This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2020.2964747, IEEE Transactions on Knowledge and Data Engineering

Robust Neural Relation Extraction via Multi-Granularity Noises Reduction

Xinsong Zhang, Tianyi Liu, Pengshuai Li, Weijia Jia, *Senior Member, IEEE* and Hai Zhao

Abstract—Distant supervision is widely used to extract relational facts with automatically labeled datasets to reduce high cost of human annotation. However, current distantly supervised methods suffer from the common problems of word-level and sentence-level noises, which come from a large proportion of irrelevant words in a sentence and inaccurate relation labels for numerous sentences. The problems lead to unacceptable precision in relation extraction and are critical for the success of using distant supervision. In this paper, we propose a novel and robust neural approach to deal with both problems by reducing influences of the multi-granularity noises. Three levels of noises from word, sentence until knowledge type are carefully considered in this work. We first initiate a question-answering based relation extractor (QARE) to remove noisy words in a sentence. Then we use multi-focus multi-instance learning (MMIL) to alleviate the effects of sentence-level noise by utilizing wrongly labeled sentences properly. Finally, to enhance our method against all the noises, we initialize parameters in our method with a priori knowledge learned from the relevant task of entity type classification by transfer learning. Extensive experiments on both existing benchmark and an improved larger dataset demonstrate that our proposed approach remarkably achieves new state-of-the-art performance.

Index Terms—Neural Relation Extraction, Distant Supervision, Multi-instance Learning, Transfer Learning.

1 Introduction

To extract relations from large corpora, one may often face the challenge that training datasets have not been well-labeled. The traditional human-labeling way is costly for constructing a large-scale training set. Therefore, distant supervision [1] has been proposed for relation extraction by automatically constructing datasets with knowledge bases. There are amounts of relation triples such as [*Steve Jobs, Founder, Apple*] in knowledge bases. Distant supervision assumes that if a sentence contains entities in a relation triple, the sentence can probably describe the relation. Apparently, this assumption is too strong, since a sentence that mentions two entities does not necessarily express their relation contained in a known knowledge base. As described in [2], the assumption leads to a wrongly labeling problem. In order to tackle the problem, various multi-instance learning methods are adopted by mitigating noisy sentences with incorrect relation labels [3], [4], [5], [6]. Apart from the sentence-level noise, distantly supervised methods also suffer from word-level noise which derives from a large proportion of irrelevant words in a sentence. The word-level noise inside a sentence weakens the importance of significant words which contain key relation features, and the sentence-level noise misleads relation extractors to a poor convergence. To handle these multi-granularity noises, we face three major challenges in both extracting relation features and dealing with wrongly labeled sentences.

- *X. Zhang is with the ByteDance AI Lab, China. E-mail: zhangxinsong.0320@bytedance.com. Work done while X. Zhang was a PhD candidate in Shanghai Jiao Tong University.*
- *T. Liu, P. Li and H. Zhao are with the Department of Computer and Information Science, Shanghai Jiao Tong University. E-mail: {liutianyi, pengshuai.li}@sjtu.edu.cn and zhaohai@cs.sjtu.edu.cn.*
- *W. Jia is with the State Key Lab of IoT for Smart City, University of Macau and the Department of Computer Science and Engineering, Shanghai Jiao Tong University. E-mail: jia-wj@cs.sjtu.edu.cn.*

First, it is challenging to alleviate the impact of noisy words and select significant ones for relation extraction. For example, as shown in Fig. 1(a), these sentences indicate three relations. We can deduce that only a few words spotted in the sentences are useful for identifying relations. The other parts are all irrelevant words that can be seen as word-level noise. The sub-sentence *[Paul Malignaggi, an Italian American from Brooklyn.]* is much shorter but sufficiently express the relation *person/place_of_birth* for the sentence *S1*. Furthermore, we compute the distribution of sentence length in NYT-10 [2], which is a widely used benchmark dataset for distantly supervised relation extraction. As shown in Fig. 1(b), half of the original sentences are longer than 40 words while the corresponding parsed data [7] containing relation features is much shorter, which means that there are many irrelevant words inside sentences. To be more detail, there are more than twelve noisy words in each sentence on average, and 99.4% of sentences in NYT-10 have noise. Although a few methods have been proposed to get rid of irrelevant words for relation extraction such as dependency tree parser [7], [8] and word-level attention [9], they either are limited by the fixed syntactic patterns or weaken the importance of relation features contained in entities and other significant words. Moreover, current neural methods are tendentiously overused in relation extraction for their complicated structures applied to the entire sequences, as relation features do not distribute all around sequences and are usually discrete and sparse such as *S2* and *S3*. Therefore, modeling entire sequences including noisy words with previous convolutional or recurrent neural networks not only weakens relation features but also increases amounts of extra computation.

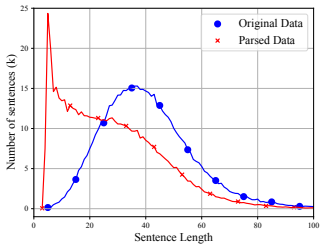
Second, to tackle with the problem of sentence-level noise, previous multi-instance learning approaches [5], [6] cannot make full use of wrongly labeled training sentences,

1041-4347 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on September 20,2020 at 14:48:13 UTC from IEEE Xplore. Restrictions apply.

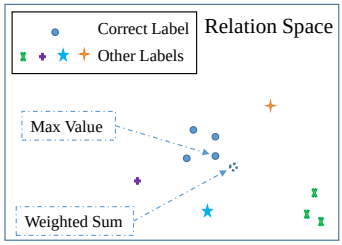
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2020.2964747, IEEE Transactions on Knowledge and Data Engineering

Sentences	Labels
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]*** .	PPB
The politicians want to block a proposal for the ***[Goshute] Indian reservation*** to gain income by storing nuclear waste produced ***in other states***, while ***[Utah] has no*** nuclear power ***plants***.	LC
[Steve Jobs] was the ***co-founder*** and CEO of ***[Apple]*** and formerly Pixar.	CF

(a)



(b)



(c)

Fig. 1. (a) presents the illustration of key words in sentences for relation extraction. Words in brackets are entities and the italic red parts are key words relevant to the relations. **PPB**, **LC** and **CF** represent relation types *person/place_of_birth*, *location/contain* and *company/founders* respectively. (b) sentence length distributions on NYT-10. (c) shows the sentence distribution of a bag in relation space. Previous multi-instance learning algorithms generate relation representations such as **Max Value** or **Weighted Sum** to represent the bag.

but only represent a bag label by choosing the most possible sentence or computing the weighted sum of the sentences in the bag as shown in Fig. 1(c). However, such approaches may miss potentially useful information in the wrongly labeled sentences. The former [5] neglects amounts of correctly labeled sentences, and the latter [6] treats wrongly labeled sentences as noise by assigning small weights which ignores positive impacts of them. Besides, the weighted sum of a bag possibly introduces deviation to represent relation features because of wrongly labeled sentences in the bag. Therefore, both correctly labeled sentences and wrongly labeled ones should be utilized in proper ways separately.

Third, a robust method is supposed to extract precise relation features even from noisy datasets and have strong capability of noise immunity. However, nearly all existing neural methods are lacking in robustness because parameters are initialized randomly and hard to tune on noisy training data, resulting in an inevitable poor performance. Inspired by [10], initializing neural networks with a priori knowledge by transfer learning on relevant tasks could improve the robustness of the target task. Entity type classification can be used as a relevant task for relation extraction since entity types provide abundant background knowledge. For instance, the sentence *S3* in Fig. 1(a) has a relation *company/founders*, which is uneasily disclosed without knowing that *Steve Jobs* is a person and *Apple* is a company. This demonstrates that entity types are useful a priori knowledge to initialize relation extractors.

In this paper, we propose a novel and robust method for distantly supervised relation extraction to jointly tackle the challenges above. We first initiate a question-answering based relation extractor to reduce word-level noise. Then, a multi-focus multi-instance learning is proposed for sentence-level noise. Finally, to enhance noise immunity, we initialize our model parameters with a priori knowledge learned on entity type by transfer learning.

For the first challenge, we initiate a Question-Answering based Relation Extractor (QARE), which effectively tackles word-level noise with much less computational cost. Compared with noisy words in a sentence, entities contain significant information to extract relation features. Therefore, we utilize two entities as a query question to search salient words containing relation features in a sentence. Specifically, given an entity pair $[e_1, e_2]$, a *question* is considered as *what is the relationship between* e_1 *and* e_2*?*. Its *answer* will be relation features which are queried discretely from the target sentence. In QARE, a few significant words are first filtered according to their relevance to the entity pair. The *answer* vector is then computed with these relevant words and represented as final relation representation of the sentence. With the process of question-answering inside a sentence, QARE will obtain high-quality relation features effectively by reducing word-level noise.

Moreover, QARE does not model entire sequences like traditional neural networks such as CNN/RNN in consideration of the efficiency of neural relation extraction. We select a few significant words and only maintain *question* and *answer* vectors instead of plenty hidden vectors in CNN/RNN. Therefore, QARE can also extract relation features efficiently by saving amounts of computational cost.

