the same relation should have similar bag representations, we intend to enclose these similar bags into a bag group, which is thought to provide enhanced features for building RE models. Therefore, the goal of Bag-Level Attention is to generate bag-group

presentations, during which bags in a bag group will be assigned different weights.

Suppose we have N bags $\{\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_N\}$. (In practice, N is also the batch size in training.) We consider each bag one after another. For bag $\mathbf{b}_i$, we calculate the inner product similarity of the other bags against it:

$$sim_j(\mathbf{b}_j, \mathbf{b}_i) = \mathbf{b}_j {\mathbf{b}_i}^{\mathrm{T}}, j = 1, \cdots, N \wedge j \neq i. \tag{6}$$

These $N-1$ similarities are sorted in descending order. The bags (indexed by j in Eq.(6)) corresponding to the top-$(n-1)$ similarities and the base bag (indexed by i) are enclosed to a bag group, denoted by $\{\mathbf{b}_1^{(q)}, \mathbf{b}_2^{(q)}, \ldots, \mathbf{b}_n^{(q)}\}$, where $q = 1, \cdots, N$ is the index of the bag group. The label (relation) of the group is the same as that of the base bag.

The bag-label attention mechanism assigns different weights to the bags in a bag group, indicating their contributions to the bag-group presentation. Similarly, bag-group presentation $\mathbf{g}_q$ is computed as follows:

$$\mathbf{g}_q = \sum_{i=1}^{n} \gamma_i^{(q)} \mathbf{b}_i^{(q)}. \tag{7}$$

Note that $\mathbf{g}_q$ is the q-th row of matrix $\mathbf{G}$ and all vectors $\boldsymbol{\gamma}^{(q)} (q = 1, \cdots, N)$ compose an attention weight matrix $\boldsymbol{\Gamma}$, as shown in Figure 1.

To obtain $\gamma_i^{(q)}$, we first evaluate the confidence of labeling bag $\mathbf{b}_i^{(q)}$ with the label of bag group $\mathbf{g}_k$ as follows:

$$u_i^{(q)} = \mathbf{r}_q {\mathbf{b}_i^{(q)}}^{\mathrm{T}}, \tag{8}$$

where $\mathbf{r}_q$ is the vector representation of the relation label of $\mathbf{g}_q$. (Note that all vectors here are H-dimensional.) Then, $\gamma_i^{(q)}$ is a normalized confidence of $u_i^{(q)}$, calculated by applying the softmax function as follows:

$$\gamma_i^{(q)} = \frac{\exp\left(u_i^{(q)}\right)}{\sum_{i'=1}^{n} \exp\left(u_{i'}^{(q)}\right)}. \tag{9}$$

Finally, the output of Bag-Level Attention is the representation ($\mathbf{G}$) for all bag groups.

3.5 Model Training and Prediction

We can add a full connection layer at the tail of the above model to realize model training. The objective function is set to the negative log likelihood at the bag-group level as follows:

$$\mathrm{J}(\boldsymbol{\theta}) = -\sum_{q=1}^{N} \log \mathrm{p}(\mathbf{r}_q \mid \mathbf{g}_q; \boldsymbol{\theta}), \tag{10}$$

	Sentences	Entity Pairs	Triplets
train	522,611	281,270	18,252
test	172,448	96,678	1,950

Table 2: Details of the NYT dataset

where N is the number of the bag groups in the training set, $\mathbf{r}_q$ is the label of a bag group $\mathbf{g}_q$, and $\boldsymbol{\theta}$ is the set of model parameters. The training process minimizes the objective function $\mathrm{J}(\boldsymbol{\theta})$) through mini-batch stochastic gradient descent (SGD).

When making prediction, the score o_k of classifying bag group g into relation $\mathbf{r}_k$ is calculated as follows:

$$o_k = \mathbf{g} {\mathbf{r}_k}^{\mathrm{T}} + d, \tag{11}$$

where d is a bias term. Finally, a softmax function is employed to obtain the probability that the bag group $\mathbf{g}$ is classified into relation $\mathbf{r}_k$ as follows:

$$\mathrm{p}(\mathbf{r}_k \mid \mathbf{g}) = \frac{\exp\left(o_k\right)}{\sum_{k'=1}^{K} \exp\left(o_{k'}\right)}, \tag{12}$$

where K is the total number of relation types.

