Hierarchical constrained attention for distantly supervised relation extraction

Bao Liu
bliu187@foxmail.com
School of Computer Science and Engineering, Southeast University
Nanjing, China
Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education
China

Guilin Qi[*]
gqi@seu.edu.cn
School of Computer Science and Engineering, Southeast University
Nanjing, China
Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education
China

Yuxin Zhang
zzyx_cs@seu.edu.cn
School of Computer Science and Engineering, Southeast University
Nanjing, China
Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education
China

ABSTRACT

Distantly supervised relation extraction (DSRE) can automatically collect training data with existing knowledge bases. However, label noise and long-tailed distributions severely affect the performance of DSRE. Most previous studies alleviate these problems by the multi-instance learning to construct a sentence bag as the input of the classifier, while the background information related to DSRE is not fully utilized. In this paper, we design a hierarchical constrained attention to exploit background information such as relation hierarchy and entity types. Furthermore, a hierarchical constrained attention-based distantly supervised relation extraction framework (HCAT) is proposed. Specifically, HCAT employs a hierarchical relation extraction framework to propagate information from data-rich top-layer relations to data-poor long-tailed relations. To further facilitate the information sharing between different relations, graph attention networks are used to encode all the relations connected by entity types. In addition, for the label noise problem, at each level of the relation hierarchy, entity types are concatenated into corresponding sentences and relations to better identify the valid sentences for bag representations. Substantial experimental results demonstrate that our model HCAT achieves significant improvement over the previous methods for both denoising and long-tailed relation extraction[1].

CCS CONCEPTS

• **Computing methodologies → Natural language processing**.

KEYWORDS

Distant Supervision, Relation Extraction, Label Noise, Long-tailed Distributions, Hierarchical Constrained Attention

ACM Reference Format:
Bao Liu, Guilin Qi, and Yuxin Zhang. 2023. Hierarchical constrained attention for distantly supervised relation extraction. In *Proceedings of The 12th International Joint Conference on Knowledge Graph (IJCKG'23).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Relation extraction (RE) aims to extract relation facts between two given entities from unstructured text, which plays an important role in many natural language processing applications [14, 5, 15]. For example, during the knowledge graph construction process, there are still many missing facts. As a key technology for knowledge graph completion, relation extraction can effectively identify semantic relationships between existing entity pairs [28, 7, 13]. Conventional supervised approaches have shown their capabilities in tackling RE problems, but they require large-scale labeled training data which is time-consuming and labor-intensive. In this case, distantly supervised relation extraction (DSRE) [23] is proposed to automatically generate large-scale labeled data by aligning an existing knowledge base (KB) and plain text. It assumes that if one entity pair holds a relation in the existing KB, then all sentences that mention the entity pair will express this relation. Although DSRE can bring abundant training data, it still suffers from two major challenges.

The first challenge is label noise caused by the aforementioned strong assumption. As illustrated in Figure 1, there is a relation triple < *Kevin Durant, Birthplace, Washington* > in the knowledge base, and sentence *S1* and *S2* both contain the same entity pair < *Kevin Durant, Washington* >. According to the distant supervision assumption, sentence *S1* and *S2* are both labeled with the relation *Birthplace* and used as training data. However, sentence *S2* does not express the relation *Birthplace* and obtains a wrong label. To mitigate the interference of label noise, the multi-instance learning (MIL) framework [27, 11, 31] was employed to identify the relation of a common entity pair for a bag of sentences. In addition, to take advantage of all sentences in the bag, the attention mechanism was

[*] Corresponding author.
[1] https://github.com/bliu187/HCAT

IJCKG'23, December 08–09, 2023, Tokyo, Japan

ACM ISBN 978-1-4503-XXXX-X/23/12…$15.00
https://doi.org/XXXXXXX.XXXXXXX

introduced into DSRE [20], which can dynamically assign different weights to sentences to mitigate the influence of noisy sentences.

The second challenge is long-tailed relation extraction. Although distant supervision methods can generate a large amount of training data, there is a large gap in these training data with different labels. For example, if we treat relations with less than 100 training sentences as long-tailed relations, then there are more than 60% long-tail relations suffering from data deficiency in the popular New York Times (NYT) [27] dataset. To accomplish this, the hierarchy-based strategy [9, 40, 18] has been extensively studied in the last few years, which developed a relation hierarchical tree and tried to mine latent correlation between relations on different layers, information can propagated from data-rich top-layer relations to data-poor long-tailed ones.

To alleviate the label noise and long-tailed distributions, some researchers attempt to exploit the background information to improve the relation extraction performance of distant supervision model, e.g., entity types [19], entity descriptions [12], and relation hierarchy [9], which can mitigate the interference of noisy sentences in the bag, and connect different relations to facilitate the transfer of information between data-rich relations and long-tailed relations. However, previous works separately exploit background information, focusing only on a specific aspect of the background information, which fails to realize its full potential value. Therefore, we suspect that employing appropriate methods to simultaneously use multiple background information may better promote the performance of distantly supervised relation extraction.

To fully exploit the background information, we design a hierarchical constrained attention to employ both background information: relation hierarchy and entity types. Furthermore, a hierarchical constrained attention-based distantly supervised relation extraction framework (HCAT) is proposed for both label noise and long-tailed distributions. In detail, HCAT employs a hierarchical relation extraction framework [9] to propagate information from data-rich relations to long-tailed relations, and graph attention networks (GAT) [32] are used to encode all relations and entity types, which can further promote the information sharing between different relations. Moreover, for the label noise problem, entity types are connected to corresponding sentences and relation representations at each level of the relation hierarchy as the external information constraints, which can improve the selective attention mechanism's ability to identify valid sentences.

The main contributions are summarized as follows:

- We design a hierarchical constrained attention network to make full use of background information such as relation hierarchy and entity types, which can facilitate valid sentence recognition in bags and information transfer between different relations.
- We propose a hierarchical constrained attention-based distantly supervised relation extraction framework, which can alleviate both label noise and long-tailed relation problems.
- We conduct comprehensive experiments on the popular New York Times (NYT) [27] dataset. Our model HCAT receives state-of-the-art performance in terms of multiple metrics.

