Improving Distantly-Supervised Relation Extraction with Joint Label Embedding

Linmei Hu[1], **Luhao Zhang**[1], **Chuan Shi**[1] [*], **Liqiang Nie**[2], **Weili Guan**[3], **Cheng Yang**[1]
[1]Beijing University of Posts and Telecommunications, China
[2]Shan Dong University, China
[3]Hewlett Packard Enterprise Singapore, Singapore
{hulinmei,zhangluhao,shichuan}@bupt.edu.cn
{nieliqiang, honeyguan, albertyang33}@gmail.com

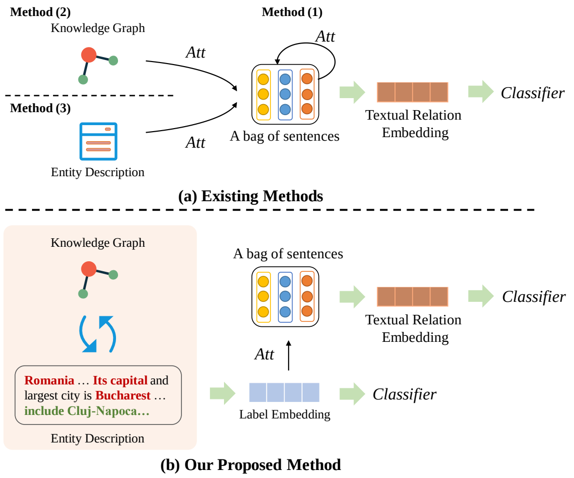
Abstract

Distantly-supervised relation extraction has proven to be effective to find relational facts from texts. However, the existing approaches treat labels as independent and meaningless one-hot vectors, which cause a loss of potential label information for selecting valid instances. In this paper, we propose a novel multi-layer attention-based model to improve relation extraction with joint label embedding. The model makes full use of both structural information from Knowledge Graphs and textual information from entity descriptions to learn label embeddings through gating integration, while avoiding the imposed noise with an *attention mechanism*. Then the learned label embeddings are used as *another attention* over the instances (whose embeddings are also enhanced with the entity descriptions) for improving relation extraction. Extensive experiments demonstrate that our model significantly outperforms state-of-the-art methods.

1 Introduction

Knowledge Graphs (KGs) such as Freebase and DBpedia have shown their strong power in many natural language processing tasks including question answering (Zhang et al., 2018) and dialog generation (Zhou et al., 2018). However, these KGs are far from complete. Relation extraction, which aims to fill this gap by extracting semantic relationships between entity pairs from plain texts, is thus of great importance.

Most existing supervised relation extraction methods require a large number of labeled training data, which is time-consuming and laborious. Distant supervision has been proposed by (Mintz et al., 2009) to address the challenge. It assumes that if two entities have a relation in KGs, then all sentences mentioning the two entities express this relation. Thus, distant supervision can automatically generate a large number of labeled data without labor cost. Simultaneously, it often suffers from wrong labeling problem (Surdeanu et al., 2012; Zeng et al., 2015).



(a) Existing Methods

(b) Our Proposed Method

Figure 1: Illustration of comparison of existing methods and our proposed method.

Recently, significant progress has been made in applying deep neural networks for relation extraction under distant supervision (Zeng et al., 2014, 2015; Feng et al., 2017). To alleviate the wrong labeling problem in distant supervision, attention models have been proposed to select valid instances (Ji et al., 2017). As shown in Figure 1 (a), they can be divided into three categories: (1) typical attention models without external information (Lin et al., 2016; Du et al., 2018), (2) attention models using KGs (Han et al., 2018a), (3) attention models using side information such as entity descriptions (Vashishth et al., 2018; Ji et al., 2017). However, they all have flaws in selecting valid instances due to failing to exploit potential label information. They treat labels as independent and meaningless one-hot vectors, which cause a loss

[*]Corresponding author: Chuan Shi.

of potential label information. Label embeddings aim to learn representations of labels based on related information. The label embeddings can be used to attend over the bag of instances for relation classification. Additionally, they don't take advantage of both structural information from KGs and textual information from entity descriptions and ignore the imposed noise.

