Neural Relation Extraction via Inner-Sentence Noise Reduction and Transfer Learning

Tianyi Liu[1], **Xinsong Zhang**[1], **Wanhao Zhou**[1] **and Weijia Jia**[2,1]
[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2]Department of Computer and Information Science, University of Macau
{liutianyi, xszhang0320, whzhou, jiawj}@sjtu.edu.cn

Abstract

Extracting relations is critical for knowledge base completion and construction in which distant supervised methods are widely used to extract relational facts automatically with the existing knowledge bases. However, the automatically constructed datasets comprise amounts of low-quality sentences containing noisy words, which is neglected by current distant supervised methods resulting in unacceptable precisions. To mitigate this problem, we propose a novel word-level distant supervised approach for relation extraction. We first build Sub-Tree Parse (STP) to remove noisy words that are irrelevant to relations. Then we construct a neural network inputting the subtree while applying the entity-wise attention to identify the important semantic features of relational words in each instance. To make our model more robust against noisy words, we initialize our network with a priori knowledge learned from the relevant task of entity classification by transfer learning. We conduct extensive experiments using the corpora of New York Times (NYT) and Freebase. Experiments show that our approach is effective and improves the area of Precision/Recall (PR) from 0.35 to 0.39 over the state-of-the-art work.

1 Introduction

Relation extraction aims to extract relations between pairs of marked entities in raw texts. Traditional supervised methods are time-consuming for the requirement of large-scale manually labeled data. Thus, Mintz et al. (2009) propose the distant supervised relation extraction, in which amounts of sentences are crawled from web pages of New York Times (NYT) and labeled with a known knowledge base automatically. The method assumes that if two entities have a relation in a known knowledge base, all instances that mention these two entities will express the same relation. Obviously, this assumption is too strong, since a sentence that mentions the two entities does not necessarily express the relation contained in a known knowledge base. As described in Riedel et al. (2010), the assumption leads to the wrong labeling problem. In order to tackle the wrong labeling problem, various multi-instance learning methods are adopted by mitigating noise between sentences (Hoffmann et al., 2011; Surdeanu et al., 2012; Zeng et al., 2015; Lin et al., 2016). Despite the wrong labeling problem, distant supervised methods may suffer from the low quality of sentences which derive from the large-scale automatically constructed dataset by crawling web pages (Yang et al., 2017). To handle the problem of low-quality sentences, we have to face two major challenges: (1) Reduce word-level noise within sentences; (2) Improve the robustness of relation extraction against noise.

To explain the influence of word-level noise within sentences, we consider the following sentence as an example: *[It is no accident that the main event will feature the junior welterweight champion miguel cotto, a puerto rican, against Paul Malignaggi, an Italian American from Brooklyn.]*, where *Paul Malignaggi* and *Brooklyn* are two corresponding entities. The subsentence *[Paul Malignaggi, an Italian American from Brooklyn.]* keeps enough words to express the relation */people/person/place_of_birth*, and the other words could be regarded as noise that may hamper the extractor's performance. Meanwhile, as shown in Figure 1, half of the original sentences are longer than 40 words, which means that there are many irrelevant words inside sentences. To be more detail, there are about 12 noisy words in each sentence on average, and 99.4% of sentences in the NYT-10 dataset have noise. Although the Shortest Dependency Path (SDP) proposed by Xu et al. (2015) tries to get rid of irrelevant words for relation extraction, it is not suitable to handle such informal sentences. Moreover, word-level attention has been leveraged to alleviate the impact of noisy

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2195–2204
Brussels, Belgium, October 31 - November 4, 2018.

words (Zhou et al., 2016), but it weakens the importance of entity features for relation extraction.

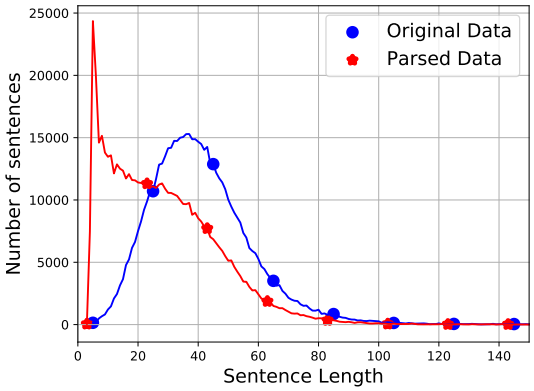

Figure 1: Comparison of sentence length distribution between original data and parsed data.

