# Distant Supervision Relation Extraction with Intra-Bag and Inter-Bag Attentions

**Zhi-Xiu Ye**
University of Science and Technology of China
zxye@mail.ustc.edu.cn

**Zhen-Hua Ling**
University of Science and Technology of China
zhling@ustc.edu.cn

## Abstract

This paper presents a neural relation extraction method to deal with the noisy training data generated by distant supervision. Previous studies mainly focus on sentence-level de-noising by designing neural networks with intra-bag attentions. In this paper, both intra-bag and inter-bag attentions are considered in order to deal with the noise at sentence-level and bag-level respectively. First, relation-aware bag representations are calculated by weighting sentence embeddings using intra-bag attentions. Here, each possible relation is utilized as the query for attention calculation instead of only using the target relation in conventional methods. Furthermore, the representation of a group of bags in the training set which share the same relation label is calculated by weighting bag representations using a similarity-based inter-bag attention module. Finally, a bag group is utilized as a training sample when building our relation extractor. Experimental results on the New York Times dataset demonstrate the effectiveness of our proposed intra-bag and inter-bag attention modules. Our method also achieves better relation extraction accuracy than state-of-the-art methods on this dataset[1].

## 1 Introduction

Relation Extraction is a fundamental task in natural language processing (NLP), which aims to extract semantic relations between entities. For example, sentence "[*Barack Obama*]$_{e1}$ was born in [*Hawaii*]$_{e2}$" expresses the relation *BornIn* between entity pair **Barack Obama** and **Hawaii**.

Conventional relation extraction methods, such as (Zelenko et al., 2002; Culotta and Sorensen, 2004; Mooney and Bunescu, 2006), adopted supervised training and suffered from the lack of large-scale manually labeled data. To address this issue, the distant supervision method (Mintz et al., 2009) was proposed, which generated the data for training relation extraction models automatically. The distant supervision assumption says that if two entities participate in a relation, **all** sentences that mention these two entities express that relation. It is inevitable that there exists noise in the data labeled by distant supervision. For example, the precision of aligning the relations in Freebase to the New York Times corpus was only about 70% (Riedel et al., 2010).

| bag | sentence | correct? |
|---|---|---|
| B1 | S1. **Barack Obamna** was born in **the United States**. | **Yes** |
| | S2. **Barack Obamna** was the 44th president of **the United States** | **No** |
| B2 | S3. **Kyle Busch** , a **Las Vegas** resident who ran second to Johnson last year, finished third, followed by Kasey Kahne, Jeff Gordon and mark martin . | **No** |
| | S4. Hendrick drivers finished in three of the top four spots at **Las Vegas** , including **Kyle Busch** in second and ... | **No** |

Table 1: Examples of sentences with relation *place_of_birth* annotated by distant supervision, where "Yes" and "No" indicate whether or not each sentence actually expresses this relation.

Thus, the relation extraction method proposed in (Riedel et al., 2010) argued that the distant supervision assumption was too strong and relaxed it to *expressed-at-least-once* assumption. This assumption says that if two entities participate in a relation, **at least one sentence** that mentions these two entities might express that relation. An example is shown by sentences S1 and S2 in Table 1. This relation extraction method first divided the training data given by distant supervision into bags where each bag was a set of sentences containing the same entity pair. Then, bag representations were derived by weighting sentences within

[1]The code is available at https://github.com/ZhixiuYe/Intra-Bag-and-Inter-Bag-Attentions.

each bag. It was expected that the weights of the sentences with incorrect labels were reduced and the bag representations were calculated mainly using the sentences with correct labels. Finally, bags were utilized as the samples for training relation extraction models instead of sentences.

In recent years, many relation extraction methods using neural networks with attention mechanism (Lin et al., 2016; Ji et al., 2017; Jat et al., 2018) have been proposed to alleviate the influence of noisy training data under the *expressed-at-least-once* assumption. However, these methods still have two deficiencies. First, only the target relation of each bag is utilized to calculate the attention weights for deriving bag representations from sentence embeddings at training stage. Here we argue that the bag representations should be calculated in a relation-aware way. For example, the bag B1 in Table 1 contains two sentences S1 and S2. When this bag is classified to relation *BornIn*, the sentence S1 should have higher weight than S2, but when classified to relation *PresidentOf*, the weight of S2 should be higher. Second, the *expressed-at-least-once* assumption ignores the **noisy bag problem** which means that all sentences in one bag are incorrectly labeled. An example is shown by bag B2 in Table 1.

In order to deal with these two deficiencies of previous methods, this paper proposes a neural network with multi-level attentions for distant supervision relation extraction. At the instance/sentence-level, i.e., intra-bag level, all possible relations are employed as queries to calculate the relation-aware bag representations instead of only using the target relation of each bag. To address the noisy bag problem, a bag group is adopted as a training sample instead of a single bag. Here, a bag group is composed of bags in the training set which share the same relation label. The representation of a bag group is calculated by weighting bag representations using a similarity-based inter-bag attention module.