4 Experiment

4.1 Dataset

The New York Times (NYT) dataset was used in our experiments. This dataset was first released by Riedel et al. (2010). It was widely used in the studies of distant supervision RE. The details of the dataset are listed in Table 2. The dataset contains a total of 52+1 ("1" for N/A) relation types and is divided into a training set and a test set. The training set contains 522,611 sentences, 281,270 entity pairs, and 18,252 relational facts [1]. The test set contains 172,448 sentences, 96,678 entity pairs, and 1,950 relations. The division of training and test sets in our experiments is the same as they were in previous studies (Mintz et al., 2009; Hoffmann et al., 2011; Surdeanu et al., 2012; Lin et al., 2016; Liu et al., 2017; Qin et al., 2018a,b; Vashishth et al., 2018; Christou and Tsoumakas, 2021).

4.2 Evaluation Metrics

Following the previous work (Lin et al., 2016; Ji et al., 2017; Christou and Tsoumakas, 2021), we evaluate our model through three metrics: Precision/Recall (PR) curve, Area Under the ROC Curve (AUC), and Precision@N (P@N).

[1] A relational fact is also called a triplet (in Table 2), which is the combination of a relation and an entity pair.

Parameter Name	Value	Candidate set
max _ seq _ length	64	{32,64,128}
batch size	32	{8,16,32,64}
epochs	3	{2,3,4,5}
learning rate	$2e^{-5}$	$\{2e^{-5}, 2e^{-4}\}$
dropout	0.4	{0.2,0.3,0.4,0.5}
weight _ decay	0.001	{0.01,0.001}
bag group size	5	{3,4,5,6}

Table 3: Parameter settings for our proposed method.

4.3 Parameter Settings

We referred to (Christou and Tsoumakas, 2021) to set the parameters for our proposed model. The values of the parameters were selected from their candidates by grid searching, shown in Table 3. For the existing methods used in comparison, we set their parameters the same as the values deriving the best performance reported in the original articles.

4.4 Variants of the Proposed Method

Since our proposed framework can be implemented in different ways, we will have four variants of the proposed method: 1) BERTenc+HSATT+HBATT: This is the sophisticated version of the proposed method, which a fine-turned BERT (BERTenc) (Devlin et al., 2018) is used as the sentence encoder, our proposed sentence-level attention (HSATT) and bag-level attention (HBATT) are used. 2) PCNN+HSATT+HBATT: BERTenc is substituted by a piece-CNNs encoder (Zeng et al., 2015). 3) BERTenc+HSATT: The bag-level attention is not implemented. 4) PCNN+HSATT: The bag-level attention is not implemented and the sentence encoder is PCNN. In addition, we add two methods BERTenc+ATT and PCNN+ATT as the baselines, where ATT is the attention method proposed by Lin et al. (2016). The comparison of these six methods serves as an ablation experiment.

The experimental results are shown in Figures 2, 3, and Table 4. We have the following observations: 1) No matter what sentence encoders are used, our HSATT performs better than ATT. This reveals that it is essential to consider the correlation among sentences in the bag. 2) For both sentence encoders, the models with HBATT achieve better performance than the ones without HBATT. It can be attributed to the reason as follow: our HSATT can only utilize the correlation among sentences in a bag to assign higher weights to those correctly labeled sentences. However, for the bag with a few sentences, the sentence weights still may be incorrectly assigned because of the insufficient features of the bag. Therefore, it is effective to introduce

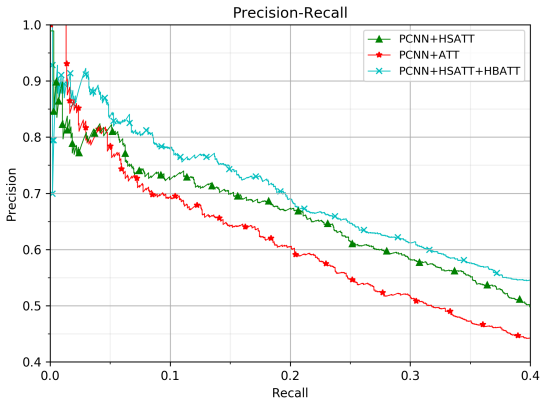

Figure 2: PR curves of the three models using PCNN.