The remainder of our paper is organized as follows. An overview of related work is introduced in Section 2. Section 3 details the proposed HCAT model. We present the extensive experiments for performance comparison and ablation studies in Section 4. Section 5 concludes the paper and gives some important future directions.

2 RELATED WORK

Distant supervision can automatically generate a large amount of labeled training data for relation extraction tasks. However, the aforementioned strong assumption also brings two important challenges: label noise and long-tailed distributions.

To alleviate the label noise problem, multi-instance learning (MIL) was employed to alleviate the strong assumption of DSRE [27, 11, 31], which extracts relation from a sentence bag instead of a single sentence for an entity pair. Then, various denoising strategies under the MIL framework were proposed. Lin et al. [20] designed sentence-level selective attention to get the bag representation by giving sentences different weights. Qu et al. [26] proposed a word-level attention to increase the attention weights of those critical words. Yu et al. [35] elaborated segment-level attention to mine the information of continuous words in a sentence, which can capture dependencies between different entity pairs and corresponding relations. In addition, Yuan et al. [37] developed bag-level selective attention to pay more attention to entity pairs with higher quality. Ye et al. [34] designed an intra-bag- and inter-bag-level attention to alleviate the noise that existed in both sentences and bags. In addition, many other denoising techniques have been introduced to DSRE, such as consensus-enhanced training [21], cross-stitch bi-encoders [4], deep clustering [33], interaction-and-response networks [29], etc.

To tackle the long-tailed distributions problem, existing methods are mostly based on relation hierarchy. Although long-tail relations contain less training data, sufficient training data likely exists in their ancestor or sibling relations. Han et al. [9] first incorporated the hierarchical information of relations to DSRE and proposed a hierarchical instance-level attention, which can transfer knowledge from data-rich relations to data-poor ones. Then, many other strategies based on relation hierarchy were proposed, such as top-down classification [36], global hierarchy embeddings [24], and recursive hierarchy interactive attention [8]. In addition, Zhang et al. [40] developed a coarse-to-fine knowledge-aware attention network to learn relation information through knowledge graph embeddings. Li et al. [18] designed a relation-augmented attention network to enrich the sentence representations, which incorporate relation information to sentence embeddings at each level of the relation hierarchy. Contrastive learning was also introduced to DSRE to alleviate the long-tail problem [16]. Recently, a novel constraint graph-based relation extraction framework (CGRE) [19] was proposed, which uses graph convolution networks to propagate information between different relations and designs constraint-aware attention for bag representation.

All the aforementioned works try to alleviate the influence of noisy sentences and facilitate information transfer between different relations, but the background information is not fully utilized. Therefore, we exploit both relation hierarchy and entity types information and propose a hierarchical constrained attention-based distantly supervised relation extraction framework (HCAT) for both label noise and long-tailed distributions.

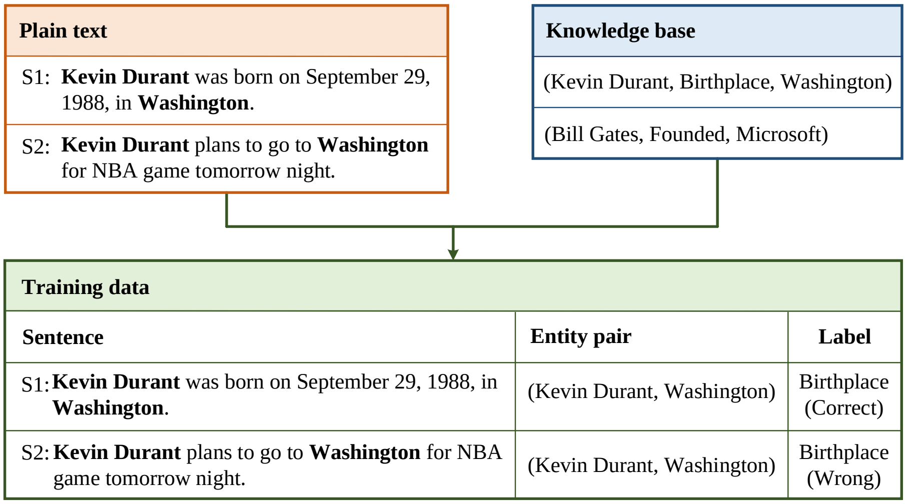

Plain text
S1: **Kevin Durant** was born on September 29, 1988, in **Washington**.
S2: **Kevin Durant** plans to go to **Washington** for NBA game tomorrow night.
Knowledge base
(Kevin Durant, Birthplace, Washington)
(Bill Gates, Founded, Microsoft)
Training data
Sentence
Entity pair
Label
S1: **Kevin Durant** was born on September 29, 1988, in **Washington**.
(Kevin Durant, Washington)
Birthplace (Correct)
S2: **Kevin Durant** plans to go to **Washington** for NBA game tomorrow night.
(Kevin Durant, Washington)
Birthplace (Wrong)

Figure 1: The illustration of DSRE process.

3 APPROACH

3.1 Task definition

Distantly supervised relation extraction (DSRE) can automatically obtain a large amount of training data, but it is also accompanied by the problem of noisy label sentences and long-tailed distributions. Therefore, DSRE employs the multi-instance learning (MIL) framework, all sentences are divided into multiple bags $\{\mathcal{B}_1, \mathcal{B}_2, ..., \mathcal{B}_k\}$, each bag $\mathcal{B} = \{S_1, S_2, ..., S_n\}$ groups some sentences with the same target entity pair <e^1, e^2>, and each sentence is defined as a word sequence $S = \{w_1, w_2, ..., w_m\}$. In addition, all relations are predefined as a set $\mathcal{R} = \{r_1, r_2, ..., r_l\}$. The goal of distantly supervised relation extraction is to predict the relations for all the given entity pairs.

