Complex & Intelligent Systems (2024) 10:1551–1563
https://doi.org/10.1007/s40747-023-01226-w

ORIGINAL ARTICLE

LTACL: long-tail awareness contrastive learning for distantly supervised relation extraction

Tianwei Yan[1] · Xiang Zhang[1] · Zhigang Luo[1]

Received: 6 May 2023 / Accepted: 18 August 2023 / Published online: 28 September 2023

Abstract

Distantly supervised relation extraction is an automatically annotating method for large corpora by classifying a bound of sentences with two same entities and the relation. Recent works exploit sound performance by adopting contrastive learning to efficiently obtain instance representations under the multi-instance learning framework. Though these methods weaken the impact of noisy labels, it ignores the long-tail distribution problem in distantly supervised sets and fails to capture the mutual information of different parts. We are thus motivated to tackle these issues and establishing a long-tail awareness contrastive learning method for efficiently utilizing the long-tail data. Our model treats major and tail parts differently by adopting hyper-augmentation strategies. Moreover, the model provides various views by constructing novel positive and negative pairs in contrastive learning for gaining a better representation between different parts. The experimental results on the NYT10 dataset demonstrate our model surpasses the existing SOTA by more than 2.61% AUC score on relation extraction. In manual evaluation datasets including NYT10m and Wiki20m, our method obtains competitive results by achieving 59.42% and 79.19% AUC scores on relation extraction, respectively. Extensive discussions further confirm the effectiveness of our approach.

Keywords Distantly supervised learning · Information extraction · Relation extraction · Contrastive learning

Introduction

Relation extraction (RE) is a crucial task in natural language processing (NLP) [1–5], which aims to identify the relationship between entities in a sentence. Recently, there are various research directions in relation extraction, including cross-modal relation extraction [6, 7], multilingual relation extraction [8, 9], unified structure based relation extraction [1, 2] and large language model-based relation extractions [10]. Most existing supervised RE methods rely on manual annotations [11, 12]. While the annotation of large training data is a time-consuming and laborious job in real-world scenarios. To automatically obtain the large-scale labeled corpus, Mintz et al. [13] propose a DS approach to generate abundant data for RE by aligning knowledge bases (KB) with raw texts. The assumption is that if two entities have the same relation in KB, all the sentences containing target entities would be labeled as this relation.

However, this assumption inevitably introduces massive wrong labeling data. As shown in Fig. 1, the DS incorrectly labels the relation for the first and third sentences with the given entity pairs ⟨Amazon.com, Jeffrey P. Bezos⟩, on account of the limited-scale KB and the strong assumption. In response, recent studies [14–17] adopt multi-instance learning (MIL) on improving the robustness against the noisy label. These approaches apply the piece-wise convolutional neural networks (PCNN) with attention to alleviate the noise. Furthermore, with the help of pretrained model [18–20], Bert-based methods gain better performance in learning instance representations under MIL. However, as pointed out in [21], MIL only forms accurate bag-level representations, but fails to effectively utilize abundant instances inside MIL bags, which turns out a significant limitation. To overcome insufficient learning of instance representations, the contrastive learning [21–23] is used to enhance the sentence-

✉ Xiang Zhang
zhangxiang08@nudt.edu.cn

Tianwei Yan
yantianwei20@nudt.edu.cn

Zhigang Luo
zgluo@nudt.edu.cn

[1] College of Computer Science, National University of Defense Technology, Changsha, China

Sentence | Entity Pair <Amazon.com, Jeffrey P. Bezos>	Relation from DS
A roster of Linden Labs investors that includes Jeffrey P. Bezos, the founder of Amazon.com says that the entire Internet is moving toward being a three-dimensional experience.	founders
At the 2003 TED conference for Technology Entertainment and Design, Jeffrey P. Bezos, the Amazon.com chief executive, gave a speech about the early stages of technology	major_shareholders
Among the speakers will be Jeffrey P. Bezos of Amazon.com, who is expected to talk about selling more Web services to business customers.	unknow