As for the second challenge, a robust model could extract precise relation features even from low-quality sentences containing noisy words. However, previous neural methods are always lacking in robustness because parameters are initialized randomly and hard to tune with noisy training data, resulting in the poor performance of extractors. Inspired by Kumagai (2016), initializing neural networks with a priori knowledge learned from relevant tasks by transfer learning could improve the robustness of the target task. For the relation extraction, entity type classification can be used as the relevant task since entity types provide abundant background knowledge. For instance, the sentence *[Alfead Kahn, the Cornell-University economist who led the fight to deregulate airplanes.]* has a relation *business/person/company*, which is hard to decide without the information that *Alfead Kahn* is a person and *Cornell-University* is a company. Therefore, type features learned from entity type classification are proper a priori knowledge to initialize the relation extractor.

In this paper, we propose a novel word-level approach for distant supervised relation extraction by reducing inner-sentence noise and improving robustness against noisy words. To reduce inner-sentence noise, we utilize a novel Sub-Tree Parse (STP) method to remove irrelevant words by intercepting a subtree under the parent of entities' lowest common ancestor. As shown in Figure 1, the average length of the parsed sentences is much shorter. Furthermore, the entity-wise attention is adopted to alleviate the influence of noisy words in the subtree and emphasize the task-relevant features. To tackle the second challenge, we initialize our model parameters with a priori knowledge learned from the entity type classification task by transfer learning. The experimental results show that our model can achieve satisfactory performance among the state-of-the-art works. Our contributions are summarized as follows:

- To handle the problem of low-quality sentences, we propose the STP to remove noisy words of sentences and the entity-wise attention mechanism to enhance semantic features of relational words.
- We first propose to initialize the neural relation extractor with a priori knowledge learned from entity type classification, which strengthens its robustness against low-quality corpus.
- Our model achieves significant results for distant supervised relation extraction, which improves the Precision/Recall (PR) curve area from 0.35 to 0.39 and increases top 100 predictions by 6.3% over the state-of-the-art work.

2 Related Work

The distant supervised method plays an increasingly essential role in relation extraction due to its less requirement of human labor (Mintz et al., 2009). However, an evident drawback of the method is the wrong labeling problem. Thus, multi-instance and multi-label learning methods are proposed to address this issue (Riedel et al., 2010; Hoffmann et al., 2011; Surdeanu et al., 2012). Meanwhile, other researches (Angeli et al., 2014; Han and Sun, 2016) incorporate human-designed features and leverage Natural Language Processing (NLP) tools.

As neural networks have been widely used, an increasing number of researches have been proposed. Zeng et al. (2015) use a piecewise convolutional neural network with multi-instance learning. Furthermore, selective attention over instances with the neural network is proposed (Lin et al., 2016). Making use of entity description, Ji et al. (2017) assign more precise attention weights. Focused on the imbalance of datasets, a soft label method has been proposed by Liu et al. (2017).

Recently, reinforcement learning and adversarial learning are widely used to select the valid instances for relation extraction (Feng et al., 2018; Qin et al., 2018b,a).

However, above methods ignore inner-sentence noise. To better remove irrelevant words, the SDP between entities is proved to be effective (De Marneffe and Manning, 2008; Chen and Manning, 2014; Xu et al., 2015; Miwa and Bansal, 2016). Nevertheless, in our observation, the SDP deals with informal texts difficultly (See Section 3.1 for details). Furthermore, word-level attention is adopted to focus on relational words for relation extraction (Zhou et al., 2016), but it hinders the effect of entity words.

Transfer learning proposed by Pratt (1993) provides a new approach to leverage knoweldge extracted by related tasks to enhance the performance of the target task. Furthermore, parameter transfer learning is proved to be effective to improve the stability of models by initializing model parameters reasonably (Pan and Yang, 2010; Kumagai, 2016).

3 Methodology

In this section, we present our methodology for distant supervised relation extraction. Figure 2 shows the overall architecture of our model. Our model is divided into three parts:

Sub-Tree Parser. Input instances are parsed to dependency parse trees by the Stanford parser[1] (Chen and Manning, 2014) at first. Then words in the STP and relative positions are transformed to distributed representations.

Entity-Wise Neural Extractor. Given the representation of each subtree, Bidirectional Gated Recurrent Unit (BGRU) extracts specific features. Then, entity-wise attention combined with word-level attention is applied to reducing irrelevant features for relation extraction. Finally, the sentence-level attention is used to alleviate the influence of wrong labeling sentences.

Parameter-Transfer Initializer. The transfer learning method pre-trains our model parameters from the task of entity type classification aiming at boosting the performance of relation extraction.

3.1 Sub-Tree Parser

Each instance is put into the dependency parse module to build the dependency parse tree in the first place. Then we can tailor the sentences based on the STP method. Finally, we transform word tokens and position tokens of each instance to distributed representations by embedding matrixes.

