ToHRE: A Top-Down Classification Strategy with Hierarchical Bag Representation for Distantly Supervised Relation Extraction

Erxin Yu[1,2], **Wenjuan Han**[5], **Yuan Tian**[1,2,4*], **Yi Chang**[1,2,3*]
[1]School of Artificial Intelligence, Jilin University
[2]Key Laboratory of Symbolic Computation and Knowledge Engineering, Jilin University
[3]International Center of Future Science, Jilin University
[4]College of Biological and Agricultural Engineering, Jilin University
[5]Department of Computer Science, National University of Singapore
yuex19@mails.jlu.edu.cn, dcshanw@nus.edu.sg,
yuantian@jlu.edu.cn, yichang@jlu.edu.cn

Abstract

Distantly Supervised Relation Extraction (DSRE) has proven to be effective to find relational facts from texts, but it still suffers from two main problems: the wrong labeling problem and the long-tail problem. Most of the existing approaches address these two problems through flat classification, which lacks hierarchical information of relations. To leverage the informative relation hierarchies, we formulate DSRE as a hierarchical classification task and propose a novel hierarchical classification framework, which extracts the relation in a top-down manner. Specifically, in our proposed framework, 1) we use a hierarchically-refined representation method to achieve hierarchy-specific representation; 2) a top-down classification strategy is introduced instead of training a set of local classifiers. The experiments on NYT dataset demonstrate that our approach significantly outperforms other state-of-the-art approaches, especially for the long-tail problem.

1 Introduction

Knowledge bases (KBs) such as Freebase (Bollacker et al., 2008), DBpedia (Auer et al., 2007), and NELL (Carlson et al., 2010) currently play an essential role in NLP tasks including information retrieval and question answering. As current KBs are still far from complete compared with the infinite real-world facts, relation extraction (RE) which aims to extract relations between two entities in texts and enrich KBs has attracted a surge of research interest. Most existing supervised RE methods require high quality labeled data, which is time-consuming and labor-intensive. Therefore, Mintz et al. (2009) propose the distant supervision (DS) approach to automatically generate a large amount of training data for RE by aligning KBs with large-scale unlabeled corpora.

Two problems have to be addressed in building an efficient DS model. The main problem, namely the wrong labeling problem, comes from the assumption of DS. DS is based on an assumption: if two entities have a relation in KBs, all sentences that contain these two entities will express this relation. This strong and unrealistic assumption inevitably bring some noise. For example, DS method may mistakenly label the sentence "*Steve Jobs* passed away the day before *Apple* unveiled iPhone 4s in late 2011" with the relation *business/company/founders*. Secondly, although DS method can automatically generate large-scale training data, this training data can only cover a limited part of relations. Nearly 70% of the relations are long-tail and still suffer from data deficiency in the widely used New York Times (NYT) dataset. To alleviate the wrong labeling problem, Riedel et al. (2010) and Hoffmann et al. (2011) propose the multi-instance learning (MIL) framework, which assigns a label to a bag of sentences containing a common entity pair. Based on the MIL framework, many efforts (Zeng et al., 2015; Lin et al., 2016; Du et al., 2018; Ye and Ling, 2019) have been devoted to identifying valid sentences from labeled bags. However, they ignore the hierarchical relation information. For example, the relations in Freebases are labeled as shown in Figure 1 (LEFT). Considering the correlations among relations,

[*]Joint Corresponding Author

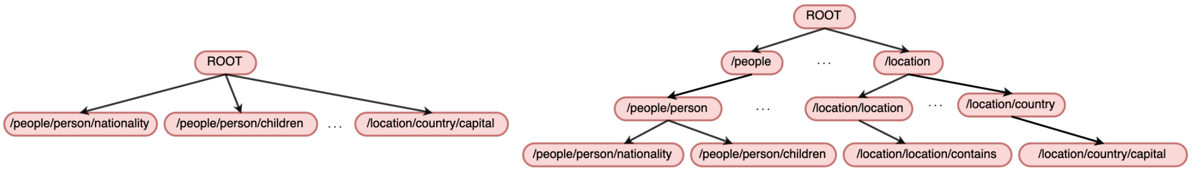

Figure 1: Illustration of the relations. Left: an example in Freebases. Right: the corresponding class hierarchy.

relations can also be naturally organized as a class hierarchy (Han et al., 2018)–typically like a tree shown in Figure 1 (RIGHT). Each layer in the relation hierarchies has its semantic information. Han et al. (2018) and Zhang et al. (2019) take advantage of this relation hierarchies and propose a hierarchical attention scheme to simultaneously solve the noise problem and achieve state-of-the-art performance in extracting long-tail relations. Nevertheless, they are based on a flat classification approach without fully exploring the informative relation hierarchies.

To leverage the inherent hierarchical structure of relations, we formulate DSRE as a hierarchical classification task, which extracts relations in a top-down manner. Intuitively, coarse-grained relations in the top level are easy whereas fine-grained relations in the bottom layer are harder to classify. In this way, we can preferentially extract the relation in the top level and then use the top-level relation to boost the performance of the relation in the bottom level. There exist two challenges when conducting a hierarchical top-down classification: *capturing specific bag representations in different levels* and *training large amounts of classifiers*. For the challenge of the bag representation, a bag of sentences expresses different relations in different levels. Hence, it is necessary to dynamically adjust the bag representations in different relation levels. To capture the hierarchy-specific bag representation, we propose a hierarchical bag representation method, which incorporates the hierarchically-refined selective attention mechanism to dynamically adjust the bag representation in different levels. For the challenge of massive classifiers, traditional hierarchical classification methods which adopt the top-down manner need to train a set of local classifiers (D'Alessio et al., 2000; Clare and King, 2003; Holden and Freitas, 2009). The number of local classifiers depends on the size of the label hierarchies, making hierarchical classification infeasible to scale. To handle the massive local classifiers problem, we introduce a top-down classification strategy which shares most of its parameters in different levels to avoid training massive local classifiers, thus making our methods available for various relation hierarchies.