To alleviate the influence of sentence-level noise, we propose Multi-focus Multi-Instance Learning (MMIL) to utilize wrongly labeled sentences properly. For the sake of simplicity, we call sentences labeled with right relations as true instances, and false instances represent wrongly labeled ones. For a sentence bag, we jointly consider the bag label and actual relations of false instances. In MMIL, the sentences in a bag will be split into true and false sets according to the possibility of satisfying the bag label. We select the most likely sentence and its nearest neighbors in relation space as true instances, and the others are treated as false ones. Then, the true instances are used to train our model with the bag label, while the false ones are computed as consistency regularization. Inspired by [11], [12], we add a perturbation to the relation representations of false instances and make the predictions of them consistent to the perturbation. In MMIL, we not only focus on the bag label but also pay close attention to the predictions of false instances. Therefore, the false instances which have been treated as noisy sentences in previous works have less influence on true instances in our method, and both true instances and false ones contribute to strengthening relation features collaboratively.

Finally, to enhance noise immunity and improve the robustness of our method, we initialize parameters with a priori knowledge from an entity type classification task by transfer learning [13]. The entity types are helpful to distinguish relations. Besides, transfer learning from multiple

1041-4347 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on September 20,2020 at 14:48:13 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2020.2964747, IEEE Transactions on Knowledge and Data Engineering

tasks can improve generalization of all the tasks by using domain information contained in the training signals of related tasks as an inductive bias [14]. Our main contributions in this work are summarized as follows:

- We propose a robust neural approach for relation extraction by tackling the multi-granularity noises jointly. QARE is initiated to remove noisy words in sentences. To our best knowledge, QARE is the first efficient neural network for relation extraction in the way of question-answering inside a sentence.
- MMIL is presented to alleviate the influence of sentence-level noise by utilizing wrongly labeled training instances properly.
- We initialize neural relation extractors with a priori knowledge from entity type classification, which enhances immunity against multi-granularity noises.
- Our approach can achieve solid performance improvement over existing state-of-the-art works on both a widely used benchmark and a new larger dataset.

2 Related Work

Neural Relation Extraction. Relation extraction [15], [16], [17] is a critical task for Natural Language Processing (NLP) in which neural models have been widely used for their capability of extracting semantic meanings without hand-designed features. CNN has been proved effective for relation extraction [18], [19], [20]. Then, Zhou *et al.* classified relations based on long short-term memory network with attention mechanism [9]. Wang *et al.* improved relation classification with multi-level attentive CNN [21]. In addition, a few more complicated neural methods have been proposed [22], [23], [24], [25], which pay no attention to word-level noise and model efficiency though. To better remove irrelevant words, the dependency path between entities was proposed and verified effectiveness in [7], [8], [26], [27], [28] but lacked the ability to deal with large amounts of relations because of the limitation of fixed syntactic patterns. Besides, all the previous methods consume a large amount of computation and suffer from the overuse of complicated neural network structures.

Multi-instance Learning. To alleviate the influence of wrongly labeled instances in automatically constructed datasets for distantly supervised relation extraction, multi-instance learning algorithms were integrated with neural relation extractors [5], [6]. Previous works computed representations of sentence bags by selecting important instances. Zeng *et al.* [5] chose the most possible instance to represent the bag, and other works computed the weighted sum of all the instances with the selective attention [6] and non-IID relevance embedding [29]. The work of Ji *et al.* [30] introduced another kind of attention weights computed with external entity descriptions. Ye *et al.* proposed a multi-level attention mechanism to deal with wrongly labeled sentences and bags [31]. Besides, there are a few methods handling the wrongly labeling problem by changing training sets in recent years. Liu *et al.* [32] relabeled the datasets with soft labels generated in training. Feng *et al.* [33] and Qin *et al.* [34], [35] selected true instances for training by reinforcement learning or generative adversarial nets which will lead to a huge training cost increase. All the existing multi-instance learning solutions never attempt to extract relation features from false instances which could also provide helpful signal.

Transfer Learning. Transfer learning [13] provides a new approach to leverage knowledge extracted by related tasks to enhance the target task. For example, a recurrent neural network based architecture is introduced to model text sequence with multi-task learning for relation classification [36]. Cooperated with multi-task learning and attention mechanism, a multi-lingual neural relation extraction framework was introduced to utilize the information within mono-lingual texts [37]. Furthermore, parameter transfer learning has shown effectiveness to improve the robustness of models by initializing model parameters reasonably [10], [38].

All the existing neural methods for distantly supervised relation extraction cannot sufficiently reduce the negative impacts of multi-granularity noises. Moreover, they are lacking in robustness with randomly initialized parameters. In contrast, the robust approach proposed in this paper focuses on multi-granularity noises and achieves impressive improvements for large-scale relation extraction in dealing with different levels of noises.

3 Methodology

In distantly supervised relation extraction paradigm, all sentences labeled by a relation triple constitute a bag, and each sentence is called an instance. The relation triple is described as [*head, relation, tail*], where $head$ and $tail$ are both entities. Suppose that there are N bags $\{B_1, \cdots, B_N\}$ in training set and that the i-th bag contains q_i instances $B_i = \{b_1^i, \cdots, b_{q_i}^i\}(i = 1, \cdots, N)$. The objective of relation extraction is to predict the labels of unseen bags. As shown in Fig. 2, our model is divided into three parts:

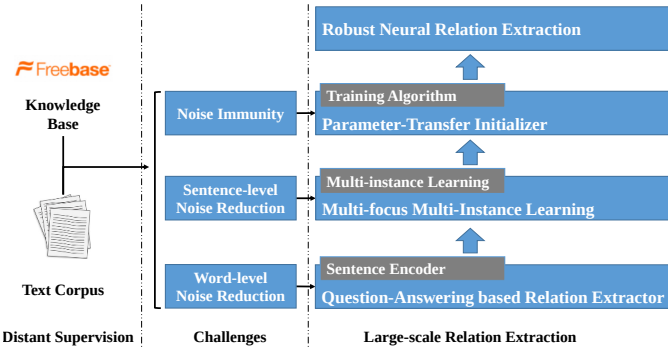


Fig. 2. Overview of the challenges and the framework of our proposed approach.

Question-Answering based Relation Extractor. Given an instance b^* and two target entities, QARE encodes it to a relation representation from salient words with a novel neural network which is more accurate and efficient than previous sentence encoders such as CNN/RNN.

Multi-focus Multi-Instance Learning. Given a bag of instances B^* and two target entities, we handle true instances and false ones in different ways to fully use wrongly labeled instances.

Parameter-Transfer Initializer. After dealing with word-level and sentence-level noises, we initialize parameters

1041-4347 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on September 20,2020 at 14:48:13 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2020.2964747, IEEE Transactions on Knowledge and Data Engineering

with a prior knowledge learned from related tasks to improve the robustness of our model.

3.1 Question-Answering based Relation Extractor

The proposed question-answering based relation extractor (QARE) extracts relation features in the perspective of question-answering as shown in Fig. 3. The question is an entity pair which can be described as *what is the relationship between* q_1 *and* q_2*?*, and the answering is relation features queried from all the word tokens in a sentence. By transforming relation extraction to a task of question-answering inside a sentence, we can reduce word-level noise and achieve better performance efficiently.

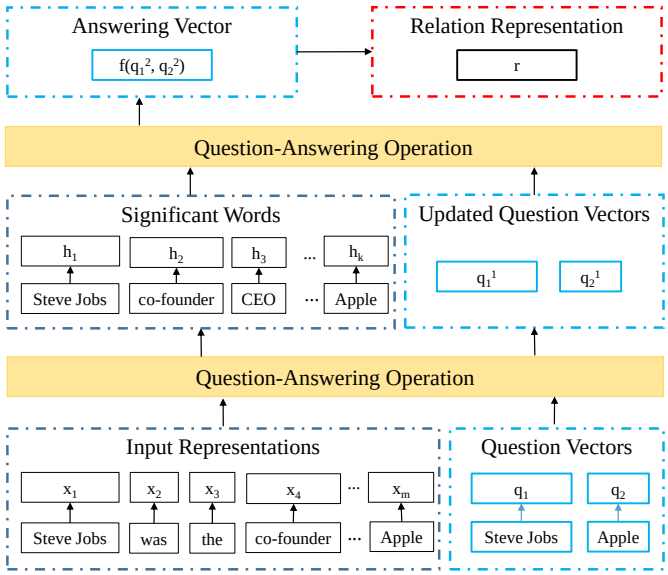


Fig. 3. The architecture of question-answering based relation extractor, illustrating the procedure for handling one sentence and predicting the relation between [Steve Jobs] and [Apple].

3.1.1 Network Architecture

The inputs of our network are word representation X and question embeddings $[q_1, q_2]$. The question embeddings are initialized with the word embeddings of entities. The first question-answering operation module selects top k significant words and updates question vectors with all the input representations. The second question-answering operation module updates question vectors for the second time with the filtered significant words. With the final updated question vectors, we compute the answering vector which contains all the relation features related to the initial entities. Inspired by [39], we present translating embedding function $f(\cdot)$ to compute the answering vector with two updated question vectors $[{q_1}^2, {q_2}^2]$. Finally, the answering vector represents relation features.

Input Representation. Tokens in sentences should be embedded to distributed representations for mathematical operations in neural networks [40]. For the input tokens $\{t_1, t_2, \cdots, t_m\}$ in a sentence, we train the token t_i to vector $word_i \in \mathcal{R}^{d_w}$ in a priori with the setting of $skip\text{-}gram$ [41]. The parameter d_w indicates the dimension of the $word$.

In addition, in consideration of the fact that relative positions between input tokens and entity words are really important for predicting relations between the entities, we encode the relative distances to position embedding of each token. For example, the relative distances from the token *co-founder* to the head entity [*Steve Jobs*] and the tail entity [*Apple*] are respectively 3 and -4 in the sentence *[Steve Jobs] was the co-founder and CEO of [Apple] and formerly Pixar*. We encode distances of token t_i to vector $position_i \in \mathcal{R}^{d_p}$, where d_p is the dimension of position embedding. The position embeddings are initialized randomly and updated in training.

