Cross-Relation Cross-Bag Attention for Distantly-Supervised Relation Extraction

Yujin Yuan,[1] Liyuan Liu,[2] Siliang Tang,[1]* Zhongfei Zhang,[1] Yueting Zhuang,[1]
Shiliang Pu,[3] Fei Wu,[1] Xiang Ren[4]

[1]Zhejiang University [2]University of Illinois at Urbana Champaign [3]Hikvision Research Institute
[4]University of Southern California
{yujin, siliang, zhongfei, yzhuang, wufei}@zju.edu.cn, llychinalz@gmail.com,
pushiliang@hikvision.com, xiangren@usc.edu

Abstract

Distant supervision leverages knowledge bases to automatically label instances, thus allowing us to train relation extractor without human annotations. However, the generated training data typically contain massive noise, and may result in poor performances with the vanilla supervised learning. In this paper, we propose to conduct multi-instance learning with a novel Cross-relation Cross-bag Selective Attention (C^2SA), which leads to noise-robust training for distant supervised relation extractor. Specifically, we employ the sentence-level selective attention to reduce the effect of noisy or mismatched sentences, while the correlation among relations were captured to improve the quality of attention weights. Moreover, instead of treating all entity-pairs equally, we try to pay more attention to entity-pairs with a higher quality. Similarly, we adopt the selective attention mechanism to achieve this goal. Experiments with two types of relation extractor demonstrate the superiority of the proposed approach over the state-of-the-art, while further ablation studies verify our intuitions and demonstrate the effectiveness of our proposed two techniques.

Introduction

Aiming to detect and classify the relation between an entity pair in the given sentences, Relation Extraction (RE) plays a vital role in natural language understanding (Etzioni et al. 2004; Mintz et al. 2009; Liu et al. 2017a). The typical methods follow the supervised learning paradigm and require extensive human annotations, which are costly and time-consuming. To alleviate such reliance, attempts have been made to build relation extractor with distant supervision, i.e., automatically generating training data by knowledge base (KB). For example, with the KB fact <*Jimi Hendrix*, died_in, *London*> in Table 1, distant supervision would annotate all sentences containing *Jimi Hendrix* and *London* as the relation type died_in.

Despite its efficiency in cost and time, distant supervision is context-agnostic thus containing massive noise for sentence-level RE. Accordingly, it may lead to an unsatisfying performance before noise-robust training paradigms are developed (Lin et al. 2016). Multi-instance learning (MIL)

---

*Corresponding author.

KB Fact		<*Jimi Hendrix*, died_in, *London*>
Sentence Bag with Distant Supervision	S1	**Jimi Hendrix** died in 1970 in **London** at 27.
	S2	George Frideric Handel and **Jimi Hendrix** lived at adjacent addresses in **London**
	...	...

Table 1: Distant Supervision and Sentence Bag.

has been employed to reduce the noise and bring such robustness. As shown in Table 1, it treats sentence bag as the basic training instance, instead of individual sentence. Each bag would contain a group of sentences labeled by the same KB fact. By selecting from such bags, it allows a model to focus more on sentences of a higher quality and reduces the effect of noisy ones. Specifically, some methods (Riedel, Yao, and McCallum 2010; Hoffmann et al. 2011; Surdeanu et al. 2012) try to pick only one sentence from one bag, while more improvements (Lin et al. 2016; Liu et al. 2017b) have been observed by replacing the hard selection with Sentence-level Selective attention (ATT). ATT tries to assign attention weights to sentences and combines all sentences in the bag for the training.

However, ATT generates the attention weight for each relation type independently and overlooked their correlation. For example, in Table 1, by identifying S2 as a high-quality sentence for live_in, we are also able to recognize it as a low-quality sentence for die_in. Based on this intuition, we propose a novel attention mechanism, Cross-relation Attention, which generates the attention weight after examining their relatednesses to all relation types.

Moreover, we go beyond ATT and construct training instances at a higher level. We relax the constraint that one training instance only contains one entity pair. Specifically, we propose the Cross-bag Attention to combine different sentence bags, refer the combined structure as superbag, and set superbag as the training instance instead of sentence bag. This allows us to focus more on sentence bags of a higher quality, and reduce the noise brought by KB Facts which are outdated or unexpressed in the corpus.

Combining these two mechanisms, we refer our method as Cross-relation Cross-bag Selective Attention (C^2SA). Applying such attention to two types of relation extractor,

we observe consistent improvements over the vanilla ATT. Extensive ablation studies are further conducted to verify both our intuitions and the effectiveness of both components.

Related work

Relation extraction is one of the most important tasks in NLP. Over the years, many efforts have been invested in relation extraction, especially in supervised relation extraction (Mooney and Bunescu 2006; Zelenko, Aone, and Richardella 2003; Rink and Harabagiu 2010). However, most of them are based on extra NLP systems to derive lexical features.