We name our approach A **To**p-Down Classification Strategy for **H**ierarchical **R**elation **E**xtraction (**ToHRE**). Our main contributions are summarized as follows:

- To the best of our knowledge, we are the first to explore the feasibility of the hierarchical classification in Distantly Supervised Relation Extraction.
- We design a hierarchically-refined representation method to enhance the bag representation in different relation levels and a top-down classification strategy to avoid training massive local classifiers.
- We conduct thorough experiments on the widely-used NYT dataset and achieve significant improvements over state-of-the-art models, especially for long-tail relations.

2 Methodology

2.1 Overview

Problem Definition Following the MIL setting, we split entire sentences into multiple entity-pair bags $\{B_{h_1,t_1}, B_{h_2,t_2}, ...\}$. Each entity-pair bag B_{h_i,t_i} contains m sentences $\{s_1, s_2, ..., s_m\}$ mentioning both the entities h_i and t_i. Each sentence is a sequence of tokens, i.e., $s = \{w_1, w_2, ..., w_n\}$, where n is the length of the sentence. Besides, we define the relation classes as $\mathcal{R} = \{r_1, r_2, ...\}$. Given an entity-pair bag and two corresponding entities, the previous works are all focusing on flat classification, which directly label the bag with a predicted relation r from pre-defined relation classes $\mathcal{R}$.

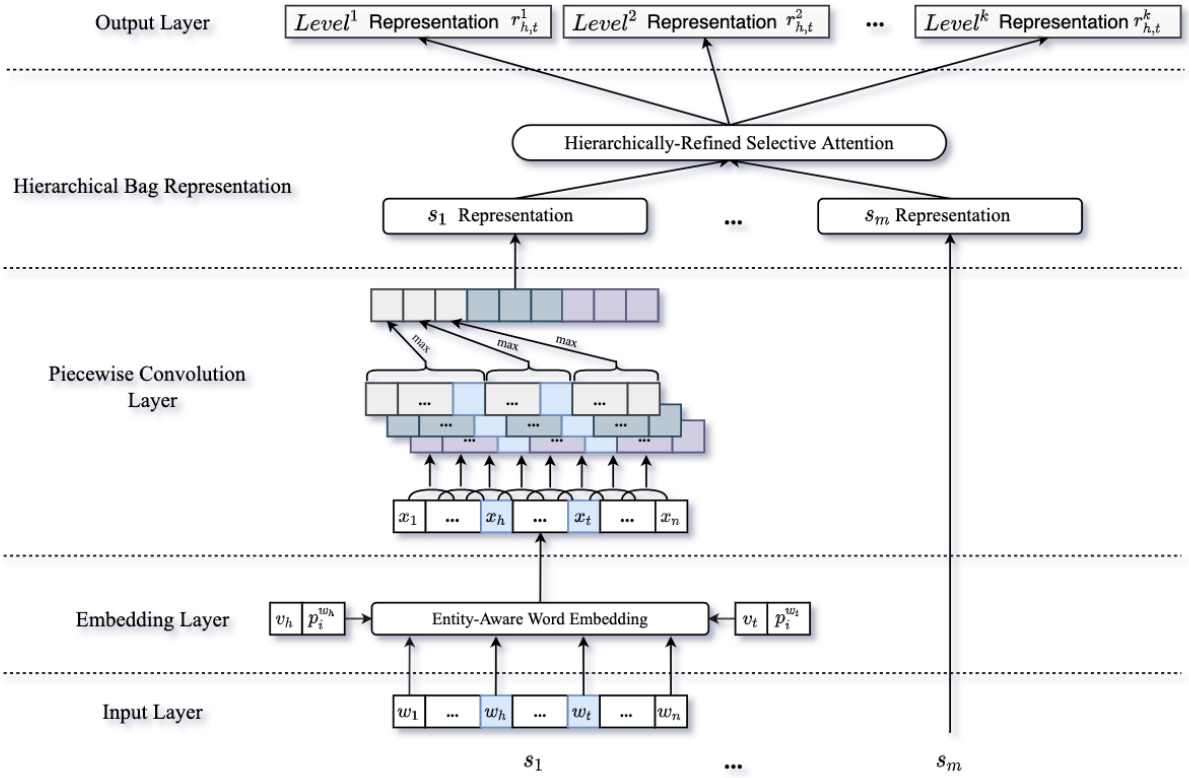
Output Layer
$Level^1$ Representation $r^1_{h,t}$
$Level^2$ Representation $r^2_{h,t}$
...
$Level^k$ Representation $r^k_{h,t}$
Hierarchically-Refined Selective Attention
Hierarchical Bag Representation
s_1 Representation
...
s_m Representation
Piecewise Convolution Layer
max
max
max
x_1 ... x_h ... x_t ... x_n
Embedding Layer
v_h $p_i^{w_h}$
Entity-Aware Word Embedding
v_t $p_i^{w_t}$
Input Layer
w_1 ... w_h ... w_t ... w_n
s_1
...
s_m

Figure 2: Illustration of hierarchical bag presentation.

