Knowledge-Based Systems 235 (2022) 107637

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Distantly Supervised Relation Extraction using Global Hierarchy Embeddings and Local Probability Constraints

Tao Peng [a,b,c], Ridong Han [a,c], Hai Cui [a,c], Lin Yue [d,*], Jiayu Han [e], Lu Liu [a,b,c,*]

[a] College of Computer Science and Technology, Jilin University, Changchun, Jilin 130012, China
[b] College of Software, Jilin University, Changchun, Jilin 130012, China
[c] Key Laboratory of Symbol Computation and Knowledge Engineer of the Ministry of Education, Changchun, Jilin 130012, China
[d] School of Information Technology and Electrical Engineering, University of Queensland, Brisbane, QLD 4072, Australia
[e] Department of Linguistics, University of Washington, Seattle, WA 98195, United States

ARTICLE INFO

Article history:
Received 17 June 2021
Received in revised form 26 September 2021
Accepted 21 October 2021
Available online 25 October 2021

MSC:
00-01
99-00

Keywords:
Distant Supervision
Relation Extraction
Relation Hierarchies
Global Hierarchy Embedding
Local Probability Constraint

ABSTRACT

To find relational facts of interest from plain texts, distantly supervised relation extraction (DSRE) has drawn significant attention. Recent works exploit relation hierarchies to mine more clues for long-tail relations and achieve good performance. However, they ignore or underutilize the correlation of relations in the hierarchical structure. Empirically, the correlation facilitates knowledge transfer between different relations to further handle long-tail relations and improves inter-relational discrimination. In this paper, we devise an end-to-end network to model the correlation of relations from two perspectives. Globally, we construct an undirected connected graph according to the relation hierarchies, and employ Graph Attention Networks (GATs) to aggregate node information and generate correlation-aware Global Hierarchy Embeddings (GHE). Locally, we assume that *along the relation hierarchies, the classification results of adjacent levels should be highly interdependent*, and introduce a constraint called Local Probability Constraints (LPC) to take it into account. LPC is then combined with a branch network for both sentence-level and bag-level classification. Experimental results on the popular *New York Times* (NYT) dataset show that, our model GHE-LPC outperforms other state-of-the-art baselines in terms of AUC, Top-N precision, accuracy of Hits@K, etc.

1. Introduction

Various large-scale knowledge bases (KBs) such as YAGO [1], Freebase [2] and DBpedia [3] have been proven to play an important role in many natural language processing (NLP) tasks, yet current KBs are still far from complete compared with real-word facts. In this case, learning to extract facts of interest from plain texts (i.e., relation extraction, RE) is a very important task. Recently, supervised methods are widely used to solve it due to their relatively high performance. Such methods, however, always require large-scale training data which is time-consuming and laborious to obtain. One common technique for coping with this difficulty is distant supervision (DS) [4] which generates training data via aligning KBs and massive plain texts. It assumes that *if two entities have a relation in KBs, then all sentences that mention these two entities will be labeled as training sentences for this relation.*

In distant supervision scenario, two problems have to be addressed. The main one is wrong labeling problem caused by its strong assumption. For example, <**Google**, **Sergey Brin**>expresses the */business/company/founders* relation in Freebase. So, the sentence "*Sergey Brin, Google's president for technology, said the rate of hiring had slowed.*" will be incorrectly labeled as a training instance. The other one is long-tail problem. Training data generated by DS can only cover a limited part of real-word relations. For example, if we treat relations with training instances less than 1000 as long-tail relations, over 70% of the relations in NYT dataset are long-tail and still suffer from data deficiency. Over the past few years, to alleviate the effects of mislabeled sentences, Riedel et al. [5] and Hoffmann et al. [6] develop the multi-instance learning (MIL) framework, which identifies a label between two entities for a bag of sentences. Following the MIL framework, many efforts have been devoted to alleviate the impact of noisy supervised signals during the training phase, including attention mechanism [7–9], soft-labeling [10], reinforcement learning [11,12], etc. For the long-tail problem, a natural idea is to exploit the relation hierarchies to transfer knowledge between data-rich relations and long-tail ones [13–16]. Take the relation */people/family/members* in Freebase as an example, it has two ancestor relations, i.e., */people* and */people/family*. Like this, all relations form a tree-like hierarchical structure, i.e., relation hierarchies.

* Corresponding authors.
E-mail addresses: l.yue@uq.edu.au (L. Yue), liulu@jlu.edu.cn (L. Liu).

https://doi.org/10.1016/j.knosys.2021.107637
0950-7051/

However, the above methods ignore or underutilize the correlation between relations, which is defined as the "relation of relations (ROR)" phenomenon by Jin et al. [17]. Some of them assume that relations are discrete and independent of each other [7–12]. Others simply define relation embeddings for different levels of relation hierarchies as query vectors to augment the sentence representations or to obtain multi-level representations for sentence bags [13–16], in which relation embeddings are randomly initialized and the different levels do not affect each other in the calculation. Here the correlation of relations refers to two aspects: (1) *the interdependence between different levels of relation hierarchies* and (2) *the mutual heuristic effect between sibling relations within the same levels*. Empirically, modeling the correlation facilitates knowledge transfer between different relations to further alleviate the long-tail problem and improves inter-relational discrimination. In this paper, we fully exploit the correlation of relations with two strategies. In the global view, we treat the relation hierarchies as an undirected connected graph, and employ Graph Attention Networks (GATs) to aggregate node information. Specifically, each relation of relation hierarchies constitutes a node of the input graph. During the calculation, each relation receives information from itself, its father relation and its children relations. Then correlation-aware relation embeddings, called Global Hierarchy Embeddings (GHE) are generated. In the local view, we introduce a constraint called Local Probability Constraints (LPC). It assumes that *along the relation hierarchies, the classification results of adjacent levels should be highly interdependent.* Specifically, along the relation hierarchies, we first use the current level's classification probability to construct the expected classification probability for other adjacent levels. Then, for each pair of adjacent levels, we compute the similarity of classification probabilities to take the interactions between adjacent levels into account. Finally, LPC is combined with a branch network for sentence-level and bag-level classification. Unlike existing works that utilizes pre-trained relation embeddings, our model is an end-to-end network. The main contributions of this paper are summarized as follows:

- To the best of our knowledge, our model is the first approach to explicitly model "the correlation of relations" on DSRE task, which fully leverages the relation hierarchies and addresses both wrong labeling problem and long-tail problem in distant supervision scenario.
- We model the correlation of relations from two perspectives. Globally, we construct an undirected connected graph according to the relation hierarchies, and employ Graph Attention Networks as Hierarchy Structure Encoder to aggregate relation information in order to obtain the global relation embeddings.
- Locally, we introduce a constraint called Local Probability Constraints (LPC), which is combined with a branch network for hierarchical classification. LPC aims to take the local correlation of relations into account by modeling the interdependencies between the classification probabilities of adjacent levels.
- Substantial experiments on the popular *New York Times* (NYT) dataset are conducted. Our method achieves state-of-the-art performance in terms of multiple metrics. The source code of this work will be released at https://github.com/RidongHan/GHE-LPC.