The contributions of this paper are threefold. First, an improved intra-bag attention mechanism is proposed to derive relation-aware bag representations for relation extraction. Second, an inter-bag attention module is introduced to deal with the noisy bag problem which is ignored by the expressed-at-least-once assumption. Third, our methods achieve better relation extraction accuracy than state-of-the-art models on the widely used New York Times (NYT) dataset (Riedel et al., 2010).

## 2 Related Work

Some previous work (Zelenko et al., 2002; Mooney and Bunescu, 2006) treated relation extraction as a supervised learning task and designed hand-crafted features to train kernel-based models. Due to the lack of large-scale manually labeled data for supervised training, the distant supervision approach (Mintz et al., 2009) was proposed, which aligned raw texts toward knowledge bases automatically to generate relation labels for entity pairs. However, this approach suffered from the issue of noisy labels. Therefore, some subsequent studies (Riedel et al., 2010; Hoffmann et al., 2011; Surdeanu et al., 2012) considered distant supervision relation extraction as a multi-instance learning problem, which extracted relation from a bag of sentences instead of a single sentence.

With the development of deep learning techniques (LeCun et al., 2015), many neural-network-based models have been developed for distant supervision relation extraction. Zeng et al. (2015) proposed piecewise convolutional neural networks (PCNNs) to model sentence representations and chose the most reliable sentence as the bag representation. Lin et al. (2016) employed PCNNs as sentence encoders and proposed an intra-bag attention mechanism to compute the bag representation via a weighted sum of all sentence representations in the bag. Ji et al. (2017) adopted a similar attention strategy and combined entity descriptions to calculate the weights. Liu et al. (2017) proposed a soft-label method to reduce the influence of noisy instances. All these methods represented a bag with a weighted sum of sentence embeddings, and calculated the probability of the bag being classified into each relation using the same bag representation at training stage. In our proposed method, intra-bag attentions are computed in a relation-aware way, which means that different bag representations are utilized to calculate the probabilities for different relation types. Besides, these existing methods focused on intra-bag attentions and ignored the noisy bag problem.

Some data filtering strategies for robust distant supervision relation extraction have also been proposed. Feng et al. (2018) and Qin et al. (2018b) both employed reinforcement learning to

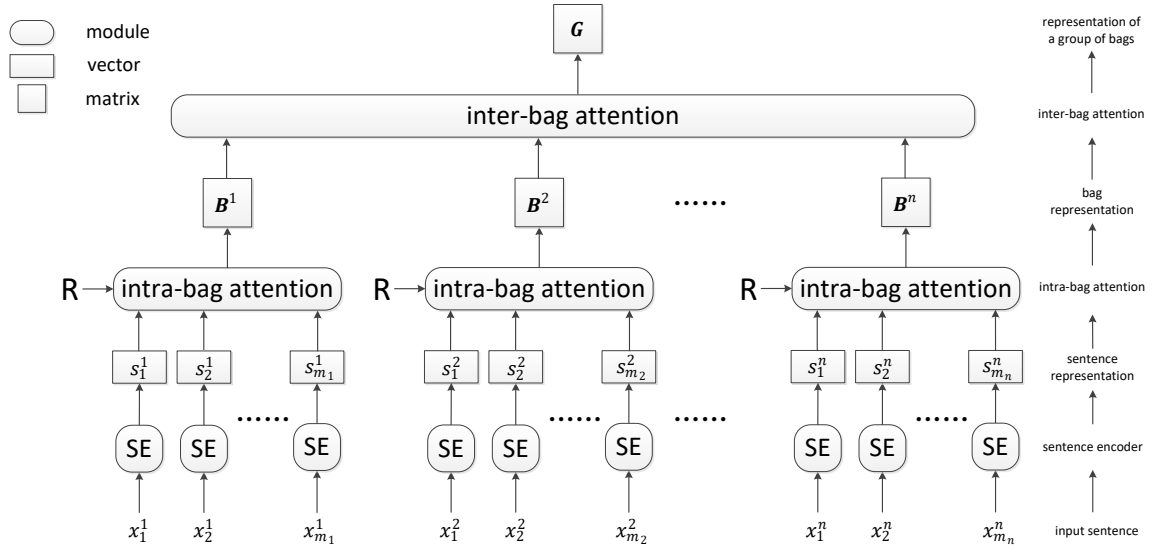


Figure 1: The framework of our proposed neural network with intra-bag and inter-bag attentions for relation extraction.

train instance selector and to filter out the samples with wrong labels. Their rewards were calculated from the prediction probabilities and the performance change of the relation classifier respectively. Qin et al. (2018a) designed an adversarial learning process to build a sentence-level generator via policy-gradient-based reinforcement learning. These methods were proposed to filter out the noisy data at sentence-level and also failed to deal with the noisy bag problem explicitly.