To leverage the inherent hierarchical structure of relations to conduct hierarchical relation extraction, we define a relation hierarchy $H = (L, E)$ as a tree-structured hierarchy with a set of nodes (i.e., relations) L and a set of edges E indicating the relationship between the parent node and its child node. As illustrated in Figure 1, the leaf nodes in H are made up of the pre-defined relation classes $\mathcal{R}$. Hence, all leaf nodes are base-level relations (e.g., */people/person/place_of_birth*). We generate the corresponding higher-level relations (e.g., */people/person* and */people*) as their parent nodes. Specifically, for a relation r in the leaf node, we generate k times to construct its hierarchical chain of parent relations $\{r^0, r^1, ..., r^k\}$, where r^{i-1} is the parent relation of r^i. It is worth noting that r^0 is the virtual root relation and r^k is the base-level relation, namely r.

Different from previous methods, our proposed model aims to explore the relation hierarchy H in a top-down manner and output relation probability in each relation level. The entity-pair bag along with the history of parent relations are integrated to predict a relation from pre-defined relation classes $\mathcal{R}$.

Framework Our model consists of two key components: hierarchical bag representation module and a top-down classification strategy. The hierarchical bag representation module is shown in Figure 2. It takes an entity-pair bag as input and outputs the bag representations in different relation levels. First, each sentence in the entity-pair bag is transformed to a matrix with the entity-aware embedding. Then, a Piecewise Convolution layer (Zeng et al., 2015) is used to obtain the sentence representation. After that, a hierarchically-refined selective attention is leveraged to select sentences in the bag which actually expresses the corresponding hierarchical relation and output hierarchy-specific bag representation. The top-down classification strategy is illustrated in Figure 3. The designed strategy takes the hierarchical bag representation and corresponding relation embeddings as input to discriminate its child relations, then walks down the relation hierarchies to predict the next relation until the leaf node. The details of the two components are introduced in Section 2.2 and Section 2.3 respectively.

2.2 Hierarchical Bag Representation

As mentioned in Section 2.1, the hierarchically-refined representation method aims to obtain bag representations in different relation levels. Specifically, given a bag of sentences $B_{h,t} = \{s_1, s_2, ..., s_m\}$ and two corresponding entities w_h and w_t [1], we aim to obtain k-level bag representations $\left\{\mathbf{r}^1_{h,t}, \mathbf{r}^2_{h,t}, ..., \mathbf{r}^k_{h,t}\right\}$.

Entity-Aware Embedding For a sentence $s = \{w_1, w_2, ..., w_n\}$ in the bag, each word w_i is transformed into a low-dimensional dense-vector representation, i.e., $[\mathbf{v}_1, ..., \mathbf{v}_n] \in \mathbb{R}^{d_a \times n}$, where d_a denotes the dimension of word embedding. Besides, relative position (Zeng et al., 2015) is a crucial feature for relation extraction model to specify the target entity pair and make the model pay more attention to the words close to the target entities. For i-th word, the relative position features can be represented by two dense-vectors $\mathbf{p}_i^{w_h}$ and $\mathbf{p}_i^{w_t} \in \mathbb{R}^{d_b}$. Previous works concatenate the word embedding and position embedding for i-th word, i.e., $\mathbf{x}_i = [\mathbf{v}_i, \mathbf{p}_i^{w_h}, \mathbf{p}_i^{w_t}]$.

In addition to relative position features, Li et al. (2020) verify both the head entity embedding $\mathbf{v}_h$ and tail entity embedding $\mathbf{v}_t$ are vitally important for the relation extraction task. They use a position-wise gate mechanism to dynamically select features between relative position embeddings and entity embeddings. To go a step further, we argue that head entity and tail entity are not equally important in extracting hierarchical relations. To this end, we propose an entity-aware embedding method to dynamically enhance entity information in word representation. Formally, we denote the head-entity-enhanced embedding as $\mathbf{X}^{(h)} = [\mathbf{x}_1^{(h)}, \mathbf{x}_2^{(h)}, \cdots, \mathbf{x}_n^{(h)}] \in \mathbb{R}^{d_w \times n}$ in which $\mathbf{x}_i^{(h)} = [\mathbf{v}_i, \mathbf{v}_h, \mathbf{p}_i^{w_h}]$ and $d_w = d_b + 2 \times d_a$. The tail-entity-enhanced embedding is denoted as $\mathbf{X}^{(t)}$ in a similar way. Then we use a selective mechanism to dynamically select features between $\mathbf{X}^{(h)}$ and $\mathbf{X}^{(t)}$. The selective mechanism is formulated as:

$$\alpha = \mathrm{sigmoid}(\mathbf{W}^{(s)}\mathbf{X}^{(h)} + \mathbf{b}^{(s)}) \tag{1}$$

$$\mathbf{X} = \alpha \cdot \mathbf{X}^{(h)} + (1 - \alpha) \cdot \mathbf{X}^{(\mathbf{t})} \tag{2}$$

in which $\mathbf{W}^{(s)} \in \mathbb{R}^{d_h \times d_w}$ and bias vector $\mathbf{b}^{(s)}$ are parameters, and $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbb{R}^{d_h \times n}$ is the final input representation for sentence s.