Recently, deep neural networks can learn underlying features automatically and have been used in the literature. (Socher et al. 2012) uses a recursive neural network in relation extraction. (Zeng et al. 2014; Santos, Xiang, and Zhou 2015; Zeng et al. 2015) adopts an end-to-end convolutional neural network for relation extraction. (Zhou et al. 2016; Li et al. 2017; Zhang et al. 2017) uses the attention-based LSTM network to mitigate the weakness of the CNN network in processing long-span information. Based on CNN or RNN, there are still many efforts (Xu et al. 2015a; 2015b; Vu et al. 2016; Wang et al. 2016; Jiang et al. 2016; Huang and Wang 2017) to improve the network structures for more suitable for RE tasks. (Zeng et al. 2018) trains a relation extractor using Reinforcement learning.

Although reasonably good performances are reported in the above models, training these models requires a large amount of annotation data, which are difficult and expensive to obtain. To address this issue, distant supervision (DS) was proposed (Mintz et al. 2009) by assuming that all the sentences that mention two entities of a fact triple describe the relation in the triple. In order to suppress the large amount of noise introduced by DS, many studies formulate the problem of relation classification as a multiple instance learning (MIL) problem (Riedel, Yao, and McCallum 2010; Hoffmann et al. 2011; Surdeanu et al. 2012; Zeng et al. 2015). All sentences containing the same entity pair are taken as a bag in MIL. (Lin et al. 2016) proposes the selective attention to select high quality sentence features in the bag as the bag feature and train the model by the bag feature. (Luo et al. 2017) proposes a transition matrix based method to dynamically characterize the noise. (Feng et al. 2018) uses reinforcement learning to select a more reliable subset on the DS dataset and uses it to train the classifier. In order to solve the bag level noisy label problem, (Liu et al. 2017b) uses a posterior probability constraint to correct potentially incorrect bag labels.

The selective attention method proposed by (Lin et al. 2016) is widely used in many recent efforts (Liu et al. 2017b; Li et al. 2017; Ji et al. 2017). The main differences between our approach and that selective attention are: 1. Our approach takes into account the interplay between multiple relations. 2. Our approach assesses the quality of the bag feature and reduces the impact of bag-level noisy label problem while the existing selective attention in the literature fails when processing a completely incorrect bag.

Methodology

Here, we develop a novel selective attention to reduce the noise of distant supervision for training relation extractors. We present the Cross-relation Cross-bag Selective Attention(C^2SA). It improves the sentence-level attention by considering the correlation among relations, and conducts the selection at the bag level with another attention layer.
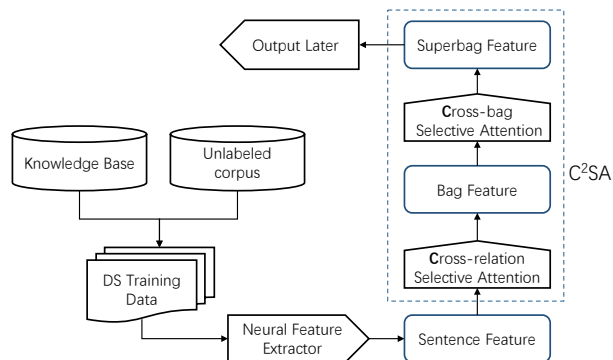

Figure 1: Distant Training with C^2SA

As in Figure 1, the relation extractor has two components i.e., a neural feature extractor and an output layer. As for the neural feature extractor, it extracts useful features for relation classification and can be embodied by any popular neural structures including CNNs and RNNs. Based on the extracted features, the output layer makes the prediction on the relation type.

At the same time, the distant training pipeline with C^2SA has four steps. The neural feature extractor is first employed to construct representations for each sentence. Then, the cross-relation selective attention combines the sentence representations and generate the representation for a sentence bag. Similarly, the cross-bag selective attention combines representations for the sentence bags and generates the representation for the superbag. At the end, the loss is calculated based on the superbag feature that guides the learning of the relation extractor.

We now proceed by introducing these components in further details.

Relation Extractor

Typically, the neural feature extractor can be considered as a neural sentence encoder, which encodes sentences into low-dimensional, fixed-length vectors. It can be employed as any neural encoder, such as RNNs (e.g., LSTMs, GRUs) or CNNs. Since the CNNs-based models achieve the best results in our experiments, we take them as the defaults.

Input Representation   To encode the sufficient sentence information in an entity-aware manner, we formulate the input for neural networks as follows: for a word at the position i in a sentence, we first transform it into a pre-trained word vector $\mathbf{v}_i$ (Mikolov et al. 2013). Then we calculate its relative distances to the target entities in the sentence (i.e., d_1 and d_2), and find their position embedding vectors $\mathbf{p}_{d_1}$ and
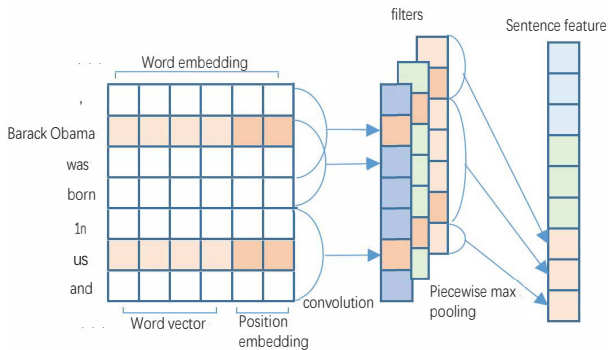
Figure 2: P-CNN based Neural Relation Extractor

$\mathbf{p}_{d_2}$ by looking up a position embedding table (Zeng et al. 2014). The position embedding table is randomly initialized and is updated during the model training. After that, we concatenate $\mathbf{v}_i$ with $\mathbf{p}_{d_1}$ and $\mathbf{p}_{d_2}$ as the word representation $\mathbf{w}_i$. By repeating these steps, we transform each sentence into a fixed-sized matrix $C = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_m]^T$, where $\mathbf{w}_i$ is a fixed-length vector and m is the maximum length of a sentence in the whole data corpus. For shorter sentences, we pad them with zeros.

