Neural Networks 152 (2022) 191–200

Contents lists available at ScienceDirect

Neural Networks

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

Distantly Supervised Relation Extraction via Recursive Hierarchy-Interactive Attention and Entity-Order Perception

Ridong Han [a,c], Tao Peng [a,b,c,*], Jiayu Han [d], Hai Cui [a,c], 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 Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin 130012, China*
[d] *Department of Linguistics, University of Washington, Seattle, WA 98195, United States*

A R T I C L E I N F O

Article history:
Received 5 December 2021
Received in revised form 24 February 2022
Accepted 14 April 2022
Available online 21 April 2022

Keywords:
Distant Supervision
Relation Extraction
Relation Hierarchies
Entity Order
Long-tail Relations
Attention

A B S T R A C T

Wrong-labeling problem and long-tail relations severely affect the performance of distantly supervised relation extraction task. Many studies mitigate the effect of wrong-labeling through selective attention mechanism and handle long-tail relations by introducing relation hierarchies to share knowledge. However, almost all existing studies ignore the fact that, in a sentence, the appearance order of two entities contributes to the understanding of its semantics. Furthermore, they only utilize each relation level of relation hierarchies separately, but do not exploit the heuristic effect between relation levels, i.e., higher-level relations can give useful information to the lower ones. Based on the above, in this paper, we design a novel Recursive Hierarchy-Interactive Attention network (RHIA) to further handle long-tail relations, which models the heuristic effect between relation levels. From the top down, it passes relation-related information layer by layer, which is the most significant difference from existing models, and generates relation-augmented sentence representations for each relation level in a recursive structure. Besides, we introduce a newfangled training objective, called Entity-Order Perception (EOP), to make the sentence encoder retain more entity appearance information. Substantial experiments on the popular *New York Times* (NYT) dataset are conducted. Compared to prior baselines, our RHIA-EOP achieves state-of-the-art performance in terms of precision–recall (P–R) curves, AUC, Top-N precision and other evaluation metrics. Insightful analysis also demonstrates the necessity and effectiveness of each component of RHIA-EOP.

1. Introduction

Various large-scale knowledge bases (KBs), including YAGO (Suchanek, Kasneci, & Weikum, 2007), Freebase (Bollacker, Evans, Paritosh, Sturge, & Taylor, 2008) and DBpedia (Lehmann et al., 2015), are extremely supportive of many sub-tasks in the field of natural language processing (NLP). However, although existing KBs contain a large number of facts, they are still far from complete compared to the real-world facts, which is infinite. To enrich KBs, many methods have been proposed to automatically extract fact triples from unstructured texts, i.e., relation extraction (RE). Among these, supervised approaches are the most commonly used methods and yield relatively high performance. But existing supervised RE systems require massive training data, especially when using neural networks. Furthermore, obtaining high-quality and large-scale training data is very time-consuming and labor-intensive. In this case, distant supervision (DS) is proposed to automatically label training instances by matching KBs to the corpus (Mintz, Bills, Snow, & Jurafsky, 2009). It assumes that given a pair of entities, all sentences contain these two entities will express the relation between them in KBs. The assumption is too strong and results in some problems.

First, wrong labeling problem is inevitable and has become a bottleneck limiting models' performance by introducing noisy supervision signals. For instance, <**Phil Amicone**, **Yonkers**>expresses the */people/person/place_of_birth* relation in Freebase. So, the sentence "*Mayor Phil Amicone of Yonkers and the board of Education have supported Mr. Petrone during the controversy.*" will be automatically labeled as a training sentence, even though it does not express this type of relation. To mitigate the impact of noise labels, multi-instance learning (MIL) (Hoffmann, Zhang, Ling, Zettlemoyer, & Weld, 2011; Riedel, Yao, & McCallum, 2010) is proposed to identify a relation label for a sentence-bag containing two common entities. Then, many techniques have been introduced into distantly supervised relation extraction (DSRE) task,

* Corresponding author at: College of Computer Science and Technology, Jilin University, Changchun, Jilin 130012, China.
** Corresponding author at: College of Software, Jilin University, Changchun, Jilin 130012, China.
E-mail addresses: tpeng@jlu.edu.cn (T. Peng), liulu@jlu.edu.cn (L. Liu).

https://doi.org/10.1016/j.neunet.2022.04.019
0893-6080/

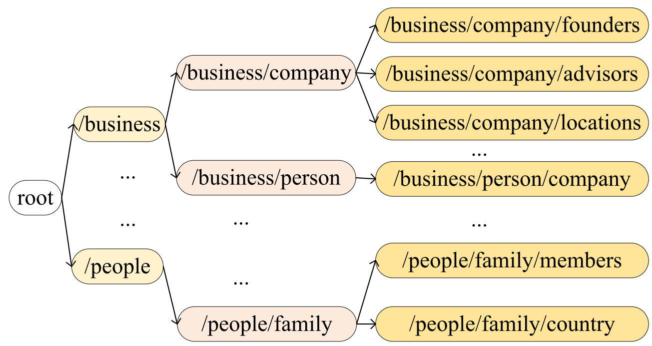


Fig. 1. Illustration of the relation hierarchies in Freebase.

such as multiple perspectives' attention (Du, Han, Way, & Wan, 2018; Lin, Shen, Liu, Luan, & Sun, 2016; Qu, Ouyang, Hua, Ye, & Li, 2018), soft-labeling (Liu, Wang, Chang, & Sui, 2017), reinforcement learning (Xiao, Tan, Fan, Xu, & Zhu, 2020; Yang, Wang, Su, & Wang, 2020), etc. Among these, the attention mechanism is the most common and successful, so we also extend this technique.

Second, although DS method can generate large-scale training data, it can only cover a limited part of real-world relations and causes long-tail problem. For example, in NYT dataset, nearly 77% of the relations are long-tail. Here a relation is long-tail if the number of corresponding training instances is less 1000. In this case, data imbalance severely limits the performance of RE systems. Recently, some approaches naturally share the knowledge from data-rich relations to long-tail ones by leveraging relation hierarchies (Han, Yu, Liu, Sun, & Li, 2018; Li, Shen, Long, Jiang, Zhou, & Zhang, 2020; Yu, Han, Tian, & Chang, 2020; Zhang et al., 2019). It is based on the observation that although some relations are long-tailed, their ancestor or sibling relations are not. Therefore long-tail relations can benefit from their ancestors or siblings. For example, we select some relations from Freebase and illustrate them in Fig. 1. Note that we add a *root* node to facilitate the narrative. It can be seen that the relation */business/company/founders* has two ancestor relations (i.e., */business* and */business/company*) and several siblings (i.e., */business/company/advisors*, */business/company/locations*, etc.). And all relations form a taxonomic structure. With it, Han et al. (2018) propose a hierarchical attention scheme to generate extra bag-level features for each relation level of relation hierarchies separately. Then Zhang et al. (2019) enrich the relation embeddings with TransE (Bordes, Usunier, García-Durán, Weston, & Yakhnenko, 2013) and graph convolutional networks (Defferrard, Bresson, & Vandergheynst, 2016). After that, Li, Shen et al. (2020) propose an attention-based method to enhance sentence representations for each relation level independently, while Yu et al. (2020) design a Top-Down classification strategy along the relation hierarchies.