## 3 Methodology

In this section, we introduce a neural network with intra-bag and inter-bag attentions for distant supervision relation extraction. Let $g = \{b^1, b^2, ..., b^n\}$ denote a group of bags which have the same relation label given by distant supervision, and $n$ is the number of bags within this group. Let $b^i = \{x_1^i, x_2^i, ..., x_{m_i}^i\}$ denote all sentences in bag $b^i$, and $m_i$ is the number of sentences in bag $b^i$. Let $x_j^i = \{w_{j1}^i, w_{j2}^i, ..., w_{jl_{ij}}^i\}$ denote the $j$-th sentence in the $i$-th bag and $l_{ij}$ is its length (i.e., number of words). The framework of our model is shown in Fig. 1, which contains three main modules.

- **Sentence Encoder** Given a sentence $x_j^i$ and the positions of two entities within this sentence, CNNs or PCNNs (Zeng et al., 2015) are adopted to derive the sentence representation $\mathbf{s}_j^i$.

- **Intra-Bag Attention** Given the sentence representations of all sentences within a bag $b^i$ and a relation embedding matrix $\mathbf{R}$, attention weight vectors $\boldsymbol{\alpha}_k^i$ and bag representations $\mathbf{b}_k^i$ are calculated for all relations, where $k$ is the relation index.

- **Inter-Bag Attention** Given the representations of all bags with the group $g$, a weight matrix $\boldsymbol{\beta}$ is further calculated via similarity-based attention mechanism to obtain the representation of the bag group.

More details of these three modules will be introduced in the following subsections.

### 3.1 Sentence Encoder

#### 3.1.1 Word Representation

Each word $w_{jk}^i$ within the sentence $x_j^i$ is first mapped into a $d_w$-dimensional word embedding. To describe the position information of two entities, the position features (PFs) proposed in (Zeng et al., 2014) are also adopted in our work. For each word, the PFs describe the relative distances between current word and the two entities and are further mapped into two vectors $\mathbf{p}_{jk}^i$ and $\mathbf{q}_{jk}^i$ of $d_p$ dimensions. Finally, these three vectors are concatenated to get the word representation $\mathbf{w}_{jk}^i = [\mathbf{e}_{jk}^i; \mathbf{p}_{jk}^i; \mathbf{q}_{jk}^i]$ of $d_w + 2d_p$ dimensions.

#### 3.1.2 Piecewise CNN

For sentence $x_j^i$, the matrix of word representations $\mathbf{W}_j^i \in \mathbb{R}^{l_{ij} \times (d_w + 2d_p)}$ is first input into a CNN with $d_c$ filters. Then, piecewise max pooling (Zeng et al., 2015) is employed to extract features from the three segments of CNN outputs, and the segment boundaries are determined by the positions of the two entities. Finally, the sentence representation $\mathbf{s}_j^i \in \mathbb{R}^{3d_c}$ can be obtained.

### 3.2 Intra-Bag Attention

Let $\mathbf{S}^i \in \mathbb{R}^{m_i \times 3d_c}$ represent the representations of all sentences within bag $b^i$, and $\mathbf{R} \in \mathbb{R}^{h \times 3d_c}$ denote a relation embedding matrix where $h$ is the number of relations.

Different from conventional methods (Lin et al., 2016; Ji et al., 2017) where a unified bag representation was derived for relation classification, our method calculates bag representations $\mathbf{b}_k^i$ for bag $b^i$ on the condition of all possible relations as

$$\mathbf{b}_k^i = \sum_{j=1}^{m_i} \alpha_{kj}^i \mathbf{s}_j^i, \tag{1}$$

where $k \in \{1, 2, ..., h\}$ is the relation index. The attention weights $\boldsymbol{\alpha}_k^i$ can be further defined as

$$\alpha_{kj}^i = \frac{exp(e_{kj}^i)}{\sum_{j'=1}^{m_i} exp(e_{kj'}^i)}, \tag{2}$$

where $e_{kj}^i$ is the matching degree between the $k$-th relation query and $j$-th sentence in bag $b^i$. In our implementation, a simple dot product between vectors is adopted to calculate the matching degree as

$$e_{kj}^i = \mathbf{r}_k \mathbf{s}_j^{i\top}, ^2 \tag{3}$$

where $\mathbf{r}_k$ is the $k$-th row of the relation embedding matrix $\mathbf{R}$.