Sub-Tree Parse

In order to reduce inner-sentence noise and extract relational words, we propose the STP method which intercepts the subtree of each instance under the parent of entities' lowest common ancestor. For instance, in Figure 2(b), *China* and *Shanghai* are entities connected directly with the appositive relation. The instance *[In 1990, he lives in Shanghai, China.]* will be transformed to *[in Shanghai, China.]* on the basis of the STP, where *in* is the parent of *Shanghai* and *China* lowest common ancestor and kept as important information for expressing the relation *location/location/contain*. Words connected by the imaginary line indicating the extracted subtree are reorganized into their original sequence order to form network inputs.

Among the parse tree, the SDP has been widely used by Chen and Manning (2014) and Xu et al. (2015) to help models focus on relational words. However, in our observation, the SDP is not appropriate in the condition that key relation words are not in the SDP. Although additional information (dependency relations between words) is adopted to enhance the performance of SDP, we found they have the minor effect through our experiment. Thus, we do not make use of other types of linguistic information. As Figure 2(b) shows, in the SDP method, the original sentence will be transformed to *[Shanghai China]* because *Shanghai* and *China* are connected with each other directly in the dependency parse tree, which results in deleting the keyword *in* and may confuse the model when extracting relations. Compared with SDP, the STP method is more appropriate to extract useful information in informal sentences where relational words are always not in the SDP.

Word and Position Embeddings

The inputs of the network are word and position tokens, which are transformed to the distributed representations before they are fed into the neural model. We map j^{th} word in the i^{th} instance to a vector of k dimensions denoted as $x_{ij}^{w} \in R^{k}$ through the Skip-Gram model (Mikolov et al., 2013). Like Zeng et al. (2014), we leverage Pos1 and Pos2 to specify entity pairs, which are defined as the relative distances of current word from head

[1] https://nlp.stanford.edu/software/lex-parser.shtml

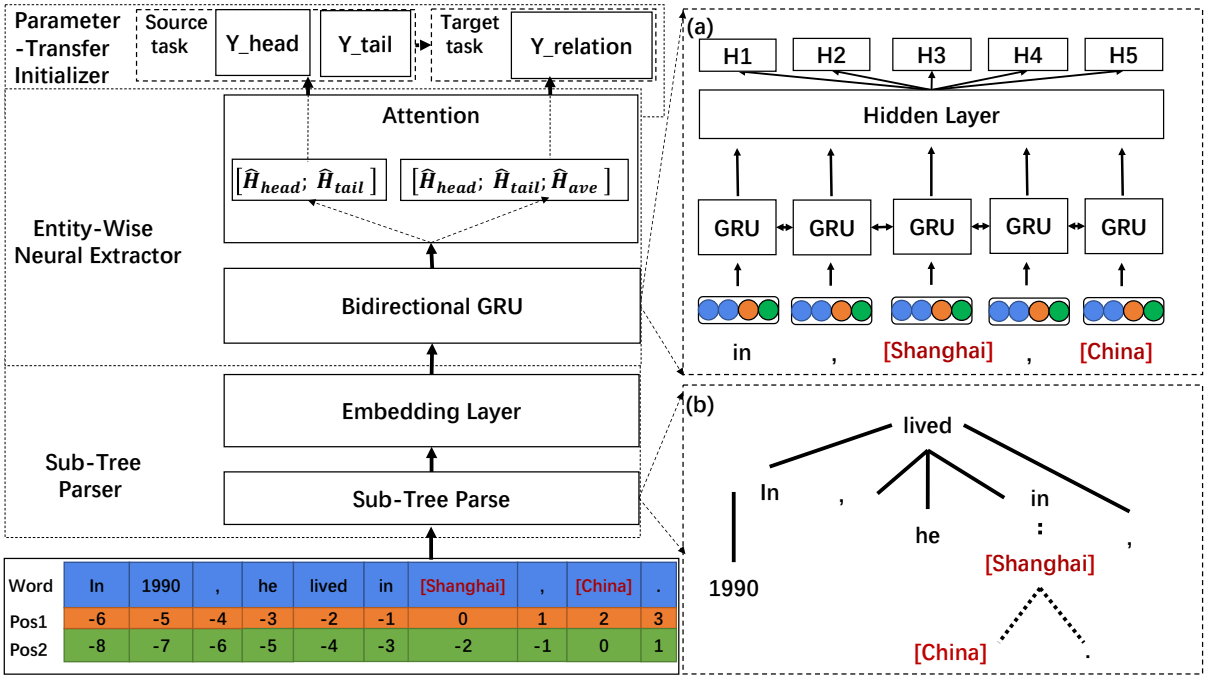

Figure 2: Overall architecture of our model is used for distant supervised relation extraction, expressing the process of handling instances. There are two modules described in detail: (a) One is the BGRU; (b) Another is the STP, where words in the red brackets represent entities (better viewed in color).