Neural Feature Extractor As in Figure 2, we adopt piecewise-CNN (P-CNN) as the neural feature extractor. It is composed of a Convolution layer and a Piecewise Max-pooling layer.

In the Convolution layer, the output $\mathbf{c}$ is calculated as:

$$\mathbf{c}_{i,j} = P_i \circ C_{j,j+l-1}.$$

where P_i is the i-th convolutional kernel (filter), l is the width of the kernel, and $C_{i,j}$ is a sliding window on sentence C that starts from $\mathbf{w}_i$ and ends at $\mathbf{w}_j$, i.e., $C_{i,j} = \{\mathbf{w}_i, \mathbf{w}_{i+1}, \cdots, \mathbf{w}_j\}$.

The Piecewise Max-pooling (Zeng et al. 2015) is a variant of the traditional max-pooling layer by considering the specific situation in relation extraction. For a sentence that contains an entity pair, the corresponding $\mathbf{c}$ is divided by such pair into three pieces. After that, the max-pooling operation is applied to each piece respectively, yielding three different output features. We then concatenate them into one feature vector $\mathbf{x}_i$, where $\mathbf{x}_i \in R^{3 \cdot n}$ and n is the number of filters. Finally, we apply hyperbolic tangent function at the output vector $\mathbf{x}_i$.

Output Layer To compute the confidence of each relation, we employ the linear projection and softmax function to calculate the conditional probability:

$$\mathbf{o} = W \cdot \mathbf{f} \qquad (1)$$

$$P(r|\mathbf{f}) = \frac{e^{\mathbf{o}_r}}{\sum_{k=1}^{n_r} e^{\mathbf{o}_k}}$$

where $\mathbf{f}$ is the extracted feature and W are the weights of the transformation.

In the experiments, we adopt the dropout strategy (Hinton et al. 2012) on the output layer to prevent overfitting.
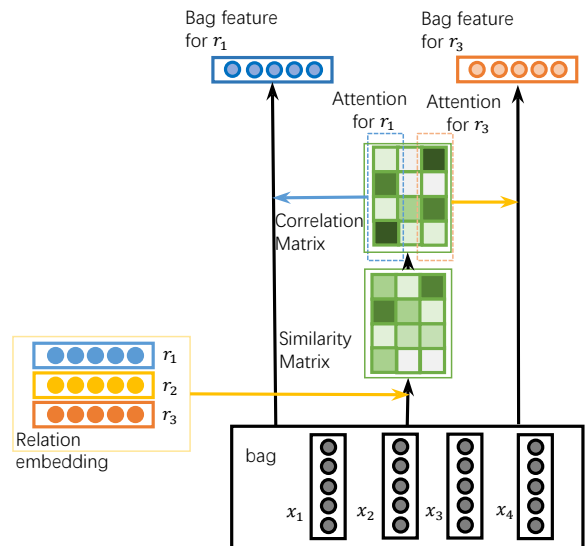

Figure 3: Structure of Sentence-level Cross-relation Selective Attention

Dropout prevents co-adaptation of hidden units by randomly setting them to zero for a proportion p. Thus, we revise Equation 1 to Equation 2:

$$\mathbf{o} = W \cdot (\mathbf{f} \odot \mathbf{h}) \qquad (2)$$

where $\mathbf{h}$ is a vector of Bernoulli random variables with probability p of being 1.

Cross-relation Cross-bag Selective Attention

Now we describe the proposed Cross-relation Cross-bag Selective Attention for relation extractor training. As introduced before, we follow the standard MIL and construct the sentence bags, i.e., $B_i = \{x_{i,1}, x_{i,2}, \cdots, x_{i,n_b}\}$, where $x_{i,*}$ contains the same entity pair and n_b is the number of sentences in the bag. With the distant supervision, each sentence bag is annotated by relations existing between the entity pair.

With the sentence bag, we first leverage the cross-relation attention to combine sentences in the same bag (as shown in Figure 3); we then employ the cross-bag attention to integrate different bags into the superbag (as shown in Figure 5). The first attention attempts to reduce the effect of noisy or mismatched sentences, and the second aims to focus more attention on the high quality sentence bags.