Piecewise Convolution Neural Network We employ the Piecewise Convolution Neural Network (PCNN) as the sentence encoder to map the aforementioned input representation $\mathbf{X}$ into a sentence representation. Note that, our hierarchically-refined representation method is designed independently of the sentence encoder. Hence, other neural networks such as RNN can be the alternative and used as the sentence encoder in our approach. Since previous works show that PCNN can achieve state-of-the-art performance and give time efficiency, we select it as the sentence encoder in this paper. PCNN mainly consists of two parts: the convolution layer and piecewise max-pooling.

The convolution layer applies a kernel of window size l to slide over the input representation $[\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]$ and output the d_c dimensional hidden embedding $\mathbf{h}$, where $\mathbf{h} \in \mathbb{R}^{d_c \times n}$ and d_c is the number of feature maps.

Then, a piecewise max-pooling method (Zeng et al., 2015) is applied on the hidden embeddings. The hidden embedding $\mathbf{h}$ is firstly divided into three parts $\left\{\mathbf{h}^{(1)}, \mathbf{h}^{(2)}, \mathbf{h}^{(3)}\right\}$ by the position of head and tail entities. After that, we perform max-pooling on each part respectively and concatenate the pooling results to get the final embedding $\mathbf{u}$:

$$\mathbf{u} = [\max(\mathbf{h}^{(1)}), \max(\mathbf{h}^{(2)}), \max(\mathbf{h}^{(3)})] \tag{3}$$

where $\mathbf{u} \in \mathbb{R}^{3d_c}$ is the final sentence representation.

Hierarchically-Refined Selective Attention Given a bag of sentences $B_{h,t} = \{s_1, s_2, ..., s_m\}$, we already achieve the sentence embeddings $\mathbf{U}_{h,t} = \{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_m\}$ through input representation and PCNN encoder. For a relation $r \in \mathcal{R}$, we generate its k-level chain of parent relations $\left\{r^0, r^1, ..., r^k\right\}$. In this section, we aim to identify valid sentences in different relation levels and obtain hierarchy-specific bag representations, i.e., $\left\{\mathbf{r}^1_{h,t}, \mathbf{r}^2_{h,t}, ..., \mathbf{r}^k_{h,t}\right\}$

We propose a hierarchically-refined selective attention mechanism to identify valid sentences in k-level relation hierarchies. Specifically, the bag presentation of j-level ($1 \leq j \leq k$) relation level is

[1] For a clear demonstration, we omit indices of examples.

formulated as:

$$\mathbf{r}^j_{h,t} = \sum_{i=1}^{m} \beta^j_i \mathbf{u}_i, \qquad \mathbf{u}_1, ..., \mathbf{u}_m \in \mathbf{U}_{h,t} \tag{4}$$

Here β^j_i is the hierarchically-refined attention weight for i-th sentence in j-th relation level, which is further defined as:

$$\beta^j_i = \frac{\exp(e^j_i)}{\sum_{g=1}^{m} \exp(e^j_g)} \tag{5}$$

where e^j_i is referred as a query-based function which scores how well the input sentence $\mathbf{u}_i$ and the j-th level of predicted relation r matches. We denote $\mathbf{q}_{r^j}$ as layer-wise query vector associated with j-th level of relation r to compute e^j_i:

$$e^j_i = \mathbf{u}_i \mathbf{A} \mathbf{q}_{r^j} \tag{6}$$

where $\mathbf{A}$ is a weighted diagonal matrix. Finally, the hierarchical bag representation $\left\{\mathbf{r}^1_{h,t}, \mathbf{r}^2_{h,t}, ..., \mathbf{r}^k_{h,t}\right\}$ is obtained. Different from works of Han et al. (2018) and Zhang et al. (2019), we directly employ bag representations of different levels as critical features for hierarchical relation extraction instead of concatenating them to conduct flat classification.

2.3 Top-Down Classification Strategy

In this section, a novel top-down classification strategy for DSRE is proposed to explore the relation hierarchy H step by step. We first define some notations used and then introduce the novel strategy in detail.

Definition For each relation in the hierarchy H, the relation embedding $\mathbf{l} \in \mathbb{R}^{\mathcal{C}}$ is randomly initialized and updated during training. In level j, we define the current parent node as r^j and the child nodes of r^j as $C(r^j)$. Furthermore, an embedding matrix consists of $C(r^j)$ is denoted as $\mathbf{C}_j \in \mathbb{R}^{|C(r^j)| \times \mathcal{C}}$.

Top-Down Manner The proposed top-down manner starts from the virtual root node and goes down one level on the hierarchy and stops at a leaf node. Notably, only the ground-truth hierarchical relations are visited during the training phase, e.g., for a bag labeled with relation r, only the nodes in its hierarchical chain $\left\{r^0, r^1, ..., r^k\right\}$ are visited. While in the testing phase, the proposed top-down manner visits each node in the relation hierarchy H and outputs local probabilities for each node.

Local Classification Strategy In each level j, we aim to conduct local classification, which outputs the probability distribution of $C(r^j)$, i.e., the child node probabilities of r_j. Traditional methods train a set of massive classifiers for each node or each parent node in relation hierarchy H, making them infeasible to scale. Inspired by (Mao et al., 2019), who propose a top-down supervised method to pre-train a reinforcement learning model for hierarchical classification, we design a local classification strategy for DSRE which calculates the matching score between the hierarchical bag representation and the candidate embedding matrix.