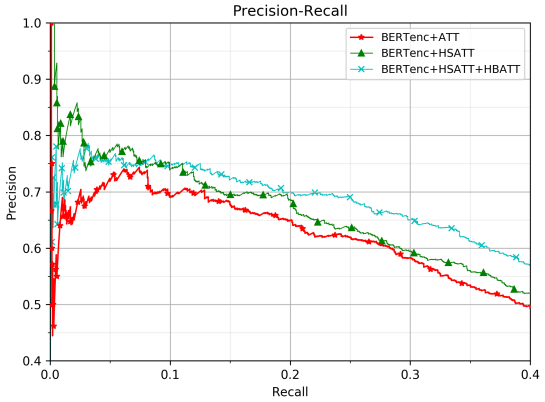

Figure 3: PR curves of three models using BERTenc.

our HBATT method. 3) Comparing Figures 2 and 3, BERTenc works better than PCNN, which may be attributed to the bidirectionality of BERTenc so that it can efficiently capture head and tail interaction. 4) BERTenc+HSATT+HBATT achieves the best AUC of 0.454. Compared with BERTenc+ATT, the AUC increased by 4.8%.

4.5 Comparison with Previous Work

Our sophisticated BERTenc+HSATT+HBATT model is further compared with eight state-of-the-art models. The eight models are briefly summarized as follows: 1) Mintz (Mintz et al., 2009) is the earliest distant supervision model to solve the relation extraction problem. 2) MultiR (Hoffmann et al., 2011) is a multi-instance learning model that combines a sentence-level extraction model with a simple corpus-level component to aggregate single facts. 3) MIML (Surdeanu et al., 2012) is the first RE method to jointly model multiple sentences (by modeling the potential labels assigned to the sentences) and multiple labels (by providing a simple method to capture the dependencies between labels). 4) PCNN+ATT (Lin et al., 2016)

Model	AUC
PCNN+ATT	0.386
PCNN+HSATT	0.414
PCNN+HSATT+HBATT	0.428
BERTenc+ATT	0.406
BERTenc+HSATT	0.440
BERTenc+HSATT+HBATT	**0.454**

Table 4: AUC of our four variants and two baselines.

employs a selective attention mechanism over multiple sentences to alleviate the mislabeling problem, which serves as a principle baseline of our proposed model. 5) PCNN+ATT+soft-label (Liu et al., 2017) introduces an entity pair level de-noising method, i.e., the soft label method, which can dynamically correct incorrect labels during the training process. 6) PCNN+ATT+RL (Qin et al., 2018b) introduces a deep reinforcement learning strategy to generate the false-positive indicator. 7) RESIDE (Vashishth et al., 2018) utilizes additional side information from knowledge bases to improve relation extraction. 8) REDSandT (Christou and Tsoumakas, 2021) is a recent transformer-based relation extraction model for distant supervision, which can recognize relations that other methods fail to detect, including the long-tail relations.

The evaluation metrics for comparison are still RP curve and AUC.

4.5.1 Performance in terms of PR curve

The comparison results of nine models in terms of PR curve are shown in Figure 4. Because in the articles of PCNN+ATT (Lin et al., 2016) and PCNN+ATT+soft-label (Liu et al., 2017) the authors only plotted the first 2,000 points on the PR curves, for a fair comparison, we also plot the first 2,000 points. We have the following observations: 1) Mintz, MultiR, and MIMLRE are probabilistic methods, the others are all NN-based methods. Obviously, the NN-based methods outperform the probabilistic ones, which indicates that human-designed features are usually worse than the features automatically extracted by neural networks. 2) Our BERTenc+HSATT+HBATT model performs the best against the other models, which shows the effectiveness of our hierarchical attention networks. 3) Our model shows a more stable performance. When recall is small, PCNN+ATT+soft-label performs better. When recall is higher than 0.18, our model always performs the best.

4.5.2 Peformance in terms of AUC

Comparisons of PR curves above have shown that methods with older ages have significantly

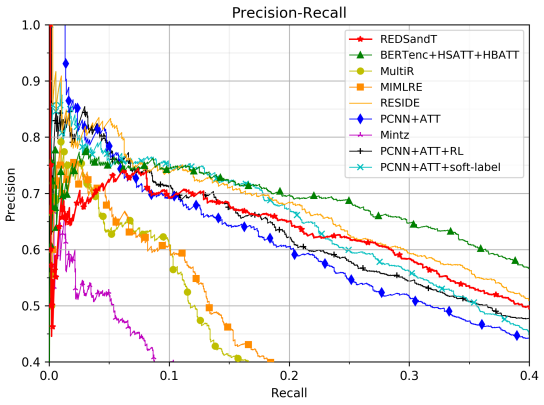

Figure 4: PR curves of nine models in comparison.