entity and tail entity. For instance, in Figure 2 relative distances of *lived* from *Shanghai* and *China* are -2 and -4 respectively. Then, the position token of each word is transformed to a vector in l dimensions. Position embeddings are denoted as $x_{ij}^{p1} \in R^l$ and $x_{ij}^{p2} \in R^l$ respectively. Finally, the input representation for x_{ij} is concatenated by word embedding x_{ij}^{w}, position embeddings x_{ij}^{p1} and x_{ij}^{p2} , which is denoted as $x_{ij} = [x_{ij}^{w}; x_{ij}^{p1}; x_{ij}^{p2}]$ where $x_{ij} \in R^{k+2l}$.

3.2 Entity-Wise Neural Extractor

As shown in Figure 2, we transform the STP into feature vectors by BGRU at first. Next, entity-wise attention combined with the hierarchical-level attention mechanism is applied to enhancing semantic features of each instance.

BGRU over STP

Since the transfer learning and entity-wise attention require the specific features of entities in tree parsed instances as their input, we adopt Gated Recurrent Unit (GRU) (Cho et al., 2014) to be our based relation extractor, which can extract global information of each word by pointing out its corresponding position in the sequence. It can be briefly described as below:

$$h_{it} = GRU(x_{it}) \tag{1}$$

where x_{it} is the t^{th} word representation in the i^{th} parsed instance as described in the input layer, and $h_{it} \in R^m$ is the hidden state of GRU in m dimensions.

Furthermore, BGRU implementing GRU in a different direction can access future as well as past context. Under our circumstance, BGRU combined with the STP can extract semantic and syntactic features adequately. Figure 2(a) shows the processing of BGRU over STP. The following equation defines the operation mathematically.

$$h_{it} = [\overrightarrow{h_{it}} \oplus \overleftarrow{h_{it}}] \tag{2}$$

In above equation, the t^{th} word output $h_{it} \in R^m$ of BGRU is the element-wise addition of the t^{th} hidden states of forward GRU and backward one.

Entity-wise Attention

To reduce noise within sentences, we propose the entity-wise attention mechanism to help our model focus on relational words, especially entity words for relation extraction. Assume that H_i is the i^{th} instance matrix consisting of T word vectors $[h_{i1}, h_{i2}, \cdots, h_{iT}]$ produced by BGRU.

Not all words contribute equally to the representation of the sentence. Entity words are of great importance because they are significantly beneficial to relation extraction. In our model, entity-wise attention assigns the weight α_{it}^{e} to focus on the target entity and removes noise further. It is defined as follows:

$$\alpha_{it}^{e} = \begin{cases} 1 & t = head, tail \\ 0 & others \end{cases} \tag{3}$$

In the above equation, $\alpha_{it}^{e} = 1$ if t^{th} word belongs to the head or tail entity.

Hierarchical-level Attention

To reduce inner-sentence noise further and de-emphasize noisy sentences, we incorporate word-level attention and sentence-level attention as hierarchical-level attention which is introduced in Yang et al. (2016).

Word-level Attention. It assigns an additional weight α_{it}^{w} to relational word h_{it} due to its relevance to the relation as described by Zhou et al. (2016). It can be described as follows:

$$\alpha_{it}^{w} = \frac{exp(h_{it} A^{w} r^{w})}{\sum_{t=1}^{T} exp(h_{it} A^{w} r^{w})} \tag{4}$$

where A^{w} is a weighted matrix, and vector r^{w} can be seen as a high level representation in a fixed query what is the informative word over the other words.

The i^{th} sentence representation $S_i \in R^m$ is computed as a weighted sum of h_{it}:

$$S_i = \sum_{t=1}^{T} (\alpha_{it}^{w} + \alpha_{it}^{e}) h_{it} \tag{5}$$

Sentence-level Attention. After we get the instance representation S_i, we adopt the selective attention mechanism over instances to de-emphasize the noisy sentence (Lin et al., 2016), which is described as follows:

$$S = \sum_{i} \alpha_i^{s} S_i \tag{6}$$

$$\alpha_i^{s} = \frac{exp(S_i A^{s} r^{s})}{\sum_{i} exp(S_i A^{s} r^{s})} \tag{7}$$

where A^{s} is a weighted matrix, r^{s} is the query vector associated with the relation, and $S \in R^m$ is the output of the sentence-level attention layer.

3.3 Parameter-Transfer Initializer

The transfer learning method pre-trains our model parameters in the entity type classification task, which in turn contributes to the relation extraction.