However, the above methods all use each relation level of relation hierarchies independently, i.e., although they use multi-grained relation levels to generate extra features or enrich sentence representations, these levels are discrete, independent and do not affect each other in the calculation. In fact, along the hierarchical relation chains, higher-level relations must be instructive to the lower ones. For example, the sentence "***Google*** *was founded by Larry Page and* ***Sergey Brin*** *on September 4, 1998.*" expresses the relation */business/company/founders*. When classifying along the hierarchical relation chains, if it is identified as */business* at the first level, then at the second level, we can select labels only from the child relations of */business*. Keep going down until the end of chains, the final label is obtained. During this process, the heuristic effect between relation levels is reflected in layer-by-layer narrowing the scope of the current level's labels based on the classification probability of previous level. It can also be seen as a continuous refinement from coarse to fine granularity. Distinguishing from existing studies, we aim to implicitly model the heuristic effect through interactions between relation levels, which is one of the most prominent contributions of this paper. Considering further, modeling these interactions is somewhat equivalent to exploiting the taxonomic structure of relations to uncover the correlation of relations, which can improve inter-relational discrimination from the side.

Besides, as we all know, the fact triple $\langle e_1, r, e_2\rangle$ is not equal with $\langle e_2, r, e_1\rangle$, therefore the appearance order of two entities is extremely crucial. But during sentence encoding, each feature map generated by CNN/PCNN only retains 1 or 3 maximums along the word sequence through pooling operation. This process ignores the importance of entity order and causes the loss of entity order information, i.e., entity order features are underutilized in the deep learning paradigm.

In this paper, we firstly propose a novel network, named as **R**ecursive **H**ierarchy-**I**nteractive **A**ttention (RHIA), to fully exploit the relation hierarchies. It assumes that *along the hierarchical relation chains, lower-level relations are influenced by the higher-level ones and current known information*. Based on this, we leverage a recursive structure along the chains to deliver heuristic information about higher-level relations, and obtain the relation-augmented sentence representations. Then we design a new attention pooling module by using the final hidden state to generate bag-level representations. Besides, to retain more entity-order information in sentence representations, an effective training objective, called **E**ntity-**O**rder **P**erception (EOP), is introduced. Our key contributions are summarized as follows:

- We take the heuristic effect between relation hierarchies into account, then a novel network, called **RHIA**, is proposed to model this heuristics. It is the first approach in DSRE to uncover the heuristic effect between relation levels in relation hierarchies.
- A newfangled training objective, called **EOP**, is introduced to improve the expressive ability of sentence encoder. It enables the sentence representations to retain more entity-order information.
- We conduct substantial evaluation on the widely-used benchmark NYT, and receive state-of-the-art performance in multiple metrics. Insightful analysis also verifies the capability and effectiveness of our RHIA-EOP. The code is released at https://github.com/RidongHan/RHIA-EOP.

2. Related work

As an important subtask of natural language processing (NLP), relation extraction (RE) task can be divided into sentence-level extraction (Chen et al., 2021; Geng, Chen, Han, Lu, & Li, 2020; Zeng, Liu, Lai, Zhou, & Zhao, 2014), document-level extraction (Huang et al., 2021; Xu, Chen, & Zhao, 2021), few-shot extraction (Yang, Zhang, Niu, Zhao, & Pu, 2021), distantly supervised extraction (Deng, Yang, Kang, Yang, & Wu, 2021; Li, Shen et al., 2020; Qu et al., 2018; Zhou, Pan, Bai, Luo, & Wu, 2021), etc. Here we concentrate on the sentence-level distantly supervised relation extraction (DSRE) scenario.

For wrong-labeling problem, some people (Hoffmann et al., 2011; Riedel et al., 2010; Surdeanu, Tibshirani, Nallapati, & Manning, 2012) relax the assumption behind DS and develop multi-instance learning (MIL) framework. After that, many techniques have been applied to DSRE. First, Zeng, Liu, Chen, and Zhao

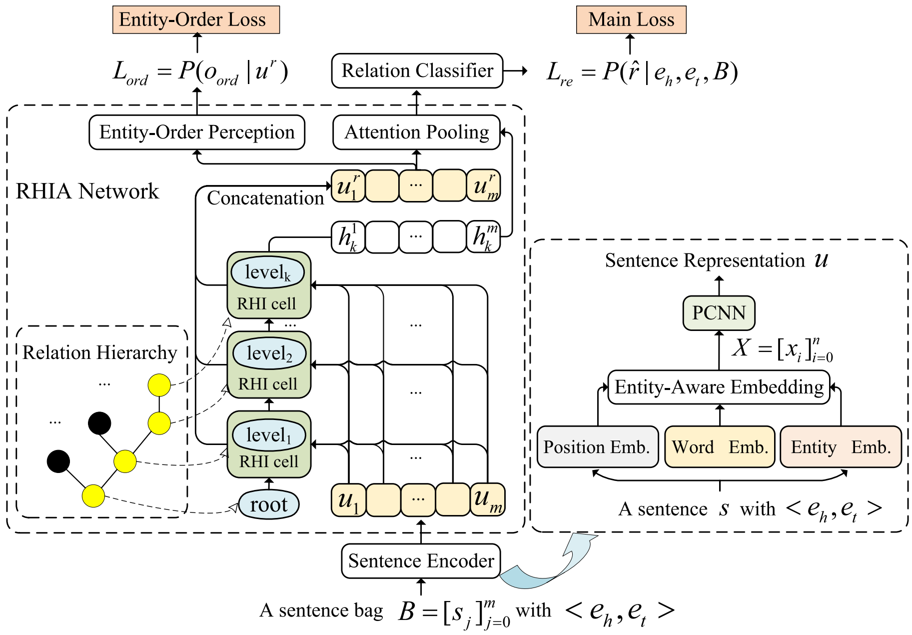


Fig. 2. The overview of our proposed RE model, RHIA-EOP.

(2015) improve the sentence encoder with Piecewise Convolutional Neural Networks (PCNN), and identify the instance that most likely expresses the corresponding relation in sentence-bag level. Then, Lin et al. (2016) design a selective attention among sentences/instances in the sentence level. Inspired by this work, attention models in different aspects are proposed, including word-level attention (Qu et al., 2018), self-attention (Du et al., 2018), bag-level attention (Ye & Ling, 2019; Yuan et al., 2019), feature-level attention (Dai, Xu, & Song, 2019), segment-level attention (Yu et al., 2019), etc. Besides, in order to pick out correct training instances to train the model efficiently, reinforcement learning is introduced into DSRE (Xiao et al., 2020; Yang et al., 2020). Since the attention mechanism is the most commonly used and successful, we also extend this technique.

Relation hierarchies contain taxonomic structure of relations. Although some relations are long-tail, their sibling or ancestor relations are not. To handle long-tail relations, existing practices take relation hierarchies into account to share knowledge from higher-level data-sufficient relations to the lower-level long-tail ones, which is intended for long-tail relations to benefit from the training phase of sibling/ancestor relations, i.e., supervised signals come from the relation hierarchies. Han et al. (2018) get advanced performance by introducing a hierarchical attention scheme to derive extra bag-level features. Zhang et al. (2019) use TransE (Bordes et al., 2013) and graph convolutional network (Defferrard et al., 2016) to obtain relation embeddings, and design a novel attention network along relation hierarchies. After that, Li, Shen et al. (2020) augment the sentence representations with relation embeddings at each level of relation hierarchies to provide more clues to the classifier, while Yu et al. (2020) exploit the relation hierarchies to design a top-down classification strategy. Recently, Peng et al. (2022) explore the correlation of relations in the relation hierarchies from both global and local perspectives, aiming to make long-tail relations benefit from their sibling or ancestor relations. The shortcoming of these models is that the relation levels are discrete, independent and do not affect each other during the calculation, i.e., the heuristic effect between relation levels is ignored. As stated in Section 1, higher-level relations are instructive to lower-level ones when classifying. This paper aims to address this flaw, and it is our contribution to differentiating from existing studies. Based on the models of Han et al. (2018) and Li, Shen et al. (2020), although our model also generates extra bag-level features and relation-augmented sentence representations, we further design a recursive interaction method to pass the relation-related heuristic information along the relational hierarchical chains.