Finally, the representations of bag $b^i$ compose the matrix $\mathbf{B}^i \in \mathbb{R}^{h \times 3d_c}$ in Fig. 1, where each row corresponds to a possible relation type of this bag.

### 3.3 Inter-Bag Attention

In order to deal with the noisy bag problem, a similarity-based inter-bag attention module is designed to reduce the weights of noisy bags dynamically. Intuitively, if two bags $b^{i_1}$ and $b^{i_2}$ are both labeled as relation $k$, their representations $\mathbf{b}_k^{i_1}$ and $\mathbf{b}_k^{i_2}$ should be close to each other. Given a group of bags with the same relation label, we assign higher weights to those bags which are close to other bags in this group. As a result, the representation of bag group $g$ can be formulated as

$$\mathbf{g}_k = \sum_{i=1}^{n} \beta_{ik} \mathbf{b}_k^i, \tag{4}$$

where $\mathbf{g}_k$ is the $k$-th row of the matrix $\mathbf{G} \in \mathbb{R}^{h \times 3d_c}$ in Fig. 1, $k$ is the relation index and $\beta_{ik}$ compose the attention weight matrix $\boldsymbol{\beta} \in \mathbb{R}^{n \times h}$. Each weight is defined as

$$\beta_{ik} = \frac{exp(\gamma_{ik})}{\sum_{i=1}^{n} exp(\gamma_{ik})}, \tag{5}$$

where $\gamma_{ik}$ describes the confidence of labeling bag $b^i$ with the $k$-th relation.

Inspired by the self-attention algorithm (Vaswani et al., 2017) which calculates the attention weights for a group of vectors using the vectors themselves, we calculate the weights of bags according to their own representations. Mathematically, $\gamma_{ik}$ is defined as

$$\gamma_{ik} = \sum_{i'=1,...,n, i' \neq i} \text{similarity}(\mathbf{b}_k^i, \mathbf{b}_k^{i'}), \tag{6}$$

where the function *similarity* is a simple dot product in our implementation as

$$\text{similarity}(\mathbf{b}_k^i, \mathbf{b}_k^{i'}) = \mathbf{b}_k^i \mathbf{b}_k^{i'\top}. \tag{7}$$

And also, in order to prevent the influence of vector length, all bag representations $\mathbf{b}_k^i$ are normalized to unit length as $\overline{\mathbf{b}_k^i} = \mathbf{b}_k^i / ||\mathbf{b}_k^i||_2$ before calculating Eq.(4)-(7).

Then, the score $o_k$ of classifying bag group $g$ into relation $k$ is calculated via $\mathbf{g}_k$ and relation embedding $\mathbf{r}_k$ as

$$o_k = \mathbf{r}_k \mathbf{g}_k^\top + d_k, \tag{8}$$

where $d_k$ is a bias term. Finally, a softmax function is employed to obtain the probability that the bag group $g$ is classified into the $k$-th relation as

$$p(k|g) = \frac{exp(o_k)}{\sum_{k'=1}^{h} exp(o_{k'})}. \tag{9}$$

It should be noticed that the same relation embedding matrix $\mathbf{R}$ is used for calculating Eq.(3) and Eq.(8). Similar to Lin et al. (2016), the dropout strategy (Srivastava et al., 2014) is applied to bag representation $\mathbf{B}^i$ to prevent overfitting.

[2] We also tried $\mathbf{r}_k \mathbf{A} \mathbf{s}_j^{i\top}$, where $\mathbf{A}$ was a diagonal matrix, in experiments and achieved similar performance.

### 3.4 Implementation Details

#### 3.4.1 Data Packing

First of all, all sentences in the training set that contain the same two entities are accumulated into one bag. Then, we tie up every $n$ bags that share the same relation label into a group. It should be noticed that a bag group is one training sample in our method. Therefore, the model can also be trained in mini-batch mode by packing multiple bag groups into one batch.

#### 3.4.2 Objective Function and Optimization

In our implementation, the objective function is defined as

$$\mathrm{J}(\theta) = -\sum_{(g,k)\in T} \log p(k|g;\theta) \qquad (10)$$

where $T$ is the set of all training samples and $\theta$ is the set of model parameters, including word embedding matrix, position feature embedding matrix, CNN weight matrix and relation embedding matrix. The model parameters are estimated by minimizing the objective function J($\theta$) through mini-batch stochastic gradient descent (SGD).

#### 3.4.3 Training and Test

As introduced above, at the training phase of our proposed method, $n$ bags which have the same relation label are accumulated into one bag group and the weighted sum of bag representations is calculated to obtain the representation $\mathbf{G}$ of the bag group. Due to the fact that the label of each bag is unknown at test stage, each single bag is treated as a bag group (i.e., $n$=1) when processing the test set.