2. Related work

Distantly supervised relation extraction (DSRE). Although distant supervision (DS) can generate large-scale training data and address the drawback of supervised methods that rely on large amounts of training instances, it may bring some mislabeled sentences. To alleviate this, Riedel et al. [5], Hoffmann et al. [6] and Surdeanu et al. [18] relax the assumption behind DS and remodel DSRE by multi-instance learning (MIL). Following the MIL setting, Zeng et al. [19] design the piecewise convolutional neural networks (PCNNs) and select the sentence that is most likely to correctly express the specified relation during the training phase. Lin et al. [7] utilize attention mechanism to obtain the bags' representations by giving sentences different weights. Inspired by the above methods, many subsequent efforts have been devoted to alleviate the impact of noisy supervised signals, including word attention [8], multi-level structured self-attention [9], bag-level attention [20,21], feature-level attention [22], segment attention [23], etc. To fully exploit entity information, multi-feature fusion with entity sense [24] and dynamic adjustment of parameters according to entity types [25] are good ideas. For relation information, semantic scenarios are also very vital [26]. Besides, reinforcement learning-based approaches to select correctly labeled sentences from noisy sentence bags also achieve good performance [11,12].

Hierarchical relation extraction. Hierarchical relation extraction focuses on exploiting relation hierarchies for knowledge transfer between data-rich relations and long-tail ones to handle long-tail relations. Han et al. [13] use coarse-to-fine grained relation embeddings as queries to perform a hierarchical attention along the relation hierarchies. Then Zhang et al. [14] enhance the above multi-granular relation embeddings by merging the embeddings from both pre-trained TransE [27] and graph convolutional networks [28]. Most recently, Li et al. [15] design a collaborating relation-augmented attention network to augment sentence representations, while Yu et al. [16] adopt a top-down classification strategy along the hierarchical relation chains.

Graph representation learning. Graph representation learning is an important research topic because graph data is widely available in the real world and many current tasks involve the processing of graph data. Specific types of graphs include social networks [29], knowledge graphs [30], protein–protein interaction networks [31], and so on. In this paper, relation hierarchies can also be seen as a graph. To obtain the representations of nodes, graph neural networks (GNNs) are widely used, such as, graph convolutional networks (GCNs) [32] and graph attention networks (GATs) [33]. For more complex graphs, Chen et al. [34] focus on solving the "oversmoothing" problem in attributed network representation learning, while Li et al. [35] model the coupling and interaction phenomena. More information can be found in [36].

3. Our proposed approach

3.1. Task definition

Following the MIL setting, all sentences can be split into multiple bags, i.e., $\{B_1, B_2, \ldots\}$, according to the common entity-pairs. Each bag B_i contains some sentences $\{s_1, s_2, \ldots, s_m\}$ mentioning the same entity pair $\langle h_i, t_i \rangle$. Each sentence is a sequence of tokens of length n obtained by truncating or padding, i.e., $s_j = [w_1, w_2, \ldots, w_n]$. Based on the above definitions, distantly supervised relation extraction aims to select a relation label for a sentence bag from the pre-defined relation set $\mathcal{R} = \{r_1, r_2, \ldots\}$.

3.2. Model architecture

As shown in Fig. 1, our model includes four components: (1) The Entity-Aware Embedding module [37] for highlighting the essence of entities by merging entity embeddings and position

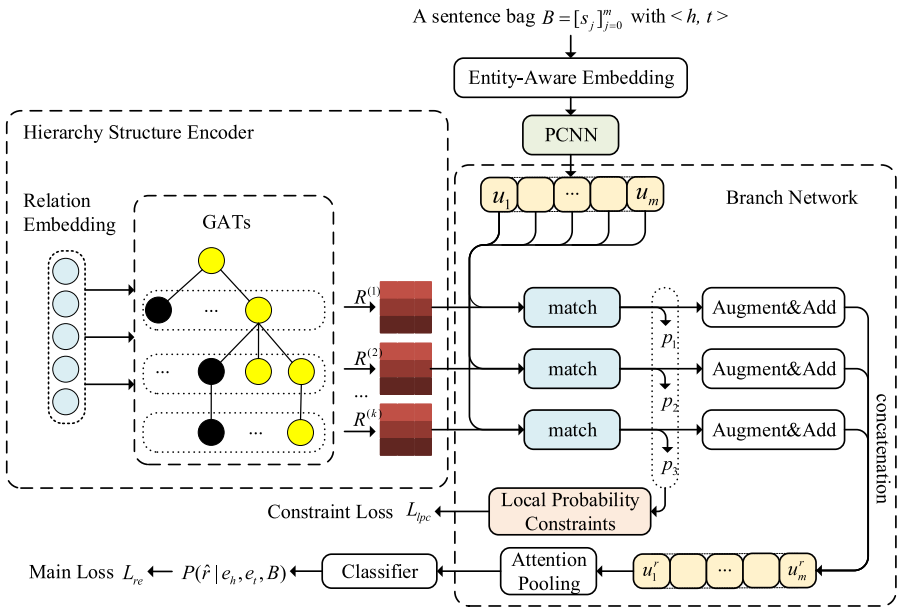


Fig. 1. The overview of our proposed RE model, GHE-LPC.

embeddings into word embeddings. (2) A sentence encoder based on piecewise convolutional neural networks (PCNNs) [19] for generating sentence representations. (3) The Hierarchy Structure Encoder for obtaining correlation-aware relation embeddings, i.e., Global Hierarchy Embeddings (GHE). (4) A Branch Network with Local Probability Constraints for sentence-level and bag-level classification. It generates more valuable bag representations and meanwhile captures local correlation of relations to alleviate long-tail problem.

3.3. Entity-aware embedding

Following Li et al. [37], to yield more expressively-powerful representations for downstream modules, we integrate word embeddings [38], position embeddings [39] and entity embeddings [37]. The integration has been proven useful and powerful. The details are as follows.

Given a bag of sentences $B = \{s_1, s_2, \ldots, s_m\}$, for $j \in [1, 2, \ldots, m]$, each sentence $s_j = [w_1, w_2, \ldots, w_n]$, can be converted into low-dimensional, real-valued vector embeddings using a pre-trained word2vec model [38], i.e., $V = [v_1, v_2, \ldots, v_n] \in \mathbb{R}^{d_w \times n}$, where d_w denotes the dimension of word embedding.

Relative position information is introduced to RE task by Zeng et al. [39]. This feature can be modeled by the relative distances from the current word to head entity h and tail entity t. For instance, in the sentence "***Alberto Lattuada*** *was born in* ***Milan*** *in 1914.*", the relative distance from *born* to entity h (*Alberto Lattuada*) and entity t (*Milan*) are 2 and -2, respectively. Then, two distances are transformed into low-dimensional vectors, x_i^{ph} and $x_i^{pt} \in \mathbb{R}^{d_p}$, where d_p is the dimension of position embedding. Consequently, position-aware embeddings can be denoted as $F^{(p)} = [x_1^p, x_2^p, \ldots, x_n^p] \in \mathbb{R}^{(d_w+2d_p)\times n}$, where $x_i^p = [v_i; x_i^{ph}; x_i^{pt}]$, $i \in [1, 2, \ldots, n]$, ";" denotes the operation of vector concatenation.