3. Our proposed RE approach

Our model, RE via Recursive Hierarchy-Interactive Attention and Entity-Order Perception (RHIA-EOP), consists of three cascaded components: (1) A sentence encoder based on Entity-Aware Embedding and Piecewise Convolutional Neural Networks (PCNNs). (2) A Recursive Hierarchy-Interactive Attention (RHIA) module for generating more valuable bag representations by fully leveraging relation hierarchies. (3) A relation classifier module with MultiLayer Perceptron (MLP). The overall model inputs a bag of sentences and outputs the relation label in bag level. Fig. 2 shows the overall architecture.

3.1. Task definition

Given a sentence, the goal of relation extraction (RE) is to identify the relation between a pair of entities in this sentence. To facilitate this in distant supervision scenario, we split all sentences into multiple entity-pair bags $\{B_1, B_2, \ldots\}$. Each bag B_i contains some sentences $\{s_1, s_2, \ldots, s_m\}$ mentioning the same entity pair $\langle e_{hi}, e_{ti}\rangle$. Each sentence is a sequence of words, i.e., $s = [w_1, w_2, \ldots, w_n]$, and the maximum length is set to n. Besides, we have a set of pre-defined relation classes $\mathcal{R} = \{r_1, r_2, \ldots\}$. In this case, the goal of DSRE is to distinguish the relation between two given entities based on an entity-pair bag.

3.2. Sentence encoder

In this part, three kinds of features are taken into account, including word embedding (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), position embedding (Zeng et al., 2014) and entity embedding (Li, Long, Shen, Zhou, Yao, Huo, & Jiang, 2020). For each sentence s_j in bag $B = \{s_1, s_2, \ldots, s_m\}$,[1] we remove the index j for brevity in the following narrative.

Word embedding. Each sentence $s = [w_1, w_2, \ldots, w_n]$ is translated into low-dimensional embeddings, 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.

Position embedding. Relative position information is very vital to RE task (Zeng et al., 2014), which is defined as the combination of relative distances from each word to entity e_h and entity e_t. Take the sentence "*It showed that **Sergey Brin**, a co-founder of **Google**, not his partner, Larry Page, who is speaking at the conference.*" as an example, the relative distance from *co-founder* to entity e_h (*Google*) and entity e_t (*Sergey Brin*) are -2 and 3, respectively. Then, two low-dimensional vectors, $p_i^{e_h}$ and $p_i^{e_t} \in \mathbb{R}^{d_p}$ are converted from these two distances. In this way, we can define position-aware embeddings as $X^p = [x_1^p, x_2^p, \ldots, x_n^p] \in \mathbb{R}^{(d_w+2d_p)\times n}$, where $x_i^p = [v_i; p_i^{e_h}; p_i^{e_t}]$, $i \in [1, 2, \ldots, n]$, ";" is vector concatenation operation.

Entity embedding. The sequence of entity embeddings is represented as $X^e = [x_1^e, x_2^e, \ldots, x_n^e] \in \mathbb{R}^{3d_w \times n}$, where $x_i^e = [v_i; v_{e_h}; v_{e_t}]$, $i \in [1, 2, \ldots, n]$, v_{e_h} and v_{e_t} are the embeddings of two entities. We employ the same way as Li, Shen et al. (2020) to obtain the embeddings of entities. Each entity is one entry in the vocabulary of word embedding even if it is usually composed of multiple words. To achieve this, if an entity consists of multiple words, all words are connected with "_" to denote it.

Entity-aware embedding. To make better use of the above features, a position-wise gate (Li, Long et al., 2020) is used to integrate them, i.e.,

$$A^e = sigmoid(\lambda \cdot (W^e X^e + b^e)) \tag{1}$$

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

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

where "$\circ$" indicates the element-wise product, $W^e \in \mathbb{R}^{d_x \times 3d_w}$, $W^p \in \mathbb{R}^{d_x \times (d_w+2d_p)}$, λ is a trade-off weight. And $X = [x_1, x_2, \ldots, x_n] \in \mathbb{R}^{d_x \times n}$ is the resulting representation of words for encoding.

Piecewise Convolutional Neural Networks. We select the Piecewise Convolutional Neural Networks (PCNNs) as the sentence encoder (Zeng et al., 2015) because of its high performance and efficiency. Given the input representation X, PCNN applies a kernel of window size ω to slide over X, and output feature representation f, where $f \in \mathbb{R}^{d_c \times n}$ and d_c is the number of filters. After that, the feature f is firstly divided into three segments $\{f^{(1)}, f^{(2)}, f^{(3)}\}$ based on the position of two entities. And then, the max-pooling operation is employed on each segment, respectively. The results are concatenated as the final sentence representation u:

$$u = [max(f^{(1)}); max(f^{(2)}); max(f^{(3)})] \tag{4}$$

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

3.3. Recursive Hierarchy-Interactive Attention Network

To fully exploit the taxonomic structure of relations and relation embeddings, an attention network with a recursive structure along the relation hierarchies, called RHIA, is proposed. RHIA consists of several RHI cells. Each cell completes the calculation for one relation level. See Fig. 3 for the details of the RHI cell.

For a bag $B = \{s_1, s_2, \ldots, s_m\}$, the sentence representations $U = \{u_1, u_2, \ldots, u_m\}$ are derived via PCNNs encoder. For each relation $r \in \mathcal{R}$, its hierarchical chain of relations $\{r^0, r^1, \ldots, r^k\}$ can be generated, where r^0 is the *root* relation node and r^k is identical to r. Connected by the *root*, all chains compose the relation hierarchies, a kind of tree-like structure. Then, for the i-th level of relation hierarchies ($i \in [1, 2, \ldots, k]$), we define a learnable relation embedding matrix $R^{(i)} \in \mathbb{R}^{d_f \times N^i}$, where N^i denotes the number of relations at level i.

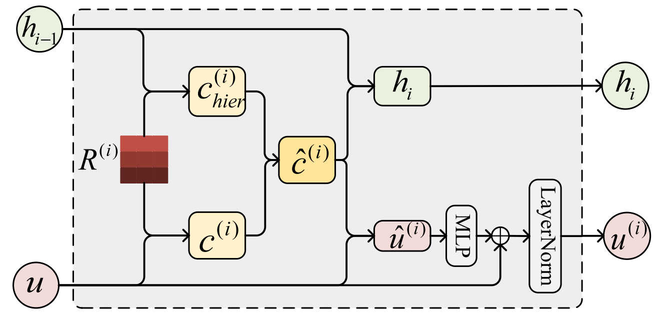


Fig. 3. The process of recursive computing unit, RHI cell.

For a sentence representation $u \in U$,[2] we aim to augment it with the above relation embeddings for each level, and all levels' augmented representations are concatenated as the relation-augmented sentence representation u^r. Then an attention-pooling module is employed to generate the bag-level representation. For i-th level, it is assumed that *the augmented representation of i-th level is determined by the input sentence representation u and the heuristic information about relations h_{i-1} from the higher/previous level* .