Cross-relation Selective Attention For sentence bag B_i, we refer the extracted feature representation for $x_{i,j}$ as to $\mathbf{x}_{i,j}$. Then we calculate the selective attention based on the similarity between the sentence and the relation:

$$S_{i,j,k} = \frac{\mathbf{x}_{i,j} \cdot \mathbf{r}_k}{\|\mathbf{x}_{i,j}\| \|\mathbf{r}_k\|} \qquad (3)$$

where $\mathbf{r}_k$ is the attention parameter corresponding to the k-th relation.
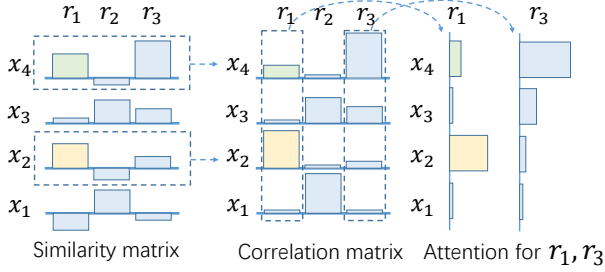
Figure 4: Similarity matrix and Correlation matrix

In order to capture the correlation among relations, we calculate the expected attention weight by the Bayes' rule:

$$P(\text{j}_{th}\text{sentence}|\text{k}_{th}\text{relation})$$
$$= \frac{P(\text{k}_{th}\text{relation}|\text{j}_{th}\text{sentence})P(\text{j}_{th}\text{sentence})}{\sum_{\tilde{j}=1}^{n_b} P(\text{k}_{th}\text{relation}|\tilde{\text{j}}_{th}\text{sentence})P(\tilde{\text{j}}_{th}\text{sentence})} \quad (4)$$

Specifically, we assume $P(\text{j}_{th}\text{sentence})$ to be the uniform distribution, and calculate $P(\text{k}_{th}\text{relation}|\text{j}_{th}\text{sentence})$ with the softmax function:

$$P(\text{k}_{th}\text{relation}|\text{j}_{th}\text{sentence}) = \frac{e^{S_{i,j,k}}}{\sum_{\tilde{k}=1}^{n_r} e^{S_{i,j,\tilde{k}}}} \quad (5)$$

To simplify the notion, we refer the calculated value of $P(\text{k}_{th}\text{relation}|\text{j}_{th}\text{sentence})$ as to $\alpha_{j,k}$, and the value of $P(\text{j}_{th}\text{sentence}|\text{k}_{th}\text{relation})$ as to $\beta_{j,k}$. Then we rewrite Equation 4 as

$$\beta_{j,k} = \frac{\alpha_{j,k}}{\sum_{\tilde{j}=1}^{n_b} \alpha_{\tilde{j},k}} \quad (6)$$

Accordingly, the bag feature for B_i for the k-th relation can be calculated as

$$\mathbf{b}_{i,k} = \sum_{\tilde{j}=1}^{n_b} \beta_{\tilde{j},k}\mathbf{x}_{i,\tilde{j}} \quad (7)$$

As shown in Figure 4, the cross-relation selective attention depends on the similarities from the sentence to not only the target relation, but also the other relations. For example, in Figure 4, x_2 and x_4 have similar similarities to r_1, but since x_4 is more inclined to expressing r_3, the model tends to use features of x_2 to generate the bag feature and use this feature to predict r_1.

Cross-bag Selective Attention The sentence level attention assumes that at least one sentence in a bag expresses the relation between entity pairs. However, due to the noisy nature of distant supervision, there may still exist noise in the sentence bag level. For example, there are a large number of entity pairs that cannot find any expression about their relations in the given corpus. Such entity pairs would result in the mismatched and noisy training instances for the sentence level relation extraction.

As shown in Figure 5, we go beyond the existing setting and propose to establish an additional selective attention layer, i.e., Cross-bag Selective Attention. Specifically,
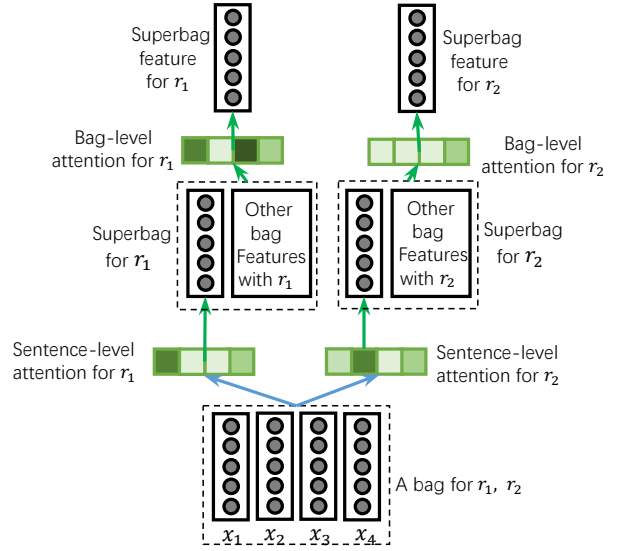

Figure 5: Cross-bag Selective Attention

we intend to combine several sentence bags of the same relation type and put more attention to the higher quality ones. We refer the higher-level structure that contains a group of sentence bags as to *superbag*, and denote it as $\mathcal{B} = \{B_1, B_2, ..., B_{n_s}\}$, where n_s is the size of the superbag and all B_i are labelled with the k-th relation type.