3.2 Framework

To alleviate the interference of label noise and long-tailed distributions, we employ hierarchical relation extraction framework [9] and graph attention networks [32] to facilitate the transfer of information between different relations, and long-tailed relations and benefit from the data-rich relations. Furthermore, the encoded entity types are used as additional constraints for the bag representation, which can better identify valid sentences to mitigate the impact of label noise. As illustrated in Figure 2, our proposed model HCAT consists of three key modules:

- **Sentence encoder**. Given a bag of sentences with a target entity pair, the sentence encoder is employed to extract features of each sentence in the bag.
- **Graph encoder**. Given all relations and corresponding entity types, we first conduct the hierarchy constraint graph and then adopt the graph attention networks (GAT) [32] to extract the interactive features of the relations and entity types.
- **Hierarchical constrained attention**. This module is designed to get bag embeddings at each level of the relation hierarchy, the entity types are used as the external information constraint to identify valid sentences.

3.3 Sentence encoder

In this module, we first employ an entity-aware embedding [17] to encode each word in a sentence, and then adopt the piecewise convolutional neural network (PCNN) [38] to get the entire sentence representation.

3.3.1 *Entity-aware embeding.*

Given a sentence $S = \{w_1, w_2, ..., w_m\}$, We firstly use a pre-trained word2vec model [22] to map each word w_i into a d_w-dimensional vectors $\mathbf{w}_i$. Then following Li et al. [17], the position and entity information are incorporated into each word vector $\mathbf{w}_i$ to capture their semantic and syntactic information.

The entity information is represented as $\mathbf{w}^{e^1}$ and $\mathbf{w}^{e^2}$, they are the word vectors of target entity pair <e^1, e^2>. The position information [39] is the relative distances between each word w_i and the target entity pair. For example, in the sentence "Kevin Durant was born on September 29, 1988, in Washington.", the relative distance from *born* to entity e^1 ($KevinDurant$) and entity e^2 ($Washington$) are 2 and -8, respectively. Then, these two distances are embedded as two d_p-dimensional vectors $\mathbf{p}_i^{e^1}$ and $\mathbf{p}_i^{e^2}$. We concatenate the entity and position information to each word and get two types of word representations as follows:

$$\begin{aligned} \mathbf{f}_i^e &= [\mathbf{w}_i; \mathbf{w}^{e^1}; \mathbf{w}^{e^2}] \in \mathbb{R}^{3d_w}, \\ \mathbf{f}_i^p &= [\mathbf{w}_i; \mathbf{p}_i^{e^1}; \mathbf{p}_i^{e^2}] \in \mathbb{R}^{d_w+2d_p}, \end{aligned} \tag{1}$$

where the word dimension d_w and position dimension d_p are both pre-defined. And two types of word vectors compose the corresponding sentence representations $\mathbf{F}_e = \{\mathbf{f}_1^e, ..., \mathbf{f}_m^e\}$ and $\mathbf{F}_p = \{\mathbf{f}_1^p, ..., \mathbf{f}_m^p\}$. Consequently, we employ the entity-aware embedding

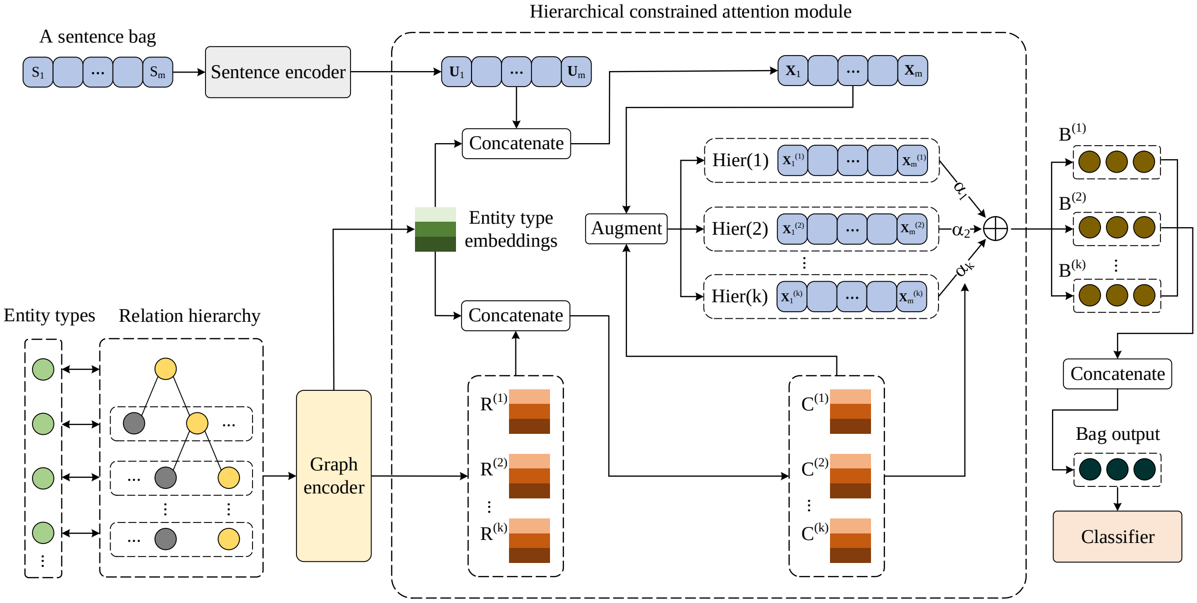


Figure 2: The overview of our proposed model, HCAT.

to get the sentence representation as follows:

$$\begin{aligned} \alpha &= sigmoid(\lambda \cdot (\mathbf{W}_e \mathbf{F}_e + \mathbf{b}_e)), \\ \tilde{\mathbf{F}}_p &= tanh(\mathbf{W}_p \mathbf{F}_p + \mathbf{b}_p), \\ \mathbf{S} &= \alpha \odot \mathbf{F}_e + (1-\alpha) \odot \tilde{\mathbf{F}}_p, \end{aligned} \tag{2}$$

Where "$\odot$" indicates the element-wise product, λ is a trade-off weight. $\mathbf{W}_e$, $\mathbf{W}_p$, $\mathbf{b}_e$ and $\mathbf{b}_p$ are all learnable parameters.