Model	AUC
PCNN+ATT+Gan	0.264
PCNN+ATT+RL	0.271
REDSandT	0.312
BERT-based+HSATT+HBATT	**0.334**

Table 5: AUC of four models in comparison.

lower performance. Therefore, we only list the comparison results in terms of AUC between our method and other three recently published methods (i.e., PCNN+ATT+GAN (Qin et al., 2018a), PCNN+ATT+RL (Qin et al., 2018b), and REDSandT (Christou and Tsoumakas, 2021)) in Table 5. Again, our proposed BERTenc+HSATT+HBATT model significantly outperforms the others.

4.6 Effect of the Number of Sentences

In the original test dataset, there are 74,857 bags with only one sentence, accounting for almost 3/4 of all bags. To evaluate the effect of the number of sentences in a bag in the test set, we compare our four variants of the proposed method with four baselines (PCNN+ATT, PCNN+ATT+soft-label, RESIDE, REDSandT) under three different number-of-sentence settings. First, we group the sentences with the same entity pair together. Then, we randomly select a certain number of sentences to form bags in the test set. The numbers of sentences are set to One, Two, and All sentences in the group. Thus, we have three test sets. In this experiment, we use P@100, P@200, P@300, and their mean value as the evaluation metrics. The metric P@N measures the precision of the top-N results in the test set with highest probabilities of belonging some class of relations. The experimental results are listed in Table 6.

We have the following observations: 1) Our proposed model achieves the highest P@N against

Testing set	One				Two				All			
P@N (%)	**100**	**200**	**300**	**mean**	**100**	**200**	**300**	**mean**	**100**	**200**	**300**	**mean**
PCNN+ATT	73.3	69.2	60.8	67.8	77.2	71.6	66.1	71.6	76.2	73.1	67.4	72.2
PCNN+ATT+soft-label	84.0	75.5	68.3	75.9	86.0	77.0	73.3	78.8	87.0	84.5	77.0	82.8
PCNN+HSATT	85.9	74.8	69.1	76.6	89.0	83.2	76.7	83.0	89.2	86.3	79.1	84.9
PCNN+HSATT+HBATT	**87.1**	76.2	70.2	77.8	87.9	85.1	**77.3**	**83.4**	90.1	87.0	79.5	85.5
RESIDE	80.0	75.5	69.3	74.9	83.0	73.5	70.6	75.7	84.0	78.5	75.6	79.4
REDSandT	78.0	74.2	72.5	74.9	80.6	75.3	72.1	76.0	81.2	72.5	67.8	73.8
BERTenc+HSATT	86.1	78.2	70.2	78.2	87.3	**86.7**	75.7	83.2	88.0	83.5	78.6	83.4
BERTenc+HSATT+HBATT	86.0	**81.3**	**78.7**	**82.0**	85.3	83.2	76.1	81.5	**91.0**	**87.3**	**90.1**	**89.5**

Table 6: P@N values of the entity pairs with different numbers of test sentences.

Bag-group Label: /location/location/contains			
Triple (Bag)	**Sentence**	**Sentence Attention**	**Bag Attention**
<queens,contains, Belle Harber>	She is a daughter of Marion I. Rabbin and Dr. Murvin Rabbin of Belle Harbor , Queens	high	high
	....at St. Francis de Sales Roman Catholic Church , in Belle Harbor , Queens , the parish of his birth	high	
	...St. Francis de Sales Roman Catholic Church in Belle Harbor ; another board studded with ... , Queens	low	
<Tennessee,contains, White House>	When he won the White House in 1844 , James K. Polk did not carry ...governor , Tennessee	high	low

Table 7: A case study for two bags and their corresponding bag group.

the other models. 2) On test set One, our BERTenc+HSATT+HBATT achieves a great improvement than the ones without HBATT. Because the feature fusion mechanism of HBATT is specially designed for few-sentence bags. Even if there is only one sentence for each test entity pair in a bag, the model can still have the desired effect. 3) Comparing BERTenc-based methods with PCNN-based methods, BERTenc-based one always performs better, which indicates that BERTenc can learn better semantic information. 4) As the number of sentences increases, the performance of all models improves. This is because test sets One and Two randomly select one and two sentences from each bag, respectively, which greatly reduces the information contained in each bag. Therefore, the performance of selecting all sentences for a test bag significantly improves.

