Distant Supervised Relation Extraction with Separate Head-Tail CNN

Rui Xing, Jie Luo*
State Key Laboratory of Software Development Environment
School of Computer Science and Engineering
Beihang University, China
{xingrui,luojie}@nlsde.buaa.edu.cn

Abstract

Distant supervised relation extraction is an efficient and effective strategy to find relations between entities in texts. However, it inevitably suffers from mislabeling problem and the noisy data will hinder the performance. In this paper, we propose the Separate Head-Tail Convolution Neural Network (SHTCNN), a novel neural relation extraction framework to alleviate this issue. In this method, we apply separate convolution and pooling to the head and tail entity respectively for extracting better semantic features of sentences, and coarse-to-fine strategy to filter out instances which do not have actual relations in order to alleviate noisy data issues. Experiments on a widely used dataset show that our model achieves significant and consistent improvements in relation extraction compared to statistical and vanilla CNN-based methods.

1 Introduction

Relation extraction is a fundamental task in information extraction, which aims to extract relations between entities. For example, "Bill Gates is the CEO of Microsoft." holds the relationship **/business/company/founders** between the head entity **Bill Gates** and tail entity **Microsoft**.

Traditional supervised relation extraction systems require a large amount of manually well-labeled relation data (Walker et al., 2005; Doddington et al., 2004; Gábor et al., 2018), which is extremely labor intensive and time-consuming. (Mintz et al., 2009) instead proposes distant supervision which exploits relational facts in knowledge bases. Distant supervision aligns entity mentions in plain texts with those in knowledge base and assumes that if two entities have a relation there, then all sentences containing these two entities will express that relation. If there is no relation link between a certain entity pair in knowledge base, the sentence will be labeled as a Not A relation (NA) instance. Although distant supervision is an efficient and effective strategy for automatically labeling large-scale training data, it inevitably suffers from mislabeling problems due to its strong assumption. As a result, the dataset created by distant supervision is usually very noisy. According to (Riedel et al., 2010), the precision of using distant supervision aligning Freebase to New York Times corpus is about 70%, an example of labeled sentences in New York Times corpus is shown in Table 1. Therefore, many efforts have been devoted to alleviate noise in distant supervised relation extraction.

Bag	Sentence	Correct
b_1	**Barack Obama** was born in the **United States**.	True
	Barack Obama was the 44th president of the **United States**.	False
b_2	**Bill Gates** is the CEO of **Microsoft**.	True
	Bill Gates announced that he would be transitioning to a part-time role at **Microsoft** and full-time work in June 2006.	False

Table 1: Examples of relations annotated by distant supervision. Sentences in b_1 are annotated with the *place_of_birth* relation and sentences in b_2 the *business_company_founders* relation.

With the development of deep learning techniques (LeCun et al., 2015), large amount of work using deep neural networks has been proposed for distant supervised relation extraction (Zeng et al., 2014, 2015; Lin et al., 2016; Liu et al., 2017; Jat et al., 2018; Ji et al., 2017; Han et al., 2018;

*Corresponding author.

Proceedings of the 2019 EMNLP Workshop W-NUT: The 5th Workshop on Noisy User-generated Text, pages 249–258
Hong Kong, Nov 4, 2019. ©2019 Association for Computational Linguistics

Du et al., 2018; Vashishth et al., 2018; Lei et al., 2018; Qin et al., 2018a,b; Ye and Ling, 2019; Xu and Barbosa, 2019). Various previous work also used well-designed attention mechanism (Lin et al., 2016; Jat et al., 2018; Ji et al., 2017; Su et al., 2018; Du et al., 2018) which have achieved significant results. Besides, knowledge-based methods (Lei et al., 2018; Han et al., 2018; Vashishth et al., 2018; Ren et al., 2018) incorporated external knowledge base information with deep neural network, obtaining impressive performance.