3.3.2 Piecewise convolutional neural network (PCNN). To effectively extract features of the given sentence $\mathbf{S} = \{\mathbf{s}_1, ..., \mathbf{s}_m\}$, we employ the piecewise convolutional neural network (PCNN) [38] as sentence encoder and get a high-dimensional representation. Firstly, k convolutional kernels $\mathbf{W}^c = \{\mathbf{w}_1^c, ..., \mathbf{w}_k^c\}, \mathbf{w}_i^c \in \mathbb{R}^w$ are slide on the sentence sequence as follows:

$$\mathbf{h}_{ij} = \mathbf{w}_i^c \mathbf{s}_{j-w+1:j} \quad 1 \leqslant i \leqslant k, 1 \leqslant j \leqslant m, \tag{3}$$

where $\mathbf{s}_{i:j}$ operates the concatenating from $\mathbf{s}_i$ to $\mathbf{s}_j$. Then the hidden representation is divided into three segments $\{\mathbf{h}_{ij}^{(1)}, \mathbf{h}_{ij}^{(2)}, \mathbf{h}_{ij}^{(3)}\}$ according to the positions of the target entity pair $<e^1, e^2>$. In addition, the piecewise max pooling operation is conducted as follows:

$$\begin{aligned} \mathbf{q}_i^{(1)} &= \max_{1 \leqslant j \leqslant p_1} (\mathbf{h}_{ij}^{(1)}) \\ \mathbf{q}_i^{(2)} &= \max_{p_1 < j \leqslant p_2} (\mathbf{h}_{ij}^{(2)}) \quad 1 \leqslant i \leqslant k, \\ \mathbf{q}_i^{(3)} &= \max_{p_2 < j \leqslant m} (\mathbf{h}_{ij}^{(3)}) \end{aligned} \tag{4}$$

where p_1 and p_2 are the positions of e^1 and e^2, respectively. The piecewise max pooling layer can effectively capture structural information between two entities.

Finally, all the pooling results are concatenated to get the sentence feature representation $\mathbf{U}$ as follows:

$$\mathbf{U} = tanh(\mathbf{q}_{1:k}) \in \mathbb{R}^{3k} \tag{5}$$

3.4 Graph encoder

To facilitate the transfer of information between different relations, we exploit the background information such as relation hierarchy and entity types. we firstly combine the hierarchical relation extraction framework [9] and entity type constraints [19] to construct hierarchical constraint graph. Then, graph attention networks (GAT) [32] are employed to encode the relations and entity types for information exchange between them.

3.4.1 Hierarchical constraint graph construction. Most relations in knowledge bases are composed of hierarchical structures. Given a relation set $\mathcal{R} = \{r_1, r_2, ..., r_l\}$, for each relation $r \in \mathcal{R}$, we can obtain a hierarchical chain of its ancestors $\{r^{(1)}, r^{(2)}, ..., r^{(h)}\}$. For example, given a relation */sports/sports_team/location* in the knowledge base Freebase [2], the obtained hierarchy chain is denoted as {*/sports*, */sports/sports_team*, */sports/sports_team/location*}. In addition, a root relation is defined as the unique common ancestor of all relations, and all chains form a tree-like hierarchical structure as shown in Figure 3, each relation at a different level is treated as a separate node in the relation hierarchical tree.

Under the hierarchical relation extraction framework, long-tailed relations can benefit from their ancestors or siblings. For example, if */people/deceased_person/place_of_death* is a long-tailed relation with a small number of training sentences, it is difficult to train an effective model to identify this relation without sufficient training data. The hierarchical relation extraction framework computes scores for those sentences containing the same entity pair on

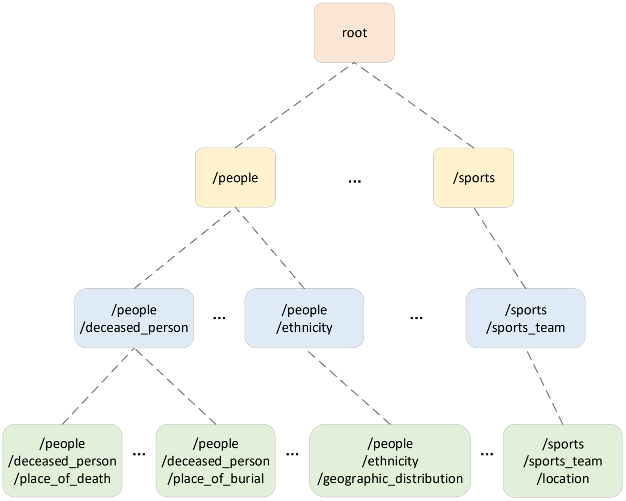


Figure 3: Relation hierarchy.

each layer of the hierarchies. For instance, as the father of the long-tailed relation $/people/deceased\_person/place\_of\_death$, relation $/people/deceased\_person$ contains many other children relations such as $/people/deceased\_person/place\_of\_burial$. All the training sentences of children relations construct the training data of $/people/deceased\_person$, which can effectively alleviate the problem of insufficient training data.

Furthermore, we employ entity types to connect different relations. Firstly, for these relations in the bottom level of the hierarchical tree, we can find the entity type constraint information from the corresponding knowledge base. Due to the large number of entity types contained in the original constraint information, we adopt 18 entity types $\mathcal{T} = \{t_1, ..., t_{18}\}$ defined in *OntoNotes* 5.0 [25], and a unified type *Others* is defined to represent entities that do not belong to the 18 types following [19]. Then, for other relations at higher levels of the hierarchical tree, because the semantic information of these relations is too broad to be constrained by specific types, we use a special type *None* to represent them. Finally, we denote the raw hierarchical constraint graph as $\mathcal{G} = \{\mathcal{T}, \mathcal{R}, C\}$, where $\mathcal{T}$ and $\mathcal{R}$ are the entity type set and relation set respectively, and they form the node set $\mathcal{V} = \mathcal{T} \cup \mathcal{R}$. For each constraint $(t_r^{e^1}, r, t_r^{e^2}) \in C$, $t_r^{e^1}$ and $t_r^{e^2}$ denote as the head entity type and tail entity type of the relation r, respectively. $(t_r^{e^1}, r)$ and $(r, t_r^{e^2})$ form the edge set $\mathcal{E}$.