Finally, the representation of an input token is the concatenation of word embedding $word$ and position embedding $position$. We denote all the input tokens in a sentence as an input matrix $X = [x_1, \cdots, x_i, \cdots, x_m]$, where representation $x_i \in \mathcal{R}^{d_x}$ ($d_x = d_w + d_p$) and m is the number of tokens.

Question-answering operation accomplishes two tasks, 1) select the significant tokens according to their relevancy to the entities and 2) update the question vectors with relation features. The relevancy matrix $A \in \mathcal{R}^{2\times m}$ of input tokens are computed with the following equation,

$$E = [q_1, q_2]^T X, \tag{1}$$

$$A_{ij} = \frac{\exp(E_{ij})}{\sum_{j'=1}^{m} \exp(E_{ij'})}, \tag{2}$$

where T is the transpose of the matrix. For the task of selecting top k word tokens, we compute the weights α for word tokens with relevance matrix by the equation $\alpha_j = \sum_{i=1}^{2} A_{ij}$. Meanwhile, we update the question vectors with the relevancy matrix and all the input embeddings to integrate relation features,

$$\left[q_1^1, q_2^1\right]^T = AX^T \tag{3}$$

Following [42], we use multi-head mechanism and feed-forward method to make the relation extractor efficient and stable. The multi-head mechanism computes question vectors for l different linear projections of the original A and X in parallel,

$$\left[q_1^1, q_2^1\right]^T_{multi-head} = [A_1 X_1^T; \cdots; A_l X_l^T], \tag{4}$$

where $[x; y]$ denotes the horizontal concatenation of x and y, and X_l represents the l-th linear projection of original X. The multi-head mechanism performs better than single-head one by allowing the models jointly attend to information from different representation subspaces at different positions. The feed-forward method actives each question-answering operation with linear transformations. $Q \in \mathcal{R}^{2\times d_r}$ represents question vectors of each question-answering operation, which is activated by two linear transformations with a ReLU activation in between,

$$\sigma(Q) = max(0, a_1 Q + b_1)a_2 + b_2, \tag{5}$$

where a_1, a_2, b_1 and b_2 are learnable parameters. The input matrix and output matrix $\sigma(\cdot)$ keep the same shape. In addition, we apply residual information [43] to avoid the vanishing of features between the two question-answering operations,

$$Q^{j+1} = A^j {X^j}^T + Q^j, \tag{6}$$

1041-4347 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on September 20,2020 at 14:48:13 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2020.2964747, IEEE Transactions on Knowledge and Data Engineering

where Q^j and Q^{j+1} are question vectors at the j-th, $(j+1)$-th question-answering operation, X^j are the input embeddings of j-th operation. The residual information prevents QARE from converging at earlier layers.

Answering Vector is a translation of the two question vectors integrated with entity information and relation features. In [39], it has been proved that relations can be represented as translations in the feature space: if [*h, r, t*] holds, then the embedding of the head entity h should be closed to the embedding of the tail entity t minus the relation feature vector that depends on *r*. The equation can be expressed as $h \approx t - r$. However, the minus operation may be too simple to capture the complex cases of relation semantics. Our relation vector r is fitted by a neural layer using the following equation,

$$r = f(q_1^2, q_2^2) = q_1^2 + a_t q_2^2 + b_t, \tag{7}$$

where q_1^2 and q_2^2 are final question vectors after the second question-answering module. $a_t \in \mathcal{R}^{d_r}$ and $b_t \in \mathcal{R}^{d_r}$ represent parameters for translating operation. $r \in \mathcal{R}^{d_r}$ represents the answering vector which is also the relation representation.

3.1.2 Complexity Analysis

We theoretically compare the complexity of QARE to that of CNN and RNN used for relation extraction. First, we introduce two former neural structures in detail. We then compare the performance of the three networks from three aspects which are executing time, memory[1] occupied and minimum number of sequential operations per layer. Time and memory cost shows the computational complexity, while the sequential operations indicate the amount of computation that can be parallelized.

CNN Network. Convolutional Neural Network (CNN) is a widely used structure for sentence encoding in many relation extractors [5], [6]. With the input embeddings, the convolutional layer extracts local features with a sliding window of length w over the input tokens. Local features h from w adjacent word tokens are extracted with dot production between convolutional kernels and input embeddings. The convolutional kernels are weight vectors represented by $W \in \mathcal{R}^{d_r \times w d_x}$ and the number of kernels is d_r. In summary, the convolutional operation follows the equation,

$$h_{ij} = W_i \cdot [x_{j-1}; x_j; x_{j+1}], \tag{8}$$

where $[x; y]$ denotes the vertical concatenation of x and y. h_{ij} presents j-th value of the i-th filter, where i and j are in range $[1, d_r]$ and $[1, m]$ respectively. Out-of-range input values x_j, where $j < 1$ or $j > m$, are taken to be zero. A max-pooling operation selects the most important features of each h_i with $h_i^* = max(h_{ij})$, where $h^* \in \mathcal{R}_r^d$. Then, we summarize h^* to relation representation r by a non-linear function such as the hyperbolic tangent.

BGRU Network. Given the input representations of each instance, Gated Recurrent Unit (GRU) [44] can extract global information of each word by pointing out its corresponding position in the sequence, which has been applied to relation extraction successfully [7], [9]. GRU consists of two key components: 1) the update gate u_t with the corresponding weight matrix, and 2) the reset gate r_t with the corresponding weight matrix. The update gate decides how much the unit updates its state or content, and the reset gate effectively makes the unit act as if it is reading the first symbol of an input sequence, allowing it to forget the previously computed state. Both of the two gates are set to generate candidate state $\tilde{h_t}$, using current input x_t and the state h_{t-1} that the previous time step generated. Finally, the state h_t at time t is a linear interpolation of the previous state h_{t-1} and the candidate state $\tilde{h_t}$. The whole procedure is demonstrated with the following equations,

$$u_t = \sigma(W_u[x_t] + U_u h_{t-1} + b_u), \tag{9}$$

$$r_t = \sigma(W_r[x_t] + U_r h_{t-1} + b_r), \tag{10}$$

$$\tilde{h_t} = \tanh(W[x_t] + U(r_t \odot h_{t-1}) + b), \tag{11}$$

$$h_t = u_t \odot \tilde{h_t} + (1 - u_t) \odot h_{t-1}, \tag{12}$$

where $W_u, U_u, b_u, W_r, U_r, b_r, W, U, b$ are neural parameters for the update gate, the reset gate and the candidate state respectively. σ is the sigmoid function and $\odot$ is the element-wise multiplication. Furthermore, BGRU implementing GRU in bidirectional directions can access future as well as past context. The following equation defines the operation mathematically,

$$h_t = [\overrightarrow{h_t} \oplus \overleftarrow{h_t}], \tag{13}$$

In above equation, the t-th word output $h_t \in \mathcal{R}^{d_r}$ of BGRU is the element-wise addition of the t-th hidden states of forward GRU and backward one. With all the hidden states h, the sentence representation $r \in \mathcal{R}_r^d$ is computed as a weighted sum of h_t,

$$r = \sum_{t=1}^{T} \alpha_t^w h_t, \tag{14}$$

where weights α_t^w are from attention mechanisms [7], [9].

Time and Space Complexity. For the time cost, CNN keeps $O(d_r \times w d_x)$ computation for each word token when multiplying kernels with the word representation. The computation of BGRU for one word token is $O(d_r \times d_x)$, because it transforms word representation to a hidden state with variables $W_u(W_r, W) \in \mathcal{R}^{d_r \times d_x}$. In QARE, only two query vectors are applied on m or k words, the costs thus are extremely low. Besides, shown as the equations in QARE, all the word tokens only do linear computation, and the representation dimension d_r is equal to that of word representation d_x. Therefore, the time cost of QARE is $O(d_x(m+k))$.

In memory analysis, the space cost of word representations Φ is same for the three neural networks, and the parameters of them are not major consumption of memory especially for long sentences. Therefore, we focus on the major extra memory consumption parts which are outputs of each neural layer for a sentence. Due to the need of back-propagation, all the three networks store the hidden states of work tokens which are treated as the major memory cost in this paper. The number of stored hidden states for CNN and RNN are m, while that for QARE is only 2 because of the limited question vectors. Finally, the space complexity for the CNN, BGRU and QARE are $O(md_r)$, $O(md_r)$ and $O(2d_x)$ respectively.

1. All the memory referred in this paper is GPU memory.

1041-4347 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on September 20,2020 at 14:48:13 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2020.2964747, IEEE Transactions on Knowledge and Data Engineering

As shown in Table 1, we compare all the time and space complexity of the three neural structures for relation extraction. In the table, w is the kernel size of convolutions, m is the sequence length, k is the number of selected significant words, d_r is the relation representation dimension and d_x is the dimension of input representation. Besides, Φ is common memory consumption for input representations. RNN is the slowest of all the three methods for its requirement of $O(m)$ sequential operations, which makes it difficult to be executed in parallel. QARE achieves the best efficiency in both time and space for its simple operations and focusing on significant words only.

TABLE 1
Complexity and the minimum number of sequential operations for the three neural architectures.