Based on this, we can build the network in a recursive structure. In details, the sentence-to-relation (sent2rel) attention (Li, Shen et al., 2020) is employed firstly. The sentence representation u and the heuristic information h_{i-1} are used as the query to calculate attention scores by dot product with the relation embedding matrix $R^{(i)}$, respectively,

$$\alpha^{(i)} = softmax(u^T R^{(i)}) \tag{5}$$

$$c^{(i)} = R^{(i)} \alpha^{(i)} \tag{6}$$

$$\alpha^{(i)}_{hier} = softmax({h_{i-1}}^T R^{(i)}) \tag{7}$$

$$c^{(i)}_{hier} = R^{(i)} \alpha^{(i)}_{hier} \tag{8}$$

where $h_{i-1} \in \mathbb{R}^{d_f}$, $softmax(\cdot)$ is an activation function for the last dimension, $c^{(i)}$ and $c^{(i)}_{hier}$ are the relation-aware information.

Since the importance of u and h_{i-1} is different, then, we leverage an element-wise gate mechanism to integrate the relation-aware information $c^{(i)}$ and $c^{(i)}_{hier}$:

$$\beta^{(i)}_1 = sigmoid(W^{g1}[u; h_{i-1}] + b^{g1}) \tag{9}$$

$$\hat{c}^{(i)} = \beta^{(i)}_1 \circ c^{(i)} + (1 - \beta^{(i)}_1) \circ c^{(i)}_{hier} \tag{10}$$

where $W^{g1} \in \mathbb{R}^{2d_f \times d_f}$, $\hat{c}^{(i)}$ is the resulting relation-aware representation corresponding to i-th level.

After that, to prevent the information loss of the original representation u, we leverage an element-wise gate to inject the information of $\hat{c}^{(i)}$ into u. In this process, residual connection (He, Zhang, Ren, & Sun, 2016) and layer normalization (Ba, Kiros, & Hinton, 2016) are also applied. Then, the augmented representation $u^{(i)}$ at level i is generated,

$$\beta^{(i)}_2 = sigmoid(W^{g2}[u; \hat{c}^{(i)}] + b^{g2}) \tag{11}$$

$$\hat{u}^{(i)} = \beta^{(i)}_2 \circ u + (1 - \beta^{(i)}_2) \circ \hat{c}^{(i)} \tag{12}$$

1 "sentence" and "instance" are semantically identical.

2 For convenience, we remove indices in the remaining parts.

$$u^{(i)} = LayerNorm(u + MLP(\hat{u}^{(i)})) \tag{13}$$

where $W^{g2} \in \mathbb{R}^{2d_f \times d_f}$, $MLP(\cdot)$ is a multi-layer perceptron that aims to increase nonlinearity.

Finally, we update h_{i-1} to obtain the current heuristic information about relations h_i for the next level's calculation, which is achieved by merging relation-aware information $\hat{c}^{(i)}$ into h_{i-1} with an element-wise gate,

$$\beta_3^{(i)} = sigmoid(W^{g3}[h_{i-1}; \hat{c}^{(i)}] + b^{g3}) \tag{14}$$

$$h_i = \beta_3^{(i)} \circ h_{i-1} + (1 - \beta_3^{(i)}) \circ \hat{c}^{(i)} \tag{15}$$

where $W^{g3} \in \mathbb{R}^{2d_f \times d_f}$.

During the calculation, h_0 is randomly initialized. Along the relation hierarchies, all levels' augmented representations are generated, i.e., $\{u^{(1)}, u^{(2)}, \dots, u^{(k)}\}$. Then the relation-augmented sentence representation u^r is generated by concatenation operation,

$$u^r = [u^{(1)}; u^{(2)}; \dots; u^{(k)}] \tag{16}$$

For the bag B, all relation-augmented sentence representations can be denoted as $B^r = [u_1^r, u_2^r, \dots, u_m^r]$. Next, to alleviate wrong labeling problem, we use the attention-pooling (Lin et al., 2017; Shen et al., 2018) to select the correctly labelled sentences from the bag in order to facilitate the generation of an accurate bag-level representation. The attention score for each sentence is calculated from its original representation u and its final hidden state h_k (i.e., the k-th level's heuristic information about relations). Then a weighted sum over the bag is employed,

$$b = B^r softmax(W_{att}^T[U; H]) \in \mathbb{R}^{kd_f} \tag{17}$$

where H is the matrix consisting of the final hidden state h_k of all sentences in B.

Finally, a softmax classifier based on the MultiLayer Perceptron is used to classify the bag representation b,

$$o_b = P(\hat{r}|e_h, e_t, B) = softmax(MLP(b)) \tag{18}$$

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

3.4. Entity-Order Perception and training objectives

To retain more entity-order information, we design a classification sub-task in the multi-task paradigm. Specifically, a Multi-Layer Perceptron is employed to bicategorize relation-augmented representations u^r, i.e., whether entity e_h appears before entity e_t or not. That is,

$$P(o_{ord}|u^r) = softmax(MLP(u^r)) \tag{19}$$

To optimize our model, three objectives are introduced: (1) The main objective is bag-level classification and is defined as minimizing cross-entropy loss,

$$L_{re} = -\frac{1}{|D|}\sum_{B \in D} \log P(\hat{r}|e_h, e_t, B) \tag{20}$$

where D is the train set consisting of sentence bags. (2) The hierarchical auxiliary objective is designed to guide RHIA module in choosing appropriate relation embeddings to augment each sentence representation. That is,

$$L_{hier} = -\frac{1}{|D| \times |B| \times k}\sum_{B \in D}\sum_{s \in B}\sum_{l=1}^{k} \log \alpha_{[r^l]}^{(l)} \tag{21}$$

where $[\cdot]$ denotes indexing operation. (3) The entity-order perception objective is introduced to take entity-order information into account:

$$L_{ord} = -\frac{1}{|D| \times |B|}\sum_{B \in D}\sum_{s \in B} \log P(o_{ord}|u^r) \tag{22}$$

Eventually, these three objectives are integrated into a whole. The final loss function can be represented as:

$$L = L_{re} + \mu L_{hier} + \xi L_{ord} + \delta \|\theta\|_2^2 \tag{23}$$

where μ, ξ and δ are weight scores of different losses. $\|\theta\|_2^2$ is the L_2 regularizer.

4. Experiments and results

Since this paper focuses on the sentence-level distantly supervised relation extraction task, we choose the *New York Times* (NYT) dataset (Riedel et al., 2010) for evaluation,[3] which is the only widely-adopted distantly supervised relation extraction benchmark and freely available.[4] The wrong labeling problem and long-tail problem are extremely serious on it, which is why we choose it. As for other available DS datasets, GIDS (Jat, Khandelwal, & Talukdar, 2018) has no the long-tail phenomenon; NYT-H (Zhu et al., 2020) is subset variants of NYT, and its test set contains only 9955 sentences, which is so small that the results of our model and baselines are all very high and not comparable; DocRED (Yao et al., 2019) is designed for document-level relation extraction that require inter-sentence reasoning capability.

The NYT dataset was generated by aligning the corpus with Freebase, and has 53 relations. Among these relation types, there is a special *NA* class which indicates that no relation exists between two entities. Its train set consists of the corpus from the years 2005–2006, while the test set consists of the rest corpus from year 2007. For its statistics, the number of sentence and entity_pair of train set are 570088 and 293162, respectively, while the values for the test set are 172448 and 96678.

Our RHIA-EOP is programmed using the Pytorch framework and trained on the GeForce GTX 1080 Ti. During the training phase, it takes about 1.5 hours to execute 15–20 epochs before convergence. Besides, following previous work (Lin et al., 2016), three-fold cross-validation on the train set is used for hyper-parameter selection and the held-out evaluation is applied to conduct the experiments. The evaluation metrics used in this paper include **Top-N precision**, **the precision–recall curve**, **AUC**, **Max_F1** and **Hits@K**.