Specifically when conducting local classification in level j, the embedding of current relation $\mathbf{l}^j$ and the bag representation $\mathbf{r}^{j+1}_{h,t}$ are concatenated and projected to a hidden state vector $\mathbf{s}_j \in \mathbb{R}^{\mathcal{C}}$ via a two-layer feed-forward network. Then the candidate embedding matrix $\mathbf{C}_j$ is multiplied with the hidden state vector $\mathbf{s}_j$ to obtain the local probability $p(r^{j+1}|r^j, r^{j+1}_{h,t}, \theta)$, i.e., the ground-truth child relation probability of r^j. The aforementioned process is formulated as:

$$\mathbf{s}_j = \mathrm{ReLU}\Big(W^1_l \mathrm{ReLU}\big(W^2_l [\mathbf{l}_j; \mathbf{r}^{j+1}_{h,t}]\big)\Big) \tag{7}$$

$$p(r^{j+1}|r^j, r^{j+1}_{h,t}, \theta) = \mathrm{softmax}(\mathbf{C}_j \mathbf{s}_j) \tag{8}$$

2.4 Training and Testing

Here we introduce the learning and optimization details of our model. During the training process, we minimize the cross entropy loss function. Given the collection of entity-pair bags $\mathbf{B} =$

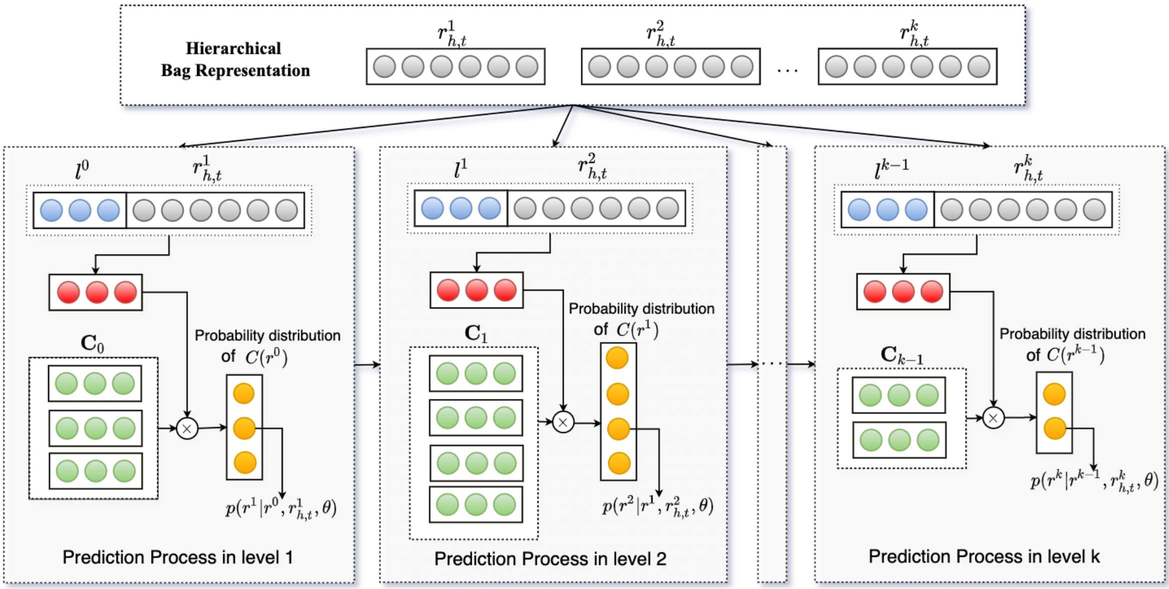
Hierarchical Bag Representation
$r^1_{h,t}$
$r^2_{h,t}$
...
$r^k_{h,t}$
l^0
$r^1_{h,t}$
$\mathbf{C}_0$
Probability distribution of $C(r^0)$
$p(r^1|r^0, r^1_{h,t}, \theta)$
Prediction Process in level 1
l^1
$r^2_{h,t}$
$\mathbf{C}_1$
Probability distribution of $C(r^1)$
$p(r^2|r^1, r^2_{h,t}, \theta)$
Prediction Process in level 2
l^{k-1}
$r^k_{h,t}$
$\mathbf{C}_{k-1}$
Probability distribution of $C(r^{k-1})$
$p(r^k|r^{k-1}, r^k_{h,t}, \theta)$
Prediction Process in level k

Figure 3: Illustration of the top-down classification strategy.

$\{B_{h_1,t_1}, B_{h_2,t_2}, ...\}$ and corresponding labeled relations $\{r_1, r_2, ...,\}$, We defined the hierarchical loss as followings:

$$J(\theta) = -\frac{1}{|B|}\sum_{i=1}^{|B|}\sum_{j=0}^{k-1}\log p(r_i^{j+1}|r_i^j, r_{h_i,t_i}^{j+1}, \theta) + \lambda\,\|\theta\|_2^2 \tag{9}$$

where λ is a harmonic factor, and $\|\theta\|_2^2$ is the regularizer defined as L_2 normalization. All models are optimized using stochastic gradient descent (SGD).

In the testing phase, we output the final probability of relation r for $B_{h,t}$ by multiplying the probabilities of its hierarchical chain of parent relations:

$$p(r|B_{h,t}) = \prod_{j=0}^{k-1} p(r^{j+1}|r^j, r_{h,t}^{j+1}, \theta) \tag{10}$$

3 Experiments

3.1 Dataset and Evaluation Metrics

We evaluate our proposed model on the New York Times (NYT) dataset (Riedel et al., 2010), which has been widely used for distantly supervised relation extraction (Zeng et al., 2015; Lin et al., 2016; Zhang et al., 2019). This dataset is constructed by aligning relation facts in Freebase with the New York Times corpus. There are 52 actual relations and a special **NA** relation indicating that there is no relation between two entities. Specially, we put **NA** relation as a leaf node in H and directly link it to the virtual root node.