Model	time	memory	operation
CNN	$O(wmd_rd_x)$	$O(md_r + \Phi)$	$O(1)$
RNN	$O(md_rd_x)$	$O(md_r + \Phi)$	$O(m)$
QARE	$O((m+k)d_x)$	$O(d_x + \Phi)$	$O(1)$

3.2 Multi-focus Multi-Instance Learning

In this section, we present our proposed Multi-focus Multi-Instance Learning (MMIL), as shown in Fig. 4, to alleviate the influence of sentence-level noise. We focus on multiple relations for a bag of sentences including the bag labels and possible relations of false instances.

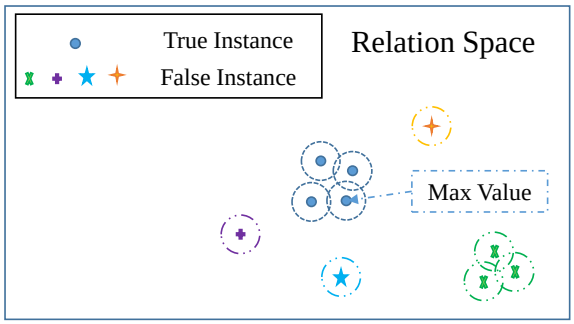


Fig. 4. The overview of multi-focus multi-instance learning. The solid circles are true instances and the other solid shapes are false ones. The overlapped hollow circles indicate a same relation and the diameters of them are all defined with a same distance η in the relation space.

With a relation representation r produced by the relation extractor, the probability $p(y|r;\theta)$ for each relation is given through softmax. We select the most confident sentence as the seed true instance, which has the highest estimated probability in a sentence bag. We assume that near sentences in relation space, whose distances are less than a threshold η, can represent same relations. Therefore, same relations can be clustered together in relation space. A tight threshold η in our assumption will select only one true instance, while loose thresholds lead to cluster too many "true instances". A proper threshold η is important in our assumption. With the assumption, true instances are selected around the seed with a greedy algorithm as shown in Algorithm 1.

With the validity classification algorithm, all the input sentences in a bag are classified into true and false sets. The sentences in true set are trained with the bag label, while false ones are used for computing consistency regularization with predictive labels and perturbations. With the method, we can utilize both true instances and false ones sufficiently shown as Fig. 5. The loss function is shown as the following equation,

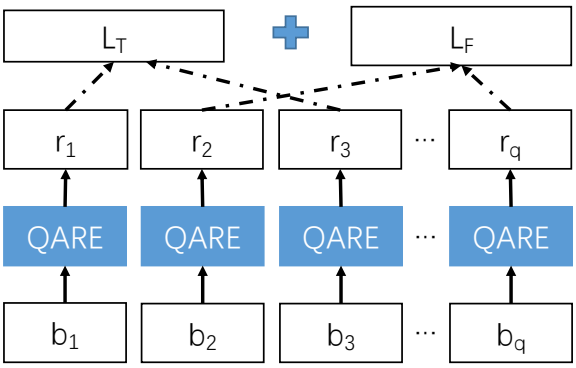


Fig. 5. The formulation of multi-focus multi-Instance learning with a sentence bag. L_T represents regular cross-entropy for true labeled instances and L_F is consistency regularization for false labeled ones. The solid lines are determinate associations, and the dotted lines are possible ones.

Algorithm 1 Validity Classification

Require: sentence representations in a bag B, threshold η
Ensure: validity sentence set V
1: Put the most possible true instance b^* to V
2: Compute shortest distance vector D from $[B-V]$ to V
3: **for** exist $d_i < \eta$ **do**
4: Put b_i to V
5: Update D
6: **end for**

$$J_m(\theta) = -\frac{1}{T}\sum_{i=1}^{T}\log p(y_i|r_i;\theta) - \frac{\beta}{zF}\sum_{j=1}^{F}\sum_{u=1}^{z} p(y_u|r_j;\hat{\theta})\log p(y_u|r_j+\phi;\theta), \tag{15}$$

where T and F are numbers of true and false instances, z is the relation number, $\hat{\theta}$ presents the constant parameters in current iteration without the gradient, ϕ is a perturbation vector and β is a linear hyper-parameter. The second term is a kind of consistency regularization which makes the predictions consistent with the noise ϕ. To accomplish the algorithm above and set an appropriate threshold, we define a distance metric in relation space and one kind of perturbations.

KL Distance. KL distance is computed from Kullback-Leibler (KL) divergence, which can evaluate differences between distributions. Given two sentence representations $a \in \mathcal{R}^{d_r}$ and $b \in \mathcal{R}^{d_r}$ in relation space, we first encode them to probability distributions in the same dimension as following,

$$p_a = softmax(a) \tag{16}$$

$$p_b = softmax(b) \tag{17}$$

1041-4347 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on September 20,2020 at 14:48:13 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2020.2964747, IEEE Transactions on Knowledge and Data Engineering

We adopt the KL divergence of (a, b) as the distance[2] from b to a,

$$KL(a, b) = \log(p_a) \log(p_a/p_b) \tag{18}$$

Perturbations. We provide random perturbations for false instances[3].

$$CE = -p(y|r; \hat{\theta}) \log p(y|r + \phi; \theta) \tag{19}$$

where CE is the cross entropy for relation extraction and ϕ is a random vector which is sampled from a Gaussian distribution with a mean of 0 and a standard deviation of 0.01.

3.3 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 by enhancing noise immunity.

3.3.1 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. The entity classification task shares the same extractor with the relation extraction task. In relation extractor, representations for the head and the tail are considered as r_{head} and r_{tail} respectively[4]. The entity representations are ultimately fed into the softmax layer for entity classification,

$$\hat{p}^i = softmax(W_i r_i + b_i);\ i \in \{head, tail\}, \tag{20}$$

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

$$J_e(\theta_0, \theta_{head}, \theta_{tail}) = \sum_i (-\frac{1}{z_i} \lambda_i \sum_{j=1}^{z_i} y_j^i log(\hat{p}_j^i)) \\ i \in \{head, tail\}, \tag{21}$$

where λ_i 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 and $y^i \in \mathcal{R}^{z_i}$ is the one-hot vector representing ground truth.

3.3.2 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. The loss function for the relation extraction task is computed in MMIL which is represented as $J_m(\theta_0, \theta_r)$.

2. Euclidean distance is an alternative distance metric, which performs slightly worse than KL distance.

3. Adversarial perturbations can also be used in our algorithm, which means that ϕ can be learned against the gradient of CE. The adversarial perturbations achieve similar results as random perturbations.

4. In QARE, they are final question answers q_1^2, q_2^2.

Based on the parameter-transfer initializer, related tasks share basic layers[5] in our relation extractor. 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 as

$$\theta = \theta_0 \cup \theta_{head} \cup \theta_{tail} \cup \theta_r, \tag{22}$$

where θ_0 represents common variables in QARE for the related tasks. θ_{head}, θ_{tail} and θ_r are extra parameters for head type classification, tail type classification and relation extraction respectively.

At first, we minimize J to obtain θ_0 at the best model state $\hat{\theta}_0$ for all the related tasks training together. Then we minimize J_m for the best performance of relation extraction under the initialization of θ_0 to be $\hat{\theta}_0$. Above process can be summarized as

$$\begin{aligned} min \quad J(\theta) =& \lambda J_e(\theta_0, \theta_{head}, \theta_{tail}) + \\ & (1 - \lambda) J_m(\theta_0, \theta_r), \end{aligned} \tag{23}$$

where $\lambda \in (0, 1)$ is the hyper-parameter to determine the importance of each task at different training steps.

4 Experiments

In this section, we conduct experiments to answer the following questions, 1) Is our method superior to the baselines in distantly supervised relation extraction? 2) Is QARE effective to reduce the word-level noise and efficient to extract relation features? 3) Is MMIL more accurate than previous multi-instance learning solutions in dealing with sentence-level noise? 4) Does transfer learning improve the robustness of neural relation extractors?

4.1 Experimental Settings

4.1.1 Datasets

We conduct experiments on two datasets. **NYT-10** is a widely used dataset that was developed in [2], and **NYT-18** is a new larger one. NYT-10 crawled three years documents on the New York Times (NYT) in 2005-2007, while NYT-18 contains ten years NYT documents from 2008 to 2017. They are both labeled with Freebase and Stanford Named Entity Recognize [45]. For the dataset NYT-10, we follow the previous works [6], [32] and split sentences from the years 2005-2006 to a training set and regard sentences from 2007 as a test set. As for NYT-18, we divide all the sentences into five parts with the same relation distribution for five-fold cross-validation. Meanwhile, all relations in Freebase are defined on head types and tail types. Therefore, we can construct datasets for type classification tasks with the same datasets. The dataset details are shown in Table 2.

TABLE 2
The datasets information. **Sen.** and **Ent.** indicate numbers of sentences and entity pairs. **Head.** and **Tail.** represent types of head entity and tail entity. **Rel.** is the number of relations.

Datasets	Training (k)		Testing (k)		Head.	Tail.	Rel.
	Sen.	Ent.	Sen.	Ent.			
NYT-10	523	281	172	97	29	26	53
NYT-18	2,446	1,234	611	394	332	277	503

5. In QARE, shared layers are under the answering vector layer.

1041-4347 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on September 20,2020 at 14:48:13 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2020.2964747, IEEE Transactions on Knowledge and Data Engineering

4.1.2 Baselines

Our method is compared with six baselines for distantly supervised relation extraction. The overall performance of all the baselines on NYT-10 is reported as their papers, while further comparisons on specific modules are evaluated by our implementations or their codes on the same platform (Tensorflow) and runtime environment (Nvidia TITAN Xp GPU) to compare the efficiency fairly.