Based on the Cross-relation Selective Attention, we construct the representation for each sentence bag while capturing the correlation among sentences. Here, we combine these representations with an attention layer. Specifically, we obtain the superbag feature $\mathbf{f}$ for $\mathcal{B}$ as:

$$\mathbf{f} = \sum_{i=1}^{n_s} \gamma_i \cdot \mathbf{b}_{i,k}$$
$$\gamma_i = \frac{e^{S(\mathbf{r}_k, \mathbf{b}_{i,k})}}{\sum_{j=1}^{n_s} e^{S(\mathbf{r}_k, \mathbf{b}_{j,k})}} \quad (8)$$

where $\mathbf{b}_{i,k}$ is the bag representation w.r.t. B_i for the k-th relation and $\mathbf{r}_k$ is the attention parameter corresponding to the j-th relation. Specifically, we tie up the $\mathbf{r}_k$ in Equation 8 with the ones in Equation 3. Also, similar to the Equation 3, we calculate $S(\mathbf{r}_k, \mathbf{b}_{i,k})$ with the cosine similarity.

Model Training

We employ the objective function as the negative log likelihood at the superbag level as follows:

$$\mathcal{L} = -\sum_{i=1}^{n_{sb}} log\,p(l_i|\mathcal{B}_i) \quad (9)$$

where n_{sb} indicates the number of the superbags in the training set and l_i indicates the label of a superbag.

Experiments

We report the performances of C^{2}SA by comparing it with the state-of-the-art relation extraction methods. Some of

Parameter Name	Value	Candidate Set
sentence embedding dimension	100	{100, 150, 200}
batch size	100	{100, 150, 200}
superbag size	2	{2, 3, 4, 5}
sliding window size	3	reused from previous work
word vector dimension	50	
position embedding dimension	5	
dropout probability	0.5	

Table 2: Hyper-parameter Settings.

Figure 6: Performance comparison on the corpus-level relation extraction.

these baselines are BLSTMs-based, while others are P-CNNs-based. Since C^2SA is model-agnostic and only used in the learning phrase, we conduct the experiments with both types of neural relation extractors. Moreover, we employ two evaluation settings for an extensive comparison. Specifically, we first follow the popular setting and evaluate the model performances on the corpus-level relation extraction task. Besides, we also conduct experiments on the sentence-level relation extraction task with a human annotated test corpus. We observe that, in all settings, our proposed C^2SA consistently outperforms the state-of-the-art.

For a better understanding, we further report a case study and ablation experiments, which further verifies our intuition and demonstrates the effectiveness of both cross-sentence and cross-bag selective attention mechanisms.

Model Training

As discussed before, we leverage C^2SA to train two types of relation extractors, i.e., P-CNNs-based and BLSTMs-based. We refer these two variants as PCNN+C^2SA and BLSTM+C^2SA.

Dataset Following the existing literature (Riedel, Yao, and McCallum 2010; Lin et al. 2016; Li et al. 2017; Feng et al. 2018; Liu et al. 2017b), we use the New York Times (NYT) dataset as the training set (Mintz et al. 2009). It uses Freebase (Bollacker et al. 2008) to provide distant supervision on the NYT corpus. Specifically, it collects sentences from 2005 to 2006 and supports 53 different relations (including NA which means no relations for an entity pair). For training set statistics, this dataset contains 522611 sentences, 281270 entity pairs and 18252 KB facts.

Model Setting In all the experiments, we use 50 dimensional word vectors that are pre-trained by the Skip-gram algorithm [1] on the NYT corpus. For hyper-parameters, we reuse part of them from the previous study (Zeng et al. 2015; Li et al. 2017), and tune the rest part by grid-search with the three-fold cross-validation (on the training set). The final hyper-parameter setting used in our experiments are summarized in Table 2.
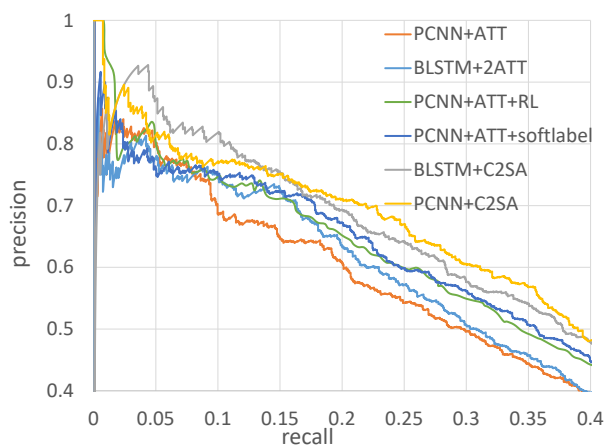
[1] https://code.google.com/p/word2vec/

Comparison on the Corpus-level Task

Here, we evaluate the performances of our method on the corpus-level relation extraction. For an entity pair, the task is to identify their relation type with regard to all the sentences that mention this entity pair. Specifically, we first feed all the sentence representations to the output layers for prediction. After that, the probability of a relation for an entity pair is the maximum probability of this relation in all the sentences mentioning this entity pair.

Test Set We use the test set of NYT for the corpus-level evaluation (Mintz et al. 2009). The test set is constructed with sentences from NYT of 2007 annotated with Freebase (Bollacker et al. 2008). Specifically, it contains 172448 sentences, 96678 entity pairs and 1950 KB facts.