And also, similar to (Qin et al., 2018b), we only apply inter-bag attentions to positive samples, i.e., the bags whose relation label is not $NA$ ($NoRelation$). The reason is that the representations of the bags that express no relations are always diverse and it's difficult to calculate suitable weights for them.

#### 3.4.4 Pre-training Strategy

In our implementation, a pre-training strategy is adopted. We first train the model with only intra-bag attentions until convergence. Then, the inter-bag attention module is added and the model parameters are further updated until convergence again. Preliminary experimental results showed that this strategy can lead to better model performance than considering inter-bag attentions from the very beginning.

| Component | Parameter | Value |
|---|---|---|
| word embedding | dimension | 50 |
| position feature | max relative distance | $\pm$30 |
| | dimension | 5 |
| CNN | window size | 3 |
| | filter number | 230 |
| dropout | dropout rate | 0.5 |
| optimization | strategy | SGD |
| | learning rate | 0.1 |
| | batch size $N_p$ | 50 |
| | batch size $N_t$ | 10 |
| | group size $n$ | 5 |
| | gradient clip | 5.0 |

Table 2: Hyper-parameters of the models built in our experiments.

## 4 Experiment

### 4.1 Dataset and Evaluation Metrics

The New York Times (NYT) dataset was adopted in our experiments. This dataset was first released by (Riedel et al., 2010) and has been widely used by previous research on distant supervision relation extraction (Liu et al., 2017; Jat et al., 2018; Qin et al., 2018a,b). This dataset was generated by aligning Freebase with the New York Times (NYT) corpus automatically. There were 52 actual relations and a special relation $NA$ which indicated there was no relation between two entities.

Following previous studies (Mintz et al., 2009; Liu et al., 2017), we evaluated our models on the held-out test set of the NYT dataset. Precision-recall (PR) curves, area under curve (AUC) values and Precision@N (P@N) values (Lin et al., 2016) were adopted as evaluation metrics in our experiments. All of the numerical results given by our experiments were the mean values of 10 repetitive trainings, and the PR curves were randomly selected from the repetitions because there was no significant visual difference among them.

### 4.2 Training Details and Hyperparameters

All of the hyperparameters used in our experiments are listed in Table 2. Most of them followed the hyperparameter settings in (Lin et al., 2016). The 50-dimensional word embeddings released by (Lin et al., 2016)[3] were also adopted for initialization. The vocabulary contained the words which appeared more than 100 times in the NYT corpus.

[3] https://github.com/thunlp/NRE.

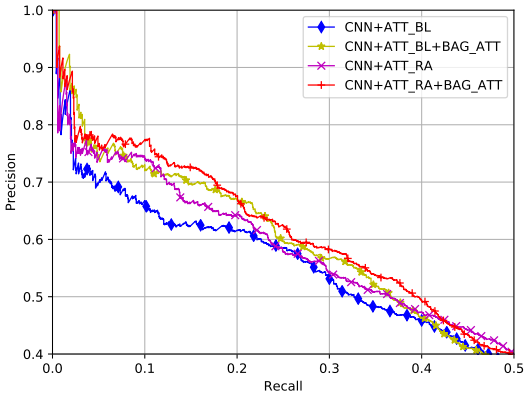


Figure 2: PR curves of different models using CNN sentence encoders.

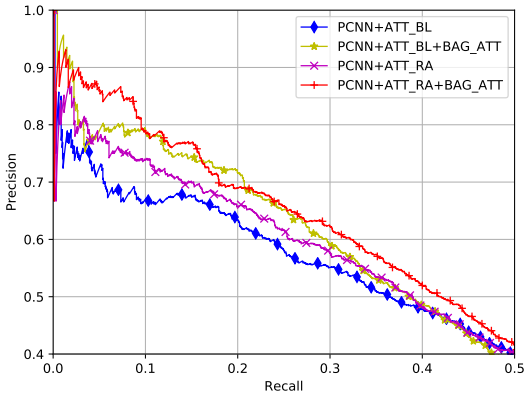


Figure 3: PR curves of different models using PCNN sentence encoders.

| No. | Model | AUC |
|---|---|---|
| 1 | CNN+ATT_BL | $0.376 \pm 0.003$ |
| 2 | CNN+ATT_BL+BAG_ATT | $0.388 \pm 0.002$ |
| 3 | CNN+ATT_RA | $0.398 \pm 0.004$ |
| 4 | CNN+ATT_RA+BAG_ATT | $0.407 \pm 0.004$ |
| 5 | PCNN+ATT_BL | $0.388 \pm 0.004$ |
| 6 | PCNN+ATT_BL+BAG_ATT | $0.403 \pm 0.002$ |
| 7 | PCNN+ATT_RA | $0.403 \pm 0.003$ |
| 8 | PCNN+ATT_RA+BAG_ATT | $\mathbf{0.422 \pm 0.004}$ |

Table 3: AUC values of different models.