Zeng *et al.* [5] extracted relation features by Piecewise CNN with the most possible sentence in a bag (PCNN+ONE).

Lin *et al.* [6] integrated the PCNN network with selective attention over instances (PCNN+ATT).

Liu *et al.* [32] changed the labels of training sentence bags with generative soft labels (PCNN+ATT+SL).

Liu *et al.* [7] shortened the training instances with the parser tree and pre-trained the word embeddings with transfer learning (STPRE).

Ye *et al.* [31] proposed a two layer attention mechanism to emphasize true labeled sentences and bags (PCNN+ATT_RA+BAG_ATT).

BGRU+ATT integrates attention based bi-directional gated recurrent unit [9] with selective attention [6].

4.1.3 Evaluation Metric

Our method is evaluated in the held-out evaluation and resource cost of the training process. The held-out evaluation is widely used for distantly supervised relation extraction by comparing relation triples discovered from the test sentences with those in Freebase. It provides an approximation of the precision without human evaluation. To be more precise in quantitative analysis, we also compute Precision at top N predictions (P@N) for the baselines. Besides, time and memory cost of the training process are significant indicators to qualify the efficiency of neural relation extractors.

To evaluate the influence of multi-instance learning algorithms, we propose three test settings which are **One**, **Two** and **All**. **One** randomly selects one instance to express the relation for each testing entity pair. **Two** fills each sentence bag with two different instances. **All** keeps at least two instances under each entity pair.

4.1.4 Parameter Settings

TABLE 3
Parameter settings for the compared methods.

Methods	QARE	PCNN	BGRU
d_w	50	50	50
d_p	10	10	10
d_r	60	230	230
k	15	-	-
l	2	-	-
η	0.03	-	-
β	0.3	-	-
$\lambda_{head}, \lambda_{tail}$	0.5,0.5	-	-
λ	0.3	-	-
s_b	50	50	50
DR	0.1	0.5	0.5
LR	0.0005	0.0005	0.0005
$L2$	0.0001	0.0001	0.0001

In our experiments, word embeddings are trained a prior with $skip\text{-}gram$ setting of word2vec [41] on the two datasets. In our work, we concatenate the words of an entity when it has multiple words. We use Adam optimizer [46] to minimize the loss function. A batch of sentences are randomly selected from the training set and fed to our method for each iteration until convergence. $L2$ regularization and the dropout method [47] are adopted to avoid overfitting. We use cross-validation and grid search to determine important parameters of our method including position dimension d_p, representation dimension d_r, selecting numbers k, multi-head number l, distance threshold η, false term weight β, entity task weights $(\lambda_{head}, \lambda_{tail})$, entity-relation task weight λ, batch size s_b, dropout rate DR, learning rate LR and regularization strength $L2$. The other parameters have little effect on the results, hence we follow the settings as the previous works [5], [6]. In Table 3, we list all hyper-parameters both for our method and the baselines[6].

4.2 Overall Performance of Our Method

Baselines are evaluated on both datasets with Precision-Recall (PR) curve as shown in Fig. 6. In Fig. 6(a), our method draws the best PR curve, and the precision is higher than the other methods at nearly all range of the recall. Specifically, we quantify the results by the area of PR curve and P@100 for three kinds of test settings as shown in Table 4. The results indicate that our method is more effective for distantly supervised relation extraction than any other baselines.

TABLE 4
P@100s (%) of different settings and PR curve areas for all the baselines. BGRU+ATT is a composite method integrated [9] with [6].

P@100	One	Two	All	Mean	PR
PCNN [5]	73.3	70.3	72.3	72.0	0.33
PCNN+ATT [6]	73.3	77.2	76.2	75.6	0.35
PCNN+ATT+SL [32]	84.0	86.0	87.0	85.7	0.34
STPRE [7]	83.0	85.0	87.0	85.0	0.39
PCNN+ATT_RA+BAG_ATT [31]	**86.8**	91.2	91.8	89.9	0.42
BGRU+ATT	78.0	82.0	82.0	80.7	0.37
Our Method	86.0	**92.0**	**93.0**	**90.3**	**0.43**

Fig. 6(b) shows the PR curves of baselines on NYT-18. The figure shows that, 1) PCNN+ATT_RA+BAG_ATT[7] obtains the worse PR curve of all the baselines. It is a complicated model which is suitable for NYT-10 containing limited relation types. A large-scale relation types in NYT-18 tend to confuse the model in the intra-bag attention. 2) STPRE [7] also performs poorly on NYT-18. One possible reason is that the shorten sentences truncated with the fixed parser pattern lose too much information to distinguish a large number of relations. In contrast, our method works well by automatically selecting significant words according to their semantic meanings. 3) The precision of PCNN+ATT+SL [32] falls fast at high recall rate, because it highly depends on the performance of labels generator and tends to converge to local optimum. 4) The RNN based method (BGRU+ATT) works better than the CNN based ones at the larger dataset. 5) Our method achieves the best PR curve among all the baselines on NYT-18 by focusing on significant words and utilizing false instances sufficiently.

6. The parameters for the previous methods are following their papers.

7. We use the code released from the authors. https://github.com/ZhixiuYe/Intra-Bag-and-Inter-Bag-Attentions.

1041-4347 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on September 20,2020 at 14:48:13 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2020.2964747, IEEE Transactions on Knowledge and Data Engineering

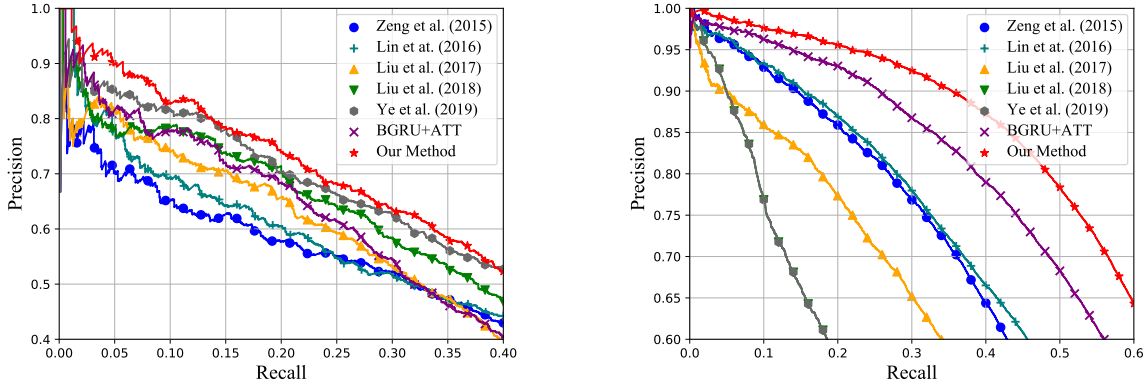


Fig. 6. PR curves of all the baselines for the dataset NYT-10 (a) and the dataset NYT-18 (b). (Better view in color.)

TABLE 5
P@Ns (%) and training resource for PCNN, BGRU and QARE. The time (in minutes) cost is for one iteration of all training sentences. The memory (in GB) here is GPU memory consumption in actual training process. The mark † indicates that our implementations with Tensorflow are slightly better than the methods in original papers.

Datasets	NYT-10					NYT-18				
Indicators	P@100	P@200	P@300	Time	Memory	P@10k	P@20k	P@30k	Time	Memory
PCNN+ONE†	78.0	72.0	67.7	16.2	4.6	81.0	63.5	51.4	75.6	5.3
+ATT†	77.0	73.0	70.3	15.8	4.6	82.2	65.6	53.2	78.3	5.3
BGRU+ONE	79.0	75.0	73.0	19.1	4.2	89.4	72.7	58.8	99.8	4.2
+ATT	83.0	77.5	**77.0**	19.5	4.2	88.1	72.4	58.9	85.7	4.2
Bi-LSTM+ONE	72.0	66.0	66.7	16.6	4.2	85.0	67.7	54.9	83.33	4.2
+ATT	73.0	67.0	63.7	15.8	4.2	88.7	73.0	59.7	102.16	4.2
QARE+ONE	83.0	78.5	75.3	**8.42**	**1.2**	**91.6**	74.9	60.0	49.7	**1.2**
+ATT	**90.0**	**80.5**	76.7	10.0	**1.2**	**91.6**	**75.7**	**61.4**	**44.2**	**1.2**

4.3 Effect of QARE

4.3.1 Performance of Top Predictions

We compare QARE with RNN/CNN relation extractors on the both datasets at P@Ns. As noted in Table 5, our QARE models achieve the best results on almost all the metrics. We select top 100, 200 and 300 predictions on NYT-10, while numbers for NYT-18 are 10, 20 and 30 thousand. From the table, we can see that BGRU based models obtain better results than PCNN based ones. BGRU and Bi-LSTM are both RNN models, which achieve similar results. QARE based models are much better than the other baselines at both NYT-10 and NYT-18, which means that modeling sentence with QARE is effective for relation extraction.

4.3.2 Performance of Complexity

Time and memory cost of training process are key indicators to quantify the efficiency of the baselines. Table 5 shows the observations that QARE is faster than CNN/RNN in the sense that it can save at least 71% memory over the RNN and CNN based approaches. In summary, QARE is not only effective for relation extraction but also efficient with the least resources.

4.3.3 Case Study

In Table 6, we present cases to show the quality of significant words. From the table we can see, 1) the selected significant words capture salient relation information, and the other words can be seen as noise. 2) the significant words are distributed discretely around a sentence, and it is hard to truncate a sub-sentence which contains all the significant words with fixed parser patterns.