3.4.2 Graph attention networks (GAT). We use the graph attention networks (GAT) [32] to encode the raw hierarchical constraint graph. Compared with other graph neural networks, GAT can effectively learn graph structure information and assign different weights to neighbor nodes. Note that all the node representations of the raw hierarchical constraint graph $\mathcal{G}$ are randomly initialized as a matrix $\mathbf{V} = \{\mathbf{v}_1, ..., \mathbf{v}_{n_v}\}$, each vector $\mathbf{v}_i \in \mathbf{V}$ denotes a node in $\mathcal{G}$. The node $\mathbf{v}_i$ of the k-th layer in the GAT is computed as follows:

$$\mathbf{v}_i^{(k)} = \sigma\left(\sum_{j \in N_i} \alpha_{ij} \mathbf{W} \mathbf{v}_j^{(k-1)}\right), \tag{6}$$

in which N_i is the set of neighbor nodes for $\mathbf{v}_i^{(k)}$ and $\mathbf{W}$ is the weight matrix. α_{ij} is the attention score of node $\mathbf{v}_i^{(k-1)}$ for $\mathbf{v}_j^{(k-1)}$, which is computed as:

$$\alpha_{ij} = \frac{exp\left(LeakyReLU\left(e_{ij}\right)\right)}{\sum_{k \in N_i} exp\left(LeakyReLU\left(e_{ik}\right)\right)}, \tag{7}$$

where $LeakyReLU(\cdot)$ is an activation function, e_{ij} is the original attention score before the above softmax operation, and is calculated as follows:

$$e_{ij} = \mathbf{a}\left(\mathbf{W}\mathbf{v}_i^{(k-1)}; \mathbf{W}\mathbf{v}_j^{(k-1)}\right), \tag{8}$$

where $\mathbf{a}$ is a weight vector to map the concatenated high-dimensional features $\left(\mathbf{W}\mathbf{v}_i^{(k-1)}; \mathbf{W}\mathbf{v}_j^{(k-1)}\right)$ to the attention score.

Finally, for the raw hierarchical constraint graph $\mathcal{G}$, we can get the output $\mathbf{V}^{(k)} = \{\mathbf{v}_1^k, ..., \mathbf{v}_{n_v}^k\}$ of GAT encoder, which can be divided into relation feature representations $\mathbf{R} = \{\mathbf{r}_1, ..., \mathbf{r}_{n_r}\}$ and entity type feature representations $\mathbf{T} = \{\mathbf{t}_1, ..., \mathbf{t}_{n_t}\}$ according to the category of each node.

3.5 Hierarchical constrained attention

In this section, we design a hierarchical constrained attention network to construct the bag representation. To further exploit the background information, we concatenate the entity types to the corresponding sentences and relations at each level of the relation hierarchy, which can improve the ability of the selective attention mechanism to identify valid sentences. For each relation $\mathbf{r_i} \in \mathbf{R}$, the corresponding entity types are obtained from the hierarchical constraint graph and concatenated to the relation $\mathbf{r_i}$ as follows:

$$\mathbf{c}_i = [\mathbf{r}_i; \mathbf{t}_r^{e^1}; \mathbf{t}_r^{e^2}], \tag{9}$$

For each sentence $\mathbf{U}$ in the sentence bag, the NER tool[2] is employed to recognize the corresponding entity pair types $(\mathbf{t}_u^{e^1}, \mathbf{t}_u^{e^2})$, which are concatenated to the sentence $\mathbf{U}$ as follows:

$$\mathbf{X} = [\mathbf{U}; \mathbf{t}_u^{e^1}; \mathbf{t}_u^{e^2}], \tag{10}$$

Then, we use the relation-augmented mechanism [18] to embed relations to sentence representations at different levels of the relation hierarchy, which can further facilitate the information transfer among different relations. Specifically, at the k-level of relation hierarchy, the sentence representation $\mathbf{X}$ is applied as a query to relation set $\mathbf{C}^{(k)} = \{\mathbf{c}_1^{(k)}, ..., \mathbf{c}_{n_r}^{(k)}\}$ with a dot product:

$$\begin{aligned} \beta^{(k)} &= softmax(\mathbf{X}^{\mathrm{T}}\mathbf{C}^{(k)}), \\ \mathbf{g}^{(k)} &= \mathbf{C}^{(k)}\beta^{(k)}, \end{aligned} \tag{11}$$

where $softmax(\cdot)$ denotes a normalization function along the last dimension. In addition, an element-wise gate with residual connection [10] and layer normalization [1] are used to merge the

[2] https://github.com/flairNLP/flair

relation-aware representation $\mathbf{g}^{(k)}$ into the sentence representation $\mathbf{X}$:

$$\begin{aligned}\gamma^{(k)} &= sigmoid(\mathbf{W}_g^{(k)}[\mathbf{X};\mathbf{g}^{(k)}] + \mathbf{b}_g^{(k)}),\\ \widetilde{\mathbf{X}}^{(k)} &= \gamma^{(k)} \circ \mathbf{X} + (1-\gamma^{(k)}) \circ \mathbf{g}^{(k)},\\ \mathbf{X}^{(k)} &= LayerNorm(\mathbf{X} + MLP(\widetilde{\mathbf{X}}^{(k)})),\end{aligned} \tag{12}$$

where $\mathbf{W}_g^{(k)}$ and $\mathbf{b}_g^{(k)}$ are both learnable parameters, $MLP(\cdot)$ denotes a multi-layer perceptron to increase nonlinearity.

Next, the attention score λ_i of each sentence $\mathbf{X}^{(k)}$ and the relation $\mathbf{c}$ can be obtained as

$$\begin{aligned}e_i &= \mathbf{X}_i^{(k)}\mathbf{c},\\ \lambda_i &= \frac{exp(e_i)}{\sum_{j=1}^{n} exp(e_j)},\end{aligned} \tag{13}$$

Then, we obtain the bag representation at each level of the relation hierarchy as follows:

$$\mathbf{z}^{(k)} = \sum_{i=1}^{n} \lambda_i \mathbf{X}_i^{(k)}, \tag{14}$$

Finally, we concatenate all the bag representations at each level of the relation hierarchy as the ultimate representation of sentence bag, and a softmax classifier is used to obtain a categorical distribution over all relations as bag prediction:

$$P = (r|\mathcal{B};\mathcal{G};\theta) = softmax(\mathbf{W}\mathbf{z} + \mathbf{b}) \tag{15}$$

where θ is the set of parameters, $\mathbf{W}$ and $\mathbf{b}$ are the weight and bias, respectively.