In this paper, we propose a novel multi-layer attention-based model RELE (**R**elation **E**xtraction with Joint **L**abel **E**mbedding) to improve relation extraction. Our model integrates both structural information from KGs and textual information from entity descriptions with a gating mechanism to learn label embeddings, while avoiding the imposed noise (highlighted in green) in entity descriptions with an *attention mechanism*. Then the label embeddings are used as *another antention* over the bag of instances to select valid ones for improving relation extraction. Note that we also enhance the instance embedding with entity descriptions. The contributions of this paper can be summarized as follows:

- We propose a novel multi-layer attention-based model RELE to improve distantly supervised relation extraction with joint label embedding. The label embeddings can be used to attend over the bag of instances for relation classification.
- RELE makes full use of both structural information from KGs and textual information of entity descriptions to learn label embeddings through gating integration, while avoiding the imposed noise with attention.
- Extensive experiments on two benchmark datasets have demonstrated that our model significantly outperforms state-of-the-art methods on distantly-supervised relation extraction.

2 Related Work

Our work is mainly related to distant supervision, neural relation extraction and label embedding.

Distant Supervision. Most supervised relation extraction methods require large-scale labeled training data which are expensive. Distant Supervision proposed by (Mintz et al., 2009) is an effective method to automatically label large-scale training data under the assumption that if two entities have a relation in a KG, then all sentences mentioning those entities express this relation. The assumption does not work in all cases and causes the mislabeling problem.

MultiR (Hoffmann et al., 2011) and MIMLRE (Surdeanu et al., 2012) introduce multi-instance learning where the instances mentioning the same entity pair are processed at a bag level. However, these methods rely heavily on handcrafted features.

Neural Relation Extraction. With the development of deep learning, neural networks have proven to be efficient to automatically extract valid features from sentences in recent years. Some researches (Zeng et al., 2014, 2015) adopt Convolution Neural Networks (CNN) to learn sentence representations automatically. To alleviate the mislabeling problem, attention mechanisms (Lin et al., 2016; Du et al., 2018) have been employed. Apart from that, some studies apply other relevant information to improve relation extraction (Zeng et al., 2017; Vashishth et al., 2018; Han et al., 2018b,a; Ji et al., 2017). For example, RESIDE (Vashishth et al., 2018) utilizes the available side information from knowledge bases, including entity types and relation alias information. Han et al. (2018a) proposed a joint representation learning framework of KGs and instances, which leverages the KG embeddings to select valid instances. APCNN+D (Ji et al., 2017) exploits the entity descriptions as background knowledge for selection of valid instances, and ignores the imposed noise.

Different from the existing works, we propose a novel multi-layer attention-based model RELE with joint label embedding. Our model makes full use of both structural information from KGs and textual information from entity descriptions to learn label embeddings through gating integration, while avoiding the imposed noise with an *attention mechanism*. The label embeddings are then used as *another attention* over the bag of instances to select valid instances for relation classification.

Label Embedding. Label embedding has been widely exploited in computer vision including image classification (Akata et al., 2016) and text recognition (Rodriguez-Serrano et al., 2015). Recently, LEAM (Wang et al., 2018) successfully applies label embedding in text classification, which obtains each label's embedding by its corresponding text descriptions. In this work, we are the first to apply it for relation extraction. We propose a novel multi-layer attention-based model to improve relation extraction with joint label embedding.

3 Preliminaries

In this section, we briefly introduce some notations and concepts used in this paper.

For convenience, we denote a KG as $G = \{(h, r, t)\}$, which contains considerable triplets (h, r, t) where h and t are respectively head entity and tail entity, and r denotes the relation. Their embeddings are denoted as $(\mathbf{h}, \mathbf{r}, \mathbf{t})$.

Formally, given a pair of entities (h, t) in a KG G and a bag of instances (sentences) $B = \{s_1, s_2, \cdots, s_m\}$, where each instance s_i contains (h, t), the task of *relation extraction* is to train a classifier based on B to predict the relation label y of (h, t) from a predefined relation set. If no relation exists, we simply assign NA to it.

To improve relation extraction, we make full use of both the KG G and entity descriptions $D = \{d_1, d_2, \cdots, d_n\}$ to learn label embeddings which can benefit selection of valid instances. For each entity e_i, we take the first paragraph of its corresponding Wikipedia page as its description text $d_i = \{w_1, w_2, \cdots, w_l\}$, where $w_i \in V$ denotes the description word, l is the length and V is the vocabulary.

4 Our Proposed Model

In this section, we will detail our proposed multi-layer attention-based model RELE for relation extraction with joint label embedding. Existing methods for relation extraction take labels as independent and meaningless one-hot vectors, which cause a loss of potential label information for selecting valid instances. Additionally, they don't take full advantage of both structural information from KGs and textual information from entity descriptions and ignore the imposed noisy information.