Following the previous works from Zeng et al. (2015) and Lin et al. (2016), we evaluate our model using the held-out evaluation. In the held-out evaluation, the relations extracted from test data are automatically compared with those in Freebase. It is an approximate measure of the model without requiring costly human evaluation. We report the precision-recall (PR) curves, top-N precision (P@N) and accuracy of Hits@K in our experiments.

Hyper-Parameter	Value
Word Dimension d_a	50
Position Dimension d_b	5
Window Size l	3
Sentence Embedding Size d_h	300
Relation Embedding Size $\mathcal{C}$	50
Learning Rate λ	0.5
Dropout Probability	0.5
Batch Size	160

Table 1: Detailed hyper-parameters.

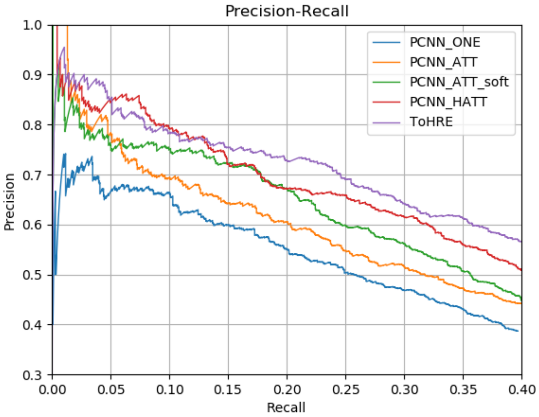

Figure 4: Precision-recall curves.

3.2 Setup

We use the pre-trained word embeddings released by Lin et al. (2016)[2] for initialization. The vocabulary contains the words which appeared more than 100 times in the NYT corpus. We apply the dropout strategy (Srivastava et al., 2014) to the hidden state vector $\boldsymbol{s}_j$ to prevent overfitting. Besides, we pre-train the PCNN before training our hierarchical classification model. Table 1 shows all the hyper-parameters used in our experiments.

3.3 Overall Performance

Precision-Recall Curves We compare the precision-recall curves of our model with several major baselines to evaluate the overall performance in Figure 4. We report the results of previous RE models which adopt the PCNN architecture as sentence encoder: the original PCNN and its variants with different attention-based modules. More specifically, **+ONE** (Zeng et al., 2015) indicates the original PCNN under MIL setting; **+ATT** (Lin et al., 2016) is a plain selective attention over sentences; **+ATT+SL** (Liu et al., 2017) combines the ATT scheme with soft-label method to solve the bag-level noise; **+HATT** (Han et al., 2018) leverages relation hierarchies to calculate a coarse-to-fine grained attention, which is the principal baseline of our work; **ToHRE** is the abbreviation of our proposed framework.

As shown in Figure 4, our model achieves the best result among all attention-based models. Even compared with PCNN+HATT, our model achieves higher precision over most part of the entire range of recall. It indicates the ability of our proposed model to handle the noise problem in the RE task.

P@N To further verify the effectiveness of our proposed model, we compare our model with previous state-of-the-art approaches on the Top-N precision. We briefly introduce these SOTA models: **RESIDE** (Vashishth et al., 2018) utilizes the available side information from knowledge bases, including entity types and relation alias information; **DISTRE** (Alt et al., 2019) is a Transformer which combines an attentive selection mechanism for the multi-instance learning scenario; **PCNN+BAG_ATT** (Ye and Ling, 2019) combines both intra-bag and inter-bag attentions to cope with the noisy sentence and noisy bag problems in DSRE and achieve the SOTA performance in terms of top-n precision metric. The evaluation results are shown in Table 2. It can be observed that: (1) ToHRE outperforms previous methods in most cases from P@100 to P@2000 indicating that our model have a consistent performance. (2) ToHRE has the highest precision in P@100, which is critical to some NLP tasks like Knowledge Base Completion that need convincing prediction at top-100 to obtain high-quality relational triple.

3.4 Evaluation Result for Long-tail Relations

To demonstrate the improvements in performance for long-tail relations, we follow the work from Han et al. (2018) to evaluate our model on a subset of test dataset in which all the relations have fewer than

[2] https://github.com/thunlp/OpenNRE.

Approach	P@100	P@200	P@300	P@500	P@1000	P@2000	Mean
PCNN+ATT†	73.0	68.0	67.3	63.6	53.3	40.0	60.9
PCNN+HATT‡	81.0	79.5	75.7	68.0	58.7	42.1	67.5
RESIDE†	81.8	75.4	74.3	69.7	59.3	45.0	67.6
DISTRE†	68.0	67.0	65.3	65.0	60.2	47.9	62.2
PCNN+BAG_ATT‡	88.0	81.0	**80.7**	72.6	60.3	45.0	71.3
ToHRE w/o EW	81.2	80.1	76.8	**75.7**	61.9	46.4	70.4
ToHRE	**91.5**	**82.9**	79.6	74.8	**63.3**	**48.9**	**73.5**

Table 2: Precision evaluated automatically for the top rated relation instances. † indicates the baseline results reported in (Alt et al., 2019) and ‡ indicates the baseline results given by our implementation.

Training Instances		<100			<200	
Hits@(Macro)	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
PCNN+KATT	35.3	62.4	65.1	43.2	61.3	69.2
ToHRE	**62.9**	**75.9**	**81.4**	**69.7**	**80.3**	**84.8**

Table 3: Accuracy (%) of Hits@K on relations with training instances fewer than 100/200.