Pre-learn the Entity Type

As entity type information plays a significant role in detecting relation types, the entity type classification task is considered to be the source task, which is learned before the relation extraction task. According to Eq. 6, outputs of the sentence-level attention layer for the head entity and tail entity task are S_{head} and S_{tail} respectively. They are ultimately fed into the softmax layer:

$$\hat{p}^{i} = softmax(W_i S_i + b_i);\ i \in \{head, tail\} \tag{8}$$

where W_i and b_i are the weight and bias for the entity type classification task respectively, $\hat{p}^{i} \in R^{z_t}$ is the predicted probability of each class and z_t is the number of entity classes. The loss function of the source task is the negative log-likelihood of the true labels:

$$\begin{aligned} &J_e(\theta_0, \theta_{head}, \theta_{tail}) = \beta \|\theta_0\|^2 \\ &+ \sum_{t} (-\frac{1}{z_t} \lambda_t \sum_{i=1}^{z_t} y_i^{t} log(\hat{p}_i^{t}) + \beta \|\theta_t\|^2) \\ &t \in \{head, tail\} \end{aligned} \tag{9}$$

where λ_t is the weight of each task, θ_0 is the shared model parameters, θ_{head} and θ_{tail} are individual parameters for the head and tail entity classification tasks respectively, $y^{t} \in R^{z_t}$ is the one-hot vector representing ground truth, and β is the hyper-parameter for $L2$ regularization.

Train the Relation Extractor

Based on the pre-trained model in the entity type classification task, the relation extractor initializes shared parameters θ_0 within the best state of the pre-trained model and independent parameters θ_r randomly. Same as the entity type classification task, the output S_r of the attention layer for the relation extraction task is finally fed into the softmax layer and the loss is calculated by cross entropy, which is defined as follows:

$$\hat{p} = softmax(W_r S_r + b_r) \tag{10}$$

$$\begin{aligned} J_r(\theta_0, \theta_r) = &-\frac{1}{z_r} \sum_{i=1}^{z_r} y_i log(\hat{p}_i) \\ &+ \beta(\|\theta_0\|^2 + \|\theta_r\|^2) \end{aligned} \tag{11}$$

where W_r, b_r, $y \in R^{z_r}$, $\hat{p} \in R^{z_r}$, θ_r and β are defined similarly in the entity type classification task.

As shown in Figure 2, two tasks share all layers except attention and output layers. Assume that the set of total model parameters is θ. Thus, θ, θ_0, θ_r, θ_{head} and θ_{tail} have a relationship described in the following equations:

$$\theta = \theta_0 \cup \theta_{head} \cup \theta_{tail} \cup \theta_r \tag{12}$$

$$\theta_i = \{A_i^w, r_i^w, A_i^s, r_i^s, W_i, b_i\} \quad i \in \{head, tail, r\} \tag{13}$$

where A_i^w, r_i^w, A_i^s, r_i^s, W_i and b_i are parameters in attention and output layers.

Optimize the Objective Function

At first, we minimize J to obtain θ_0 at the best model state $\hat{\theta}_0$ for entity type classification. Then we minimize J_r for the best performance of relation extraction under the initialization of θ_0 to be $\hat{\theta}_0$. Above process can be summarized as the following equation:

$$min \quad J(\theta) = \lambda J_e(\theta_0, \theta_{head}, \theta_{tail}) + (1-\lambda) J_r(\theta_0, \theta_r) \tag{14}$$

where $\lambda \in (0, 1)$ is the hyperparameter to determine the importance of each task at different training steps. We use Adam (Kingma and Ba, 2014) optimizer to minimize the objective $J(\theta)$.

4 Experiments

Our experiments are designed to demonstrate that our model alleviates the influence of word-level noise arising from low-quality sentences. In this section, we first introduce the dataset and evaluation metrics. Next, we describe parameter settings. Then we evaluate effects of the STP, entity-wise attention and the parameter-transfer initializer. Finally, we compare our model with the state-of-the-art works by several evaluation metrics.

4.1 Dataset and Evaluation Metrics

To evaluate the performance of our model, we adopt a widely used dataset NYT-10 developed by Riedel et al. (2010). NYT-10 dataset is constructed by aligning relational facts in Freebase (Bollacker et al., 2008) with the NYT corpus, where sentences from 2005-2006 are used as training set, and sentences from 2007 are used for testing. For training data, there are 522,611 sentences, 281,270 entity pairs, and 18,252 relational facts; for testing data, there are 172,448 sentences, 96,678 entity pairs and 1,950 relational facts. There are 53 relations including a special relation NA, which means that there is no relation between the entity pair in the instance. Meanwhile, all relations in Freebase are defined on head types and tail types. Therefore, we can construct datasets for type prediction tasks with the same dataset. The dataset has 29 head types and 26 tail types.