As shown in Figure 2, our model is based on a multi-layer attention, containing two parts: 1) label embedding (shown in the right) and 2) neural relation classification (shown in the left) . The former makes full use of structural information from KGs and textual information from entity descriptions to learn label embeddings through gating mechanism, while avoiding the imposed noise with an attention mechanism. The latter leverages the label embeddings as another attention over the instances to select valid instances for improving neural relation extraction. Note that we also use the entity descriptions to enhance the representations of instances. We detail the two parts as follows.

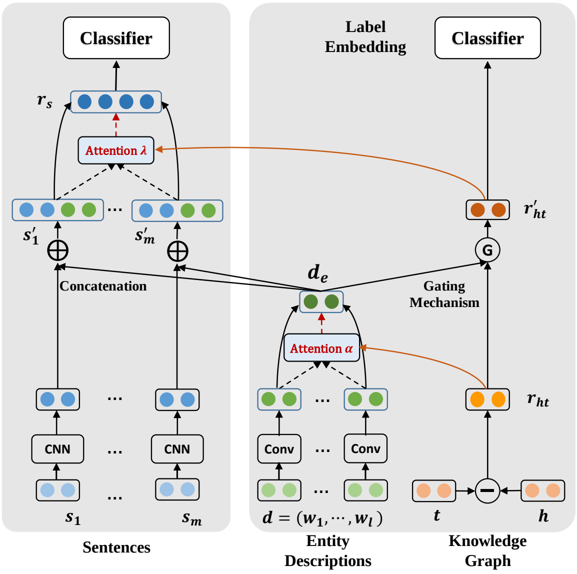


Figure 2: Illustration of our multi-layer attention based model RELE.

4.1 Joint Label Embedding

Label information plays a vital role in selecting valid instances for improving relation extraction. We make full use of both structural information from KGs (Han et al., 2018a) and textual information from entity descriptions (Ji et al., 2017) to learn label embeddings via gating integration. Entity descriptions provide rich background knowledge for entities (Newman-Griffis et al., 2018) and are supposed to benefit the label embedding and relation extraction. Nevertheless, as shown in Figure 1, entity descriptions may also contain irrelevant and even misleading information. Therefore, we propose to use KG embeddings to attend over the entity descriptions, alleviating the imposed noise. Then a gating mechanism is used to integrate both the KGs and entity descriptions for learning label embeddings.

KG Embedding. We use TransE (Bordes et al., 2013) for KG embedding. Given a triplet (h, r, t), the model aims to learn low-dimensional representations vector for entities h, t and the relationship r into the same vector space, and regards a relationship r as a translation from the head entity h to tail entity t, assuming the embedding $\mathbf{t}$ should be close to $\mathbf{h}+\mathbf{r}$ if (h, r, t) exists. The score function is defined as :

$$f(h, r, t) = -\|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_2^2. \tag{1}$$

Note that since the true relations in test set are unknown, we simply represent the relation by:

$$\mathbf{r}_{ht} = \mathbf{t} - \mathbf{h}. \tag{2}$$

In this way, we can also get the relation embeddings given the entity pairs during testing.

Entity Description Embedding. Then we use the representations of relations $\mathbf{r}_{ht}$ as attention over the words of an entity description to reduce the weights of noisy words. Formally, for each entity e, we learn the representation of its description $d = (w_1, w_2, \cdots, w_l)$ as follows:

$$\mathbf{x}_i = \text{CNN}(\mathbf{w}_{i-\frac{c-1}{2}}, \cdots, \mathbf{w}_{i+\frac{c-1}{2}}), \tag{3}$$

$$\hat{\mathbf{x}}_i = \tanh(\mathbf{W}_x \cdot \mathbf{x}_i + \mathbf{b}_x), \tag{4}$$

$$\alpha_i = \frac{\exp(\hat{\mathbf{x}}_i \cdot \mathbf{r}_{ht})}{\sum_{i=1} \exp(\hat{\mathbf{x}}_i \cdot \mathbf{r}_{ht})}, \tag{5}$$

$$\mathbf{d}_e = \sum_{i=1}^{l} \alpha_i \mathbf{x}_i, \tag{6}$$

where $\text{CNN}(\cdot)$ denotes a convolution layer with window size c over the word sequence. $\mathbf{x}_i \in \mathbb{R}^{D_h}$ is the hidden representation of the word w_i. $\mathbf{W}_x$ is the weight matrix and $\mathbf{b}_x$ is the bias vector. α_i is the attention weight of the word w_i, which is computed based on the relation embedding $\mathbf{r}_{ht}$. Finally, the text description embedding $\mathbf{d}_e$ is computed by the weighted average of words.