100/200 training instances. The Hits@K metric is introduced for evaluation. For each entity pair, the evaluation requires its corresponding golden relation in the first K candidate relations recommended by the models. Following previous work, we select K from $\{10,15,20\}$ and report the macro average Hits@K accuracies for these subsets. We compare our model with PCNN+HATT (Han et al., 2018) which is the first work to evaluate the long-tail relations under such settings and PCNN+KATT (Zhang et al., 2019), which follows Han et al. (2018) and achieves the SOTA performance in the long-tail relation extraction. From the evaluation results in Table 3, we can observe that ToHRE improves previous SOTA approach by a large margin, i.e., 27.6% and 26.5% in the aspect of Hits@10 under different training instances and has consistent performance on Hits@15 and Hits@20 settings. The result indicates that relation hierarchies have been better exploited in our hierarchical classification framework than previous methods. Although ToHRE has improved performance on the long-tail relation extracting, the results of all methods are far from satisfactory, which requires more sophisticated models to handle this problem.

3.5 Ablation Study

We conduct an ablation study to verify the effectiveness of the entity-aware word embedding module. To this end, we denote *ToHRE w/o EW* as using the position-aware word embedding (Zeng et al., 2015) instead of the proposed entity-aware embedding. From the corresponding P@N results shown in Table 2, we observe that the prediction result has an obvious drop without using the entity-aware word embedding, especially has a 10.3% decreases in top-100. It indicates that the proposed entity-aware word embedding methods can successfully capture the corrections between each word and two corresponding entities.

3.6 Case study

In this section, we conduct a case study to show the predicting process of our framework. Table 4 shows the predicted score in different relation levels for B_1 and B_2. The B_1 contains two sentences where the second sentence does not express the labeled relation */location/country/administrative_divisions*. However, we can observe that our model predicts high scores to all relation levels despite the noisy sentence, which shows that our hierarchically-refined selective attention can alleviate the noise problem in different levels. The B_2 is labeled with the relation */people/person/religion* which has few training instances. The previous flat models can not extract such relations due-to data deficiency. However, our model can

Bag	Bag of Sentences with Labeled Relation	P_1	P_2	P_3
B_1	*Sen1*. Barzin ... died at her home in **Paris**, **France** on november 27. *Sen2*. His interrogation led to arrests algeria, italy and **France**, where mr.majrar's ... armored car depot in Beauvais, north of **Paris**. *Relation.* /location/country/administrative_divisions	0.97	0.98	0.95
B_2	*Sen*. **Muhammadu buhari**, also a northern **Muslim** ... people's party. *Relation.* /people/person/religion	0.77	0.97	0.28

Table 4: A case study where the entities are mark in bold. P_1, P_2 and P_3 show the predicted relation probability in $level_1$, $level_2$ and $level_3$ respectively.

successfully extract the data-rich relations in top-level, i.e., */people* and */people/person*. Even though we have a low score in base-level relation */people/person/religion*, we combined it with the high score of top-level to assign the bag with an overall probability.

4 Related Work

4.1 Distantly Supervision Relation Extraction

Distant supervision proposed by Mintz et al. (2009) is an efficient approach that automatically generates large-scale training data for the RE task. However, the training data generated by DS usually contain amounts of wrongly labeled sentences and suffer from the long-tail problem. To alleviate the wrong labeling problem, Riedel et al. (2010) and Hoffmann et al. (2011) propose a MIL framework, where training sentences are arranged in bags and a label is provided for a bag of sentences instead of each sentence individually. This framework is well-suited for the DS setting and thus many works adopt this framework to select informative sentences from the noisy bags. For example, Zeng et al. (2015) propose PCNN to automatically extract features from sentences and select the most important sentence. Attention mechanisms (Lin et al., 2016; Du et al., 2018; Lei et al., 2018; Yuan et al., 2019; Ye and Ling, 2019) are investigated to assign high attention to informative sentences. Reinforcement learning (Zeng et al., 2018; Feng et al., 2018; Qin et al., 2018) is adopted to filter the noisy sentences. As for solving the long-tail problem in DS, Han et al. (2018) leverage relation hierarchies to calculate a coarse-to-fine attention for better extracting long-tail relations. Zhang et al. (2019) take advantage of the knowledge from data-rich relations at the head of distribution to boost the performance of the data-poor relations in the tail.

However, most previous works formulate DSRE as a flat classification problem which has not fully exploited the inherent hierarchical structure of relations. Indeed, hierarchical classification has been widely used in other tasks, such as text classification (Gopal and Yang, 2013; Peng et al., 2018; Mao et al., 2019), question answering (Qu et al., 2012) and online advertising (Agrawal et al., 2013) and demonstrates its efficiency for hierarchical structure. It is notable that in some works, such as works from Mao et al. (2019) and Silla and Freitas (2010), the flat classification is also regarded as a special circumstance of hierarchical classification where only the label at the leaf node is predicted. But in order to distinguish our model from previous works, we consider flat classification and hierarchical classification as two independent methods in this paper. To the best of our knowledge, we are the first to conduct hierarchical classification for the DSRE task.

4.2 Local Approach vs. Global Approach for Classification

We summarize hierarchical classification methods into two categories based on how the hierarchy is explored: local and global approaches. Local approaches explore the hierarchy in a top-down manner by training a set of local classifiers for each node (D'Alessio et al., 2000), or each parent node (Holden and Freitas, 2009), or each level in the hierarchy (Clare and King, 2003). The disadvantage of local approaches is that the number of local classifiers depends on the size of the label hierarchy, which makes them infeasible to scale. The global approaches (Kiritchenko et al., 2005; Cai and Hofmann, 2004) train a single classifier for all classes in the hierarchy. Compared with the local classifier, less research on the global classifier has been investigated due to the complexity of the problem (Silla and Freitas, 2010).