Besides, to highlight the essence of entities for RE task, the sequence of entity embeddings can be defined as $F^{(e)} = [x_1^e, x_2^e, \ldots, x_n^e] \in \mathbb{R}^{3d_w \times n}$, where $x_i^e = [v_i; v_h; v_t]$, $i \in [1, 2, \ldots, n]$, v_h and v_t are head and tail entity embeddings. Finally, to integrate these three features, a position-wise gate is employed [37], i.e.,

$$\alpha = sigmoid(\lambda \cdot (W^{(e)}F^{(e)} + b^{(e)})), \tag{1}$$

$$\tilde{F}^{(p)} = tanh(W^{(p)}F^{(p)} + b^{(p)}), \tag{2}$$

$$X = \alpha \circ F^{(e)} + (1 - \alpha) \circ \tilde{F}^{(p)}, \tag{3}$$

in which $W^{(e)} \in \mathbb{R}^{d_x \times 3d_w}$, $W^{(p)} \in \mathbb{R}^{d_x \times (d_w+2d_p)}$, "$\circ$" denotes Hadamard Product and λ is a hyper-parameter to highlight the importance of entities. And $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d_x \times n}$ is the resulting input representations specially for sentence s_j.

3.4. Piecewise convolutional neural networks

The piecewise convolutional neural networks (PCNNs) are first proposed by Zeng et al. [19]. Then PCNNs become the most commonly used sentence encoder in the RE task. Specifically, it firstly encodes the input representation X using the convolution operation with window size ω and generates the feature representation f, where $f \in \mathbb{R}^{d_c \times n}$ and d_c is the number of feature maps. Then, according to the position of head and tail entities, the feature f is divided into three segments $\{f^{(1)}, f^{(2)}, f^{(3)}\}$. The max-pooling procedure finally is performed in each segment separately to obtain the final sentence representation u, i.e.,

$$f = \text{1D\_CNN}(X), \tag{4}$$

$$u = [max(f^{(1)}); max(f^{(2)}); max(f^{(3)})], \tag{5}$$

where $u \in \mathbb{R}^{d_f}$, $d_f = 3d_c$.

3.5. Hierarchy structure encoder

To capture the correlation of relations from a global perspective, we construct an undirected connected graph according to the relation hierarchies and employ graph attention networks (GATs) [33] as hierarchy structure encoder to aggregate relation information on the graph. The output of GATs can be denoted as correlation-aware relation embeddings (i.e., Global Hierarchy Embeddings, GHE). Next we first introduce the relation hierarchies graph, and then the details of GATs.

3.5.1. Relation hierarchies graph

For a relation $r \in \mathcal{R}$, we can generate its hierarchical chain of parent relations $\{r^0, r^1, \ldots, r^k\}$, where r^0 denotes the *root* relation node and r^k is identical to r. The smaller the index, the higher the relation level. Due to the existence of the *root*

node, all chains form tree-like relation hierarchies. Then each relation is treated as one node, we can quickly construct the relation hierarchies graph, i.e., an undirected connected graph $\mathcal{G}$. The number of nodes is denoted as z.

3.5.2. Graph attention networks (GATs)

To process the data represented in graph domains, Kipf and Welling [32] present graph convolutional networks (GCNs) to gather information from the neighbor nodes. However, all neighbor nodes are treated equally during the calculation. To address this drawback, Veličković et al. [33] propose graph attention networks (GATs), which learn to assign different degrees of importance to neighbor nodes. Generally, GATs consist of some stacked graph attentional layers. Each layer, specifically, takes the feature representations of all nodes in the graph as input, which can be denoted as $L = [l_1, l_2, \ldots, l_z], l_i \in \mathbb{R}^{d_g}, i \in [1, 2, \ldots, z]$, and outputs transformed embeddings of all nodes. Note that the feature representations of all nodes (i.e., relation embeddings) are initialized randomly at first. The process of graph attentional layer can be described as follows.

$$\alpha_{ij} = softmax(ATT(Wl_i, Wl_j)), \tag{6}$$

where $i/j \in [1, 2, \ldots, z]$, α_{ij} is the attention score of node l_i for l_j, $W \in \mathbb{R}^{d_g \times d'_g}$ is a learnable weight matrix to increase nonlinearity by transforming the features to a higher dimensional space, and $ATT(\cdot)$ denotes any kind of attention function. In this paper, we define the $ATT(\cdot)$ function as $LeakyReLu(W_{att}^T[Wl_i; Wl_j] + b_{att})$, where $LeakyReLu(\cdot)$ is the activation function and $W_{att} \in \mathbb{R}^{2d'_g}$. See Section 4.5 for more details on $ATT(\cdot)$. Then the output features are defined as the weighted summation of transformed input features, i.e.,

$$L' = [l'_1, l'_2, \ldots, l'_z], \tag{7}$$

$$l'_i = \sigma\left(\sum_{j \in N_i} \alpha_{ij} Wl_j\right), i \in [1, 2, \ldots, z], \tag{8}$$

where N_i consists of all neighbor nodes that have edges with l_i.

The above learning process may be unstable, to address this drawback, multi-head attention can be leveraged as Vaswani et al. [40]. When using multi-head self-attention, the resulting output feature representations are defined as one of the following equations,

$$l'_i = \Big|\Big|_{h=1}^{H} \sigma\left(\sum_{j \in N_i} \alpha_{ij} Wl_j\right) \text{ or } l'_i = \sigma\left(\frac{1}{H}\sum_{h=1}^{H}\sum_{j \in N_i} \alpha_{ij} Wl_j\right), \tag{9}$$

where $\Big|\Big|$ denotes vector concatenation operation and H is the number of heads. In this paper, the second way is adopted.

After using GATs to obtain the Global Hierarchy Embeddings (GHE), we put together the embeddings of the same level in the relation hierarchies, and define global relation embedding matrix for each level, i.e., $R^{(i)} \in \mathbb{R}^{d'_g \times N_R^i}, i \in [1, 2, \ldots, k]$, where N_R^i denotes the number of distinct relations of the ith level.

3.6. Branch network with local probability constraints

For a bag of sentences $B = \{s_1, s_2, \ldots, s_m\}$, we already obtain the sentence representations $U = \{u_1, u_2, \ldots, u_m\}$ through the PCNN encoder. To further exploit the correlation of relations from a local perspective, a novel constraint called Local Probability Constraints is proposed. It is combined with a branch network for classification. Next We first describe our base branch network, and then introduce Local Probability Constraints.

3.6.1. Base branch network

In this section, we choose collaborating relation-augmented attention network (CoRA) [15] as the base branch network. It takes both sentence-level and bag-level supervised signals into consideration.

On the one hand, for a sentence, it predicts the relation label for each level in relation hierarchies. Specifically, each sentence representation $u \in U$[1] matches with each level's global relation embedding matrix described in to obtain the matching degree vector, i.e.,

$$\alpha^{(i)} = softmax(u^T R^{(i)}), i \in [1, 2, \ldots, k], \tag{10}$$

where $softmax(\cdot)$ denotes a normalization function along last dimension.

On the other hand, the base branch network identifies a relation label for a bag of sentences. In details, firstly, the relation embeddings are used to obtain relation-aware information by dot product, i.e.,

$$c^{(i)} = R^{(i)}\alpha^{(i)}, i \in [1, 2, \ldots, k], \tag{11}$$

where $c^{(i)}$ is the relation-aware information.

After that, this information is leveraged to augment the sentence representation. Specifically, for $i \in [1, 2, \ldots, k]$, we merge $c^{(i)}$ into u by an element-wise gate with residual connection [41] and layer normalization [42] to generate the ith level's augmented representation $u^{(i)}$,

$$\beta_g^{(i)} = sigmoid(W_g^T[u; c^{(i)}] + b_g), \tag{12}$$

$$\hat{u}^{(i)} = \beta_g^{(i)} \circ u + (1 - \beta_g^{(i)}) \circ W_c^T c^{(i)}, \tag{13}$$

$$u^{(i)} = LayerNorm(u + MLP(\hat{u}^{(i)})), \tag{14}$$

where $W_g \in \mathbb{R}^{(d_f + d'_g) \times d_f}$, $W_c \in \mathbb{R}^{d'_g \times d_f}$, $MLP(\cdot)$ denotes a multilayer perceptron to increase nonlinearity.