Table 1
Parameters settings.

Parameter	Value
Word/entity embedding dimension d_w	50
Position embedding dimension d_p	5
Maximum length of sentences n	120
Entity-aware smoothness λ	0.05
Entity-aware embedding dimension d_x	150
Kernel size ω	3
Hidden dimension of PCNN d_c	230
Learning rate γ	0.1
Dropout rate	0.5
Number of relation levels k	3
L_2 regularization coefficient η	1e−5
weights of loss function μ, ξ, δ	1, 1, 1

4.1. Experimental settings

For the initialization of word embeddings, we employ the pre-trained embeddings from Lin et al. (2016).[5] The relation embedding matrices are initialized randomly as Li, Shen et al. (2020). Besides, the dropout strategy (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014) is applied to the bag representations to prevent overfitting. During training phase, for optimization, we employ mini-batch SGD (Cotter, Shamir, Srebro, & Sridharan, 2011) with the initial learning rate γ.

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

Table 2
Model evaluation and ablation study on NYT. The best scores are **bolded** or **underlined** in model comparison and ablation respectively.

Approach	P@100	P@200	P@300	P@500	P@1000	P@2000	Mean	AUC	Max_F1
PCNN_ATT (Lin et al., 2016)	78.0	72.5	71.0	67.6	54.3	40.8	64.0	0.39	0.437
HNRE (Han et al., 2018)	82.0	80.5	76.0	67.8	58.3	42.1	67.8	0.42	0.455
ToHRE (Yu et al., 2020)	91.5	82.9	79.6	74.8	63.3	48.9	73.5	0.44	0.476
CoRA (Li, Shen et al., 2020)	93.0	91.0	88.0	81.2	67.6	51.4	78.7	0.53	0.525
HNRE w/ Ent. Emb.	84.7	83.1	77.3	75.5	65.6	47.9	72.4	0.47	0.467
HNRE w/ Aux. Obj.	85.0	82.5	79.3	73.9	64.2	49.8	72.5	0.49	0.485
HNRE w/ Ent. Emb. w/ Aux. Obj.	90.1	87.0	85.7	76.8	67.3	51.7	76.4	0.52	0.525
RHIA-EOP	**95.0**	**94.0**	**89.7**	**85.2**	**71.7**	**53.2**	**81.5**	**0.56**	**0.546**
Ablations									
∼ w/o EOP (RHIA)	93.0	89.0	88.3	81.0	70.8	52.1	79.0	0.546	0.531
∼ w/o RHIA (EOP)	96.0	93.0	88.3	81.2	70.3	51.8	80.1	0.545	0.529
∼ w/o Sent2rel Attention	93.0	92.5	89.0	82.2	71.2	53.0	80.2	0.548	0.539
∼ w/o Attention Pooling	89.0	87.0	83.3	79.6	68.9	52.0	76.6	0.534	0.531
∼ w/o Gating in Eqs. (11)–(12)	95.0	91.0	87.3	83.2	70.8	53.0	80.1	0.549	0.541
∼ w/o Aux. Obj. in Eq. (21)	81.5	81.0	76.0	71.6	60.8	47.2	69.7	0.451	0.481
∼ w/ uncased BERT-Base	91.5	91.0	88.1	84.4	69.9	52.5	79.6	0.593	0.584

Table 3
Model evaluation and ablation study on NYT when retaining one/two/all sentence(s) in each bag at random. The best scores are **bolded** or **underlined** in model comparison and ablation respectively.

P@N(%)	One				Two				All			
	100	200	300	Mean	100	200	300	Mean	100	200	300	Mean
PCNN_ATT (Lin et al., 2016)	73.3	69.2	60.8	67.8	77.2	71.6	66.1	71.6	76.2	73.1	67.4	72.2
HNRE (Han et al., 2018)	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 (Yu et al., 2020)	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 (Li, Shen et al., 2020)	94.0	90.5	82.0	88.8	**98.0**	91.0	86.3	91.8	**98.0**	92.5	88.3	92.9
HNRE w/ Ent. Emb.	90.2	86.5	81.3	86.0	91.3	87.2	82.4	87.0	93.1	89.0	85.8	89.3
HNRE w/ Aux. Obj.	87.7	82.0	76.8	82.2	88.0	84.4	79.2	83.9	91.0	85.5	82.7	86.4
HNRE w/ Ent. Emb. w/ Aux. Obj.	93.0	85.4	81.9	86.8	94.0	90.1	84.7	89.6	94.0	90.9	87.5	90.8
RHIA-EOP	**96.0**	**92.5**	**86.7**	**91.7**	**98.0**	**95.5**	**92.3**	**95.3**	**98.0**	**96.5**	**93.3**	**95.9**
Ablations												
∼ w/o EOP (RHIA)	95.0	89.5	84.0	89.5	96.0	95.0	90.3	93.8	97.0	96.5	91.0	94.8
∼ w/o RHIA (EOP)	95.0	91.0	86.7	90.9	99.0	92.0	87.7	92.9	99.0	94.5	90.0	94.5
∼ w/o Sent2rel Attention	93.3	91.0	85.0	89.8	95.0	92.0	89.3	92.1	96.0	94.0	91.7	93.9
∼ w/o Attention Pooling	92.0	90.0	87.7	89.9	93.0	93.0	90.7	92.2	94.0	94.0	91.3	93.1
∼ w/o Gating in Eqs. (11)–(12)	95.0	92.5	86.0	91.2	96.0	94.5	90.7	93.7	96.0	95.0	93.3	94.8
∼ w/o Aux. Obj. in Eq. (21)	83.0	81.0	74.0	79.3	89.0	83.5	78.3	83.6	89.0	85.5	80.7	85.1
∼ w/ uncased BERT-Base	94.0	90.4	85.3	89.9	95.2	93.5	90.0	92.9	96.4	93.8	91.9	94.0

The detailed parameter settings are shown in Table 1. For Entity-Aware Embedding module and PCNN module, the parameters keep consistent with those in previous work (Li, Shen et al., 2020), i.e., d_w, d_p, n, λ, ω, d_c. The learning rate and dropout rate are also the same as Han et al. (2018). For parameters of three objective functions, all are consistent with Li, Shen et al. (2020) except for the weight of L_{ord}. To better exploit Entity-Order Perception, we choose the weight of L_{ord} (i.e., ξ) from the list $[0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.5, 2.0, 3.0]$. Here we use three-fold cross-validation on the train set. Each value is trained three times since cross-validation, and the best value (i.e., 1) is determined according to the average of AUC.

For the following experiments, we apply the held-out evaluation to evaluate our model. The entire train set is used for the training phase, while the test set is used for evaluation. The results in Section 4.5 are only intended to illustrate the stability of our model, and are not relevant to cross-validation for hyper-parameter selection.