5 Conclusion

In this paper, we take advantage of the inherent hierarchical structure of relations and propose a top-down classification strategy with a hierarchical bag presentation. In this way, we formulate the DSRE as a hierarchical classification task. The experimental results indicate that our proposed model outperforms previous state-of-the-art flat methods, especially for long-tail relations. In the future, we plan to explore the following directions: (1) incorporating entity information from external knowledge graphs to enhance the hierarchical bag representation; (2) utilizing more sophisticated training strategies for the long-tail relation problem.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments. This work is supported by the National Natural Science Foundation of China (No.61976102, No.U19A2065).

References

Rahul Agrawal, Archit Gupta, Yashoteja Prabhu, and Manik Varma. 2013. Multi-label learning with millions of labels: recommending advertiser bid phrases for web pages. In *Proceedings of the 2013 World Wide Web Conference*.

Christoph Alt, Marc Hübner, and Leonhard Hennig. 2019. Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1388–1398.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Lijuan Cai and Thomas Hofmann. 2004. Hierarchical document categorization with support vector machines. In *CIKM '04*.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka, and Tom M Mitchell. 2010. Toward an architecture for never-ending language learning. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*.

Amanda Clare and Ross D. King. 2003. Predicting gene function in saccharomyces cerevisiae. *Bioinformatics*, 19 Suppl 2:ii42–9.

Stephen D'Alessio, Keitha A. Murray, Robert Schiaffino, and Aaron Kershenbaum. 2000. The effect of using hierarchical classifiers in text categorization. In *RIAO*.

Jinhua Du, Jingguang Han, Andy Way, and Dadong Wan. 2018. Multi-level structured self-attentions for distantly supervised relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2216–2225.

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

Siddharth Gopal and Yiming Yang. 2013. Recursive regularization for large-scale classification with hierarchical and graphical dependencies. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 257–265.

Xu Han, Pengfei Yu, Zhiyuan Liu, Maosong Sun, and Peng Li. 2018. Hierarchical relation extraction with coarse-to-fine grained attention. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2245.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 541–550.

Nicholas Holden and Alex Alves Freitas. 2009. Hierarchical classification of protein function with ensembles of rules and particle swarm optimisation. *Soft Computing*, 13:259–272.

Svetlana Kiritchenko, Stan Matwin, and Fazel Famili. 2005. Functional annotation of genes using hierarchical text categorization.

Kai Lei, Daoyuan Chen, Yaliang Li, Nan Du, Min Yang, Wei Fan, and Ying Shen. 2018. Cooperative denoising for distantly supervised relation extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 426–436.

Yang Li, Guodong Long, Tao Shen, Tianyi Zhou, Lina Yao, Huan Huo, and Jing Jiang. 2020. Self-attention enhanced selective gate with entity-aware embedding for distantly supervised relation extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8269–8276.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2124–2133.

Tianyu Liu, Kexiang Wang, Baobao Chang, and Zhifang Sui. 2017. A soft-label method for noise-tolerant distantly supervised relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1790–1795.

Yuning Mao, Jingjing Tian, Jiawei Han, and Xiang Ren. 2019. Hierarchical text classification with reinforced label assignment. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 445–455.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011.

Hao Peng, Jianxin Li, Yu He, Yaopeng Liu, Mengjiao Bao, Lihong Wang, Yangqiu Song, and Qiang Yang. 2018. Large-scale hierarchical text classification with recursively regularized deep graph-cnn. In *Proceedings of the 2018 World Wide Web Conference*, page 1063–1072.

Pengda Qin, Weiran Xu, and William Yang Wang. 2018. Robust distant supervision relation extraction via deep reinforcement learning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2137–2147.

Bo Qu, Gao Cong, Cuiping Li, Aixin Sun, and Hong Chen. 2012. An evaluation of classification models for question topic categorization. *J. Assoc. Inf. Sci. Technol.*, 63:889–903.

Sebastian Riedel, Limi Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*, pages 148–163.

Carlos N. Silla and Alex Alves Freitas. 2010. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22:31–72.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Shikhar Vashishth, Rishabh Joshi, Sai Suman Prayaga, Chiranjib Bhattacharyya, and Partha Talukdar. 2018. RESIDE: Improving distantly-supervised neural relation extraction using side information. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1257–1266.

Zhi-Xiu Ye and Zhen-Hua Ling. 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, Volume 1 (Long and Short Papers)*, pages 2810–2819.

Yujin Yuan, Liyuan Liu, Siliang Tang, Zhongfei Zhang, Yueting Zhuang, Shiliang Pu, Fei Wu, and Xiang Ren. 2019. Cross-relation cross-bag attention for distantly-supervised relation extraction. In *The Thirty-Third AAAI Conference on Artificial Intelligence*, pages 419–426.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762.

Xiangrong Zeng, Shizhu He, Kang Liu, and Jun Zhao. 2018. Large scaled relation extraction with reinforcement learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*.

Ningyu Zhang, Shumin Deng, Zhanlin Sun, Guanying Wang, Xi Chen, Wei Zhang, and Huajun Chen. 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, Volume 1 (Long and Short Papers)*, pages 3016–3025.