Most of previous work used vanilla Convolution Neural Network (CNN) or Piecewise Convolution Neural Network (PCNN) as sentence encoder. CNN/PCNN adopted the same group of weight-sharing filters to extract semantic feature of sentences. Though effective and efficient, there is still room to improve if we look deeper into properties of relations. We find that semantic properties of relations such as symmetry and asymmetry are often overlooked when using CNN/PCNN. For example, “Bill Gates is the CEO of Microsoft.” holds the relationship **/business/company/founders** between the head entity **Bill Gates** and tail entity **Microsoft**. While in the sentence “The most famous man in Microsoft is Bill Gates.“ where the head entity **Microsoft** and the tail **Bill Gates** do not share that relationship. It indicates that the relation **/business/company/founders** is asymmetric. Most previous work use position embedding specified by entity pairs and piecewise pooling (Zeng et al., 2015; Lin et al., 2016; Liu et al., 2017; Han et al., 2018) to predict relations. However, above examples show that they share similar position embeddings due to their similar position distances to both entities. Vanilla CNN/PCNN is not sufficient to capture such semantic features because it treats the head and tail entities equally. Thus, it tend to “memorize” certain entity pairs and may learn similar context representation when dealing with these noisy asymmetric instances.

In addition to relation properties, we also investigate some noise source in distant supervised relation extraction. NA instances usually account for a large portion in distant supervised datasets, making the data highly imbalanced. Similarly, in objection detection task (Lin et al., 2017), extreme class imbalance greatly hinders the performance.

In this paper, in order to deal with above deficiencies, we propose Separate Head-Tail CNN (SHTCNN) framework, an effective strategy for distant supervised relation extraction. The framework is composed of two ideas. First, we employ separate head-tail convolution and pooling to embed the semantics of sentences targeting head and tail entities respectively. By this means, we can capture better semantic properties of relations in the distant supervised data and further alleviate mislabeling problem. Second, relations are classified from coarse to fine. In order to do this, an extra auxiliary network is adopted for NA/Non-NA binary classification, which is expected to filter as many easy NA instances as possible while maintaining high recall of all non-NA relationships. Instances selected by binary network are treated as non-NA examples for fine-grained multi-class classification. Inspired by Retina (Lin et al., 2017), we make use of focal loss in binary classification. We evaluate our model on a real-world distant supervised dataset. Experimental results show that our model achieves significant and consistent improvements in relation extraction compared to selected baselines.

2 Related Work

Relation extraction is a crucial task and heavily studied area in Natural Language Processing (NLP). Many efforts have been devoted, especially in supervised paradigm. Conventional supervised methods require large amounts of human-annotated data, which is highly expensive and time-consuming. To deal with this issue, (Mintz et al., 2009) proposed distant supervision, which aligned Freebase relational facts with plain texts to automatically generate relation labels for entity pairs. Apparently, such assumption is too strong that inevitably accompanies with mislabeling problem.

Plenty of studies have been done to alleviate such problem. (Riedel et al., 2010; Hoffmann et al., 2011; Surdeanu et al., 2012) introduce multi-instance learning framework to the problem. (Riedel et al., 2010) and (Surdeanu et al., 2012) use a graphical model to select valid sentences in the bag to predict relations. However, the main disadvantage in conventional statistical and graphical methods is that using features explicitly derived from NLP tools will cause error propagation and low precision.

As deep learning techniques (Bengio, 2009; LeCun et al., 2015) have been widely used, plenty of work adopt deep neural network for distant su-