Then all levels' augmented representations $\{u^{(1)}, u^{(2)}, \ldots, u^{(k)}\}$ are concatenated as the resulting relation-augmented sentence representation u^r corresponding to u, i.e., $u^r = [u^{(1)}; u^{(2)}; \ldots; u^{(k)}]$. All relation-augmented sentence representations of the bag B are represented as $B^r = [u_1^r, u_2^r, \ldots, u_m^r] \in \mathbb{R}^{kd_f \times m}$.

Next moving to wrong labeling problem, the attention-pooling, a kind of self-attention [43,44], is leveraged to derive an accurate bag-level representation. It learns to assign an importance score to each sentence according to its augmented representations u^r and performs a weighted sum over all relation-augmented sentence representations of a bag,

$$b = B^r softmax(W_{att}^T B^r) \in \mathbb{R}^{kd_f}, \tag{15}$$

where $W_{att} \in \mathbb{R}^{kd_f}$ is a learnable weight matrix.

Finally, a softmax classifier is employed. It takes the bag representation b as input, and calculates the confidence score of each relation label,

$$o_b = P(\hat{r}|e_h, e_t, B) = softmax(MLP(b)), \tag{16}$$

where $o_b \in \mathbb{R}^{|\mathcal{R}|}$, $|\mathcal{R}|$ denotes the number of pre-defined relations.

3.6.2. Local probability constraints

During the above calculation, each level will have a classification probability, i.e. $\alpha^{(i)}$. We assume that *along the relation hierarchies, the classification results of adjacent levels should be highly interdependent*, and devise a novel training constraint, called Local Probability Constraints (LPC). The illustration of LPC is shown in Fig. 2.

In details, for ith level, $i \in [2, \ldots, k]$, the predicted probability vector can be denoted as $\alpha^{(i)} = [\alpha_1^{(i)}, \alpha_2^{(i)}, \ldots, \alpha_{N_R^i}^{(i)}]$. Based on

[1] For a clear demonstration, we omit indices in the following instructions.

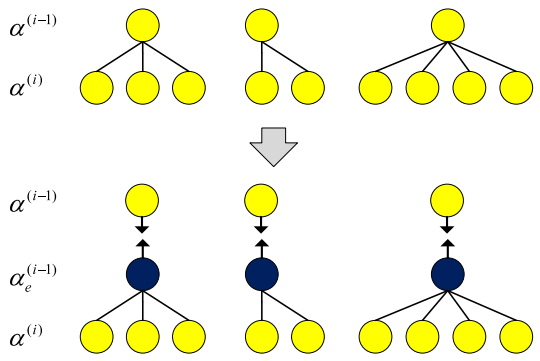


Fig. 2. The illustration of LPC.

$\alpha^{(i)}$ and the relation hierarchies, we can construct the expected probability distribution $\alpha_e^{(i-1)}$ for the previous level. Specifically, along the relation hierarchies, if category c has n_c sub-categories, we can use the probability sum of these n_c sub-categories as the probability of category c. In this way, the expected distribution $\alpha_e^{(i-1)} = [\alpha_1^{(i-1)}, \alpha_2^{(i-1)}, \ldots, \alpha_{N_R^{i-1}}^{(i-1)}]$ can be generated. The objective function is defined as follows,

$$L_{lpc} = -\frac{1}{|D| \times |B| \times (k-1)} \sum_{B \in D} \sum_{s \in B} \sum_{l=2}^{k} KL(\alpha^{(i)}, \alpha_e^{(i-1)}), \tag{17}$$

where D is the training set consisting of sentence bags and $KL(\cdot)$ denotes the Kullback–Leibler divergence.

3.7. Training objectives

The main objective for DSRE is defined to minimize a cross-entropy loss on the bag-level's prediction, i.e.,

$$L_{re} = -\frac{1}{|D|} \sum_{B \in D} \log P(\hat{r}|e_h, e_t, B). \tag{18}$$

Besides, sentence-level supervised signals have been proven to be very helpful for DSRE [15]. An auxiliary loss is introduced to leverage these signals and guide our model to augment each sentence with correct relation embeddings. That is,

$$L_{hier} = -\frac{1}{|D| \times |B| \times k} \sum_{B \in D} \sum_{s \in B} \sum_{l=1}^{k} \log \alpha_{[r^l]}^{(l)}, \tag{19}$$

where $[\cdot]$ denotes taking the value according to the index. Finally, the overall loss function can be defined as:

$$L = L_{re} + \mu L_{hier} + \xi L_{lpc} + \delta \|\theta\|_2^2, \tag{20}$$

where μ, ξ and δ are trade-off parameters. $\|\theta\|_2^2$ denotes the regularizer defined as L_2 normalization.

4. Experiments and results

We choose the widely used dataset, *New York Times* (NYT) to conduct our experiments. We choose it for the following reasons: (1) It has been used in almost all previous works [7,13,15,16,37] for DSRE task and is the only widely used DSRE dataset. (2) The wrong labeling problem and long-tail problem are extremely serious, which makes it suitable for our evaluation. (3) For other datasets generated by DS, such as GIDS [45], Wiki-KBP [46] and NYT-H [47], the GIDS dataset has no long-tail phenomenon; the Wiki-KBP dataset is specific to the joint extraction task and is not fully suitable for the task of this paper; the NYT-H dataset is simply subset variants of NYT, and the results of all approaches on it are all very high and not comparable due to its extremely small size of the test data. (4) Although our proposed approach should intuitively work on datasets containing fewer long-tail relations as long as the relation hierarchies exist, some datasets do not have taxonomic hierarchies, which is the case of the Wiki-KBP dataset above. In summary, the NYT dataset is the most suitable.

The NYT dataset [5][2] is generated by aligning the corpus of *New York Times* with Freebase [2] and has been used by Lin et al. [7] and Li et al. [15][3]. The sentences from the years 2005–2006 are used as train set, the rest sentences from the year 2007 are used as test set. In the pre-processing phase, the train set is firstly sorted by fact triples $\langle e_h, r, e_t \rangle$, and then the sentences with the same fact form a sentence-bag. While the test set is sorted by entity pairs $\langle e_h, e_t \rangle$ firstly, and then a sentence-bag consists of all sentences with the same pair. The detailed statistics of NYT are as follows ("#" indicates the number of items): *#sentence* and *#entity_pair* of train set are 570 088 and 293 162, respectively, while for the test set, the values are 172 448 and 96 678.

Besides, We exploit the held-out evaluation to evaluate our model. The evaluation metrics are divided into two categories, one is the standard metrics including precision–recall (PR) curves, the area under curve (AUC), Top-N precision (P@N) and Max_F1, the other is long-tail metrics (i.e., accuracy of Hits@K). The detailed definitions are as follows:

- **The precision–recall(PR) curve** is a curve plotted with recall values as x axis and precision values as y axis, and shows the tradeoff between precision and recall for different thresholds. The formula is as follows,

$$precision = \frac{True\_Positive}{True\_Positive + False\_Positive}, \tag{21}$$

$$recall = \frac{True\_Positive}{True\_Positive + False\_Negative}. \tag{22}$$

- **AUC** denotes the area under the precision–recall curve. The higher the AUC value, the better the performance.
- **Top-N precision** indicates precision values for the entity pairs with top-n prediction confidences.
- **Max_F1** means the maximum value of F1 score. The F1 score is the weighted average of precision and recall.

$$F1 = \frac{2 * precision * recall}{precision + recall} \tag{23}$$

- **Hits@K** is used to measure whether a test sentence bag whose gold relation label $r^{(k)}$ falls into top-K relations ranked by their prediction confidences.