TABLE 6
Cases study for selected significant words by QARE. Words in brackets are entities and the bold red parts are top words related to the relations.

He started all three games for the **U.S. at** the **1988 Olympics in** [Seoul], [South Korea], and was a member of the **U.S. team** that competed in the **1990** FIFA World Cup in **Italy**.
"Users will be able to integrate full video files in the coming **months**", **said** Mr. Mccann, **who** caught the video bug **after** a **conversation last year with** [Chad Hurley], **one of** [Youtube]'s **founders.**
The university of **Ibadan in southwest** [Nigeria], the **intellectual home of the Nobel prize-winning writer** [Wole Soyinka] , was **regarded** in **1960** as one of the **best** universities in the **British** Commonwealth.

4.4 Effect of MMIL

4.4.1 Performance of Top Predictions

We conduct experiments on four multi-instance learning algorithms which are max value (ONE), selective atten-

1041-4347 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on September 20,2020 at 14:48:13 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2020.2964747, IEEE Transactions on Knowledge and Data Engineering

TABLE 7
Performance of multi-instance learning algorithms in the metric of P@Ns under different test settings.

Test Settings	One				Two				All			
P@N (%)	100	200	300	Mean	100	200	300	Mean	100	200	300	Mean
PCNN+ONE†	72.0	68.0	59.3	66.4	79.0	69.0	63.7	70.6	77.0	71.0	66.0	71.3
PCNN+ATT†	83.0	70.0	62.3	71.8	81.0	72.5	65.0	72.8	83.0	77.0	69.0	76.3
PCNN+ATT_RA	77.0	72.0	63.7	70.9	78.0	74.5	70.7	74.4	81.0	76.0	71.0	76.0
PCNN+MMIL	82.0	71.0	63.0	72.0	84.0	78.0	70.3	77.4	86.0	76.5	72.0	78.1
BGRU+ONE	75.0	68.5	64.7	69.4	79.0	71.0	67.3	72.4	82.0	75.5	69.0	75.5
BGRU+ATT	78.0	70.5	61.0	69.8	82.0	75.0	66.7	74.6	82.0	78.5	74.3	78.3
BGRU+ATT_RA	77.0	68.5	65.3	70.3	77.0	73.0	68.0	72.7	82.0	76.0	70.6	76.2
BGRU+MMIL	78.0	75.0	67.0	73.3	87.0	75.0	69.7	77.2	86.0	76.5	72.7	78.4
QARE+ONE	86.0	72.5	67.0	75.2	85.0	74.5	68.0	75.8	85.0	75.5	70.0	76.8
QARE+ATT	84.0	75.0	67.3	75.4	86.0	77.0	70.0	77.7	86.0	81.0	73.7	80.2
QARE+ATT_RA	81.0	75.5	**72.3**	76.2	80.0	78.0	74.7	77.7	81.0	79.0	77.0	79.0
QARE+MMIL	**87.0**	**79.5**	71.3	**79.3**	**88.0**	**81.0**	**76.3**	**81.8**	**91.0**	**83.0**	**78.3**	**84.1**

TABLE 8
Performance of MMIL with or without false labeled instances in the metric of P@Ns under different test settings.

Test Settings	One				Two				All			
P@N (%)	100	200	300	Mean	100	200	300	Mean	100	200	300	Mean
PCNN+MMIL	**82.0**	71.0	63.0	72.0	**84.0**	**78.0**	**70.3**	**77.4**	**86.0**	76.5	**72.0**	**78.2**
-false	80.0	**71.5**	**64.7**	**72.1**	79.0	75.0	**70.3**	74.8	81.0	**78.5**	71.3	76.9
BGRU+MMIL	78.0	**75.0**	**67.0**	**73.3**	**87.0**	**75.0**	69.7	**77.2**	**86.0**	**76.5**	**72.7**	**78.4**
-false	**79.0**	71.5	65.3	71.9	80.0	73.5	**70.7**	74.7	82.0	74.5	71.0	75.8
QARE+MMIL	**87.0**	**79.5**	**71.3**	**79.3**	**89.0**	**81.0**	**76.3**	**81.8**	**91.0**	**83.0**	**78.3**	**84.1**
-false	84.0	77.5	69.7	77.1	85.0	80.0	71.3	78.8	86.0	79.0	72.0	79.0

tion (ATT), intra-bag attention (ATT_RA) and multi-focus (MMIL) on NYT-10. As shown in Table 7, we are aware that, 1) the selective attention is better than the max value for multiple instances by utilizing more information. 2) the intra-bag attention works as well as the selective attention. 3) MMIL is the best way to do multi-instance learning by using false instances properly. 4) MMIL can be well integrated with different relation extractors such as PCNN, BGRU and QARE.

4.4.2 Effect of False Labeled Instances

MMIL utilizes true and false labeled instances cooperatively. If the latter term in Equation (15) is dropped, we will lose rich information from false labeled instances. We conduct experiments to prove the effect of false labeled instances as shown in Table 8. The first four columns are alike because of the test setting **One** which is not much influenced by multi-instance learning. From the table, we can conclude that, 1) models with the false labeled instances achieve better performance on the test setting **Two** and **All**. 2) QARE is much better than PCNN and BGRU with or without false labeled instances.

4.4.3 Effect of Random Perturbations

We conduct experiments on NYT-10 to verify the effectiveness of random perturbations used in MMIL. Additionally, we prove that MMIL outperforms selective attention (ATT) integrating with random perturbations which can indicate that MMIL not only benefits from the perturbations. Table 9 shows that, 1) Random perturbations work better than no perturbations in MMIL. 2) MMIL based models outperform ATT based models even though they have integrated with random perturbations.

TABLE 9
P@Ns (%) for different perturbations. **RAN.** means random perturbations and **NO.** indicates no perturbations.

Methods	P@100	P@200	P@300	Mean
PCNN+ATT (**RAN.**)	80.0	77.0	71.3	76.1
PCNN+MMIL (**RAN.**)	**86.0**	76.5	**72.0**	**78.2**
PCNN+MMIL (**NO.**)	75.0	**79.0**	71.3	75.1
BGRU+ATT (**RAN.**)	**85.0**	75.0	72.0	77.3
BGRU+MMIL (**RAN.**)	83.0	**79.5**	**74.3**	**78.9**
BGRU+MMIL (**NO.**)	84.0	76.0	73.3	77.8
QARE+ATT (**RAN.**)	81.0	71.0	**77.3**	74.4
QARE+MMIL (**RAN.**)	**89.0**	**83.0**	75.0	**82.3**
QARE+MMIL (**NO.**)	84.0	79.5	74.0	79.2

4.4.4 Effect of the threshold

To verify our assumption in MMIL, we test various thresholds. Fig. 7 shows that the performance of two models will both change correspondingly to the increasing thresholds. $\eta = 0$ indicates that MMIL only chooses one true labeled instance, and more true labeled instances will be selected with a loose threshold. Compared with a proper threshold

1041-4347 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on September 20,2020 at 14:48:13 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2020.2964747, IEEE Transactions on Knowledge and Data Engineering

($\eta = 0.03$), tighter ones and looser ones both decline the performance of relation extraction. Tighter thresholds neglect true labeled instances containing rich relation information, and looser ones classify more false labeled instances to the true labeled set.

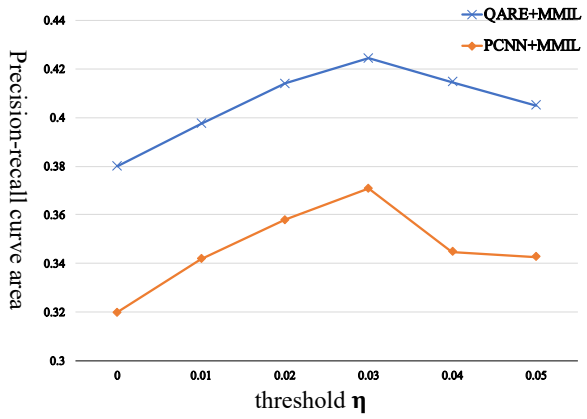


Fig. 7. The performance of different thresholds.

4.4.5 Case Study

We present realistic cases in the training set of NYT-10 to show the performance of MMIL compared with the other three multi-instance learning algorithms. The four algorithms are all implemented based on the same relation extractor. Table 10 shows three bag examples for instance selection. These bags have two correct sentences which are accord with the relation labels. The other false instances are assigned appropriate labels to compute the accurate regularization term. Therefore, we assign correct labels for false instances in training. It is clearly shown that MMIL can utilize false instances sufficiently.

4.5 Effect of the Parameter-Transfer Initializer

To evaluate the effect of the parameter-transfer initializer in our method, we implement two kinds of neural networks integrated with our Transfer Learning (TL) based initializer which are BGRU and QARE. BGRU+ATT and BGRU+ATT+TL integrate the selective attention mechanism for multi-instance learning, while QARE+MMIL and QARE+MMIL+TL use MMIL as multi-instance learning algorithm. We conduct experiments on both datasets shown as Fig. 8.

From the figures, we can conclude that, 1) Regardless of the neural networks that we use, methods with TL achieve better performance. It demonstrates that transfer learning helps neural relation extractors become more robust against noise. 2) QARE+MMIL+TL achieves the best performance and increases the area to 0.43 for the dataset NYT-10, while areas of BGRU+ATT, BGRU+ATT+TL and QARE+MMIL are 0.34, 0.37 and 0.41 respectively. It means that the TL based initializer works well with both kinds of neural networks and can resist noisy words further. 3)Transfer learning is effective either with selective attention or multi-focus multi-instance learning. 4) QARE based methods are better than BGRU based ones with or without transfer learning.