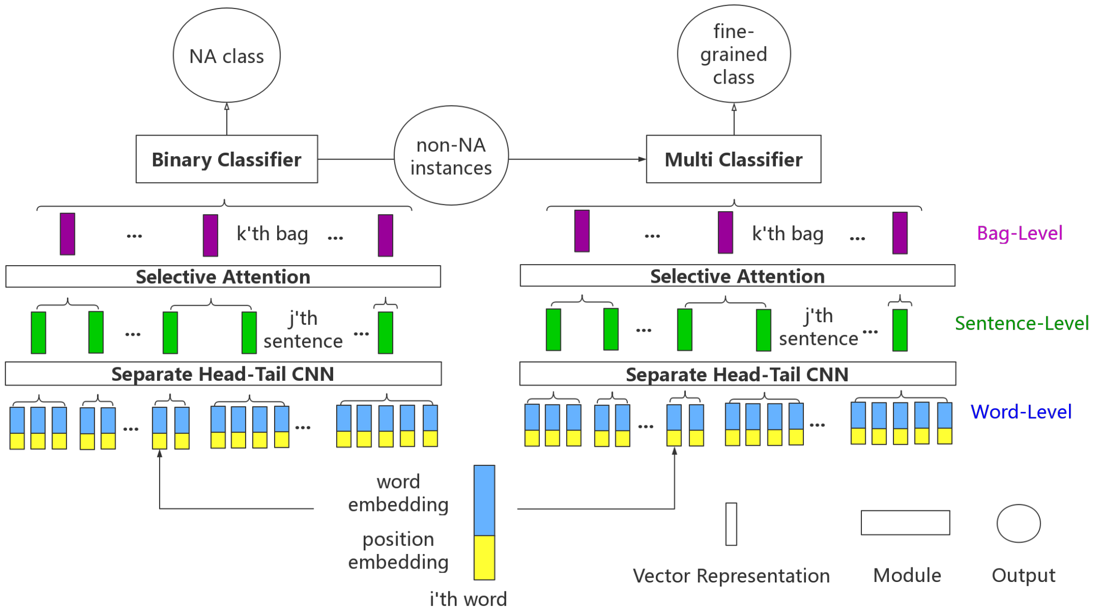

Figure 1: Separate Head-Tail CNN for distant supervised relation extraction

pervised relation extraction. (Zeng et al., 2015) proposed piecewise convolution neural network to model sentence representations under multi instance learning framework while using piecewise pooling based on entity position to capture structural information. (Lin et al., 2016) proposed sentence level attention, which is expected to dynamically reduce the weights of those noisy instances. (Ji et al., 2017) adopted similar attention strategy and combined entity descriptions to calculate weights over sentences. (Liu et al., 2017) proposed a soft-label method to reduce the influence of noisy instances on entity-level. (Jat et al., 2018) used word-level and entity-based attention for efficiently relation extraction. Due to the effectiveness of self-attention mechanism, (Du et al., 2018) proposed a structured word-level self-attention and sentence-level attention mechanism which are both 2-D matrix to learn rich aspects of data. Also, plenty of knowledge based strategies for distant supervised relation extraction have also been proposed. (Ji et al., 2017) uses hierarchical information of relations for relation extraction and achieve significant performance. (Lei et al., 2018) proposed Cooperative Denoising framework, which consists two base networks leveraging text corpus and knowledge graph respectively. (Vashishth et al., 2018) proposed RESIDE, a distantly supervised neural relation extraction method which utilizes additional side information from knowledge bases for improving relation extraction. (Han et al., 2018) aimed to incorporate the hierarchical information of relations for distantly supervised relation extraction. Although these methods achieved significant improvement in relation extraction, they tend to treat entities in sentences equally or rely more or less on knowledge base information which may be unavailable in other domains.

In order to alleviate mislabeling problem and reduce the burden of integrating external knowledge and resource, we propose SHTCNN to provide better sentence representation and reduce the impact of NA instances.

3 Methodology

In this section, we introduce our SHTCNN model. The overall framework is shown in Figure 1. Our model is built under multi-instances learning framework. It splits the training set into multiple n bags $\{\langle h_1, t_1\rangle, \langle h_2, t_2\rangle, \cdots, \langle h_n, t_n\rangle\}$, each of which contains m sentences $\{s_1, s_2, \cdots, s_m\}$ mentioning same head entity h_i and tail entity t_i. Note that sentence number m may not be the same in each bag. Each sentence consists of a sequence of k words $\{x_1, x_2, \cdots, x_k\}$. First, sentence representation s_i is acquired using our separate head-tail convolution and pooling on words $\{x_1, x_2, \cdots, x_k\}$. Next, selective attention mechanism is used to dynamically merge sentences to its bag representation $b_i = \langle h_i, t_i\rangle$. On bag level, binary classifier filters out easy NA instances with focal loss, leaving others to multi-class classifier for further fine-grained classification.

3.1 Sentence Encoder

Word Representation

First, the i-th word x_i in sentence is mapped into a d_w-dimensional word embedding e_i. Then, to keep track of head and tail entity position information, two d_p-dimensional position embeddings (Zeng et al., 2014, 2015) are also adopted for each word as p_i^1 and p_i^2 recording the distance to two entities respectively. Thus, the final word representation is the concatenation of these three vectors $w_i = [e_i, p_i^1, p_i^2]$ of $d = d_w + 2p_w$ dimensions.

Separate Head-Tail Convolution and Pooling

Convolution layer are often utilized in relation extraction to capture local features in window form and then perform relation prediction globally. In detail, convolution is an operation between a convolution matrix W and a sequence of vector q_i. We define $q_i \in R^{l \times d}$ of w words in the sentence $s_i = \{w_1, w_2, w_3, \cdots, w_n\}$ with word representations defined above.

$$q_i = w_{i-l+1:i}, \quad \text{where } 1 \le i \le m + l - 1 \quad (1)$$

Because the window may be out of the sentence boundary when sliding along. We use wide convolution technique by adding special padding tokens on both sides of sentence boundaries. Thus the i-th convolutional filter p_i computes as follows:

$$p_i = [Wq + b]_i, \quad (2)$$

where b is bias vector.