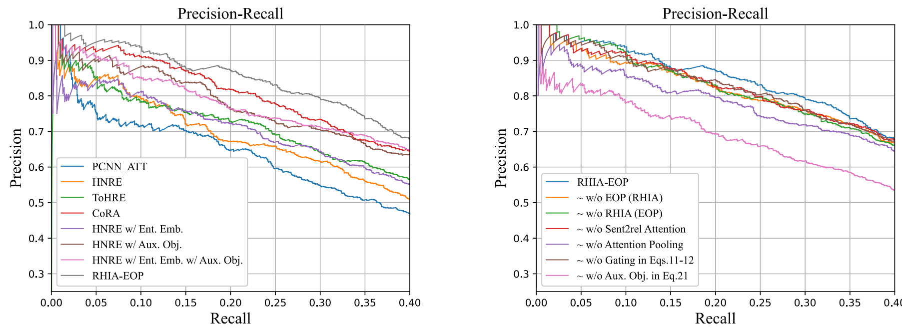
4.2. Baselines

We compare RHIA-EOP with competitive previous baselines that are summarized as follows:

- **PCNN_ATT**: Lin et al. (2016) propose a selective attention among training sentences to mitigate wrong labeling problem, which is the most classical approach in DSRE task.
- **HNRE**: Han et al. (2018) design a hierarchical attention network to enrich bag-level representations, which is the first hierarchical relation extraction baseline.
- **ToHRE**: Yu et al. (2020) introduce a top-down classification strategy and a method to enhance the bag representation in different relation levels.
- **CoRA**: Li, Shen et al. (2020) enhance sentence representations in a collaborative way across all relation levels.
- **HNRE w/ Ent. Emb.**: The model of Han et al. (2018) directly introduces the Entity-Aware Embedding layer in Eqs. (1)–(3).
- **HNRE w/ Aux. Obj.**: The model of Han et al. (2018) directly introduces the hierarchical auxiliary objective in Eq. (21).
- **HNRE w/ Ent. Emb. w/ Aux. Obj.**: The model of Han et al. (2018) introduces both the Entity-Aware Embedding layer in Eqs. (1)–(3) and the hierarchical auxiliary objective in Eq. (21).

4.3. Model comparison results

The comparison results of different models are displayed in Table 2, Table 3 and Fig. 4(a). Note that, the results in Table 2 are obtained using the full test set of the NYT dataset. While the results in Table 3 are obtained using the remaining test set with all single-sentence bags removed as Li, Shen et al. (2020), because for each single-sentence bag, the result is the same whether



(a) model comparison. (b) model comparison.

Fig. 4. Precision–recall (PR) curves.

one, two or all sentences are retained for evaluation. Here the single-sentence bag means a bag consisting of only one sentence.

It can be observed that RHIA-EOP significantly outperforms many baseline models in all metrics at the same time. Our approach achieves the AUC of 0.56, which outperforms the strong baseline CoRA (0.53) by 0.03. For Max_F1, we improve baseline approaches by at least 2.1%. For Top-N precision metric, we take relatively more unique N values in Table 2, and find that RHIA-EOP sets the best scores on all N values. And in another setting, i.e., randomly retaining one, two or all sentence(s) in each bag, RHIA-EOP achieves the best results despite the randomness of retained sentences in Table 3 (almost all values exceed 90%). Besides, the curve of RHIA-EOP is significantly higher than PCNN_ATT, HNRE, ToHRE and CoRA. The results of setups **HNRE w/ Ent. Emb.** and **HNRE w/ Aux. Obj.** indicate that both the Entity-Aware Embedding layer and the hierarchical auxiliary objective bring substantial performance improvements. However, the results of **HNRE w/ Ent. Emb. w/ Aux. Obj.** are still worse than our RHIA-EOP because the heuristic effects between relation levels and the entity order information are still not considered.

To measure RHIA-EOP's ability to handle long-tail relations, we conduct a model comparison on all long-tail relations, and the results are shown in Table 4. Here Hits@K is employed to measure this ability, which indicates whether the ground-label's probability of a bag ranking in the top-K relations. During the calculation, we use macro average regarding different relations. Very inspiringly, RHIA-EOP achieve the highest results at all values of K. These results confirm the heuristic influence of higher-level relations on the lower-level ones and the advantage of relation hierarchies.

Table 4
Hits@K on all long-tail relations. Here the relation is long-tail if its number of instances is less than 100 or 200.

#Instance	<100			<200		
Hits@K	10	15	20	10	15	20
PCNN_ATT (Lin et al., 2016)	<5.0	7.4	40.7	17.2	24.2	51.5
HNRE (Han et al., 2018)	29.6	51.9	61.1	41.4	60.6	68.2
ToHRE (Yu et al., 2020)	62.9	75.9	81.4	69.7	80.3	84.8
CoRA (Li, Shen et al., 2020)	**66.7**	72.2	87.0	**72.7**	77.3	89.3
HNRE w/ Ent. Emb.	36.4	52.2	69.6	54.5	63.6	77.8
HNRE w/ Aux. Obj.	44.4	54.5	68.2	60.6	72.4	81.8
HNRE w/ Ent. Emb. w/ Aux. Obj.	47.8	62.9	77.3	66.7	75.9	87.0
RHIA-EOP	**66.7**	**83.3**	**94.4**	**72.7**	**86.4**	**95.5**
∼ w/ uncased BERT-Base	55.6	66.7	72.2	63.6	77.3	81.8

In addition, we briefly analyze the computational overhead. Both our RHIA-EOP and the CORA of Li, Shen et al. (2020) employ multiple attention mechanisms. We further introduce the interaction effects between relation levels and entity order information, and the number of parameters grow from 13M to 18M. It is still much smaller compared to the pre-trained language models since the smallest BERT-Base has 110M parameters. The parameter growth is not really huge (i.e., 5M). The required training time does not increase dramatically, both models can converge within two hours, while the performance gains are relatively huge and substantial. Therefore, although some computational overhead is added, it is worthwhile.

Table 5
The mean and standard deviation of RHIA-EOP.

Metrics	P@N (Mean)	AUC	Max_F1
Values	80.82 ± 0.676	0.561 ± 0.003	0.546 ± 0.009

Table 6
The mean and standard deviation of RHIA-EOP on NYT when retaining one/two/all sentence(s) in each bag at random.

P@N (Mean)	ONE	TWO	ALL
Values	91.86 ± 0.873	94.36 ± 0.653	95.66 ± 0.224

4.4. Ablation study

In order to validate the effectiveness of each module in RHIA-EOP and to analyze the sources of performance improvement, we evaluate the following ablation experimental setups:

- ∼ **w/o EOP (RHIA)**: The Entity-Order Perception subtask is removed. It is equivalent to having only the Recursive Hierarchy-Interactive Attention (RHIA) module.
- ∼ **w/o RHIA (EOP)**: The module RHIA is replaced by CoRA (Li, Shen et al., 2020). It is equivalent to combining CoRA and EOP.
- ∼ **w/o Sent2rel Attention**: The sentence-to-relation attention (i.e., Eqs. (5)–(10)) is replaced by average pooling.
- ∼ **w/o Attention Pooling**: The attention pooling (i.e., Eq. (17)) is replaced by average pooling.
- ∼ **w/o Gating in Eqs. (11)–(12)**: Eqs. (11)–(12) are removed. Feature concatenation $[u; \hat{c}^{(i)}]$ is directly passed into Multi-Layer Perceptron (MLP).
- ∼ **w/o Aux. Obj. in Eq. (21)**: The hierarchical auxiliary objective L_{hier} is removed (i.e., Eq. (21)).
- ∼ **w/ uncased BERT-Base**: The Entity-Aware Embedding layer is replaced by uncased BERT-Base model.

The results are shown in the bottom of Tables 2 and 3, and Fig. 4(b). And the evaluation results reflect consistent declines in P@N, Max_F1 and AUC. For the two main modules of RHIA-EOP, in Table 2, compared to RHIA-EOP, both RHIA and EOP drop almost 0.015 on AUC, and these two models decrease by 2.5% and 1.4% on the mean of P@N, respectively. While in Table 3, the performance drop is similar obviously, the mean of P@N of RHIA decreases by

Table 7
The Top-3 attention scores in Eq. (5) at all relation levels of two examples from NYT. For the selected relation fact, Example 1 is noisy, while Example 2 is correctly labeled.