Like previous works, we evaluate our model with the held-out metrics, which compare relations found by models with those in Freebase. The held-out evaluation provides a convenient way to assess models. We report both the PR curve and Precision at top N predictions (P@N) at various numbers of instances under each entity pair:

One: For each entity pair, we randomly select one instance to represent the relation.

Two: For each entity pair, we randomly select two instances to represent the relation.

All: For each entity pair, we select all instances to represent the relation.

4.2 Experimental Settings

In the experiment, we utilize word2vec[2] to train word embeddings on NYT corpus. We use the cross-validation to tune our model and grid search to determine model parameters. The grid search approach is used to select optimal learning rate lr for Adam optimizer among $\{0.1, 0.001, 0.0005, 0.0001\}$, GRU size $m \in \{100, 160, 230, 400\}$, position embedding size $l \in \{5, 10, 15, 20\}$. Table 1 shows all parameters for our task. We follow experienced settings for other parameters because they make little influence to our model performance.

GRU size m	230
Word embedding dimension k	50
POS embedding dimension l	5
Batch size n	50
Entity-Task weights(λ_{head}, λ_{tail})	0.5,0.5
Entity-Relation Task weight λ	0.3
Learning rate lr	0.001
Dropout probability p	0.5
l_2 penalty β	0.0001

Table 1: Parameter Settings

[2]https://code.google.com/p/word2vec/

4.3 Effect of Various Model Parts

In this section, we utilize the PR curve to evaluate the effects of three main parts in our model: the STP, entity-wise attention and the parameter-transfer initializer.

Effect of the STP

To demonstrate the effect of the STP, we adopt BGRU with Word-Level Attention (WLA) proposed by Zhou et al. (2016) as our base model. We compare the performance of BGRU, BGRU+STP, and BGRU+SDP.

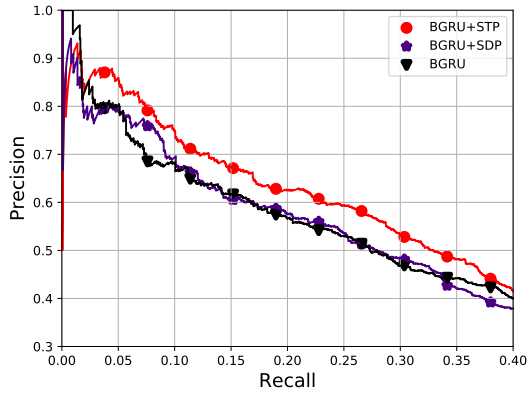

Figure 3: PR curves for BGRU, BGRU+SDP and BGRU+STP.

From Figure 3, we can observe that the model with the STP performs best, and the SDP model obtains an even worse result than the pure one. The PR curve areas of BGRU+SDP and BGRU are about 0.332 and 0.337 respectively, while BGRU+STP increases it to 0.366. The result indicates: (1) Our STP can get rid of irrelevant words in each instance and obtain more precise sentence representation for relation extraction. It proves that our STP module is effective. (2) The SDP method is not appropriate to handle low-quality sentences where key relation words are not in the SDP.

Effect of Entity-wise Attention

Test Settings	PR Curve Area	
Dataset	Original Data	STP Data
BGRU	0.337	0.366
-WLA+EWA	0.365	0.375
+EWA	**0.372**	**0.383**

Table 2: PR curve areas for BGRU, BGRU-WLA+EWA and BGRU+EWA on various datasets.

To evaluate the effect of entity-wise attention combined with word-level attention, we utilize BGRU in three settings on our tree parsed data and original data. One setting is to use WLA mechanism only (BGRU). The second one is to replace WLA with the Entity-Wise Attention (EWA) mechanism (BGRU-WLA+EWA). The third one is to incorporate two mechanisms (BGRU+EWA).

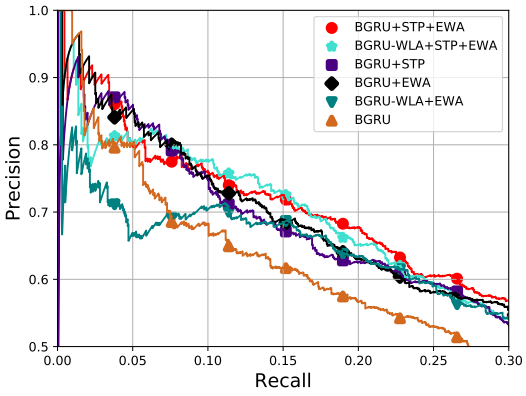

Figure 4: PR curves for BGRU, BGRU-WLA+EWA and BGRU+EWA on various datasets.