3.6 Optimization

A main objective and an auxiliary objective are introduced to optimize our model. The main objective is a bag-level prediction and is defined as minimizing cross-entropy:

$$L^{(re)} = -\frac{1}{m}\sum_{i=1}^{m} logP(r_i|\mathcal{B}_i;\mathcal{G};\theta), \tag{16}$$

where m is the number of bags and r_i is the relation of $\mathcal{B}_i$.

The auxiliary objective [18] is used to guide each sentence augmenting with correct relation representations at each level of the relation hierarchy. That is,

$$L^{(hier)} = -\frac{1}{m \cdot n \cdot l}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{k=1}^{l} log\beta^{(k)}_{[r_i^{(k)}]}, \tag{17}$$

where $[\cdot]$ is indexing operation, n denotes the number of sentences in each bag, and l is the number of relation hierarchy levels.

Eventually, the two objectives above are integrated to optimize the proposed model, i.e., $L = L^{(re)} + L^{(hier)}$.

In the test phase, for the input sentences, the ground-truth label is unknown. Therefore, we calculate posterior probabilities of all the relations and select the highest one as the prediction result [20].

4 EXPERIMENTS

4.1 Dataset and evaluation metrics

We adopt the widely used **NYT** [27] as the datasets to conduct our experiments. **NYT** is developed by aligning the corps of New York Times with relational facts in Freebase [2], which has been mainly released in the filtered version **NYT-520K** and non-filtered version **NYT-570K**. They both contain 53 relations which include a specific *NA* denoting no relation between the entity pair. The difference between the two versions is that the training set and test set of NYT-520K do not contain the same entity pairs. In the training set, the sentence numbers of NYT-520K and NYT-570K are 523 312 and 570 088, respectively, while the sentence numbers of NYT-520K and NYT-570K are both 172 448 in the test set.

We conduct the comparison experiments using four standard metrics following [9, 19, 18], which are precision-recall (PR) curves, the area under a curve (AUC), Top-N precision (P@N) and Hits@K. The first three metrics are adopted for the denoising evaluation, and the last one is used for the long-tailed evaluation. The detailed definitions are displayed as follows:

- **PR curve** exhibits the tradeoff between precision (y-axis) and recall (x-axis) for different probability thresholds.

$$\begin{aligned}precision &= \frac{N_{true\_positive}}{N_{true\_positive} + N_{false\_positive}},\\ recall &= \frac{N_{true\_positive}}{N_{true\_positive} + N_{false\_negative}},\end{aligned} \tag{18}$$

- **AUC** indicates the area under the PR curve. A model with a higher AUC score achieves better performance.
- **P@N** is designed to evaluate the relation extraction models with multi-instance learning and denotes the precision values of the sentences with top-N prediction confidences.
- **Hits@K** measures the probability that the true relation is achieved in the top-K predictions of the model for the sentences with a long-tailed distribution.

4.2 Baselines

We compare our proposed HCAT model with the following five competitive methods:

- **PCNN+ATT** [20] designs selective attention over multiple sentences to alleviate wrong labeling, which is the most classical model for the DSRE task.
- **PCNN+HATT** [9] uses hierarchical attention to exploit the correlations of relations.
- **SeG** [17] proposes an entity-aware embedding approach merging the position features and entity embeddings to the sentence representation.
- **CoRA** [18] designs a collaborating relation-augmented attention network incorporating relational information into sentence representation.
- **CGRE** [19] proposes a novel constrain graph-based DSRE model by concatenating the entity type into the sentence and relation representation.

Table 1: Parameter settings.

Component	Parameter	NYT-520K	NYT-570K
Sentence encoder	filter number	230	230
	window size	3	3
	word size	50	50
	position size	5	5
	coefficient λ	18	18
Graph encoder	embedding size	100	100
	hidden size	400	400
	layer number	2	2
	output size	1250	1250
Classifier	input size	900	900
Optimization	batch size	160	160
	learning rate	0.03	0.08
	dropout rate	0.5	0.5

4.3 Experimental settings

All the experiments are conducted on the Ubuntu 18.04 platform. We employ the pre-trained word2vec from OpenNRE[3] to initial the word embeddings, and adopt Xavier [6] to initialize all weight matrixes and position vectors, all the bias vectors are initialized to 0. We use mini-batch SGD [3] for optimization and apply dropout [30] before the classifier layer to prevent overfitting. Table 1 shows the detailed settings of our experiments.

4.4 Comparision results

The comparison results of different methods are shown for denoising and long-tail relation problems. All results of baseline CGRE[4] are from the original paper, however, the experiments on the other four baselines (i.e., PCNN+ATT[3], PCNN+HATT[5], SeG[6] and CoRA[7]) were only conducted on NYT-570K, so we experiment with the official source code to obtain these results not reported in the original papers.

To compare the denoising performance of different models, following the previous works [9, 19, 18], we conduct experiments using three widely used metrics: P@N, AUC, and PR curve. As illustrated in Table 2, our proposed HCAT achieves optimal values in all metrics at the same time. Specifically, For P@N, HCAT improves baseline approaches by at least 2.2 percentage points. Our model achieves the AUC of 0.422 and 0.546 on NYT-520K and NYT-570K, which outperforms the strong baselines by 0.02 and 0.17, respectively. The PR curves corresponding to AUC are shown in Figure 4. These comparison results indicate our proposed hierarchical constrained attention mechanism can significantly improve the model performance in handling the problem of wrong-labeled relation extraction.

For the long-tailed distribution, we conduct experiments on the NYT-570 dataset following the previous works [9, 19, 18], Hits@K [9] is employed as the evaluation metric. We extract the long tail relations with less than 100/200 training instances from the test set. Then the Hits@K metric is used to compute the probability that the true label falls into the top-K recommendations of the model. We also use macro average during the computing process. As illustrated in Table 3, our proposed HCAT achieves the best 4 out of 6 settings and the second-best value in the remaining 2 settings. In addition, compared with the other models, our proposed HCAT improves at least 7.8 percentage points on both top-15 and top-20 recommendations with training instances fewer than 100 and 200. This shows that our proposed HCAT is better at extracting relations with long-tailed distributions due to the efficient transfer of information between different relations.