Two different batch sizes $N_p$ and $N_t$ were used for pre-training and training respectively. In our experiments, a grid search is employed using training set to determine the optimal values of $n$, $N_p$ and $N_t$ among $n \in \{3, 4, ..., 10\}$, $N_p \in \{10, 20, 50, 100, 200\}$ and $N_t \in \{5, 10, 20, 50\}$. Note that increasing the bag group size $n$ may boost the effect of inter-bag attentions but lead to less training samples. The effects of inter-bag attentions would be lost when $n$=1. For optimization, we employed mini-batch SGD with the initial learning rate of 0.1. The learning rate was decayed to one tenth every 100,000 steps. The pre-trained model with only intra-bag attentions converged within 300,000 steps in our experiments. Thus, the initial learning rate for training the model with inter-bag attentions was set as 0.001.

### 4.3 Overall performance

Eight models were implemented for comparison. The names of these models are listed in Table 3, where *CNN* and *PCNN* denote using CNNs or piecewise CNNs in sentence encoders respectively, *ATT_BL* means the baseline intra-bag attention method proposed by (Lin et al., 2016), *ATT_RA* means our proposed relation-aware intra-bag attention method, and *BAG_ATT* means our proposed inter-bag attention method. At the training stage of the *ATT_BL* method, the relation query vector for attention weight calculation was fixed as the embedding vector associated with the distant supervision label for each bag. At the test stage, all relation query vectors were applied to calculate the posterior probabilities of relations respectively and the relation with the highest probability was chosen as the classification result (Lin et al., 2016). The means and standard deviations of the AUC values given by the whole PR curves of these models are shown in Table 3 for a quantitative comparison. Following (Lin et al., 2016), we also plotted the PR curves of these models in Fig. 2 and 3 with recall smaller than 0.5 for a visualized comparison.

From Table 3, Fig. 2 and Fig. 3, we have the following observations. (1) Similar to the results of previous work (Zeng et al., 2015), PCNNs worked better than CNNs as sentence encoders. (2) When using either CNN or PCNN sentence encoders, *ATT_RA* outperformed *ATT_BL*. It can be attributed to that the *ATT_BL* method only considered the target relation when deriving bag representations at training time, while the *ATT_RA* method calculated intra-bag attention weights using all relation embeddings as queries, which improved the flexibility of bag representations. (3) For both sentence encoders and both intra-bag attention methods, the models with *BAG_ATT* always achieved better performances than the ones without *BAG_ATT*. This result verified the effectiveness of our proposed inter-bag attention method for distant supervision

| # of Test Sentences | one | | | | two | | | | all | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P@N(%) | 100 | 200 | 300 | mean | 100 | 200 | 300 | mean | 100 | 200 | 300 | mean |
| (Lin et al., 2016) | 73.3 | 69.2 | 60.8 | 67.8 | 77.2 | 71.6 | 66.1 | 71.6 | 76.2 | 73.1 | 47.4 | 72.2 |
| (Liu et al., 2017) | 84.0 | 75.5 | 68.3 | 75.9 | 86.0 | 77.0 | 73.3 | 78.8 | 87.0 | 84.5 | 77.0 | 82.8 |
| CNN+ATT_BL | 74.2 | 68.9 | 65.3 | 69.5 | 77.8 | 71.5 | 68.1 | 72.5 | 79.2 | 74.9 | 70.3 | 74.8 |
| CNN+ATT_RA | 76.8 | 72.7 | 67.9 | 72.5 | 79.6 | 73.9 | 70.7 | 74.7 | 81.4 | 76.3 | 72.5 | 76.8 |
| CNN+ATT_BL+BAG_ATT | 78.6 | 74.2 | 69.7 | 74.2 | 82.4 | 76.2 | 72.1 | 76.9 | 83.0 | 78.0 | 74.0 | 78.3 |
| CNN+ATT_RA+BAG_ATT | 79.8 | 75.3 | 71.0 | 75.4 | 83.2 | 76.5 | 72.1 | 77.3 | 87.2 | 78.7 | 74.9 | 80.3 |
| PCNN+ATT_BL | 78.6 | 73.5 | 68.1 | 73.4 | 77.8 | 75.1 | 70.3 | 74.4 | 80.8 | 77.5 | 72.3 | 76.9 |
| PCNN+ATT_RA | 79.4 | 73.9 | 69.6 | 74.3 | 82.2 | 77.6 | 72.4 | 77.4 | 84.2 | 79.9 | 73.0 | 79.0 |
| PCNN+ATT_BL+BAG_ATT | 85.2 | **78.2** | 71.3 | 78.2 | 84.8 | **80.0** | 74.3 | 79.7 | 88.8 | 83.7 | 77.4 | 83.9 |
| PCNN+ATT_RA+BAG_ATT | **86.8** | 77.6 | **73.9** | **79.4** | **91.2** | 79.2 | **75.4** | **81.9** | **91.8** | **84.0** | **78.7** | **84.8** |

Table 4: P@N values of the entity pairs with different number of test sentences.