4.1. Experimental settings

Following previous works, for initialization, we use the same pre-trained word embeddings released by Lin et al. [7][4]. During training, we leverage mini-batch SGD [48] with the learning rate γ to minimize the objective functions. Besides, to prevent overfitting, we employ the dropout strategy [49] on the relation classifier layer.

In addition, the parameters of Entity-Aware Embedding layer, PCNN layer and Base Branch Network are kept consistent with [15]. So do the learning rate and dropout rate. For Hierarchy Structure Encoder and training objectives, we just use the following setting, and make little effort to select the best hyper-parameters. We have reason to believe that our model can

[2] http://iesl.cs.umass.edu/riedel/ecml/.
[3] https://github.com/thunlp/HNRE/tree/master/raw_data.
[4] https://github.com/thunlp/OpenNRE.

Table 1
Model evaluation on NYT. In the model comparison, best score is in bold. While values exceeding GHE-LPC are underlined in the ablation study.

Approach	P@100	P@200	P@300	P@500	P@1000	P@2000	Mean	AUC	Max_F1
Model comparison									
PCNN-ATT‡	78.0	72.5	71.0	67.6	54.3	40.8	64.0	0.39	0.437
PCNN-HATT†	82.0	80.5	76.0	67.8	58.3	42.1	67.8	0.42	0.455
ToHRE‡	91.5	82.9	79.6	74.8	63.3	48.9	73.5	0.44	0.476
CoRA†	93.0	91.0	88.0	81.2	67.6	51.4	78.7	0.530	0.525
GHE-LPC	**94.0**	**94.0**	**91.7**	**85.4**	**69.9**	**54.0**	**81.5**	**0.561**	**0.549**
Ablation study									
∼ w/o GHE (LPC)	94.0	90.5	87.0	81.4	71.8	52.1	79.5	0.541	0.532
∼ w/o LPC (GHE)	92.5	90.0	88.0	82.2	71.6	52.7	79.5	0.546	0.538
∼ w/o Dot Product in Eq. (11)	95.0	91.5	90.0	83.8	71.6	52.9	80.8	0.551	0.538
∼ w/o Gating in Eqs. (12)– (13)	93.0	93.0	89.0	84.8	70.3	52.6	80.5	0.549	0.536
∼ w/o Attention Pooling in Eq. (15)	92.0	89.0	87.7	79.2	69.6	53.3	78.5	0.541	0.542
∼ w/o Auxiliary Loss in Eq. (19)	83.0	80.5	77.0	69.6	61.2	46.8	69.7	0.445	0.478

Table 2
Model evaluation on NYT when randomly keeping one/two/all sentence(s) in each bag. In the model comparison, best score is in bold. While values exceeding GHE-LPC are underlined in the ablation study.

P@N (%)	One				Two				All			
	100	200	300	Mean	100	200	300	Mean	100	200	300	Mean
Model comparison												
PCNN-ATT‡	73.3	69.2	60.8	67.8	77.2	71.6	66.1	71.6	76.2	73.1	67.4	72.2
PCNN-HATT†	84.0	76.0	69.7	76.6	85.0	76.0	72.7	77.9	88.0	79.5	75.3	80.9
ToHRE‡	87.1	81.4	75.3	81.3	89.7	83.1	78.5	83.8	92.4	86.7	81.2	86.8
CoRA†	94.0	90.5	82.0	88.8	**98.0**	91.0	86.3	91.8	**98.0**	92.5	88.3	92.9
GHE-LPC	**97.0**	**94.0**	**88.7**	**93.2**	**98.0**	**95.5**	**90.3**	**94.6**	**98.0**	**96.5**	**92.3**	**95.6**
Ablation study												
∼ w/o GHE (LPC)	93.0	90.0	86.3	89.8	95.0	92.0	89.0	92.0	95.0	93.5	91.3	93.3
∼ w/o LPC (GHE)	95.0	92.5	87.3	91.6	95.0	93.5	89.0	92.5	97.0	95.0	92.0	94.5
∼ w/o dot product in Eq. (11)	91.0	89.0	85.7	88.6	94.0	93.0	90.0	92.3	97.0	95.0	91.7	94.6
∼ w/o gating in Eqs. (12)–(13)	96.0	94.0	87.3	92.4	99.0	94.2	89.9	94.4	99.0	96.1	91.7	95.6
∼ w/o attention pooling in Eq. (15)	89.0	88.5	85.7	87.7	93.0	92.5	88.3	91.3	93.0	91.0	87.7	90.6
∼ w/o auxiliary loss in Eq. (19)	84.0	72.5	65.7	74.1	87.0	83.0	75.0	81.7	90.0	85.0	78.7	84.6

achieve greater performance improvement with better parameter settings. In details, d_w, d_p, d_x, d_c, d_g, d'_g, n and ω are 50, 5, 150, 230, 690, 690, 120 and 3 respectively. λ in Section 3.3 is 0.05. For GATs, the number of heads H is set to 3. For NYT dataset, we set the number of relation levels k to 3. The numbers of distinct relations at three levels are 9, 36 and 53. For optimization, the learning rate γ is 0.1, batch size is 160, dropout rate is set to 0.5, weight decay of L2 regularization η is 1e-5. Besides, the coefficients of the loss function are set to 1, 1 and 1.

4.2. Baselines

We choose some competitive methods as the baseline models:

- **PCNN-ATT**: It is the most classical RE model proposed by Lin et al. [7], which uses attention mechanism to alleviate wrong labeling problem.
- **PCNN-HATT**: It is the first hierarchical RE model devised by Han et al. [13], which leverages relation hierarchies and design a hierarchical attention network.
- **ToHRE**: It is proposed by Yu et al. [16], which designs a Top-Down classification strategy along the relation hierarchies.
- **CoRA**: It is the most competitive method currently proposed by Li et al. [15], which designs a collaborating relation-augmented attention network to handle long-tail relations.

4.3. Overall results

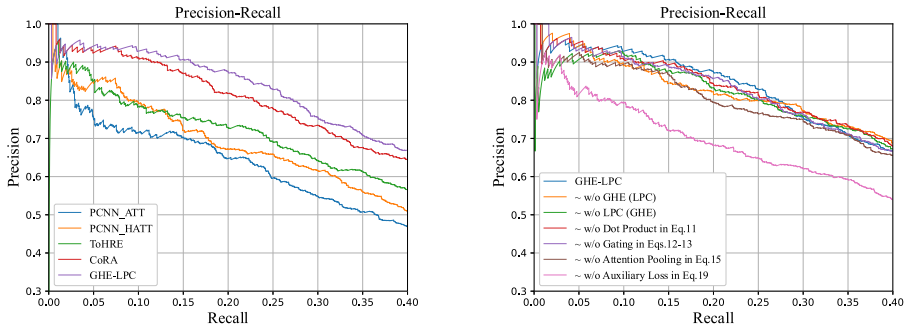
The overall performance of different methods is shown in Table 1, Table 2 and Fig. 3(a). Here † denotes that the results are from the corresponding official codes, while ‡ indicates that the results are from our own implementation.