Compared with CGRE, our model has achieved improvements in all metrics on the two widely used datasets NYT-520K and NYT-570K in terms of denoising. In terms of long-tail evaluation, our model HCAT achieves progress on Hits@15 and Hits@20 but regresses on Hits@10, a possible reason is that our model exploits the relation hierarchy information based on CGRE, which improves the relation extraction performance while increasing the model complexity, so more data is required for training. Therefore, adopting other methods to use relation hierarchy information based on a small increase in the model complexity is worth exploring in future research.

4.5 Ablation study

Our proposed distantly supervised relation extraction model HCAT exploits the background information such as relation hierarchy and entity types. To verify the effect of each element in improving the performance of the model, we separately remove them from the model as follows:

- ~ **w/o Hier**. The relation hierarchy is removed from the proposed model HCAT.
- ~ **w/o Typ**. The entity types are deleted from both the sentence and relation representations.
- ~ **w/o Gra**. The graph encoder is removed, and the embeddings of all relations and entity types are randomly generated.

The results are shown in Table 4, it can be seen that compared to HCAT, the performance of all models is degraded on multiple metrics. Experimental results show that employing relation hierarchy and entity type can effectively improve the performance of distantly supervised relation extraction models for both denoising and long-tailed relation extraction.

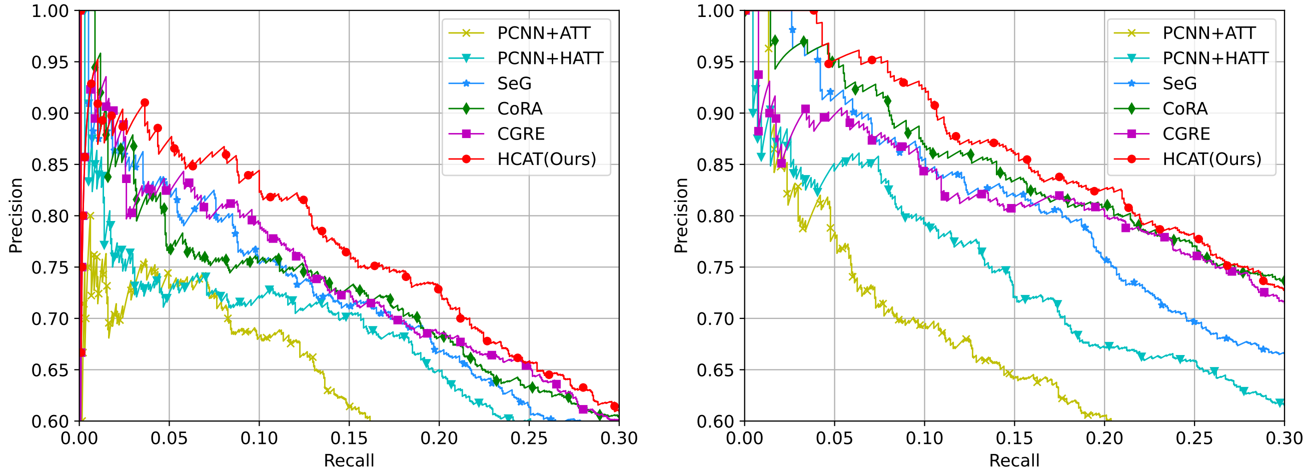
5 CONCLUSION AND FUTURE WORK

In this paper, we design a hierarchical constrained attention to exploit the background information such as relation hierarchy and entity types. Furthermore, we propose a hierarchical constrained attention-based distantly supervised relation extraction framework (HCAT), which adopts a hierarchical relation extraction framework to facilitate the transfer of information between data-rich relations and long-tailed ones. In addition, entity types are employed for constructing hierarchical constraint graphs and concatenating to the corresponding relations and sentences, which can further facilitate the information sharing between different relations and

[3] https://github.com/thunlp/OpenNRE
[4] https://github.com/tmliang/CGRE
[5] https://github.com/thunlp/HNRE
[6] https://github.com/YangLi1221/SeG
[7] https://github.com/YangLi1221/CoRa

Table 2: (%)P@N and AUC values of different models on NYT-520K and NYT-570K.

Model	NYT-520K				NYT-570K			
	P@100	P@200	P@300	AUC	P@100	P@200	P@300	AUC
PCNN+ATT [20]	74.2	72.5	68.0	34.7	73.1	74.0	71.5	39.2
PCNN+HATT [9]	73.1	72.5	72.3	37.9	85.3	80.9	79.2	42.1
SeG [17]	83.0	77.5	73.7	39.5	86.5	85.7	80.2	51.2
CoRA [18]	80.2	79.0	76.3	42.0	90.1	85.9	81.5	52.9
CGRE [19]	82.7	80.3	76.5	41.7	88.9	86.4	81.8	51.9
HCAT (Ours)	**87.0**	**82.5**	**80.0**	**42.2**	**95.0**	**93.0**	**87.3**	**54.6**



(a) NYT-520K

(b) NYT-570K

Figure 4: Precision-recall (PR) curves.

Table 3: Hits@K (Macro) on relations with training instances < 100/200.

Training instances	<100			<200		
Hits@K	10	15	20	10	15	20
PCNN+ATT [20]	<5.0	7.4	40.7	17.2	24.2	51.5
PCNN+HATT [9]	29.6	51.9	61.1	41.4	60.6	68.2
SeG [17]	62.1	70.9	83.5	70.6	75.2	88.9
CoRA [18]	66.6	72.0	87.0	72.7	77.3	89.4
CGRE [19]	**77.8**	77.8	87.0	**81.8**	81.8	89.4
HCAT (Ours)	70.0	**88.3**	**96.7**	75.0	**90.3**	**97.2**

identify the valid sentences for bag representations. Substantial experiments show that our proposed HCAT achieves state-of-the-art performance on multiple metrics, including denoising metrics (i.e., PR curve, AUC, P@N) and long tail metric (Hits@K).