Gating Integration. We apply a gating mechanism (Xu et al., 2016) to integrate the textual entity description embedding $\mathbf{d}_e$ and the structural information (entity embedding $\mathbf{e}$) from KGs:

$$\mathbf{e}' = \mathbf{g} \odot \mathbf{e} + (1 - \mathbf{g}) \odot \mathbf{d}_e, \tag{7}$$

where $\mathbf{g} \in \mathbb{R}^{D_w}$ is a gating vector for integration, $\mathbf{e}' \in \mathbb{R}^{D_w}$ represents the final integrated entity embedding and $\odot$ represents Hadamard product. Consequently, we compute the final label embedding $\mathbf{l}$:

$$\mathbf{l} = \mathbf{t}' - \mathbf{h}'. \tag{8}$$

Label Classifier. Ideally, each label embedding is supposed to act as an "anchor" for each relation class. To achieve this goal, we consider to train the learned label embeddings $\mathbf{l}$ to be easily classified as the correct relation class. Therefore, we use softmax to get the predicted probabilities of the relation classes:

$$\text{P}(y|D, G) = \text{Softmax}(\mathbf{M}_k \mathbf{l} + \mathbf{b}_k), \tag{9}$$

where $\mathbf{M}_k$ is the transformation matrix, $\mathbf{b}_k$ is the bias.

4.2 Neural Relation Extraction

After obtaining the embeddings of labels and entity descriptions, we leverage them to advance the neural relation extraction. We first use the entity descriptions to enhance instance embeddings. Then we leverage the label embeddings to attend over the instances to select valid instances for relation classification.

Instance Embedding. We enrich the representation of an instance with the pair of entity descriptions. Firstly, for each word $w \in s = \{w_1, \cdots, w_n\}$, its embedding $\hat{\mathbf{w}}_i$ is initialized as follows:

$$\hat{\mathbf{w}}_i = \mathbf{w}_i \oplus \mathbf{p}_{i1} \oplus \mathbf{p}_{i2}, \tag{10}$$

where $\mathbf{w}_i$ is the pre-trained word vector of w_i, $\mathbf{p}_{i1}$ and $\mathbf{p}_{i2}$ are its position embeddings to incorporate relative distances to two target entities into two D_p-dimensional vectors respectively (Zeng et al., 2014). The symbol $\oplus$ represents concatenation operator.

Then, we choose CNN (Zeng et al., 2014) with window size c as our encoder to learn the instance embedding considering the text of the instance itself.

$$\mathbf{z}_i = \text{CNN}(\hat{\mathbf{w}}_{i-\frac{c-1}{2}}, \cdots, \hat{\mathbf{w}}_{i+\frac{c-1}{2}}), \tag{11}$$

$$[\mathbf{s}]_j = \max\{[\mathbf{z}_1]_j, \cdots, [\mathbf{z}_n]_j\}, \tag{12}$$

where $\mathbf{s} \in \mathbb{R}^{D_h}$ is the sentence (instance) embedding, $[\cdot]_j$ is the j-th value of a vector and function $\max$ denotes max-pooling.

Finally, we concatenate the original instance embedding s with the entity descriptions $(\mathbf{d}_h, \mathbf{d}_t)$, obtaining the new instance representation $\mathbf{s}'$. Formally,

$$\mathbf{s}' = \mathbf{s} \oplus \mathbf{d}_h \oplus \mathbf{d}_t. \tag{13}$$

Attention over Instances. To alleviate the wrong labeling problem of distant supervision, we leverage the label embedding $\mathbf{l}$ as attention over instances to reduce the weights of noisy instances in the sentence bag $B = \{s_1, \cdots, s_m\}$. Then the representation of textual relation feature $\bar{\mathbf{s}}$ from the bag B can be calculated as weighted average of the instance embeddings $\mathbf{s}'_i$:

$$\hat{\mathbf{s}}_i = \tanh(\mathbf{W}_s \mathbf{s}'_i + \mathbf{b}_s), \tag{14}$$

$$\lambda_i = \frac{\exp(\mathbf{l} \cdot \hat{\mathbf{s}}_i)}{\sum_{j=1}^{m} \exp(\mathbf{l} \cdot \hat{\mathbf{s}}_i)}, \tag{15}$$

$$\bar{\mathbf{s}} = \sum_{i=1}^{m} \lambda_i \mathbf{s}'_i, \tag{16}$$

where $\mathbf{W}_s$ is the weight matrix and $\mathbf{b}_s$ is the bias vector. λ_i is the attention score of the instance s_i, computed based on the label embedding $\mathbf{l}$.