Competing Methods We choose four recent methods as baselines:

- PCNN+ATT (Lin et al. 2016) uses the vanilla sentence-level selective attention to combine sentence features for each bag. Based on this attention, the representation for each bag is obtained and is trained under the PCNN model.

- BLSTM+2ATT (Li et al. 2017) also employs the vanilla sentence-level selective attention. Different from PCNN+ATT, it is a BLSTM-based model which has an additional word-level attention module.

- PCNN+ATT+RL (Feng et al. 2018) further incorporates reinforcement learning to improve PCNN+ATT. It trains the PCNN+ATT model with a subset of the training set, which is selected by the learned policy.

- PCNN+ATT+softlabel (Liu et al. 2017b) enhances the PCNN+ATT model by using the posterior probability constraint to correct potentially incorrect bag labels.

Performance Comparison Following the existing literature (Riedel, Yao, and McCallum 2010; Lin et al. 2016; Li et al. 2017; Feng et al. 2018; Liu et al. 2017b), we eval-

uate C^2SA in the held-out setting and report the model performances with the Precision-Recall curve.

We summarize the performances of BLSTM+C^2SA, PCNN+C^2SA, and all the baselines in Figure 6. We observe that models trained with C^2SA, no matter whether they are PCNNs-based or BLSTMs-based, consistently outperform the other models. In addition, there is no clear difference between PCNN-based methods and BLSTM-based methods. For example, the PCNNs-based baseline (PCNN+ATT) performs worse than the BLSTMs-based baseline (BLSTM+2ATT). However, with our proposed C^2SA, the PCNNs-based method (PCNN+C^2SA) outperforms the BLSTMs-based method (BLSTM+C^2SA). We report the ablation studies later.

At the same time, we observe that the choice of the training setting has a large impact on the performances. That is, with the model trained more robustly to noise, there are clear improvements in performances. These observations have further verified our intuition for improving the distantly-supervised relation extraction with a more robust training.

Comparison on the Sentence-level Task

For an extensive comparison, we further evaluate the performances of our method on the sentence-level relation extraction. Different from the corpus-level task, this task aims to identify the relation type for an entity pair with regard to a specific sentence. More precisely, we feed the sentence representation to the output layer and observe the prediction.

Test Set   For the sentence-level evaluation, we adopt the dataset used in the existing literature (Hoffmann et al. 2011). It contains 395 sentences with human annotations. Compared with the test set for the corpus-level task, this set is small in size. However, considering the fact that this dataset is manually annotated, it makes sense to report this comparison study.

Competing Methods   We compare C^2SA with two major baselines, i.e., PCNN+ATT and BLSTM+2ATT. Besides, we also make comparison with two variants of C^2SA, i.e., CRSA and C^2SA-dot. CRSA only performs the cross-relation selective attention and trains the model at the sentence bag level. C^2SA-dot, on the other hand, changes the scoring function in Equation 3 and Equation 5 from the cosine similarity to the dot product.

Performance Comparison   As reported in Table 3, we observe that C^2SA outperforms both baselines and the variants. Besides, we observe that CRSA outperforms ATT and C^2SA outperforms CRSA. This observation verifies the effectiveness of the two proposed techniques. In addition, we observe that after replacing the cosine similarity with dot product in C^2SA, the performance drops significantly. This shows that when serving as the scoring function for the selective attention, cosine similarity is more effective than the dot product.

Ablation Study

To further demonstrate the effectiveness of the two proposed selective attention mechanisms, we now report the ablation

PCNNs-based	F1	BLSTMs-based	F1
PCNN+ATT	0.377	BLSTM+2ATT	0.382
PCNN+CRSA	0.411	BLSTM+CRSA	0.409
PCNN+C^2SA	0.421	BLSTM+C^2SA	0.448
PCNN +C^2SA-dot	0.400	BLSTM +C^2SA-dot	0.401

Table 3: Performance comparison on the sentence-level relation extraction.

studies.

To examine the effectiveness of the cross-relation selective attention, we compare CRSA with the vanilla selective attention (ATT). Specifically, CRSA only performs the cross-relation selective attention; both CRSA and ATT conduct the training at the sentence bag level. As to the cross-bag selective attention, it is based on the sentence bag representation. In other words, only the sentence-level attention is required. Accordingly, we demonstrate the effectiveness of the cross-bag selective attention by comparing CRSA with the full C^2SA.

For a fair comparison, we compare the two types of relation extractions separately as we did before. As shown in Figure 7, all the models are trained with the same training set introduced before and are evaluated on the corpus-level task. We employ the Precision-Recall curve for the comparison studies with the summarized performances of PCNNs-based relation extractions in Figure 7(a) and the summarized performances of BLSTMs-based ones in Figure 7(b).

We observe that C^2SA achieves the best performance in both cases. Also, we note that the PCNNs-based model achieves a better improvement than the BLSTMs-based model. For example, there is an obvious margin between the PCNN+CRSA and PCNN+ATT. We believe that this observation is due to the difference between the characteristics of the different neural feature extractors. The CNNs-type models are good at extracting local information (such as trigger words) of a sentence reflected in the dimensions of their feature vectors, leading to the observation that cosine similarity based attention mechanism delivers a better performance.