Conventional PCNN uses piecewise pooling for relation extraction which divided convolutional filter p_i into three segments based on positions of head and tail entities. Piecewise pooling is defined as follows:

$$[x]_{ij} = max(p_{ij}), \quad \text{where } 1 \le j \le 3 \quad (3)$$

where j indicates position of segments in sentence.

As mentioned in section, traditional methods get representation of each sentence using same group of convolution filters, which focuses on both head entity and tail entity equally and ignores semantic difference between them. We use two separate groups of convolution filters $W_1, W_2 \in R^{d_s \times d}$, where d_s is the sentence embedding size. Also, simply piecewise pooling can not well deal with examples of which relations are similar but asymmetric. In detail, we utilize two groups of separate head-tail entity convolution W^1, W^2 to represent the sentence s_i as p_i^1, p_i^2.

$$\begin{aligned} p_i^1 &= [Wq + b]_i^1 \\ p_i^2 &= [Wq + b]_i^2 \end{aligned} \quad (4)$$

To exploit such semantic properties of relations expressed by entity pairs, we use separate head-tail entity pooling. Targeting head and tail entities, head-entity pooling and tail-entity pooling are adopted on two convolution results respectively. p_i^1, p_i^2 are further segmented by positions of entity pair for head-tail entity pooling. Head entity pooling is defined as:

$$h_i = [max(p_{i1}^1); max([p_{i2}^1, p_{i3}^1])] \quad (5)$$

Similarly, tail pooling is defined as:

$$t_i = [max([p_{i1}^2, p_{i2}^2]); max(p_{i3}^2)] \quad (6)$$

And i-th sentence vector s_i is the concatenation of h_i and t_i:

$$s_i = [h_i; t_i] \quad (7)$$

Finally, we apply non-linear function such as ReLU as activation on the output.

3.2 Selective Attention

Bags contain sentences sharing the same entity pair. In order to alleviate mislabeling problem on sentence level, we adopted selective attention which is widely used in many works (Lin et al., 2016; Liu et al., 2017; Ji et al., 2017; LeCun et al., 2015; Han et al., 2018; Du et al., 2018). The representation of the bag $b_i = \langle h_i, t_i \rangle$ is the weighted sum of all sentence vectors in that bag.

$$\begin{aligned} b_i &= \sum_i \alpha_i s_i \\ \alpha_i &= \frac{exp(s_i A r)}{\sum_j exp(s_j A r)} \end{aligned} \quad (8)$$

where α_i is the weight of sentence representation s_i, A and r are diagonal matrix and relation query.

3.3 Coarse-to-Fine Relation Classification

Traditional methods directly predict relation classes for each bag after obtaining bag representations. However, large amount of NA instances containing mixed semantic information will hinder the performance. To alleviate such impact of

NA instances, we manually utilize a binary classifier to filter out as many NA instances as possible, while leaving hard NA instances for multi-class classification.

Binary classification can also be viewed as an auxiliary task about whether the input sentence hold an NA relation. In this method, NA is treated as negative class while all other non-NA labels are treated as positive class. In this method, we adopted focal loss (Lin et al., 2017) for NA/non-NA classification. Focal loss is designed to address class imbalance problem. When predict class label y for binary task $y \in \{0, 1\}$, we first define the prediction score p_t for positive class:

$$p_t = \begin{cases} p, & \text{if } y = 1, \\ 1 - p, & \text{otherwise} \end{cases} \tag{9}$$

Then traditional weighted cross-entropy loss can be defined as follows:

$$CE(p_t) = -\alpha log(p_t) \tag{10}$$

where α is a hyper-parameter usually set as class ratio.

Focal loss modifies it by changing α to $(1-p_t)^\gamma$ in order to dynamically adjust weights between well-classified easy instances and hard instances as:

$$CE(p_t) = -(1 - p_t)^\gamma log(p_t) \tag{11}$$

For easy instances, prediction score p_t will be high while the loss low and vise versa for hard instances. As a result, focal loss focuses on those hard NA instances. Finally, instances which are predicted as non-NA are selected for multi-class classifier for fine-grained classification. Due to existence of NA instances which are hard to handle, we also add a “NA class” in multi-class classification for further filtering those instances which do not hold an exact relationship.