From Table 2 and Figure 4, we can obtain: (1) Regardless of the dataset that we employ, BGRU-WLA(+STP)+EWA outperforms BGRU(+STP). To be more specific, the PR curve area has a relative improvement of over 2.3%, which demonstrates that entity-wise hidden states in the BGRU present more precise relational features than other word states. (2) BGRU(+STP)+EWA achieves further improvements and outperforms the baseline by over 4.6%, because it considers more information than entity or relational words alone. Thus, it indicates that entity words are essential for relation extraction, but they can not represent features of the whole sentence without other words.

Effect of Parameter-Transfer Initializer

To evaluate the effect of the parameter-transfer initializer in our model, we leverage BGRU under four circumstances. The first one is to directly apply it on the original dataset. The second one tests BGRU combined with Transfer Learning (TL) on the original dataset. The third one uses BGRU on our STP dataset. The fourth one examines BGRU+TL on our STP dataset.

From Figure 5, we can conclude: (1) Regardless of the dataset that we use, models with TL achieve better performance, which improve the PR curve area by over 4.7%. It demonstrates that

Test Settings	One				Two				All			
P@N	100	200	300	Mean	100	200	300	Mean	100	200	300	Mean
Mintz	35.0	37.5	37.3	36.6	51.0	42.0	43.3	45.4	54.0	50.5	45.3	49.9
MultiR	64.0	61.5	53.7	59.7	62.0	61.5	58.7	61.1	75.0	65.0	62.0	67.3
MIML	62.0	59.0	54.7	58.6	69.0	59.5	59.0	62.5	70.0	64.5	60.3	64.9
PCNN	73.3	64.8	56.8	65.0	70.3	67.2	63.1	66.9	72.3	69.7	64.1	68.7
PCNN+ATT	78.0	68.0	60.7	68.9	75.0	74.0	66.3	71.8	82.0	74.0	69.0	75.0
BGRU	72.0	62.5	59.0	64.5	70.0	64.0	64.7	66.2	74.0	68.0	65.0	69.0
+STP	73.0	63.0	60.7	65.6	83.0	72.5	68.0	74.5	86.0	76.0	70.3	77.4
+EWA	82.0	71.5	66.3	73.3	84.0	79.5	70.3	77.9	86.0	81.5	75.3	80.9
+TL	**83.0**	**75.5**	**67.0**	**75.2**	**85.0**	**81.0**	**72.3**	**79.4**	**87.0**	**83.0**	**78.0**	**82.7**

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

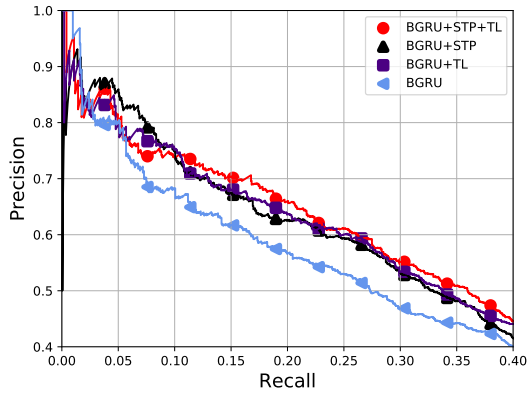

Figure 5: PR curves for BGRU, BGRU+TL, BGRU+STP and BGRU+STP+TL

transfer learning helps our model become more robust against noise. (2) BGRU+STP+TL achieves the best performance and increases the area to 0.383, while areas of BGRU, BGRU+STP and BGRU+TL are 0.337, 0.366 and 0.372 respectively. It means that the TL method works well with the STP and can resist noisy words further.

4.4 Comparison with Baselines

To evaluate our approach, we select the following six methods as our baseline:

Mintz (Mintz et al., 2009) proposes the human-designed feature model.

MultiR (Hoffmann et al., 2011) puts forward a graphical model.

MIML (Surdeanu et al., 2012) proposes a multi-instance multi-label model.

PCNN (Zeng et al., 2015) puts forward a piecewise CNN for relation extraction.

PCNN+ATT (Lin et al., 2016) proposes the selective attention mechanism with PCNN.

BGRU (Zhou et al., 2016) proposes a BGRU with the word-level attention mechanism.

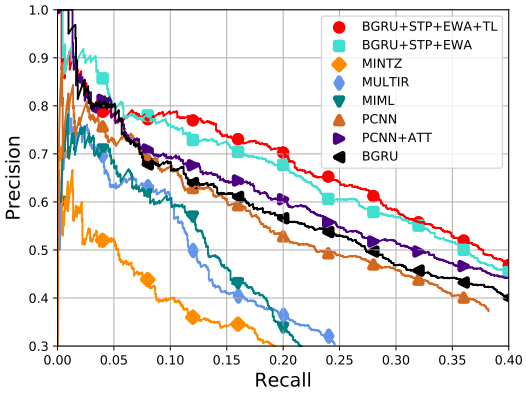

Figure 6: Performance comparison of the proposed method with baselines.