Fig. 1 A mistake label from NYT10 dataset, where the DS gives the relationship major_shareholders to all sentence in the bag with entity pairs ⟨Amazon.com, Jeffrey P. Bezos⟩

level representation. Specially, Chen et al. [21] introduce a contrastive instance learning (CIL) method to boost the distantly supervised relation extraction (DSRE) under the MIL framework. CIL adopts single augmentation term frequency-inverse document frequency (TF-IDF) in positive pairs construction and achieves competitive performance. Moreover, HiCLRE [24] proposes a multi-level hierarchical learning framework for generating the de-noising context-aware representations and obtain better representations from contrastive learning, which achieves state-of-the-art performance.

However, the above approaches treat every category in data equally, but ignore the ubiquitous long-tail distribution problem in DSRE. The long-tail data commonly consist of two parts, the major part which has few relations with rich instances, and the tail part which has much more relations but contains fewer instances. As Fig. 2 shows, the instance number for each relation on the widely used DSRE dataset NYT is clearly under long-tail distribution. And nearly 70% of the relations in NYT dataset are under long-tail distribution [25]. For example, the relation major_shareholders mentioned in Fig. 1 has only 328 instances, far less than other categories of the data. While the denotation of these tail part relations seems insignificant to the overall results, but it is critical to the model's generalization ability in real scenario [26].

To address the above issues, we introduce our long-tail awareness contrastive learning (LTACL) model. To accomplish this, we analysis and divide the long-tail dataset into the major part and the tail part according to the most numerous class. Based on the observation of different parts, we adopt a novel dual-contrastive learning framework for better capturing the mutual information between the major part and the tail part instances representation. Specially, we adopt a hyper-augmentation strategy to mitigate the impact of the long-tail distribution. In addition, we introduce a counter-intuitive rule for constructing positive pairs in contrastive learning, employing slight augmentation for the major part and more extensive augmentation for the tail part. This operation enables us to capture diverse representations and effectively address the challenges associated with the long-tail issue.

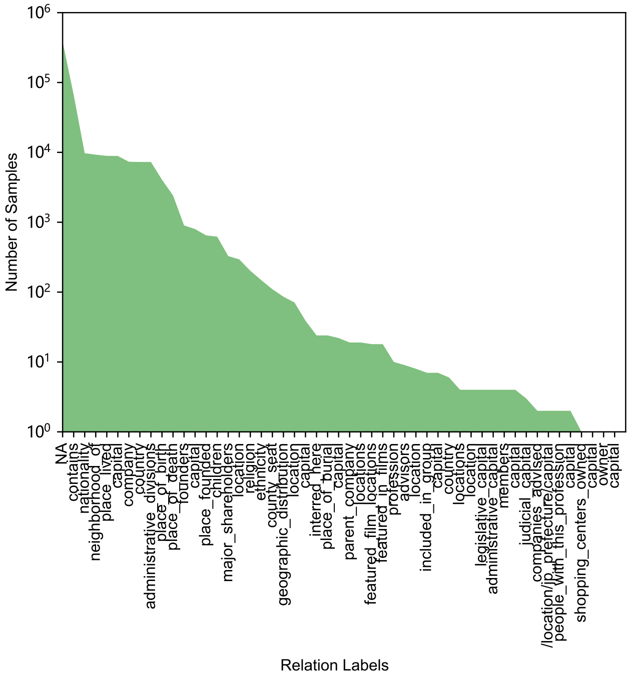


Fig. 2 The relation distribution of NYT10 trainset. With the horizontal axis representing the name of the relationship category and the vertical axis standing in for the number of samples, we can clearly conclude that the NYT10 dataset conforms to a long-tailed distribution

We carry out extensive experiments on the benchmark NYT10 and the manual annotated dataset NYT10m to evaluate our proposed model. It turns out that our LTACL model surpasses existing approaches by more than 2.61% AUC score on NYT10 dataset. In addition, our model achieves 59.42% and 79.19% AUC scores on NYT10m and Wiki20m datasets, respectively. Moreover, extensive analysis testify the effectiveness of our long-tail awareness model.