3.4 Optimization

In this section, we introduce the learning and optimization details for our SHTCNN model. As shown in Figure 1, binary and multi network share only same word representations. We define binary and multi labels as $br \in \{0, 1\}$ and $mr \in \{0, 1, 2, \cdots, n\}$ respectively. Both 0 represent NA class. In binary classification, 1 represents all non-NA classes while in multi-class classification, each non-zero number represents a certain non-NA relation. Besides, we use Θ^1, Θ^2 to denote parameters for binary and multi-class classification network respectively. The objective function for our model is:

$$\begin{aligned} J(\Theta^1, \Theta^2) = -\sum_{i=0}^{1} log(br_i|b_i, \Theta^1) \\ -\sum_{j=0}^{n} log(mr_j|b_i, \Theta^2) \end{aligned} \tag{12}$$

where n is the number of relation classes. All models are optimized using Stochastic Gradient Descent (SGD).

4 Experiments

In this section, we first introduce the dataset and evaluation metrics. Then we list our experimental parameter settings. Afterwards, we compare the performance of our method with feature-based and selected neural-based methods. Besides, case study shows our SHTCNN is an effective method to extract better semantic features.

4.1 Dataset and Evaluation Metrics

We evaluate our model on a widely used dataset New York Times (NYT) released by (Riedel et al., 2010). The dataset was generated by aligning Freebase (Bollacker et al., 2008) relations with New York Times Corpus. Sentences of year 2005 and 2006 are used for training while sentences of 2007 are used as testing. There are 52 actual relations and a special NA which indicates there was no relation between two entities. The training set contains 522,611 sentences, 281,270 entity pairs and 18,152 relational facts. The testing set contains 172,448 sentences, 96,678 entity pairs and 1950 relational facts.

4.2 Comparison with Baseline Methods

Following previous work (Mintz et al., 2009; Lin et al., 2016; Ji et al., 2017; Liu et al., 2017; Han et al., 2018; Du et al., 2018), we evaluate our model in the held-out evaluation. It evaluates models by comparing the relational facts discovered from the test articles with those in Freebase, which provides an approximate measure of precision without requiring expensive human evaluation. We draw precision-recall curves for all models and also report the Precision@N results to further verify the effort of our SHTCNN model.

For fair comparison with sentence encoders, we selected the following baselines:

- **Mintz:** Multi-class logistic regression model used by (Mintz et al., 2009) for distant supervision.
- **MultiR:** Probabilistic graphical model under multi-instance learning framework proposed by (Hoffmann et al., 2011)
- **MIMLRE:** Graphical model jointly models multiple instances and multiple labels proposed by (Surdeanu et al., 2012)
- **PCNN:** CNN based model under multi-instance learning framework for distant relation extracion proposed by (Zeng et al., 2015)
- **PCNN-ATT:** CNN based model which uses additional attention mechanism on sentence level for distant supervision proposed by (Lin et al., 2016)
- **SHTCNN:** Framework proposed in this paper, please refer to Section 3 for more details.

4.3 Experimental Settings

Word and Position Embeddings

Our model use pre-trained word embeddings for NYT corpus. Word embeddings of blank words are initialized with zero while unknown words are initialized with the normal distribution of which the standard deviation is 0.05. Position embeddings are initialized with Xavier initialization for all models. Two parts of our model share the same word and position embeddings as inputs.

Parameter Settings

We use cross-validation to determine the parameters in our model. We also use a grid search to select learning rate λ for SGD among $\{0.5, 0.1, 0.01, 0.001\}$, sliding windows size l among $\{1, 3, 5, 7\}$, sentence embedding size d_s among $\{100, 150, 200, 300, 350, 400\}$ and batch size among $\{64, 128, 256, 512\}$. Other parameters proved to have little effect on results. We show our optimal parameter settings in Table 2.

Word Embedding Size	50
Position Embedding Size	5
Sentence Embedding Size	230
Filter Window Size	3
γ in Focal Loss	2
Positive weight in Focal Loss	0.75
Threshold for Selecting non-NA	0.3
Batch Size	128
Learning rate	0.1
Dropout Probability	0.5

Table 2: Parameter Settings

4.4 Overall Performance

Figure 2 shows the overall performance of our proposed SHTCNN against baselines mentioned above. From results, we can observe that: (1) When recall is smaller than 0.05, all models have reasonable precision. When recall is higher, precision of feature-based models decrease sharply compared to neural-based methods, and the latter outperform the former over the entire range of recall. It demonstrates that human-designed features are limited and cannot concisely express semantic meaning of sentences in noisy data environment. (2) SHTCNN outperforms PCNN/PCNN-ATT over the entire range of recall, It indicates that SHTCNN is a more powerful sentence encoder which can better capture semantic features of noisy sentences. Further experimental results and case study show the effectiveness of our model.