On the other hand, it is observed that the gap between C^2SA and CRSA verifies the fact that some sentence bags are of a higher quality than others. This observation demonstrates that the proposed cross-bag selective attention allows the model training at the superbag level with a more robust performance. We report the case studies below to support this observation.

Case Study

Noise of Distant Supervision at Bag-level   Now we report the case studies to verify the intuition about the cross-bag selective attention and superbag learning. We randomly select 20 different relation types from the NYT dataset, randomly select 100 entity pairs, construct 100 sentence bags, and manually examine their qualities. Specifically, these sen-
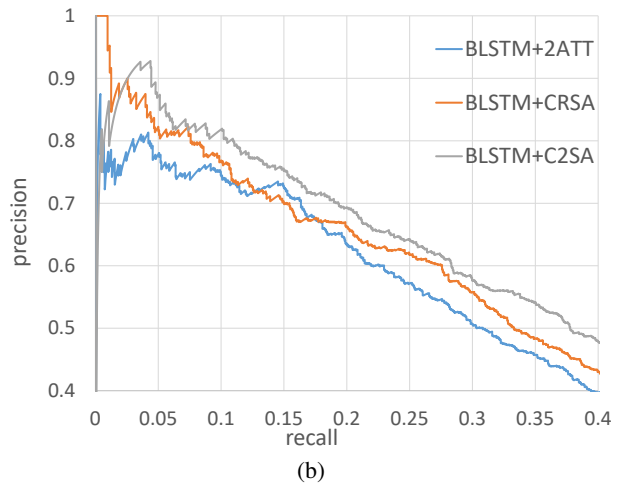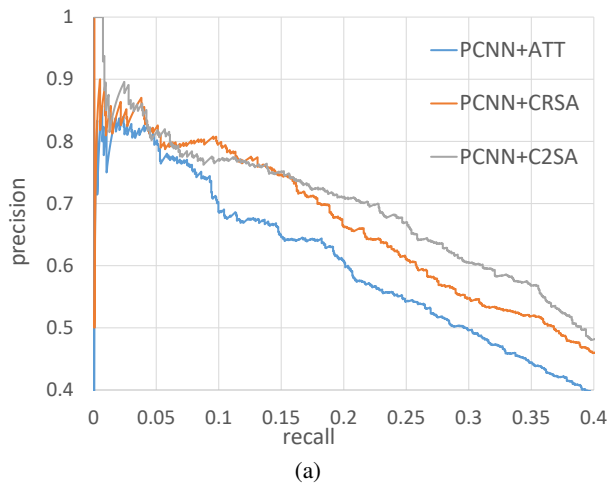
Figure 7: (a) Aggregate precision/recall curves of PCNN+ATT, PCNN+CRSA, PCNN+C^{2}SA (b) Aggregate precision/recall curves of BLSTM+2ATT, BLSTM+CRSA, BLSTM+C^{2}SA

Total count of sentence bags	100
Contain at least one sentence	69
All annotations are incorrect	31

Table 4: Manually checked qualities for the sampled sentence bags with distant supervision

tence bags contain 483 sentences and their qualities are summarized in Table 4.

We observe that, even after aggregating sentences to construct sentence bag, the distant supervision is still noisy. Actually, about 31% of all the sentence bags do not even contain one sentence that is correctly annotated. Therefore, it is fair to conclude that the superbag-level training is necessary to handle the noise of distant supervision.

Effectiveness of Cross-bag Selective Attention We further examine whether the proposed cross-bag selective attention is capable of properly handling the noise from distant supervision. Table 5 shows a superbag for the relation type `lived_in`. It contains two sentence bags. One is for the KB fact <*Hunter_s*, `lived_in`, *Colorado*>, and the other is for the KB fact <*Dan_bartlett*, `lived_in`, *Texas*>. We observe that there is only one sentence that is correctly annotated in the first sentence bag, and that all the sentences in the second bag are unmatched with the relation type of the KB fact. Clearly, the sentence-level selective attention would fail to handle the second bag. On the other hand, the proposed cross-bag selective attention allows the model to focus more on the sentence bag with a higher quality. For example, in Table 5, such selective attention reduces the effect of the distant supervision noise by assigning a smaller attention weight to the second sentence bag.

		Superbag label: person-place_lived		
KB-Facts		Sentences	Sentence attention	Bag attention
Hunter_s._thom-pson, `lived_in`, *Colo-rado*		**Hunter_s._thompson** who committed suicide last month in **Colorado** ...	low	high
		Hunter_s._thompson lived and wrote here in the high rocky mountains of central **Colorado** ...	high	
Dan_bartlett, `lived_in`, *Texas*		President is residing in **Texas**, "said **Dan_bartlett**, counselor to the president.	medium	low
		Dan_bartlett, president Bush's counselor, said ... in **Texas** ...	medium	

Table 5: A case study for two sentence bags and their corresponding superbag

Conclusion and Future Work

In this paper, we have proposed the Cross-relation Cross-bag Selective Attention to develop a better relation-entity model to effectively learn the features of true expression relations from typically noisy distant supervision data. Experiments show that the proposed attention model is capable of learning higher quality bag features than the existing literature. In addition, it is also demonstrated that Cross-bag Selective Attention is further capable of boosting the performances through the high-quality bag features.