Relation Classifier. Finally, to compute the confidence of each relation class, we feed the representation of the textual relation $\bar{\mathbf{s}}$ into a softmax classifier after being processed by a linear transformation. Formally,

$$\mathrm{P}(y|B) = \mathrm{Softmax}(\mathbf{M}_s \bar{\mathbf{s}} + \mathbf{b}_s), \tag{17}$$

where $\mathbf{M}_s$ is the transformation matrix, and $\mathbf{b}_s$ is the bias.

4.3 Model Training

The objective function of our joint model RELE includes two parts, the loss of label classifier L_1 and the loss of relation classifier L_2.

Assuming that there are N bags in training set $\{B_1, B_2, \cdots, B_N\}$, and their corresponding labels $\{y_1, y_2, \cdots, y_N\}$, we exploit cross entropy for the loss function of label classifier L_1:

$$L_1 = -\sum_{i=1}^{N} \log \mathrm{P}(y_i|D, G), \tag{18}$$

The loss L_1 aims to train the label embedding to be classified to the correct relation class.

Similarly, for the loss of relation classifier L_2, we also exploit cross entropy and get:

$$L_2 = -\sum_{i=1}^{N} \log \mathrm{P}(y_i|B_i), \tag{19}$$

Finally, we aim to minimize the loss function L with L2-norm:

$$\min L = L_1 + L_2 + \eta ||\Theta||^2, \tag{20}$$

where η is the regularization coefficient and Θ denotes the parameters in our model. Stochastic gradient descent (SGD) is used to optimize our model.

5 Experiments

5.1 Datasets

In our experiments, we evaluate our model over the NYT-FB60K dataset (Han et al., 2018a) and GIDS-FB8K dataset (Jat et al., 2018). In the following, we detail each dataset.

- NYT-FB60K. The dataset NYT-FB60K (Han et al., 2018a) includes three parts: knowledge graphs (FB60K extended from Riedel dataset (Riedel et al., 2010), containing 1,324 relations, 69,512 entities, and 335,350 facts), text corpus (whose sentences are from Riedel dataset, containing 570,088 sentences, 63,696 entities, and 56 relations) and entity descriptions (which are the first paragraphs of the entities' Wikipedia pages, containing around 80 words on average).

- GIDS-FB8K. We construct the dataset GIDS-FB8K based on GIDS dataset (Jat et al., 2018). It also contains three parts: knowledge graphs (FB8K extended from GIDS dataset, containing 208 relations, 8,917 entities, and 38,509 facts), text corpus (whose sentences are from GIDS dataset, containing 16,960 sentences, 14,261 entities, and 5 relations) and entity descriptions (which are the first paragraphs of the entities' Wikipedia pages, containing around 80 words on average).

5.2 Baselines

We compare our model with the state-of-the-art baselines:

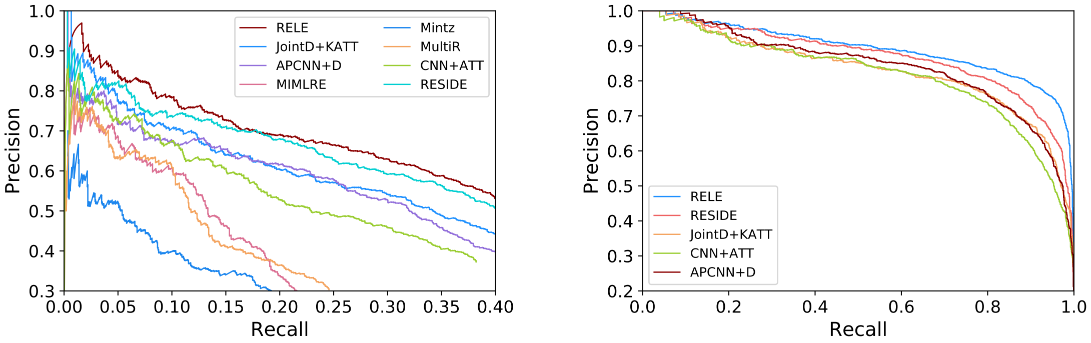
- Mintz (Mintz et al., 2009). A multi-class logistic regression model under distant supervision.

- MultiR (Hoffmann et al., 2011). A probabilistic graphical model for multi-instance learning.

- MIMLRE (Surdeanu et al., 2012). A graphical model for multi-instance multi-label learning.

- CNN+ATT (Lin et al., 2016). A CNN model with instance-level attention.

- APCNN+D (Ji et al., 2017) A Piecewise CNN model with instance-level attention using entity descriptions. As no code available, we implemented it by ourselves.