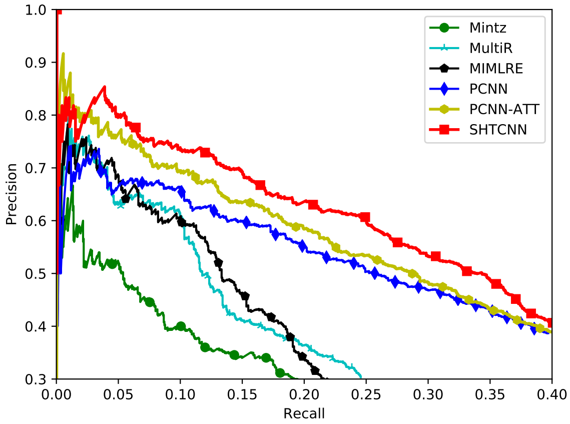

Figure 2: Precision/Recall curves of our model and previous baseline methods.

4.5 Top N Precision

We also conduct Precision@N tests on entity pairs with few instances. In our tests, three settings are used: ONE randomly select an instance in the bag; TWO randomly select two instances for each entity pair; ALL use all bag instances for evaluation. Table 3 shows the results on NYT dataset regarding P@100, P@200, P@300 and the mean of three settings for each model. From the table we can see that: (1) Performance of all methods improves as

Test Settings	One				Two				All			
P@N(%)	100	200	300	Mean	100	200	300	Mean	100	200	300	Mean
PCNN+AVE	71.3	63.7	57.8	64.3	73.3	65.2	62.1	66.9	73.3	66.7	62.8	67.6
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
SHTCNN+AVE	72.3	64.2	60.1	65.5	76.3	71.3	**68.9**	**72.2**	**77.2**	**76.6**	**71.4**	**75.1**
Coarse-to-Fine	74.3	69.6	63.2	69.0	77.7	74.4	68.2	73.4	78.6	74.3	71.2	74.7
HT+ATT	75.3	74.3	65.1	71.6	79.2	75.6	72.3	75.7	80.4	76.2	74.9	77.2
SHTCNN+ATT	**78.2**	**77.1**	**70.1**	**75.1**	**80.0**	**76.2**	**73.2**	**76.5**	**86.1**	**79.1**	**75.4**	**80.2**

Table 3: P@N for relation extraction in entity pairs with different number of sentences

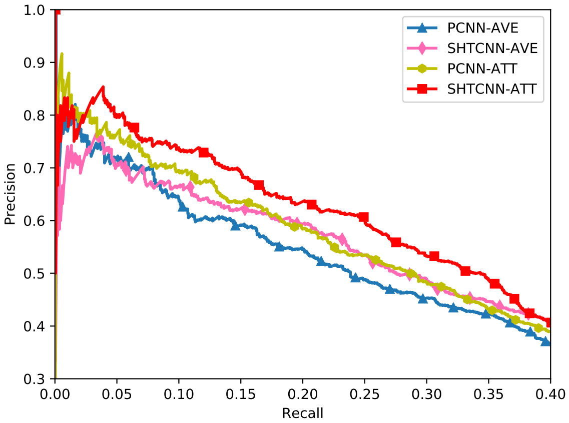

Figure 3: Precision/Recall curves of our model and selected neural based methods. PCNN-AVE and SHTCNN-AVE use Average method (AVE) while PCNN-ATT and SHTCNN-ATT use selective ATTention method (ATT) described in section 3.2 to obtain bag representation from its sentences.