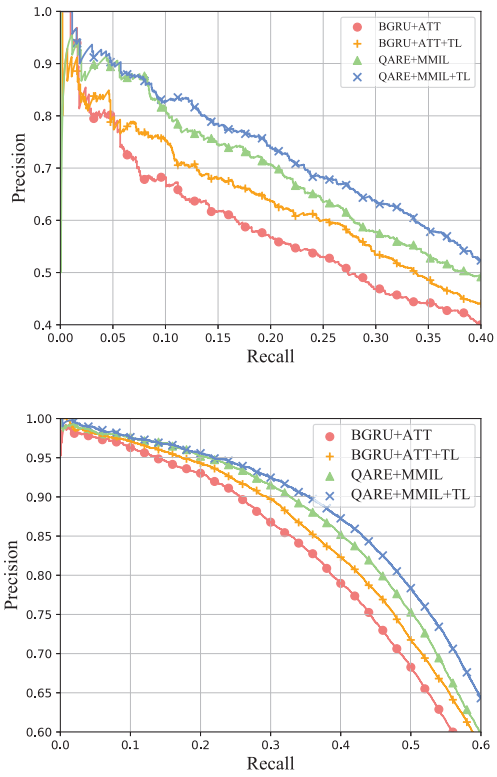


Fig. 8. PR curves for BGRU+ATT, BGRU+ATT+TL, QARE+MMIL and QARE+MMIL+TL for the dataset NYT-10 (top) and the dataset NYT-18 (bottom).

5 Discussion and Conclusion

5.1 Discussions

Utilization of Parse Information. Parse tree information over sentences has been applied to amounts of NLP applications successfully. It is a good way to truncate a sentence to sub-sentences which are shorter and easily processed for our target tasks [7], [8], [26], [27], [28]. However, the truncated path needs to be designed carefully, and it is hard to extend to large-scale relations when the relation features will be various and unpredictable. Our experimental results indicate that parse information is helpful to reduce word-level noise in a small benchmark as shown in Fig. 6(a). Whereas, the performance of sub-tree parser based extractor [7] falls far behind the other end-to-end methods on the large-scale dataset according to Fig. 6(b). Therefore, though parse information may be useful, we still need to explore a more effective way to apply it.

Efficiency of Neural Models. To our best knowledge, efficiency of neural relation extraction has not been discussed before. A few recent works design complicated models and spend amounts of computation to extract relations such as ensemble model [22] and reinforcement learning [33], [34]. Although all of these works achieve impressive results, they require too expensive computational cost to be accepted by most practical use. To reduce the threshold in the researches of large-scale relation extraction, improving the efficiency of neural models is a meaningful job which has been tried preliminarily with QARE in our work.

1041-4347 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on September 20,2020 at 14:48:13 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2020.2964747, IEEE Transactions on Knowledge and Data Engineering

TABLE 10
Cases study of selective sentences for the four multi-instance learning algorithms. The entities are labeled in the red brackets. "PN", "DL", "NA" and "LC" are all relation labels in the dataset, which are "person/nationality", "death/location", "non-relation" and "location/contain" respectively.

Labels	Sentence Bags	Weights			
		ONE	ATT	ATT_RA	MMIL
PN	**S1**: At first glance, [Álvaro García Linera] seems an unlikely vice president for the [Bolivia] of the moment.	0	0.998	0.369	1
	S2: Because of an editing error, a profile on Saturday about [Álvaro García Linera], a senior adviser to president Evo Morales of [Bolivia], misstated the month.	1	0.001	0.627	0 (NA)
	S3: Vice president [Álvaro García Linera] could not have been more explicit in a fiery speech last week calling on [Bolivia]'s indigenous groups to defend the government.	0	0.001	0.004	1
DL	**S1**: Radcliffe and [Buck O'Neil], a star player and manager with the [Kansas City] monarchs and now chairman of the Negro Leagues baseball museum in [Kansas City], were often honored as preeminent figures whose playing careers were solely in black baseball.	0	0.01	0.66	0 (NA)
	S2: [Buck O'Neil], a star first baseman and manager in the Negro Leagues, died Friday night in [Kansas City].	1	0.19	0.16	1
	S3: [Buck O'Neil], a star first baseman and manager in the Negro Leagues, died last night in [Kansas City].	0	0.80	0.17	1
NA	**S1**: A number of relief agencies who came to [Yogyakarta] more than a month ago to prepare for the eruption quickly diverted their aid to [Bantul], the district hardest hit by the quake.	0	0.005	0.87	1
	S2: A mass grave was dug in [Bantul] for unidentified people, said Sudibyo, a forensic doctor from [Yogyakarta] who uses only one name.	0	0.005	0.06	1
	S3: In the hardest hit part of the [Yogyakarta] area, [Bantul], mayor Idham Samawi said that rescuers had counted 2,200 dead and that many more people were alive but trapped under thousands of collapsed buildings.	1	0.99	0.07	0 (LC)

Flaws of Distant Supervision. The distant supervision for relation extraction is a brilliant way to generate large-scale labeled instances automatically. The distantly supervised relation extraction has drawn enough attention for nearly one decade. However, there is still a flaw coming from the incompleteness of knowledge bases. For example, over 70% of person names included in Freebase have no known place of birth [48]. Therefore, all the instances labeled as "no relation" perhaps contain a relation which is not recorded in knowledge bases. In addition, the imbalance of automatically built datasets may mislead relation extractor to neglect those rare relations. For instance, in NYT-10, there are only 0.4% instances expressing the relation *person/place_death*, while the rate of the relation *location/contains* is 13.3%. Although MMIL alleviates the influence of false labeled instances, we need more work to solve the problems of false negative instances and few-shot relations.

Future Work of Neural Relation Extraction. There are still remained future works to do for neural relation extraction. 1) In the current works of neural relation extraction, relations are independent and estimated by the softmax function, which neglects the relationships between relation types. For example, if we know [*Steven Jobs, born in, San Francisco*] and [*San Francisco, located in, United States*], we can infer [*Steven Jobs, nationality, United States*] since the relation types have a causal association. 2) Most of current relation extractors focus on the entities inside a sentence, which lack the ability to extract relations cross sentences. Actually, relation triples are not always expressed in one sentence. For instance, the relation triple [*Steven Wozniak, founder, Apple Inc.*] can be extracted from the two sentences *Steven Jobs is a founder of Apple Inc. So does Steve Wozniak.*

5.2 Conclusions

To our best knowledge, this paper for the first time proposes a robust and efficient neural relation extraction method, which aims at tackling the low-quality corpora by reducing both word-level and sentence-level noises and improving the robustness against these noises. Our model treats three levels of noises which come from word, sentence and knowledge type. For the word-level noise, it is important to query limited salient words which are corresponding to entities. Therefore, we initiate a QARE neural network to extract relation features in question-answering perspective, which enhances relation extraction on both accuracy and efficiency by focusing on significant words with a sententious structure. For the sentence-level noise, it is meaningful to utilize false labeled instances instead of ignoring them. The proposed MMIL works better than previous multi-instance learning algorithms by utilizing false instances sufficiently. Furthermore, lacking a priori knowledge hurts the performance of relation extraction, while parameter transfer learning can learn useful knowledge from other tasks. Parameter-Transfer Initializer makes our method more robust against noises by reasonable initialization of parameters. Extensive experiments show that the proposed solution outperforms previous state-of-the-art methods at a large margin.

ACKNOWLEDGMENTS

This work is supported by National China 973 Project No. 2015CB352401; Chinese National Research Fund (NSFC) Key Project No. 61532013 and No. 61872239. FDCT/0007/2018/A1, DCT-MoST Joint-project No. (025/2015/AMJ), University of Macau Grant Nos: MYRG2018-00237-RTO, CPG2018-00032-FST and SRG2018-00111-FST of SAR Macau, China. National Key Research and Development Program of China (No. 2017YFB0304100) and key projects of Natural Science Foundation of China (No. U1836222 and No. 61733011).

1041-4347 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on September 20,2020 at 14:48:13 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2020.2964747, IEEE Transactions on Knowledge and Data Engineering

REFERENCES

[1] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "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*. Singapore: Association for Computational Linguistics, 2009, pp. 1003–1011.

[2] S. Riedel, L. Yao, and A. McCallum, "Modeling relations and their mentions without labeled text," in *European Conference on Machine Learning and Knowledge Discovery in Databases*. Barcelona, Spain: Springer, 2010, pp. 148–163.

[3] R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer, and D. S. Weld, "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*. Portland, Oregon, USA: Association for Computational Linguistics, 2011, pp. 541–550.

[4] M. Surdeanu, J. Tibshirani, R. Nallapati, and C. D. Manning, "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*. Jeju, South Korea: Association for Computational Linguistics, 2012, pp. 455–465.

[5] D. Zeng, K. Liu, Y. Chen, and J. Zhao, "Distant supervision for relation extraction via piecewise convolutional neural networks," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, 2015, pp. 1753–1762.

[6] Y. Lin, S. Shen, Z. Liu, H. Luan, and M. Sun, "Neural relation extraction with selective attention over instances," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany: Association for Computational Linguistics, 2016, pp. 2124–2133.

[7] T. Liu, X. Zhang, W. Zhou, and W. Jia, "Neural relation extraction via inner-sentence noise reduction and transfer learning," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018.

[8] Y. Xu, L. Mou, G. Li, Y. Chen, H. Peng, and Z. Jin, "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*. Lisbon, Portugal: Association for Computational Linguistics, 2015, pp. 1785–1794.