(a) NYT-FB60K Dataset

(b) GIDS-FB8K Dataset

Figure 3: Precision-recall curves for different methods. RELE achieves higher precision over the entire range of recall compared to all the baselines on both datasets.

Parameter	Value
Word/Entity/Relation Dimension D_w	50
Position Dimension D_p	5
Hidden Layer Dimension D_h	230
Kernel Size c	3
Learning Rate α	0.5
Regularization Coefficient η	0.0001
Dropout Probability p	0.5

Table 1: Parameter Settings.

- JointD+KATT (Han et al., 2018a). A joint model for knowledge graph embedding and relation extraction.

- Reside (Vashishth et al., 2018). A neural network based model which makes use of relevant side information and employs Graph Convolution Networks for encoding syntactic information of instances.

5.3 Evaluation Metrics

Following previous studies (Lin et al., 2016), our model is evaluated held-out, which compares the relations discovered from test corpus with those in Freebase. We report the Precision-Recall curve and top-N precision (P@N) metric for NYT-FB60K dataset.

For GIDS-FB8K dataset, we report the Precision-Recall curve, F1 score and Mean Average Precision (MAP). We do not use the top-N precision (P@N) metric for the dataset which is small containing only 5 relation classes.

5.4 Parameter Settings

For all the models, we use the pre-trained word embeddings with word2vec tool* on NYT corpus for initialization. The embeddings of entities mentioned in datasets are pre-trained through TransE model. We select the learning rate α among $\{0.1, 0.01, 0.005, 0.001\}$ for minimizing the loss. For other parameters, we simply follow the settings used in (Lin et al., 2016; Han et al., 2018a) so that it can be fairly compared with these baselines. Table 1 shows all the parameters used in our experiment.

5.5 Experiment Results

5.5.1 Precision-Recall Curves on Both Datasets

Figure 3 shows the comparison results in terms of Precision-Recall Curves on NYT-FB60K and GIDS-FB8K datasets. Overall, we observe that: (1) As shown in Figure 3(a), the neural network based approaches have more obvious advantages than Mintz, MultiR, and MIMLRE, illustrating the limitation of human-designed features and the advancement of neural networks in relation extraction. (2) APCNN+D using entity descriptions, and JointD+KATT exploiting KGs both outperform CNN+ATT on both datasets, showing that entity descriptions and KGs are both useful for improving the performance of relation extraction under distant supervision. RESIDE achieves better performance than APCNN+D and JointD+KATT. It is probably because that RESIDE utilizes more available side information, including entity types and relation alias information. (3) Our model

*https://code.google.com/p/word2vec/

Test Setttings	One				Two				All			
P@N(%)	100	300	500	Mean	100	300	500	Mean	100	300	500	Mean
CNN+ATT	67.3	61.1	56.5	61.6	73.3	62.1	53.5	63.0	74.2	62.4	55.9	64.1
APCNN+D	78.0	67.7	60.6	68.7	77.0	63.3	55.6	65.3	77.0	66.7	63.0	68.9
JointD+KATT	80.0	70.0	64.0	71.3	81.0	67.0	59.6	69.2	84.0	68.7	63.7	71.0
RESIDE	79.0	67.6	57.4	68.0	83.0	70.6	62.6	72.0	84.0	78.5	69.8	77.4
RELE	**87.0**	**76.6**	**67.0**	**76.8**	**88.0**	**73.3**	**63.2**	**78.7**	**88.0**	**78.6**	**69.8**	**78.8**

Table 2: Evaluation results P@N of different models using different number of sentences in bags on NYT-FB60K dataset. Here, **One**, **Two** and **All** represent the number of sentences randomly selected from a bag.

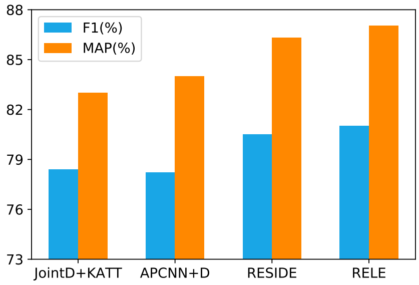


Figure 4: Comparison on GIDS-FB8K dataset.

RELE achieves the best performance compared to all the baselines. We believe the reason is that we make use of potential label information through joint label embedding. The learned label embeddings are of high quality since we fully exploit both the structural information from KGs and textual information from texts via gating integration while avoiding the imposed noise by an attention mechanism.