It can be observed that our model GHE-LPC achieves state-of-the-art performance on multiple metrics. For AUC, our model's value is 0.561, which outperforms strong baselines by at least 0.031. And we improve the Max_F1 value by at least 2.4%. For P@N metric, in Table 1, our GHE-LPC achieves the highest precision on all N values. While in Table 2, we randomly retain one, two or all sentence(s) in each bag to keep the setting consistent with Li et al. [15], and GHE-LPC gets the highest values in spite of the randomness of retained sentences. In addition, The Precision–Recall curve of our model is significantly higher than the other models, although it has a small overlap with baselines when recall values less than 0.10.

To verify the impact of GHE-LPC on long-tail relations, we filter out the instances of long-tail relations from the test set and conduct model comparison experiments. See Table 3 for detailed results. The metric is Hits@K, which denotes whether a test sentence bag whose gold label $r^{(k)}$ falls into top-K relations ranked

Table 3
Hits@K (Macro) on the relations whose number of training instances <100/200.

#Instance	<100			<200		
Hits@K	10	15	20	10	15	20
PCNN_ATT	<5.0	7.4	40.7	17.2	24.2	51.5
PCNN_HATT	29.6	51.9	61.1	41.4	60.6	68.2
ToHRE	62.9	75.9	81.4	69.7	80.3	84.8
CoRA	66.7	72.2	87.0	72.7	77.3	89.3
GHE-LPC	**72.2**	**83.3**	**88.9**	**77.3**	**86.4**	**90.9**



(a) model comparison. (b) ablation study.

Fig. 3. Precision–recall (PR) curves.

by prediction probability. Besides, macro average is applied to calculate these values. Our GHE-LPC achieves the highest values for all values of K. This shows that out GHE-LPC does place more emphasis on long-tail relations and moves them forward in the rankings.

4.4. Ablation study

To further evaluate the effectiveness of GHE and LPC, we conduct some ablation experiments. Note that, the setup $\sim$ *w/o GHE (LPC)* indicates that $R^{(i)}$ is initialized randomly and is not processed by Hierarchy Structure Encoder. The setup $\sim$ *w/o Dot Product in Eq.* (11) denotes the replacement of Eq. (11) by average pooling over $R^{(i)}$. The setup $\sim$ *w/o Gating in Eqs.* (12)–(13) indicates that feature stitching $[u; c^{(i)}]$ is processed directly by Multi-Layer perceptron. The setup $\sim$ *w/o Attention Pooling in Eq.* (15) denotes that the attention pooling is replaced by average pooling.

The results are shown at the bottom of Tables 1 and 2, and Fig. 3(b). It can be find that the performance drop is consistent in multiple metrics, i.e., P@N, Max_F1 and AUC. (1) For our two main contributions (i.e., GHE and LPC), the performance drop is noticeable when either of them is removed, which proves that they are effective and necessary. (2) In particular, when the auxiliary loss L_{hier} is removed, the results are extremely low. This also confirms the power/importance of sentence-level supervised signals. (3) Furthermore, the other setups highlight the necessity of each technique in Section 3.6.1.

4.5. Impact of ATT(·) function in Section 3.5.2

Since $ATT(\cdot)$ indicates any kind of attention function, to evaluate its impact, we implement it in the following ways:

- $ATT(l_i, l_j) = LeakyReLu(W_{att}^T[Wl_i; Wl_j] + b_{att})$: In this paper, we use this implementation, where $LeakyReLu(\cdot)$ is the activation function and $W_{att} \in \mathbb{R}^{2d_g'}$. For convenience, it is denoted by **Linear**.
- $ATT(l_i, l_j) = CosineSimilarity(Wl_i, Wl_j)$: The cosine similarity is used to calculate the distance between l_i and l_j, denoted by **Cosine**.
- $ATT(l_i, l_j) = Wl_i \cdot Wl_j$: The dot product is used to calculate the similarity between l_i and l_j, denoted by **Dot Product**.

The results are shown in Table 4. It can be seen that **Cosine** outperforms **Dot Product** in terms of multiple metrics, while **Linear** achieves the best performance. The reason may be that **Linear** can benefit more from the multi-head setting.

4.6. Case study and visualization

The probabilities derived from Eq. (10) are critical to measure the knowledge transfer between different relations. We select two examples from NYT dataset and list the top-3 matching probability values at all relation levels in Table 5. It can be seen that our model has better identification of *NA* and has better recognition of noisy sentences (i.e., Sent. 1).

As stated in Section 1, the correlation of relations is reflected in two aspects, i.e., inter- and intra-level. Along the relation hierarchies, intuitively, inter-level correlation should be more obvious than intra-level. To observe the correlation of relations, we calculate the similarity between the resulting relation embeddings by dot product, and visualize the intra- and inter-level similarity matrices in Fig. 4. Note that we only consider relations that have siblings because non-sibling relations may dilute the results. As can be seen from the figures, the correlation of relations becomes more and more obvious along the relation hierarchies, which also demonstrates that our model does capture the correlation.

4.7. Error analysis and future research directions

To find the reasons of misclassification, we manually examine the misclassified samples in the test set and summarize the following factors: (1) In most cases, the proposed method still suffers greatly from the wrong labeling problem, probably because the attention mechanism still cannot completely eliminate the effect of noisy supervised signals; (2) Some relations have similar meanings and are difficult to distinguish, which causes relation ambiguity problem. For example, relations */people/deceased_person/place_of_death* and */people/deceased_person/place_of_burial* belong to this case. (3) Inconsistent distribution of relations on the train and test sets. Some relations appear only in the test set while there are no training instances in the train set. Our model and existing baselines are not suitable for this zero-shot scenario.

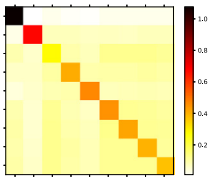
The correlation of relations in this paper is essentially the correlation between labels in the classification tasks. Since modeling the correlation of labels can intuitively improve the discrimination between categories, our ideas are applicable to all classification tasks. Since our proposed approach addresses the hierarchical classification task, future research directions include at least the following two aspects: (1) Designing methods to model correlations between labels for more general classification tasks; (2) Our approach in this paper can be applied to any classification task with taxonomic hierarchies, such as fine-grained hierarchical text/image classification, etc.

Table 4
Impact of the $ATT(\cdot)$ function.

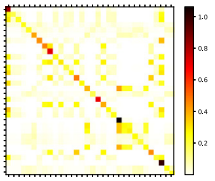
$ATT(\cdot)$	P@100	P@200	P@300	P@500	P@1000	P@2000	Mean	AUC	Max_F1
Linear	**94.0**	**94.0**	**91.7**	**85.4**	69.9	**54.0**	**81.5**	**0.561**	0.549
Cosine	94.0	89.0	86.0	82.6	69.2	53.8	79.1	0.557	**0.551**
Dot product	91.0	88.0	85.7	79.6	**71.6**	52.7	78.1	0.550	0.541

Table 5
The Top-3 matching probability values at all relation levels for two examples of NYT.