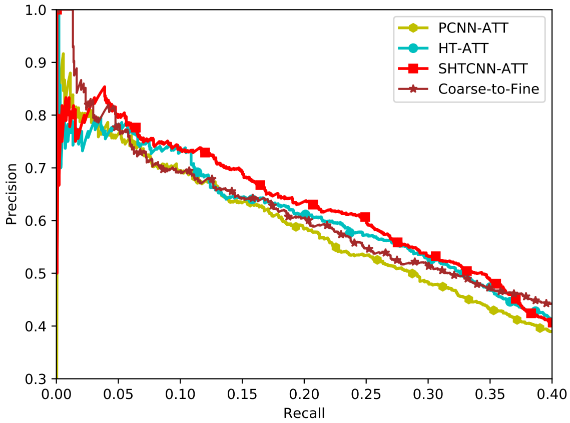

Figure 4: Precision/Recall curves of separate parts of our model. PCNN-ATT is the baseline method introduced in Section 4.2. HT-ATT replaces sentence encoder of PCNN-ATT with separate Head-Tail convolution and pooling (HT) proposed in section 3.1 without using Coarse-to-fine strategy in section 3.3. Coarse-to-Fine solely classifies relation from coarse to fine without using HT. SHTCNN-ATT is our full model combines HT and Coarse-to-Fine relation extraction together.

the instance number increases which shows that more sentences selected in the bag, more information can be utilized. (2) SHTCNN improves precision by over 8% for PCNN, PCNN-AVE and PCNN-ATT model. It indicates that in noisy textual dataset, our SHTCNN is a more powerful sentence encoder to capture better semantic features. (3) Average method improves slowly when instances number increases which indicates that it can not effectively extract relations and be easily distracted by noises in the bag.

4.6 Effectiveness of Separate Head-Tail CNN

To further verify the contribution and effectiveness of two phase of our SHTCNN, we conduct two extra experiments. First, we evaluate the ability of our model to capture better sentence semantic features under different bag representation calculation methods. PCNN-AVE (Average) assumes that all sentences in the bag contribute equally to the representation of the bag, which brings in more noise from mislabeling sentences. Compared to PCNN-ATT, PCNN-AVE hinders the performance of relation extraction as shown in Table 3. We evaluate our model using Average and Attention respectively. From results in Figure 3, we observe that: (1) Both SHTCNN-AVE and SHTCNN-ATT achieve significant performance than their compared baselines, which proves that SHTCNN offers better sentence semantic features for bag representation with or without selective attention mechanism. (2) SHTCNN-AVE achieves similar performance as PCNN-ATT when recall is between 0.15 and 0.35. (3) When recall is greater than 0.35, SHTCNN-AVE performs even better than PCNN-ATT. It demonstrates that SHTCNN is relatively more robust and stable on dealing with noisier sentences.

Second, we explore the effect of separate head-tail convolution and pooling and contribution of coarse-to-fine relation extraction. From results shown in Figure 4, we can observe that: (1) Both HT-ATT and Coarse-to-Fine improve performance of PCNN-ATT on a wide range of recall, which indicates that separate head-tail convolution and pooling, and coarse-to-fine strategy perform better on predicting relations. (2) Figure 4 and Table 3 both show that separate head-tail convolution and pooling achieve much better results than only using coarse-to-fine strategy, indicating that a better sentence encoder is more important in noisy environment. (3) Our full model SHTCNN improves performance on the entire recall compared to using separate parts (solely separate head-tail convolution and pooling or only coarse-to-fine) of our model which suggests that combining two proposed methods together can achieve better results.

/business/company/founders
That may include the chairman and chief software architect of **Microsoft**, **Bill Gates**, an otherwise infrequent television viewer.
/business/company/founders → **NA**
Bill Gates and Steve Ballmer, for example, were roommates in college, joined forces at **Microsoft** in 1980 and still work together today.
NA → **/business/shopping_center/owner**
Earlier this week, the company said it expected to sell **Madrid Xanad** and its half-interest in two other malls, Vaughan Mills in Ontario and St. Enoch Centre in Glasgow, to **Ivanhoe Cambridge**, a Montreal company that is Mills's partner in the Canadian and Scottish properties .

Table 4: Some examples of Separate Head-Tail CNN corrections compared to PCNN

4.7 Case Study

In Table 4, we show some of our SHTCNN model examples corrections compared to traditional PCNN. Left of the arrow is PCNN predicted class label on the below sentence while the right is our prediction. We can observe that the first sentence is labeled as **/business/company/founders** by both PCNN and SHTCNN since closer entities bring similar position embeddings which benefit both models. However, the second one is similar but does not hold the relationship. PCNN failed to recognize the relation but SHTCNN corrected the label. Finally, the last sentence is longer and entities are not as close as those in first two sentences. Our model outperformed PCNN by successfully giving correct label to the sentence. It indicates that SHTCNN perform better on modelling relationship in relative long sentences.