In the future, we will explore some directions as follows:

- Effective approach to exploit more background information. There is also a lot of other background information related to distantly supervised relation extraction, it is desirable to exploit them for further performance improvement.

Table 4: (%)Hits@K (Macro) and AUC values of different models on NYT-570K.

Model	<100		<200		AUC
	Hits@10	Hits@20	Hits@10	Hits@20	
~ w/o Hie	67.5	94.1	74.1	95.1	53.2
~ w/o Typ	68.9	93.9	74.3	94.9	53.5
~ w/o Gra	65.7	90.3	73.6	91.5	52.9
HCAT	**70.0**	**96.7**	**75.0**	**97.2**	**54.6**

- Application in other RE scenarios. Our proposed hierarchical constrained attention is a universal module, which can be easily applied to other RE tasks, i.e., document-level relation extraction, and multimodal relation extraction.

ACKNOWLEDGMENTS

This work is partially supported by National Nature Science Foundation of China under No. U21A20488. We thank the Big Data Computing Center of Southeast University for providing the facility support on the numerical calculations in this paper.

REFERENCES

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *CoRR, abs/1607.06450*.

[2] 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*, 1247–1250.

[3] Andrew Cotter, Ohad Shamir, Nati Srebro, and Karthik Sridharan. 2011. Better mini-batch algorithms via accelerated gradient methods. *Advances in neural information processing systems*, 24.

[4] Qin Dai, Benjamin Heinzerling, and Kentaro Inui. 2022. Cross-stitching text and knowledge graph encoders for distantly supervised relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 6947–6958.

[5] Huifang Du, Zhongwen Le, Haofen Wang, Yunwen Chen, and Jing Yu. 2022. Cokg-qa: multi-hop question answering over covid-19 knowledge graphs. *Data Intelligence*, 4, 3, 471–492.

[6] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 249–256.

[7] Lingbing Guo, Qingheng Zhang, Wei Hu, Zequn Sun, and Yuzhong Qu. 2019. Learning to complete knowledge graphs with deep sequential models. *Data Intelligence*, 1, 3, 289–308.

[8] Ridong Han, Tao Peng, Jiayu Han, Hai Cui, and Lu Liu. 2022. Distantly supervised relation extraction via recursive hierarchy-interactive attention and entity-order perception. *Neural Networks*, 152, 191–200.

[9] 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*, 2236–2245.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

[11] 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*, 541–550.

[12] Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 3060–3066.

[13] Ting Jia, Yuxia Yang, Xi Lu, Qiang Zhu, Kuo Yang, and Xuezhong Zhou. 2022. Link prediction based on tensor decomposition for the knowledge graph of covid-19 antiviral drug. *Data Intelligence*, 4, 1, 134–148.

[14] Zhengbao Jiang, Zhicheng Dou, and Ji-Rong Wen. 2016. Generating query facets using knowledge bases. *IEEE transactions on knowledge and data engineering*, 29, 2, 315–329.

[15] Junzhuo Li and Deyi Xiong. 2022. Kafsp: knowledge-aware fuzzy semantic parsing for conversational question answering over a large-scale knowledge base. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 461–473.

[16] Tingwei Li and Zhi Wang. 2023. Ldrc: long-tail distantly supervised relation extraction via contrastive learning. In *Proceedings of the 2023 7th International Conference on Machine Learning and Soft Computing*, 110–117.

[17] 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* number 05. Vol. 34, 8269–8276.

[18] Yang Li, Tao Shen, Guodong Long, Jing Jiang, Tianyi Zhou, and Chengqi Zhang. 2020. Improving long-tail relation extraction with collaborating relation-augmented attention. In *Proceedings of the 28th International Conference on Computational Linguistics*, 1653–1664.

[19] Tianming Liang, Yang Liu, Xiaoyan Liu, Hao Zhang, Gaurav Sharma, and Maozu Guo. 2023. Distantly-supervised long-tailed relation extraction using constraint graphs. *IEEE Transactions on Knowledge & Data Engineering*, 35, 07, 6852–6865.

[20] 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)*, 2124–2133.

[21] Ruri Liu, Shasha Mo, Jianwei Niu, and Shengda Fan. 2022. Ceta: a consensus enhanced training approach for denoising in distantly supervised relation extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, 2247–2258.

[22] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

[23] 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*, 1003–1011.

[24] Tao Peng, Ridong Han, Hai Cui, Lin Yue, Jiayu Han, and Lu Liu. 2022. Distantly supervised relation extraction using global hierarchy embeddings and local probability constraints. *Knowledge-Based Systems*, 235, 107637.

[25] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using ontonotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, 143–152.

[26] Jianfeng Qu, Dantong Ouyang, Wen Hua, Yuxin Ye, and Ximing Li. 2018. Distant supervision for neural relation extraction integrated with word attention and property features. *Neural Networks*, 100, 59–69.

[27] Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 148–163.

[28] Ying Shen, Ning Ding, Hai-Tao Zheng, Yaliang Li, and Min Yang. 2020. Modeling relation paths for knowledge graph completion. *IEEE Transactions on Knowledge and Data Engineering*, 33, 11, 3607–3617.

[29] Wei Song, Weishuai Gu, Fuxin Zhu, and Soon Cheol Park. 2023. Interaction-and-response network for distantly supervised relation extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 193–201.

[30] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15, 1, 1929–1958.

[31] Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 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*, 455–465.

[32] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations*.

[33] Suizhu Yang, Yanxia Liu, Yuantong Jiang, and Zhiqiang Liu. 2023. More refined superbag: distantly supervised relation extraction with deep clustering. *Neural Networks*, 157, 193–201.

[34] 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)*, 2810–2819.

[35] Bowen Yu, Zhenyu Zhang, Tingwen Liu, Bin Wang, Sujian Li, and Quangang Li. 2019. Beyond word attention: using segment attention in neural relation extraction. In *IJCAI*, 5401–5407.

[36] Erxin Yu, Wenjuan Han, Yuan Tian, and Yi Chang. 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*, 1665–1676.

[37] 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 *Proceedings of the AAAI conference on artificial intelligence* number 01. Vol. 33, 419–426.

[38] 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*, 1753–1762.

[39] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*, 2335–2344.

[40] 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)*, 3016–3025.