[9] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu, "Attention-based bidirectional long short-term memory networks for relation classification," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany: Association for Computational Linguistics, 2016, pp. 207–212.

[10] W. Kumagai, "Learning bound for parameter transfer learning," in *Advances in Neural Information Processing Systems*, Barcelona, Spain, 2016, pp. 2721–2729.

[11] T. Miyato, A. M. Dai, and I. Goodfellow, "Adversarial training methods for semi-supervised text classification," in *Processings of the fourth International Conference for Learning Representations*, 2017.

[12] Y. Wu, D. Bamman, and S. Russell, "Adversarial training for relation extraction," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, 2017, pp. 1778–1783.

[13] L. Y. Pratt, "Discriminability-based transfer between neural networks," in *Advances in Neural Information Processing Systems*, Denver, Colorado, USA, 1993, pp. 204–211.

[14] R. Caruana, "Multitask learning," in *Learning to learn*. Springer, 1998, pp. 95–133.

[15] D. Bollegala, Y. Matsuo, and M. Ishizuka, "Minimally supervised novel relation extraction using a latent relational mapping," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 2, pp. 419–432, 2013.

[16] B. Wei, J. Liu, J. Ma, Q. Zheng, W. Zhang, and B. Feng, "Motif-based hyponym relation extraction from wikipedia hyperlinks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 10, pp. 2507–2519, 2014.

[17] Z. Li, M. A. Sharaf, L. Sitbon, X. Du, and X. Zhou, "Core: A context-aware relation extraction method for relation completion," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 4, pp. 836–849, 2014.

[18] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1746–1752.

[19] C. Liu, W. Sun, W. Chao, and W. Che, "Convolution neural network for relation extraction," in *International Conference on Advanced Data Mining and Applications*. Bangkok, Thailand: National Institute of Development Administration, 2013, pp. 231–242.

[20] D. Zeng, K. Liu, S. Lai, G. Zhou, J. Zhao *et al.*, "Relation classification via convolutional deep neural network," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, Dublin, Ireland, 2014, pp. 2335–2344.

[21] L. Wang, Z. Cao, G. D. Melo, and Z. Liu, "Relation classification via multi-level attention cnns," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany: Association for Computational Linguistics, 2016, pp. 1298–1307.

[22] D. Yang, S. Wang, and Z. Li, "Ensemble neural relation extraction with adaptive boosting," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. Stockholm, Sweden: International Joint Conferences on Artificial Intelligence Organization, 2018, pp. 4532–4538.

[23] S. Su, N. Jia, X. Cheng, S. Zhu, and R. Li, "Exploring encoder-decoder model for distant supervised relation extraction." in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*. Stockholm, Sweden: International Joint Conferences on Artificial Intelligence Organization, 2018, pp. 4389–4395.

[24] X. Zhang, P. Li, W. Jia, and H. Zhao, "Multi-labeled relation extraction with attentive capsule network," in *Proceedings of the Thirty-third AAAI Conference on Artificial Intelligence*. Honolulu, Hawaii, USA: AAAI Press, 2019, pp. 7484–7491.

[25] P. Li, X. Zhang, W. Jia, and H. Zhao, "Gan driven semi-distant supervision for relation extraction," 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)*. Minneapolis, USA: Association for Computational Linguistics, 2019, pp. 3026–3035.

[26] M.-C. De Marneffe and C. D. Manning, "The stanford typed dependencies representation," in *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*. Manchester, UK: Association for Computational Linguistics, 2008, pp. 1–8.

[27] D. Chen and C. Manning, "A fast and accurate dependency parser using neural networks," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar: Association for Computational Linguistics, 2014, pp. 740–750.

[28] M. Miwa and M. Bansal, "End-to-end relation extraction using LSTMs on sequences and tree structures," in *Proceedings of the 54nd Annual Meeting of the Association for Computational Linguistics*. Berlin, Germany: Association for Computational Linguistics, 2016, pp. 1105–1116.

[29] C. Yuan, H. Huang, C. Feng, X. Liu, and X. Wei, "Distant supervision for relation extraction with linear attenuation simulation and non-iid relevance embedding," *arXiv preprint arXiv:1812.09516*, 2018.

[30] G. Ji, K. Liu, S. He, and J. Zhao, "Distant supervision for relation extraction with sentence-level attention and entity descriptions," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. San Francisco, California, USA: AAAI Press, 2017, pp. 3060–3066.

[31] Z.-X. Ye and Z.-H. Ling, "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*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 2810–2819.

[32] T. Liu, K. Wang, B. Chang, and Z. Sui, "A soft-label method for noise-tolerant distantly supervised relation extraction," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, 2017, pp. 1790–1795.

[33] J. Feng, M. Huang, L. Zhao, Y. Yang, and X. Zhu, "Reinforcement learning for relation classification from noisy data," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*. New Orleans, Louisiana, USA: AAAI Press, 2018.

[34] P. Qin, W. Xu, and W. Y. Wang, "Robust distant supervision relation extraction via deep reinforcement learning," in *Proceedings*

1041-4347 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on September 20,2020 at 14:48:13 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2020.2964747, IEEE Transactions on Knowledge and Data Engineering

of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 2137–2147.

[35] P. Qin, W. XU, and W. Y. Wang, "Dsgan: Generative adversarial training for distant supervision relation extraction," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2018, pp. 496–505.

[36] P. Liu, X. Qiu, and X. Huang, "Recurrent neural network for text classification with multi-task learning," in *Proceedings of the 25th International Joint Conference on Artificial Intelligence*. New York, USA: International Joint Conferences on Artificial Intelligence Organization, 2016, pp. 2873–2879.

[37] Y. Lin, Z. Liu, and M. Sun, "Neural relation extraction with multi-lingual attention," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 34–43.

[38] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[39] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Advances in neural information processing systems*, Lake Tahoe, USA, 2013, pp. 2787–2795.

[40] F. Huang, A. Ahuja, D. Downey, Y. Yang, Y. Guo, and A. Yates, "Learning representations for weakly supervised natural language processing tasks," *Computational Linguistics*, vol. 40, no. 1, pp. 85–120, 2014.

[41] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013. [Online]. Available: https://arxiv.org/pdf/1301.3781

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, Long Beach, California, USA, 2017, pp. 6000–6010.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Las Vegas, Nevada,USA, 2016, pp. 770–778.

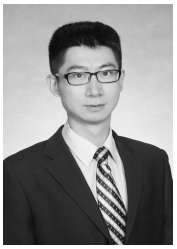
[44] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1724–1734.

[45] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proceedings of the 43rd annual meeting on association for computational linguistics*. Ann Arbor, Michigan, USA: Association for Computational Linguistics, 2005, pp. 363–370.

[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[47] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[48] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang, "Knowledge vault: A web-scale approach to probabilistic knowledge fusion," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York City, USA: ACM, 2014, pp. 601–610.



Xinsong Zhang received the BS, MS, PhD degrees in School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, China, in 2011, 2014 and 2019 respectively. He is currently working as a research fellow in ByteDance AI Lab, China. His research interests include relation extraction, question answering, natural language processing and knowledge engineering.

Tianyi Liu received the BS degree in the school of Information and Communication Engineering, University of Electronic Science and Technology of China in 2017. He is currently working toward the Ph.D. degree in computer science at Shanghai Jiao Tong University, China. His research interests include relation extraction, natural language processing and knowledge engineering.

Pengshuai Li received the BS degrees in School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, China, in 2014. He is currently working toward the MS degree in computer science at Shanghai Jiao Tong University, China. His research interests include distantly supervised relation extraction and natural language processing.

Weijia Jia is currently a Chair Professor at University of Macau while taking no-pay leave from the position of Zhiyuan Chair Prof from Shanghai Jiao Tong University, China (he received 2013 China 1000 Talent Award). He received BSc/MSc from Center South University, China in 82/84 and Master of Applied Sci./PhD from Polytechnic Faculty of Mons, Belgium in 92/93, respectively, all in computer science. For 93-95, he joined German National Research Center for Information Science (GMD) in Bonn (St. Augustine) as research fellow. From 95-13, he worked in City University of Hong Kong as a full professor in Computer Science Dept. His research interests include smart city; next generation IoT, knowledge graph constructions; multicast and any-cast QoS routing protocols, wireless sensor networks and distributed systems. In these fields, he has over 400 publications in the prestige international journals/conferences and research books and book chapters. He has served as area editor for various prestige international journals, chair and PC member/keynote speaker for many top international conferences. He is the Senior Member of IEEE and the Member of ACM.

Hai Zhao received the BEng degree in sensor and instrument engineering, and the MPhil degree in control theory and engineering from Yanshan University in 1999 and 2000, respectively, and the PhD degree in computer science from Shanghai Jiao Tong University, China in 2005. He is currently a full professor at department of computer science and engineering, Shanghai Jiao Tong University after he joined the university in 2009. He was a research fellow at the City University of Hong Kong from 2006 to 2009, a visiting scholar in Microsoft Research Asia in 2011, a visiting expert in NICT, Japan in 2012. He is an ACM professional member, and served as area co-chair in ACL 2017 on Tagging, Chunking, Syntax and Parsing, senior area chair in ACL 2018 on Phonology, Morphology and Word Segmentation. His research interests include natural language processing and related machine learning, data mining and artificial intelligence.

1041-4347 (c) 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on September 20,2020 at 14:48:13 UTC from IEEE Xplore. Restrictions apply.