As Figure 6 shows, we can observe: (1) BGRU+STP+EWA achieves the best PR curve over baselines, which improves the area to 0.38 over 0.33 of PCNN, 0.34 of BGRU and 0.35 of PCNN+ATT. At the recall rate of 0.25, our model can still achieve a precision rate above 0.6. It demonstrates that BGRU+STP+EWA is effective because the STP and entity-wise attention combined with word-level attention can reduce inner-sentence noise at a fine-grained level. (2) Integrated with transfer learning, BGRU+STP+EWA+TL performs much better and increases the PR curve area to 0.392. It means that the model is pre-trained for better parameter initialization so the TL model becomes more robust against noisy words. Parameter transfer learning can be applied in better feature extractors for further improvement.

Following previous works, we adopt P@N as

a quantitative indicator to compare our model with baselines based on various instances under each relational tuple. In Table 3, we report P@100, P@200, P@300 and the mean of them for each model in the held-out evaluation. We can find: (1) Compared with baselines, BGRU+STP+EWA+TL achieves the best performance in all test settings, which increases the performance of PCNN+ATT in three settings by 6.3%, 7.6%, and 7.7% respectively. It demonstrates that the integrated model is the most effective; (2) Our STP and entity-wise attention combined with word-level attention reduce inner-sentence noise effectively, and outperform baselines by over 5%; (3) Our neural extractor initialized with a priori knowledge learned from entity type classification is more robust against word-level noise where BGRU+STP+EWA+TL has an improvement of 2% over BGRU+STP+EWA.

5 Conclusion

In this paper, we propose a novel word-level approach for distant supervised relation extraction. It aims at tackling the low-quality corpus by reducing inner-sentence noise and improving the robustness against noisy words. To alleviate the influence of word-level noise, we propose the STP. Meanwhile, entity-wise attention combined with word-level attention helps the model focus more on relational words. Furthermore, parameter transfer learning makes our model more robust against noise by reasonable initialization of parameters. The experimental results show that our model significantly and consistently outperforms the state-of-the-art method.

In the future, we will incorporate the SDP and STP to obtain more precise shortened sentences. Furthermore, we will conduct research in how to utilize entity information to assign more appropriate initial parameters of the relation extractor.

Acknowledgments

This work is supported by FDCT 0007/2018/A1, DCT-MoST Joint-project No. (025/2015/AMJ) of SAR Macau; University of Macau Funds Nos: CPG2018-00032-FST & SRG2018-00111-FST; Chinese National Research Fund (NSFC) Key Project No. 61532013 and National China 973 Project No. 2015CB352401.

References

Gabor Angeli, Julie Tibshirani, Jean Wu, and Christopher D Manning. 2014. Combining distant and partial supervision for relation extraction. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1556–1567. Association for Computational Linguistics.

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*, pages 1247–1250. ACM.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750. Association for Computational Linguistics.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734. Association for Computational Linguistics.

Marie-Catherine De Marneffe and Christopher D Manning. 2008. The stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8. Association for Computational Linguistics.

Jun Feng, Minlie Huang, Li Zhao, Yang Yang, and Xiaoyan Zhu. 2018. Reinforcement learning for relation classification from noisy data. In *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 5779–5786.

Xianpei Han and Le Sun. 2016. Global distant supervision for relation extraction. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, pages 2950–2956.

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 - Volume 1*, pages 541–550. Association for Computational Linguistics.

Guoliang Ji, Kang Liu, Shizhu He, Jun Zhao, et al. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, pages 3060–3066.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Wataru Kumagai. 2016. Learning bound for parameter transfer learning. In *Advances in Neural Information Processing Systems*, pages 2721–2729.

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)*, volume 1, 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 (EMNLP)*, pages 1790–1795. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

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.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using lstms on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1105–1116. Association for Computational Linguistics.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Lorien Y Pratt. 1993. Discriminability-based transfer between neural networks. In *Advances in Neural Information Processing Systems*, pages 204–211.

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.

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 (EMNLP-CoNLL)*, pages 455–465. Association for Computational Linguistics.

Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. 2015. Classifying relations via long short term memory networks along shortest dependency paths. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1785–1794. Association for Computational Linguistics.

Wenmian Yang, Na Ruan, Wenyuan Gao, Kun Wang, Wensheng Ran, and Weijia Jia. 2017. Crowdsourced time-sync video tagging using semantic association graph. In *Multimedia and Expo (ICME), 2017 IEEE International Conference on*, pages 547–552. IEEE.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.

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 (EMNLP)*, 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.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 207–212. Association for Computational Linguistics.