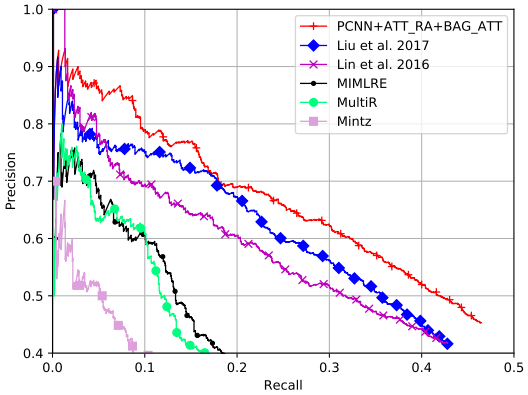


Figure 4: PR curves of several models in previous work and our best model.

relation extraction. (4) The best AUC performance was achieved by combining PCNN sentence encoders with the intra-bag and inter-bag attentions proposed in this paper.

## 4.4 Comparison with previous work

### 4.4.1 PR curves

The PR curves of several models in previous work and our best model *PCNN+ATT_RA+BAG_ATT* are compared in Fig. 4, where Mintz (Mintz et al., 2009), MultiR (Hoffmann et al., 2011) and MIMLRE (Surdeanu et al., 2012) are conventional feature-based methods, and (Lin et al., 2016) and (Liu et al., 2017) are PCNN-based ones[4]. For a fair comparison with (Lin et al., 2016) and (Liu et al., 2017), we also plotted the curves with only the top 2000 points. We can see that our model achieved the better PR performance than all the other models.

[4]All of these curve data are from `https://github.com/tyliupku/soft-label-RE`.

| model | CNN | PCNN |
|---|---|---|
| **ATT_BL**† | 0.219 | 0.253 |
| **ATT_BL+RL** | 0.229 | 0.261 |
| **ATT_BL+DSGAN** | 0.226 | 0.264 |
| **ATT_BL**‡ | 0.242 | 0.271 |
| **ATT_RA** | 0.254 | 0.297 |
| **ATT_BL+BAG_ATT** | 0.253 | 0.285 |
| **ATT_RA+BAG_ATT** | **0.262** | **0.311** |

Table 5: AUC values of previous work and our models, where *ATT_BL+DSGAN* and *ATT_BL+RL* are two models proposed in (Qin et al., 2018a) and (Qin et al., 2018b) respectively, † indicates the baseline result reported in (Qin et al., 2018a,b) and ‡ indicates the baseline result given by our implementation.

### 4.4.2 AUC values

*ATT_BL+DSGAN* (Qin et al., 2018a) and *ATT_BL+RL* (Qin et al., 2018b) are two recent studies on distant supervision relation extraction with reinforcement learning for data filtering, which reported the AUC values of PR curves composed by the top 2000 points. Table 5 compares the AUC values reported in these two papers and the results of our proposed models. We can see that introducing the proposed *ATT_RA* and *BAG_ATT* methods to baseline models achieved larger improvement than using the methods proposed in (Qin et al., 2018a,b).

## 4.5 Effects of Intra-Bag Attentions

Following (Lin et al., 2016), we evaluated our models on the entity pairs with more than one training sentence. One, two and all sentences for each test entity pair were randomly selected to construct three new test sets. The P@100, P@200, P@300 values and their means given by our proposed models on these three test sets are reported in Table 4 together with the best results of (Lin et al., 2016) and (Liu et al., 2017). Here, P@N

| bag | sentence | correct? | intra-bag weights | inter-bag weights |
|---|---|---|---|---|
| **B1** | [**Panama City Beach**]$_{e2}$ , too , has a glut of condos , but the area was one of only two in [**Florida**]$_{e1}$ where sales rose in march , compared with a year earlier. | **Yes** | 0.71 | 0.48 |
| | Like much of [**Florida**]$_{e1}$ , [**Panama City Beach**]$_{e2}$ has been hurt by the downturn in the real estate market. | **Yes** | 0.29 | |
| **B2** | Among the major rivers that overflowed were the Housatonic , Still , Saugatuck , Norwalk , Quinnipiac , Farmington , [**Naugatuck**]$_{e1}$ , Mill , Rooster and [**Connecticut**]$_{e2}$ . | **No** | 1.00 | 0.13 |
| **B3** | ..., the army chose a prominent location in [**Virginia**]$_{e1}$, at the foot of the Arlington memorial bridge , directly across the [**Potomac River**]$_{e2}$ from the Lincoln memorial . | **No** | 0.13 | 0.39 |
| | ... , none of those stars carried the giants the way barber did at Fedex field , across the [**Potomac River**]$_{e1}$ from [**Virginia**]$_{e2}$ , where he grew up as a redskins fan . | **Yes** | 0.87 | |