Relation fact:	<Grameen Bank, */business/company/founders*, Muhammad Yunus>					
Example 1:	**Muhammad Yunus**, who won the Nobel Peace Prize, last year, demonstrated with, **Grameen Bank** the power of microfinancing.					
Example 2:	On Sunday, though, there was a significant shift of the tectonic plates of Bangladeshi politics, as **Muhammad Yunus** the founder of a microfinance empire, known as the **Grameen Bank** and the winner of the 2006 Nobel Peace Prize, announced that he would start a new party and step into the electoral fray.					
Example 1	$\alpha^{(1)}$		$\alpha^{(2)}$		$\alpha^{(3)}$	
CoRA	*/business*:	0.422	*NA*:	0.383	*NA*:	0.387
	NA:	0.384	*/business/company*:	0.272	*/business/company/founders*:	0.197
	/location:	0.037	*/business/person*:	0.063	*/business/person/company*:	0.063
RHIA-EOP	*NA*:	0.524	*NA*:	0.612	*NA*:	0.559
	/business:	0.324	*/business/company*:	0.095	*/business/person/company*:	0.089
	/people:	0.049	*/business/person*:	0.072	*/business/company/founders*:	0.082
Example 2	$\alpha^{(1)}$		$\alpha^{(2)}$		$\alpha^{(3)}$	
CoRA	*/business*:	0.755	*/business/company*:	0.679	*/business/company/founders*:	0.652
	NA:	0.103	*NA*:	0.089	*NA*:	0.069
	/people:	0.031	*/business/person*:	0.059	*/business/person/company*:	0.057
RHIA-EOP	*/business*:	0.899	*/business/company*:	0.856	*/business/company/founders*:	0.741
	NA:	0.024	*/business/person*:	0.047	*/business/person/company*:	0.052
	/people:	0.022	*NA*:	0.019	*/business/company/major_shareholders*:	0.044

2.2%, 1.5% and 1.1%, respectively. For EOP, the values are 0.8%, 2.4% and 1.4%.

To further analyze whether the performance improvement comes from the new attention mechanism or from the full utilization of relation hierarchies, we set up more experimental settings. The setups $\sim$ **w/o Sent2rel Attention** and $\sim$ **w/o Attention Pooling** argue for the importance and validity of the attention mechanism. As for the relation hierarchies, firstly, for each level of the relation hierarchies, we construct the corresponding feature representation for each bag and classify it. This is the hierarchical auxiliary objective (i.e., Eq. (21)). The setup $\sim$ **w/o Aux. Obj. in Eq.** (21) has demonstrated that not using the relation hierarchies leads to a significant performance drop. Secondly, along the relation hierarchies, relational information is propagated in a recursive form (i.e., RHIA). The setup $\sim$ **w/o EOP (RHIA)** confirms its effectiveness. To sum up, the relation hierarchies are indeed contributing and the attention mechanism is indeed effective.

Interestingly, despite the Pre-trained Language Models (PLMs) are so powerful, the results achieved by the BERT-based implementation are not satisfactory. Except for the values of AUC and Max_F1, the values of all metrics have decreased. The reason may be that BERT cannot highlight two entities as the Entity-Aware Embedding layer does, i.e., it ignores the importance of the entity-pair itself and the position of entities. After all, the essence of DSRE task lies in the classification of entity pairs.

4.5. Statistical analysis of multiple training runs

We reran our model five times to calculate the respective means and standard deviations in terms of AUC, Max_F1 and the mean of P@N. Here for P@N, the randomness of the selected samples when retaining one/two/all sentence(s) in each bag at random leads to large fluctuations in its value, so we only report the mean and standard deviation of "the mean of P@N", i.e., P@N (Mean). Detailed results can be found in Tables 5 and 6. These results demonstrate that the performance gains are stable and convincing.

4.6. Case study

To visualize the heuristic effect between relation hierarchies, two example sentences from NYT are selected. And then we list/analyze the Top-3 attention scores at all relation levels of these two sentences in Table 7. Our RHIA-EOP, compared to CoRA, better handles the highly unbalanced category (i.e., **NA**), and gives greater scores to **NA** for the true instances of **NA** and less scores to **NA** for other instances. This capability can also improve the recognition of noisy sentences from the side, and alleviate the wrong labeling problem. These findings demonstrate the effectiveness of RHIA-EOP.

5. Conclusions

In this paper, we fully exploit the inherent taxonomic structure of relations and design a recursive hierarchy-interactive attention network. In this way, we model the heuristic influence of higher-level relations on the lower-level ones. Furthermore, multiple training objectives are designed to take entity-order information into account. The substantial experiments on the NYT dataset show that RHIA-EOP achieves state-of-the-art performance in multiple metrics, including standard metrics (i.e., AUC, P–R curve, Top-N precision, etc.) and long-tail metrics (i.e., Hits@K). In the future, we plan to consider the interaction between the many sibling relations within the same level and employ our recursive approach to other tasks, such as fine-grained hierarchical text classification. Besides, we will further explore the extension of traditional feature-based methods in the deep learning paradigm.

CRediT authorship contribution statement

Ridong Han: Methodology, Conceptualization, Software, Visualization, Writing, Editing. **Tao Peng:** Supervision, Funding acquisition, Reviewing, Validation. **Jiayu Han:** Editing, Reviewing. **Hai Cui:** Formal analysis, Validation. **Lu Liu:** Reviewing, Supervision, Validation, Funding acquisition.

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, China under grant No. 20210201131GX, and Jilin Provincial Education Department Project, China under grant No. JJKH20190160KJ.

References

Ba, L. J., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. CoRR, abs/1607.06450 arXiv:1607.06450.

Bollacker, K. D., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the ACM SIGMOD international conference on management of data* (pp. 1247–1250). ACM, http://dx.doi.org/10.1145/1376616.1376746.

Bordes, A., Usunier, N., García-Durán, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In *Proceedings of the 27th annual conference on neural information processing systems* (pp. 2787–2795).

Chen, Y., Yang, W., Wang, K., Qin, Y., Huang, R., & Zheng, Q. (2021). A neuralized feature engineering method for entity relation extraction. *Neural Networks*, *141*, 249–260. http://dx.doi.org/10.1016/j.neunet.2021.04.010.

Cotter, A., Shamir, O., Srebro, N., & Sridharan, K. (2011). Better mini-batch algorithms via accelerated gradient methods. In *Proceedings of the 25th annual conference on neural information processing systems* (pp. 1647–1655).

Dai, L., Xu, B., & Song, H. (2019). Feature-level attention based sentence encoding for neural relation extraction. In *Lecture notes in computer science*: *vol. 11838*, *Natural language processing and chinese computing - 8th CCF international conference* (pp. 184–196). Springer, http://dx.doi.org/10.1007/978-3-030-32233-5_15.

Defferrard, M., Bresson, X., & Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems 29: annual conference on neural information processing systems 2016* (pp. 3837–3845).

Deng, L., Yang, B., Kang, Z., Yang, S., & Wu, S. (2021). A noisy label and negative sample robust loss function for DNN-based distant supervised relation extraction. *Neural Networks*, *139*, 358–370. http://dx.doi.org/10.1016/j.neunet.2021.03.030.

Du, J., Han, J., Way, A., & Wan, D. (2018). Multi-level structured self-attentions for distantly supervised relation extraction. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 2216–2225). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/d18-1245.