4.7 Case Study

Table 7 shows a test example of bag group labeled */location/location/contains*, which contains two bags. One is for triple <Queens, contains, Belle Harbor> and the other is for triple <Tennessee, contains, White House>. We observe that two sentences are correctly labeled in the first sentence bag. The second bag has only one sentence incorrectly labeled because two entities **Tennessee** and **White House** in the sentence does not express the relation */location/location/contains*. Thus, the sentence-level attention fails to handle the second bag and assigns a high attention weight to it. However, our method designs an attention mechanism at a higher bag level, which will assign a low attention weight (the last column in Table 7) to this one-sentence bag by calculating the similarity between the bags in the group and the bag-group label.

Therefore, our proposed model can make full use of the supervision information of all correctly labeled sentences and pay more attention to those of higher quality bags, which helps improve the performance of relation extraction.

5 Conclusion

This paper proposes hierarchical attention-based networks for distant supervision RE, which employs sentence-level and bag-level attentions to address the noisy data problem. The sentence-level attention calculates the correlation among sentences in a bag and assigns higher weights to the correctly labeled sentences in the generation of bag representation. The bag-level attention encloses bags into bag groups and assigns proper weights to bags by calculating their similarity in the generation of bag-group representation. With a fine-tuned BERT as a sentence encoder in front of the above two attentions, our model can generates a better bag-group representation and exhibits the highest performance on the NYT dataset, compared with the state-of-the-art models.

References

Christoph Alt, Marc Hübner, and Leonhard Hennig. 2019. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. In *ACL*, pages 1388–1398.

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.

Razvan Bunescu and Raymond J Mooney. 2005. Subsequence kernels for relation extraction. In *NIPS*, pages 171–178. Citeseer.

Despina Christou and Grigorios Tsoumakas. 2021. Improving distantly-supervised relation extraction through bert-based label and instance embeddings. *IEEE Access*, pages 62574–62582.

Aron Culotta and Jeffrey Sorensen. 2004. Dependency tree kernels for relation extraction. In *ACL*, pages 423–429.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009.

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.

Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *AAAI*, pages 3060–3066.

Sunghee Lee, Yeongkil Song, Maengsik Choi, and Harksoo Kim. 2016. Bagging-based active learning model for named entity recognition with distant supervision. In *2016 International conference on big data and smart computing (BigComp)*, pages 321–324. IEEE.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *ACL (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 *EMNLP*, pages 1790–1795.

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.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using lstms on sequences and tree structures. In *ACL (Volume 1: Long Papers)*, pages 1105–1116.

Barbara Plank and Željko Agić. 2018. Distant supervision from disparate sources for low-resource part-of-speech tagging. In *EMNLP*, pages 614–620.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018a. Dsgan: Generative adversarial training for distant supervision relation extraction. In *ACL (Volume 1: Long Papers)*, pages 496–505.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018b. Robust distant supervision relation extraction via deep reinforcement learning. In *ACL (Volume 1: Long Papers)*, pages 2137–2147.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *ACL (Volume 1: Long Papers)*, pages 1715–1725.

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.

Shikhar Vashishth, Rishabh Joshi, Sai Suman Prayaga, Chiranjib Bhattacharyya, and Partha Talukdar. 2018. Reside: Improving distantly-supervised neural relation extraction using side information. In *EMNLP*, pages 1257–1266.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Yujin Yuan, Liyuan Liu, Siliang Tang, Zhongfei Zhang, Yueting Zhuang, Shiliang Pu, Fei Wu, and Xiang Ren. 2019. Cross-relation cross-bag attention for distantly-supervised relation extraction. In *AAAI*, pages 419–426.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *Journal of machine learning research*, 3(Feb):1083–1106.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *EMNLP*, pages 1753–1762.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *ACL (volume 2: Short papers)*, pages 207–212.