5 Conclusion

In this paper, we propose SHTCNN, a novel neural framework using separate head-tail convolution and pooling for sentence encoding and classifies relations from coarse-to-fine. Various experiments conducted show that, in our framework, separate head-tail convolution and pooling can better capture sentence semantic features compared to baseline methods, even in noisier environment. Besides, coarse-to-fine relation extraction strategy can further improve and stabilize the performance of our model.

In the future, we will explore the following directions: (1) We will explore effective separate head-tail convolution and pooling on other sentence encoders like RNN. (2) Coarse-to-fine classification is an experimental method, we plan to further investigate noisy source in distant supervised datasets. (3) It will be promising to incorporate well-designed attention and self-attention mechanisms with two parts of our framework to further improve the performance. All codes and data are available at: https://bit.ly/ds-shtcnn.

6 Acknowledgement

This work is supported by National Natural Science Foundation of China (Grant Nos. 61690202). We would like to thank the anonymous reviewers for their insightful comments.

References

Yoshua Bengio. 2009. Learning deep architectures for ai. *Found. Trends Mach. Learn.*, 2(1):1–127.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 1247–1250, New York, NY, USA. ACM.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph

Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Jinhua Du, Jingguang Han, Andy Way, and Dadong Wan. 2018. Multi-level structured self-attentions for distantly supervised relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2216–2225, Brussels, Belgium. Association for Computational Linguistics.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haïfa Zargayouna, and Thierry Charnois. 2018. SemEval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 679–688, New Orleans, Louisiana. Association for Computational Linguistics.

Xu Han, Pengfei Yu, Zhiyuan Liu, Maosong Sun, and Peng Li. 2018. Hierarchical relation extraction with coarse-to-fine grained attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2245, Brussels, Belgium. Association for Computational Linguistics.

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, Portland, Oregon, USA. Association for Computational Linguistics.

Sharmistha Jat, Siddhesh Khandelwal, and Partha Talukdar. 2018. Improving Distantly Supervised Relation Extraction using Word and Entity Based Attention. *arXiv e-prints*, page arXiv:1804.06987.

Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 3060–3066.

Yann LeCun, Yoshua Bengio, and Geoffrey E. Hinton. 2015. Deep learning. *Nature*, 521(7553):436–444.

Kai Lei, Daoyuan Chen, Yaliang Li, Nan Du, Min Yang, Wei Fan, and Ying Shen. 2018. Cooperative denoising for distantly supervised relation extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 426–436, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal Loss for Dense Object Detection. *arXiv e-prints*, page arXiv:1708.02002.

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, Berlin, Germany. 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, Copenhagen, Denmark. 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*, ACL '09, pages 1003–1011, Stroudsburg, PA, USA. Association for Computational Linguistics.

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, Melbourne, Australia. 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, Melbourne, Australia. Association for Computational Linguistics.

Feiliang Ren, Di Zhou, Zhihui Liu, Yongcheng Li, Rongsheng Zhao, Yongkang Liu, and Xiaobo Liang. 2018. Neural relation classification with text descriptions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1167–1177, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases: Part III*, ECML PKDD'10, pages 148–163, Berlin, Heidelberg. Springer-Verlag.

Sen Su, Ningning Jia, Xiang Cheng, Shuguang Zhu, and Ruiping Li. 2018. Exploring encoder-decoder model for distant supervised relation extraction. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4389–4395. International Joint Conferences on Artificial Intelligence Organization.

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, Jeju Island, Korea. Association for Computational Linguistics.

Shikhar Vashishth, Rishabh Joshi, Sai Suman Prayaga, Chiranjib Bhattacharyya, and Partha Talukdar. 2018. RESIDE: Improving distantly-supervised neural relation extraction using side information. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1257–1266, Brussels, Belgium. Association for Computational Linguistics.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2005. Ace 2005 multilingual training corpus.

Peng Xu and Denilson Barbosa. 2019. Connecting Language and Knowledge with Heterogeneous Representations for Neural Relation Extraction. *arXiv e-prints*, page arXiv:1903.10126.

Zhi-Xiu Ye and Zhen-Hua Ling. 2019. Distant supervision relation extraction with intra-bag and inter-bag attentions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2810–2819, Minneapolis, Minnesota. Association for Computational Linguistics.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762, Lisbon, Portugal. 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, Ireland. Dublin City University and Association for Computational Linguistics.