Geng, Z., Chen, G., Han, Y., Lu, G., & Li, F. (2020). Semantic relation extraction using sequential and tree-structured LSTM with attention. *Information Sciences*, *509*, 183–192. http://dx.doi.org/10.1016/j.ins.2019.09.006.

Han, X., Yu, P., Liu, Z., Sun, M., & Li, P. (2018). Hierarchical relation extraction with coarse-to-fine grained attention. In *Proceedings of the 2018 conference on empirical methods in natural language processing* (pp. 2236–2245). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/d18-1247.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE conference on computer vision and pattern recognition* (pp. 770–778). IEEE Computer Society, http://dx.doi.org/10.1109/CVPR.2016.90.

Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L. S., & Weld, D. S. (2011). Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies* (pp. 541–550). The Association for Computer Linguistics.

Huang, Q., Zhu, S., Feng, Y., Ye, Y., Lai, Y., & Zhao, D. (2021). Three sentences are all you need: Local path enhanced document relation extraction. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing* (pp. 998–1004). http://dx.doi.org/10.18653/v1/2021.acl-short.126.

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

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., et al. (2015). DBpedia - a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, *6*(2), 167–195. http://dx.doi.org/10.3233/SW-140134.

Li, Y., Long, G., Shen, T., Zhou, T., Yao, L., Huo, H., et al. (2020). Self-attention enhanced selective gate with entity-aware embedding for distantly supervised relation extraction. In *Proceedings of the thirty-fourth AAAI conference on artificial intelligence* (pp. 8269–8276). AAAI Press.

Li, Y., Shen, T., Long, G., Jiang, J., Zhou, T., & Zhang, C. (2020). Improving long-tail relation extraction with collaborating relation-augmented attention. In *Proceedings of the 28th international conference on computational linguistics* (pp. 1653–1664). International Committee on Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.coling-main.145.

Lin, Z., Feng, M., dos Santos, C. N., Yu, M., Xiang, B., Zhou, B., et al. (2017). A structured self-attentive sentence embedding. In *Proceedings of the 5th international conference on learning representations*. OpenReview.net.

Lin, Y., Shen, S., Liu, Z., Luan, H., & Sun, M. (2016). Neural relation extraction with selective attention over instances. In *Proceedings of the 54th annual meeting of the association for computational linguistics* (pp. 2124–2133). Berlin, Germany: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/P16-1200.

Liu, T., Wang, K., Chang, B., & Sui, Z. (2017). A soft-label method for noise-tolerant distantly supervised relation extraction. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 1790–1795). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/d17-1189.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 27th annual conference on neural information processing systems* (pp. 3111–3119).

Mintz, M., Bills, S., Snow, R., & Jurafsky, D. (2009). 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* (pp. 1003–1011). The Association for Computer Linguistics.

Peng, T., Han, R., Cui, H., Yue, L., Han, J., & Liu, L. (2022). Distantly supervised relation extraction using global hierarchy embeddings and local probability constraints. *Knowledge Based System*, *235*, Article 107637. http://dx.doi.org/10.1016/j.knosys.2021.107637.

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

Riedel, S., Yao, L., & McCallum, A. (2010). Modeling relations and their mentions without labeled text. In *Lecture Notes in Computer Science*: *vol. 6323*, *Machine learning and knowledge discovery in databases, european conference* (pp. 148–163). Springer, http://dx.doi.org/10.1007/978-3-642-15939-8_10.

Shen, T., Zhou, T., Long, G., Jiang, J., Pan, S., & Zhang, C. (2018). Di[san]: Directional self-attention network for RNN/CNN-Free language understanding. In *Proceedings of the thirty-second aaai conference on artificial intelligence* (pp. 5446–5455). AAAI Press.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929–1958.

Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on world wide web* (pp. 697–706). ACM, http://dx.doi.org/10.1145/1242572.1242667.

Surdeanu, M., Tibshirani, J., Nallapati, R., & Manning, C. D. (2012). Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 455–465). ACL.

Xiao, Y., Tan, C., Fan, Z., Xu, Q., & Zhu, W. (2020). Joint entity and relation extraction with a hybrid transformer and reinforcement learning based model. In *Proceedings of the thirty-fourth AAAI conference on artificial intelligence* (pp. 9314–9321). AAAI Press.

Xu, W., Chen, K., & Zhao, T. (2021). Document-level relation extraction with reconstruction. In *Proceedings of the thirty-fifth AAAI conference on artificial intelligence* (pp. 14167–14175). URL https://ojs.aaai.org/index.php/AAAI/article/view/17667.

Yang, J., Wang, Q., Su, C., & Wang, X. (2020). Threat intelligence relationship extraction based on distant supervision and reinforcement learning (s). In *Proceedings of the 32nd international conference on software engineering and knowledge engineering* (pp. 572–576). KSI Research Inc., http://dx.doi.org/10.18293/SEKE2020-149.

Yang, S., Zhang, Y., Niu, G., Zhao, Q., & Pu, S. (2021). Entity concept-enhanced few-shot relation extraction. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing* ACL/IJCNLP, (pp. 987–991). http://dx.doi.org/10.18653/v1/2021.acl-short.124.

Yao, Y., Ye, D., Li, P., Han, X., Lin, Y., Liu, Z., et al. (2019). DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th conference of the association for computational linguistics* (pp. 764–777). http://dx.doi.org/10.18653/v1/p19-1074.

Ye, Z., & Ling, Z. (2019). Distant supervision relation extraction with intra-bag and inter-bag attentions. In *Proceedings of the 2019 conference of the north american chapter of the association for computational linguistics: human language technologies* (pp. 2810–2819). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/n19-1288.

Yu, E., Han, W., Tian, Y., & Chang, Y. (2020). 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* (pp. 1665–1676). International Committee on Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.coling-main.146.

Yu, B., Zhang, Z., Liu, T., Wang, B., Li, S., & Li, Q. (2019). Beyond word attention: Using segment attention in neural relation extraction. In *Proceedings of the twenty-eighth international joint conference on artificial intelligence* (pp. 5401–5407). ijcai.org, http://dx.doi.org/10.24963/ijcai.2019/750.

Yuan, Y., Liu, L., Tang, S., Zhang, Z., Zhuang, Y., Pu, S., et al. (2019). Cross-relation cross-bag attention for distantly-supervised relation extraction. In *Proceedings of the thirty-third AAAI conference on artificial intelligence* (pp. 419–426). AAAI Press, http://dx.doi.org/10.1609/aaai.v33i01.3301419.

Zeng, D., Liu, K., Chen, Y., & Zhao, J. (2015). Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1753–1762). The Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/d15-1203.

Zeng, D., Liu, K., Lai, S., Zhou, G., & Zhao, J. (2014). Relation classification via convolutional deep neural network. In *Proceedings of the 25th international conference on computational linguistics* (pp. 2335–2344).

Zhang, N., Deng, S., Sun, Z., Wang, G., Chen, X., Zhang, W., et al. (2019). 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* (pp. 3016–3025). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/n19-1306.

Zhou, Y., Pan, L., Bai, C., Luo, S., & Wu, Z. (2021). Self-selective attention using correlation between instances for distant supervision relation extraction. *Neural Networks*, *142*, 213–220. http://dx.doi.org/10.1016/j.neunet.2021.04.032.

Zhu, T., Wang, H., Yu, J., Zhou, X., Chen, W., Zhang, W., et al. (2020). Towards accurate and consistent evaluation: A dataset for distantly-supervised relation extraction. In *Proceedings of the 28th international conference on computational linguistics* (pp. 6436–6447). International Committee on Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.coling-main.566.