Acknowledgments

References

Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; and Taylor, J. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of SIGMOD*, 1247–1250. AcM.

Etzioni, O.; Cafarella, M.; Downey, D.; Kok, S.; Popescu, A.-M.; Shaked, T.; Soderland, S.; Weld, D. S.; and Yates, A. 2004. Web-scale information extraction in knowitall:(preliminary results). In *Proceedings of WWW*, 100–110. ACM.

Feng, J.; Huang, M.; Zhao, L.; Yang, Y.; and Zhu, X. 2018. Reinforcement learning for relation classification from noisy data. In *Proceedings of AAAI*.

Hinton, G. E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. R. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Hoffmann, R.; Zhang, C.; Ling, X.; Zettlemoyer, L.; and Weld, D. S. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of ACL-HLT*, 541–550. ACL.

Huang, Y., and Wang, W. Y. 2017. Deep residual learning for weakly-supervised relation extraction. In *Proceedings of EMNLP*, 1803–1807.

Ji, G.; Liu, K.; He, S.; and Zhao, J. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *AAAI*.

Jiang, X.; Wang, Q.; Li, P.; and Wang, B. 2016. Relation extraction with multi-instance multi-label convolutional neural networks. In *Proceedings of COLING*.

Li, L.; Nie, Y.; Han, W.; and Huang, J. 2017. A multi-attention-based bidirectional long short-term memory network for relation extraction. In *International Conference on Neural Information Processing*, 216–227. Springer.

Lin, Y.; Shen, S.; Liu, Z.; Luan, H.; and Sun, M. 2016. Neural relation extraction with selective attention over instances. In *ACL (1)*.

Liu, L.; Ren, X.; Zhu, Q.; Zhi, S.; Gui, H.; Ji, H.; and Han, J. 2017a. Heterogeneous supervision for relation extraction: A representation learning approach. *arXiv preprint arXiv:1707.00166*.

Liu, T.; Wang, K.; Chang, B.; and Sui, Z. 2017b. A soft-label method for noise-tolerant distantly supervised relation extraction. In *Proceedings of EMNLP*, 1790–1795.

Luo, B.; Feng, Y.; Wang, Z.; Zhu, Z.; Huang, S.; Yan, R.; and Zhao, D. 2017. Learning with noise: Enhance distantly supervised relation extraction with dynamic transition matrix. *Transformation*.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mintz, M.; Bills, S.; Snow, R.; and Jurafsky, D. 2009. Distant supervision for relation extraction without labeled data. In *ACL-ICJNLP*, 1003–1011. ACL.

Mooney, R. J., and Bunescu, R. C. 2006. Subsequence kernels for relation extraction. In *NIPS*.

Riedel, S.; Yao, L.; and McCallum, A. 2010. Modeling relations and their mentions without labeled text. *MLKDD* 148–163.

Rink, B., and Harabagiu, S. 2010. Utd: Classifying semantic relations by combining lexical and semantic resources. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, 256–259. ACL.

Santos, C. N. d.; Xiang, B.; and Zhou, B. 2015. Classifying relations by ranking with convolutional neural networks. *arXiv preprint arXiv:1504.06580*.

Socher, R.; Huval, B.; Manning, C. D.; and Ng, A. Y. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of EMNLP-CoNLL*, 1201–1211. ACL.

Surdeanu, M.; Tibshirani, J.; Nallapati, R.; and Manning, C. D. 2012. Multi-instance multi-label learning for relation extraction. In *EMNLP-CoNLL*. ACL.

Vu, N. T.; Adel, H.; Gupta, P.; and Schütze, H. 2016. Combining recurrent and convolutional neural networks for relation classification. *arXiv preprint arXiv:1605.07333*.

Wang, L.; Cao, Z.; de Melo, G.; and Liu, Z. 2016. Relation classification via multi-level attention cnns. In *ACL*, volume 1, 1298–1307.

Xu, K.; Feng, Y.; Huang, S.; and Zhao, D. 2015a. Semantic relation classification via convolutional neural networks with simple negative sampling. *arXiv preprint arXiv:1506.07650*.

Xu, Y.; Mou, L.; Li, G.; Chen, Y.; Peng, H.; and Jin, Z. 2015b. Classifying relations via long short term memory networks along shortest dependency paths. In *Proceedings of EMNLP*, 1785–1794.

Zelenko, D.; Aone, C.; and Richardella, A. 2003. Kernel methods for relation extraction. *JMLR* 3(Feb):1083–1106.

Zeng, D.; Liu, K.; Lai, S.; Zhou, G.; Zhao, J.; et al. 2014. Relation classification via convolutional deep neural network. In *COLING*, 2335–2344.

Zeng, D.; Liu, K.; Chen, Y.; and Zhao, J. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *EMNLP*.

Zeng, X.; He, S.; Liu, K.; and Zhao, J. 2018. Large scaled relation extraction with reinforcement learning. *Relation* 2:3.

Zhang, Y.; Zhong, V.; Chen, D.; Angeli, G.; and Manning, C. D. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of EMNLP*, 35–45.

Zhou, P.; Shi, W.; Tian, J.; Qi, Z.; Li, B.; Hao, H.; and Xu, B. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of ACL*, volume 2, 207–212.