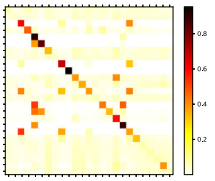
Sent. 1:	The cat-and-mouse game of the news media is something Mr. Langella has had a chance to study recently, for his film roles as **William S. Paley**, the chief executive of **CBS**, in "Good Night, Good Luck" and, for that matter, Perry White, the editor of a major metropolitan newspaper, in "Superman Returns".					
	$\alpha^{(1)}$		$\alpha^{(2)}$		$\alpha^{(3)}$	
GHE-LPC	*NA*:	0.679	*NA*:	0.677	*NA*:	0.676
	/business:	0.236	*/business/company*:	0.209	*/business/company/founders*:	0.129
	/people:	0.019	*/business/person*:	0.015	*/business/company/major_shareholders*:	0.051
Sent. 2:	Now, in the Internet and cellphone era, that name seems out of date as well, so the museum is renaming itself again, this time as the Paley Center for media, after the late **CBS** founder **William S. Paley**.					
	$\alpha^{(1)}$		$\alpha^{(2)}$		$\alpha^{(3)}$	
GHE-LPC	*/business*:	0.755	*/business/company*:	0.619	*/business/company/founders*:	0.449
	NA:	0.160	*NA*:	0.168	*NA*:	0.167
	/location:	0.024	*/business/person*:	0.085	*/business/company/major_shareholders*:	0.127



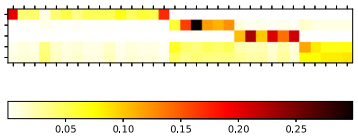
(a) Relation level 1.



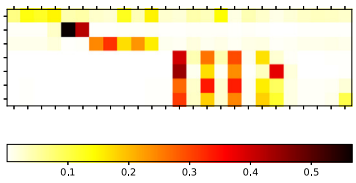
(b) Relation level 2.



(c) Relation level 3.



(d) Correlation between level 1 and level 2.



(e) Correlation between level 2 and level 3.

Fig. 4. Visualization of the correlation of relations on NYT. Note that, (a), (b) and (c) show the intra-level correlations, where the relations without siblings are removed. (d) and (e) show the inter-level correlations, which are more obvious than intra-level ones.

5. Conclusions

In this paper, we introduce the correlation of relations to the DSRE task and model it from two perspectives. Globally, we utilize Hierarchy Structure Encoder to aggregate relation information on the relation hierarchies graph and obtain Global Hierarchy Embeddings. Locally, a novel constraint called Local Probability Constraints is proposed, which captures the similarity of classification probabilities of adjacent relation levels. Compared with the competitive baselines, our proposed method achieves state-of-the-art performance using the NYT dataset, which demonstrates the importance of correlation between relations. In addition, there is still room for improvement, especially for long-tail problem. In the future, we plan to utilize more sophisticated strategies to further handle long-tail relations.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under grant No. 61872163 and 61806084, Jilin Province Key Scientific and Technological Research and Development Project under grant No. 20210201131GX, and Jilin Provincial Education Department Project under grant No. JJKH20190160KJ.

References

[1] F.M. Suchanek, G. Kasneci, G. Weikum, Yago: a core of semantic knowledge, in: Proceedings of the 16th International Conference on World Wide Web, WWW, 2007, pp. 697–706, http://dx.doi.org/10.1145/1242572.1242667.

[2] K.D. Bollacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor, Freebase: a collaboratively created graph database for structuring human knowledge, in: Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD, 2008, pp. 1247–1250, http://dx.doi.org/10.1145/1376616.1376746.

[3] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P.N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, C. Bizer, Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia, Semant. Web 6 (2) (2015) 167–195, http://dx.doi.org/10.3233/SW-140134.

[4] M. Mintz, S. Bills, R. Snow, D. Jurafsky, Distant supervision for relation extraction without labeled data, in: Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, ACL-AFNLP, 2009, pp. 1003–1011.

[5] S. Riedel, L. Yao, A. McCallum, Modeling relations and their mentions without labeled text, in: Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases, ECML-PKDD, in: Lecture Notes in Computer Science, vol. 6323, 2010, pp. 148–163, http://dx.doi.org/10.1007/978-3-642-15939-8_10.

[6] R. Hoffmann, C. Zhang, X. Ling, L.S. Zettlemoyer, 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, ACL, 2011, pp. 541–550.

[7] Y. Lin, S. Shen, Z. Liu, H. Luan, M. Sun, Neural relation extraction with selective attention over instances, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL, 2016, pp. 2124–2133, http://dx.doi.org/10.18653/v1/P16-1200.

[8] J. Qu, D. Ouyang, W. Hua, Y. Ye, X. Li, Distant supervision for neural relation extraction integrated with word attention and property features, Neural Netw. 100 (2018) 59–69, http://dx.doi.org/10.1016/j.neunet.2018.01.006.

[9] J. Du, J. Han, A. Way, D. Wan, Multi-level structured self-attentions for distantly supervised relation extraction, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2018, pp. 2216–2225, http://dx.doi.org/10.18653/v1/d18-1245.

[10] T. Liu, K. Wang, B. Chang, 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, EMNLP, 2017, pp. 1790–1795, http://dx.doi.org/10.18653/v1/d17-1189.

[11] J. Yang, Q. Wang, C. Su, X. Wang, Threat intelligence relationship extraction based on distant supervision and reinforcement learning, in: Proceedings of the 32nd International Conference on Software Engineering and Knowledge Engineering, SEKE, 2020, pp. 572–576, http://dx.doi.org/10.18293/SEKE2020-149.

[12] Y. Xiao, C. Tan, Z. Fan, Q. Xu, W. Zhu, Joint entity and relation extraction with a hybrid transformer and reinforcement learning based model, in: Proceedings of the 34th AAAI Conference on Artificial Intelligence, AAAI, 2020, pp. 9314–9321.

[13] X. Han, P. Yu, Z. Liu, M. Sun, P. Li, Hierarchical relation extraction with coarse-to-fine grained attention, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2018, pp. 2236–2245, http://dx.doi.org/10.18653/v1/d18-1247.

[14] N. Zhang, S. Deng, Z. Sun, G. Wang, X. Chen, W. Zhang, H. Chen, Long-tail relation extraction via knowledge graph embeddings and graph convolution networks, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, 2019, pp. 3016–3025, http://dx.doi.org/10.18653/v1/n19-1306.

[15] Y. Li, T. Shen, G. Long, J. Jiang, T. Zhou, C. Zhang, Improving long-tail relation extraction with collaborating relation-augmented attention, in: Proceedings of the 28th International Conference on Computational Linguistics, COLING, 2020, pp. 1653–1664, http://dx.doi.org/10.18653/v1/2020.coling-main.145.

[16] E. Yu, W. Han, Y. Tian, Y. Chang, Tohre: A top-down classification strategy with hierarchical bag representation for distantly supervised relation extraction, in: Proceedings of the 28th International Conference on Computational Linguistics, COLING, 2020, pp. 1665–1676, http://dx.doi.org/10.18653/v1/2020.coling-main.146.

[17] Z. Jin, Y. Yang, X. Qiu, Z. Zhang, Relation of the relations: A new paradigm of the relation extraction problem, 2020, CoRR abs/2006.03719 arXiv:2006.03719.

[18] M. Surdeanu, J. Tibshirani, R. Nallapati, 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, EMNLP-CoNLL, 2012, pp. 455–465.

[19] D. Zeng, K. Liu, Y. Chen, 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, EMNLP, 2015, pp. 1753–1762, http://dx.doi.org/10.18653/v1/d15-1203.