Table 6: A test set example of relation */location/location/contains* from the NYT corpus.

| sentence number | mean$\pm$std |
|---|---|
| **1** | $0.163 \pm 0.029$ |
| **2** | $0.187 \pm 0.033$ |
| **3** | $0.210 \pm 0.034$ |
| **4** | $0.212 \pm 0.037$ |
| $\geq$ **5** | $0.256 \pm 0.043$ |

Table 7: The distributions of inter-bag attention weights for the bags with different number of sentences.

means the precision of the relation classification results with the top $N$ highest probabilities in the test set.

We can see our proposed methods achieved higher P@N values than previous work. Furthermore, no matter whether *PCNN* or *BAG_ATT* were adopted, the *ATT_RA* method outperformed the *ATT_BL* method on the test set with only one sentence for each entity pair. Note that the decoding procedures of *ATT_BL* and *ATT_RA* were equivalent when there was only one sentence in a bag. Therefore, the improvements from *ATT_BL* to *ATT_RA* can be attributed to that *ATT_RA* calculated intra-bag attention weights in a relation-aware way at the training stage.

### 4.6 Distributions of Inter-Bag Attention Weights

We divided the training set into 5 parts according to the number of sentences in each bag. For each bag, the inter-bag attention weights given by the *PCNN+ATT_RA+BAG_ATT* model were recorded. Then, the mean and standard deviation of inter-bag attention weights for each part of the training set were calculated and are shown in Table 7. From this table, we can see that the bag with smaller number of training sentences were usually assigned with lower inter-bag attention weights. This result was consistent with the finding in (Qin et al., 2018b) that the entity pairs with fewer training sentences were more likely to have incorrect relation labels.

### 4.7 Case Study

A test set example of relation */location/location/contains* is shown in Table 6. The bag group contained 3 bags, which consisted of 2, 1, and 2 sentences respectively. We calculated the intra-bag and inter-bag attentions for this bag group using our *PCNN+ATT_RA+BAG_ATT* model and the weights of the target relation are also shown in Table 6.

In this example, the second bag was a noisy bag because the only sentence in this bag didn't express the relation */location/location/contains* between the two entities **Naugatuck** and **Connecticut**. In conventional methods, these three bags were treated equally for model training. After introducing inter-bag attention mechanism, the weight of this noisy bag was reduced significantly as shown in the last column of Table 6.

## 5 Conclusion

In this paper, we have proposed a neural network with intra-bag and inter-bag attentions to cope with the noisy sentence and noisy bag problems in distant supervision relation extraction. First, relation-aware bag representations are calculated by a weighted sum of sentence embeddings where the noisy sentences are expected to have smaller weights. Further, an inter-bag attention module is designed to deal with the noisy bag problem by calculating the bag-level attention weights dynamically during model training. Experimental results

on New York Times dataset show that our models achieved significant and consistent improvements compared with the models using only conventional intra-bag attentions. To deal with the multi-label problem of relation extraction and to integrate external knowledge into our model will be the tasks of our future work.

## Acknowledgments

We thank the anonymous reviewers for their valuable comments.

## References

Aron Culotta and Jeffrey Sorensen. 2004. Dependency tree kernels for relation extraction. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*.

Jun Feng, Minlie Huang, Li Zhao, Yang Yang, and Xiaoyan Zhu. 2018. Reinforcement learning for relation classification from noisy data. In *Proceedings of AAAI*.

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. Association for Computational Linguistics.

Sharmistha Jat, Siddhesh Khandelwal, and Partha Talukdar. 2018. Improving distantly supervised relation extraction using word and entity based attention. *arXiv preprint arXiv:1804.06987*.

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

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436.

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

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

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: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.

Raymond J Mooney and Razvan C Bunescu. 2006. Subsequence kernels for relation extraction. In *Advances in neural information processing systems*, pages 171–178.

Pengda Qin, Weiran XU, and William Yang Wang. 2018a. Dsgan: Generative adversarial training for distant supervision relation extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505. Association for Computational Linguistics.

Pengda Qin, Weiran XU, and William Yang Wang. 2018b. Robust distant supervision relation extraction via deep reinforcement learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2137–2147. Association for Computational Linguistics.

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.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

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. Association for Computational Linguistics.

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.

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2002. Kernel methods for relation extraction. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via a piecewise convolutional neural networks. In *Pro-*

*ceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762. Association for Computational Linguistics.

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