5.5.2 P@N Evaluation on NYT-FB60K Dataset

As shown in Table 2, we report Precision@N of different neural network based approaches on NYT-FB60K dataset. To verify the performance of our model on those entity pairs with few instances, we randomly select one, two and all instances for each entity pair, following previous studies (Du et al., 2018; Vashishth et al., 2018). As we can observe: (1) APCNN+D and JointD+KATT both outperform CNN+ATT in all cases, demonstrating the effectiveness of entity descriptions and KGs. RESIDE achieves better performance than APCNN+D and JointD+KATT by incorporating more side information (e.g., entity types). (2) Our model RELE significantly outperforms all the baselines. The reason is that our model learns high quality of label embeddings which play a critical role in relation extraction.

P@N(%)	100	300	500	Mean
RELE w/o LE	74.2	62.4	55.9	64.1
RELE w/o ATT$_e$	82.0	75.6	69.6	75.7
RELE w/o LC	81.0	74.0	69.0	74.7
RELE	**88.0**	**78.6**	**69.8**	**78.8**

Table 3: Evaluation results P@N of variant models on NYT-FB60K dataset.

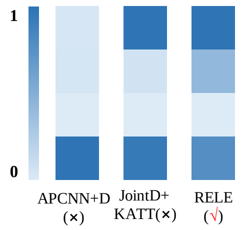
5.5.3 Results on GIDS-FB8K Dataset

Based on the results on the NYT-FB60K dataset, we choose APCNN+D, JointD+KATT and RESIDE as representative baselines to compare against our models on the GIDS-FB8K dataset in terms of F1 and MAP. As shown in Figure 4, our model consistently achieves better performance, which verifies the effectiveness of our model with joint label embedding. The overall results on GIDS-FB8K dataset show that our model can be well applied to smaller-scale datasets.

5.5.4 Comparison of Variant Models

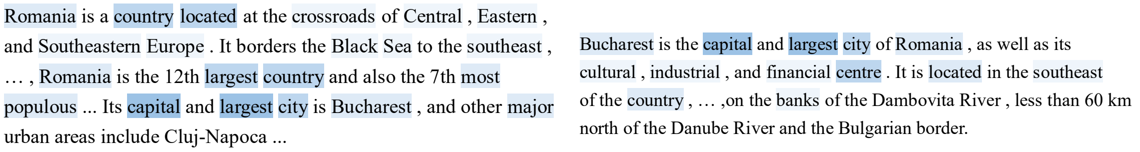
In order to verify the effectiveness of different modules of our model, we design three variant models:

- **RELE w/o LE** removes the label embedding from RELE, which degenerates to CNN+ATT.
- **RELE w/o ATT$_e$** removes the attention of the KG over entity descriptions during label embedding. We use max-pooling instead, which does not consider the noise in the entity descriptions.



s_1: In **Bucharest** , **Romania** , president Traian Basescu said Sunday night that the entry into …
s_2: Last year, it opened offices in Warsaw and **Bucharest** , the capital of **Romania**.
s_3: … by the presence of two films from **Romania** , the way I spent … east of **Bucharest** …
s_4:… Ervin a of New York City , on April 18th , age 98 , formerly of **Bucharest** , **Romania** .
Triplet(*Romania, Bucharest, /location/country/capital*)

Figure 5: A case study for predicting the relation between "Bucharest" and "Romania". The baselines all predict wrongly while our model gives the right result. The left shows the weights assigned to different sentences by different models. Our model always gives higher weights to correct sentences (shown in red).



Romania is a country located at the crossroads of Central , Eastern , and Southeastern Europe . It borders the Black Sea to the southeast , … , Romania is the 12th largest country and also the 7th most populous ... Its capital and largest city is Bucharest , and other major urban areas include Cluj-Napoca ...

(a) Description text of "Romania"

Bucharest is the capital and largest city of Romania , as well as its cultural , industrial , and financial centre . It is located in the southeast of the country , … ,on the banks of the Dambovita River , less than 60 km north of the Danube River and the Bulgarian border.

(b) Description text of "Bucharest"

Figure 6: Visualization of attention values of words in the descriptions of the entities "Romania" and "Bucharest".

- **RELE w/o LC** removes the label classifier. It does not train the label embeddings to be classified to the correct classes.

As shown in Table 3, without label embedding, the performance of RELE w/o LE drops significantly (more than 10%). It demonstrates the effectiveness of our joint label embedding. If we do not consider the imposed noise in the entity descriptions, the performance of RELE w/o ATT$_e$ decreases by around 3% on mean P@N. It demonstrates that the attention of KGs over entity descriptions is important for learning high-quality label embeddings. We also explore the performance of RELE w/o LC which does not train the label embeddings to be classified to the correct classes and find that label classifier plays a critical role in label embedding. The performance decreased by around 4% on mean P@N, without label classifier.