[20] Y. Yuan, L. Liu, S. Tang, Z. Zhang, Y. Zhuang, S. Pu, F. Wu, X. Ren, Cross-relation cross-bag attention for distantly-supervised relation extraction, in: Proceedings of the 33rd AAAI Conference on Artificial Intelligence, AAAI, 2019, pp. 419–426, http://dx.doi.org/10.1609/aaai.v33i01.3301419.

[21] Z. Ye, Z. 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, NAACL-HLT, 2019, pp. 2810–2819, http://dx.doi.org/10.18653/v1/n19-1288.

[22] L. Dai, B. Xu, H. Song, Feature-level attention based sentence encoding for neural relation extraction, in: Proceedings of the 8th CCF International Conference of Natural Language Processing and Chinese Computing, NLPCC, in: Lecture Notes in Computer Science, 11838, 2019, pp. 184–196, http://dx.doi.org/10.1007/978-3-030-32233-5_15.

[23] B. Yu, Z. Zhang, T. Liu, B. Wang, S. Li, Q. Li, Beyond word attention: Using segment attention in neural relation extraction, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI, 2019, pp. 5401–5407, http://dx.doi.org/10.24963/ijcai.2019/750.

[24] J. Zhang, K. Hao, X. Tang, X. Cai, Y. Xiao, T. Wang, A multi-feature fusion model for Chinese relation extraction with entity sense, Knowl. Based Syst. 206 (2020) 106348, http://dx.doi.org/10.1016/j.knosys.2020.106348, URL https://doi.org/10.1016/j.knosys.2020.106348.

[25] Y. Gou, Y. Lei, L. Liu, P. Zhang, X. Peng, A dynamic parameter enhanced network for distant supervised relation extraction, Knowl. Based Syst. 197 (2020) 105912, http://dx.doi.org/10.1016/j.knosys.2020.105912, URL https://doi.org/10.1016/j.knosys.2020.105912.

[26] H. Zhao, R. Li, X. Li, H. Tan, CFSRE: context-aware based on frame-semantics for distantly supervised relation extraction, Knowl. Based Syst. 210 (2020) 106480, http://dx.doi.org/10.1016/j.knosys.2020.106480, URL https://doi.org/10.1016/j.knosys.2020.106480.

[27] A. Bordes, N. Usunier, A. García-Durán, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, in: Proceedings of the 27th Annual Conference on Neural Information Processing Systems, NIPS, 2013, pp. 2787–2795.

[28] M. Defferrard, X. Bresson, P. Vandergheynst, Convolutional neural networks on graphs with fast localized spectral filtering, in: Proceedings of the 30th Annual Conference on Neural Information Processing Systems, NIPS, 2016, pp. 3837–3845.

[29] W.L. Hamilton, Z. Ying, J. Leskovec, Inductive representation learning on large graphs, in: Proceedings of the Annual Conference on Neural Information Processing Systems, NIPS, 2017, pp. 1024–1034, URL https://proceedings.neurips.cc/paper/2017/hash/5dd9db5e033da9c6fb5ba83c7a7ebea9-Abstract.html.

[30] T. Hamaguchi, H. Oiwa, M. Shimbo, Y. Matsumoto, Knowledge transfer for out-of-knowledge-base entities : A graph neural network approach, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI, 2017, pp. 1802–1808, http://dx.doi.org/10.24963/ijcai.2017/250, URL https://doi.org/10.24963/ijcai.2017/250.

[31] A. Fout, J. Byrd, B. Shariat, A. Ben-Hur, Protein interface prediction using graph convolutional networks, in: Proceedings of the Annual Conference on Neural Information Processing Systems, NIPS, 2017, pp. 6530–6539, URL https://proceedings.neurips.cc/paper/2017/hash/f507783927f2ec2737ba40afbd17efb5-Abstract.html.

[32] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: Proceedings of the 5th International Conference on Learning Representations, ICLR, 2017.

[33] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio, Graph attention networks, in: Proceedings of the 6th International Conference on Learning Representations, ICLR, 2018.

[34] J. Chen, M. Zhong, J. Li, D. Wang, T. Qian, H. Tu, Effective deep attributed network representation learning with topology adapted smoothing, IEEE Trans. Cybern. (2021) 1–12, http://dx.doi.org/10.1109/TCYB.2021.3064092.

[35] Z. Li, X. Wang, J. Li, Q. Zhang, Deep attributed network representation learning of complex coupling and interaction, Knowl. Based Syst. 212 (2021) 106618, http://dx.doi.org/10.1016/j.knosys.2020.106618.

[36] G. Xue, M. Zhong, J. Li, J. Chen, C. Zhai, R. Kong, Dynamic network embedding survey, 2021, CoRR abs/2103.15447 arXiv:2103.15447 URL https://arxiv.org/abs/2103.15447.

[37] Y. Li, G. Long, T. Shen, T. Zhou, L. Yao, H. Huo, J. Jiang, Self-attention enhanced selective gate with entity-aware embedding for distantly supervised relation extraction, in: Proceedings of the 34th AAAI Conference on Artificial Intelligence, AAAI, 2020, pp. 8269–8276.

[38] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Proceedings of the 27th Annual Conference on Neural Information Processing Systems, NIPS, 2013, pp. 3111–3119.

[39] D. Zeng, K. Liu, S. Lai, G. Zhou, J. Zhao, Relation classification via convolutional deep neural network, in: Proceedings of the 25th International Conference on Computational Linguistics, COLING, 2014, pp. 2335–2344.

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the Annual Conference on Neural Information Processing Systems, NIPS, 2017, pp. 5998–6008.

[41] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2016, pp. 770–778, http://dx.doi.org/10.1109/CVPR.2016.90.

[42] L.J. Ba, J.R. Kiros, G.E. Hinton, Layer normalization, 2016, CoRR abs/1607.06450 arXiv:1607.06450.

[43] Z. Lin, M. Feng, C.N. dos Santos, M. Yu, B. Xiang, B. Zhou, Y. Bengio, A structured self-attentive sentence embedding, in: Proceedings of the 5th International Conference on Learning Representations, ICLR, 2017.

[44] T. Shen, T. Zhou, G. Long, J. Jiang, S. Pan, C. Zhang, DiSAN: Directional self-attention network for rnn/cnn-free language understanding, in: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, AAAI, 2018, pp. 5446–5455.

[45] S. Jat, S. Khandelwal, P.P. Talukdar, Improving distantly supervised relation extraction using word and entity based attention, 2018, CoRR abs/1804.06987 arXiv:1804.06987.

[46] X. Ling, D.S. Weld, Fine-grained entity recognition, in: Proceedings of the 26th AAAI Conference on Artificial Intelligence, AAAI, Jörg Hoffmann and Bart Selman (eds.), 2012.

[47] T. Zhu, H. Wang, J. Yu, X. Zhou, W. Chen, W. Zhang, M. Zhang, Towards accurate and consistent evaluation: A dataset for distantly-supervised relation extraction, in: Proceedings of the 28th International Conference on Computational Linguistics, COLING, 2020, pp. 6436–6447, http://dx.doi.org/10.18653/v1/2020.coling-main.566.

[48] A. Cotter, O. Shamir, N. Srebro, K. Sridharan, Better mini-batch algorithms via accelerated gradient methods, in: Proceedings of the 25th Annual Conference on Neural Information Processing Systems, NIPS, 2011, pp. 1647–1655.

[49] N. Srivastava, G.E. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.