5.5.5 Case Study

Figure 5 illustrates an example from the test set of NYT-FB60K. As we can observe, representative attention-based baselines APCNN+D and JointD+KATT both predict wrong relation labels for the entities, while our model RELE correctly predicts the relation label "*/location/country/capital*". That is because the baselines assign relatively low weights to correct sentences (s_1 and s_2) and achieve inferior representations of the textual relation for classification. Our model RELE assigns higher weights to all the correct sentences s_1, s_2 and s_4, demonstrating that RELE learns high-quality label embeddings.

We also provide the insights of the attention of KGs over entity descriptions about "Romania" and "Bucharest". As shown in Figure 6, the words "capital", "largest". and "centre" which are related to the relation are given higher weights. It demonstrates that the attention of KGs over entity descriptions can help reduce the noise in entity descriptions and thus improve the label embeddings.

6 Conclusion

In this work, we consider leveraging potential label information to select valid instances for distantly-supervised relation extraction. We propose a novel multi-layer attention-based model RELE to improve relation extraction with joint label embedding. Our model takes full advantage of both structural information from KGs and textual information from entity descriptions to learn label embeddings, while avoiding the imposed noise with an attention mechanism. The label embeddings are trained to be classified to the correct relation classes. Then, the learned label embeddings are used as another attention over the bag to select valid instances for relation extraction. Extensive experiments have demonstrated that our model significantly outperforms state-of-the-art methods.

In the future, we will explore other useful information (e.g., correlations among the relations from KGs) available to improve the label embeddings.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (No. 61806020, 61772082, 61702296), the National Key Research and Development Program of China (2017YFB0803304), the Beijing Municipal Natural Science Foundation (4182043), the CCF-Tencent Open Fund, and the Fundamental Research Funds for the Central Universities.

References

Zeynep Akata, Florent Perronnin, Zaïd Harchaoui, and Cordelia Schmid. 2016. Label-embedding for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:1425–1438.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *NIPS*, pages 2787–2795.

Jinhua Du, Jingguang Han, Andy Way, and Dadong Wan. 2018. Multi-level structured self-attentions for distantly supervised relation extraction. In *EMNLP*, pages 2216–2225.

Xiaocheng Feng, Jiang Guo, Bing Qin, Ting Liu, and Yongjie Liu. 2017. Effective deep memory networks for distant supervised relation extraction. In *IJCAI*, pages 4002–4008.

Xu Han, Zhiyuan Liu, and Maosong Sun. 2018a. Neural knowledge acquisition via mutual attention between knowledge graph and text. In *AAAI*, pages 4832–4839.

Xu Han, Pengfei Yu, Zhiyuan Liu, Maosong Sun, and Peng Li. 2018b. Hierarchical relation extraction with coarse-to-fine grained attention. In *EMNLP*, pages 2236–2245.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *ACL*, pages 541–550.

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

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

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *ACL*, volume 1, pages 2124–2133.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *ACL/IJCNLP*, pages 1003–1011.

Denis Newman-Griffis, Albert M Lai, and Eric Fosler-Lussier. 2018. Jointly embedding entities and text with distant supervision. *ACL 2018*, page 195.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *ECML/PKDD*, pages 148–163.

Jose A Rodriguez-Serrano, Albert Gordo, and Florent Perronnin. 2015. Label embedding: A frugal baseline for text recognition. *International Journal of Computer Vision*, 113(3):193–207.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. Multi-instance multi-label learning for relation extraction. In *EMNLP-CoNLL*, pages 455–465.

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

Guoyin Wang, Chunyuan Li, Wenlin Wang, Yizhe Zhang, Dinghan Shen, Xinyuan Zhang, Ricardo Henao, and Lawrence Carin. 2018. Joint embedding of words and labels for text classification. In *ACL*, volume 1, pages 2321–2331.

Jiacheng Xu, Kan Chen, Xipeng Qiu, and Xuanjing Huang. 2016. Knowledge graph representation with jointly structural and textual encoding. In *IJCAI*, pages 1318–1324.

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

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via conlutional deep neural network. In *COLING*, pages 2335–2344.

Wenyuan Zeng, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2017. Incorporating relation paths in neural relation extraction. In *EMNLP*, pages 1768–1777.

Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alexander J Smola, and Le Song. 2018. Variational reasoning for question answering with knowledge graph. In *AAAI*, pages 6069–6076.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *IJCAI*, pages 4623–4629.