Our main contributions of proposed LTACL model can be summarized as follows:

1. A novel dual-contrastive learning architecture based on different parts of long-tail distribution data is proposed, which can enhance mutual information learning of sentence-level representations between the major part

and the tail part, and improves the performance of unbalanced categories relation extraction.

2. A hyper-augmentation strategy based on unsupervised contrastive learning is proposed, which uses different level augmentation for long-tail instances representation to construct positive pairs. The various view generated by hyper-augmentation strategy can be beneficial for the proposed model to learn meaningful representations to alleviate unbalanced distribution affection.
3. We evaluate on the widely used NYT dataset and two more accurately annotated sets, LTACL achieves significant improvements over previous SOTA models.

Related work

DSRE

Recently, burgeoning distantly supervised (DS) learning approaches [13] have raised particular interest in RE tasks, due to fact that the DS can generate relation labels for given entity automatically via aligning source corpus with a knowledge base. However, it suffers from data noise and long-tail distribution problems which is caused by its heuristic assumption. To alleviate noisy data, Lin et al. [15] focus on collecting the high-quality samples by adopting selective attention mechanism. Ye et al. [17] only use soft attention on intra-bag and inter-bag to deal with the noise at sentence-level and bag-level respectively. Qin et al. [27] employ reinforcement learning to construct an instance selector to denoise input data iteratively. In addition, Vashishth et al. [28] leverage graph convolution networks with side information to improve DSRE. Sui et al. [29] introduce a federated de-noising framework to suppress label noise in federated settings.

Long-tail

There are only few studies available on the long-tail relation extraction task. Gui et al. [30] follow the approach of explanation-based learning and learns rules to utilize unlabeled. Zhang et al. [31] propose a hierarchical attention to convert data-rich information to data-poor class at the tail of the distribution without explicit external data. Wang et al. [32] employ a hierarchical relational searching module to preprocess the relation labels and alleviate noisy simultaneously via RL. Unlike merging long-tail relations to the data-rich ones, we create discriminatory augmentations for different parts without extending the dataset.

Contrastive learning

As a very popular technique in unsupervised method, contrastive learning is widely explored in computer vision (CV) and natural language processing (NLP). In CV, Chen et al. [22] make use of two random augmentations for each image to generate different views of representations. Besides, it requires a large batch size of negative samples. Same as above, MoCo [23] gets two different augmented representations as positive pairs, while introducing the momentum update with a vast queue of negative samples. In NLP tasks, i.e., sentence representations, Yan et al. [33] propose a data augmentation module including adversarial attack, token shuffling, cutoff and dropout to generate views for contrastive learning. In DSRE, CIL [21] first combines multi-instance learning with contrastive learning, and gains the competitive results without any additional knowledge bases or relation information. The HiCLRE model extends contrastive learning by exploring representations across three levels and bolstering positive example pairs using the a dynamic gradient adversarial perturbation module. This enhancement enables the model to effectively mitigate the noise associate with DSRE. We argue that CIL and HiCLRE are efficient models under data noise, but lack the consideration of long-tail distribution in DS. Tracking of the problem, we provide a mix-augmentation module against different parts in the long-tail data. Our inspiration comes from SimCLR [22], which reveals that mixed augmentations would improve the performance of learning image representation.

Methodology

In this paper, we argue that the contrastive framework is effective to leverage sentence-level representations with bag-level ones, but deficient to handle long-tail distributions. Along the intention, we present an approach for simply and intelligently reducing the impact of long-tail problem in integrating into contrastive pairs construction.

The overall model architecture is shown in Fig. 3, on the left side is pipeline of our LTACL, and on the right side is detail of hyper-augmentation strategy module. First, we leverage long-tail analysis module to divide the bag-level data into different parts. Then, various dotted arrows point to different augmented views, which are generated by the hyper-augmentation strategy module. On top of model, we build a contrastive learning layer. The solid arrow connects the representation with its corresponding augmented from same instance to minimize their difference. The dotted arrow connects the instance with another bag representation in the same batch while keeping it distant from each other.

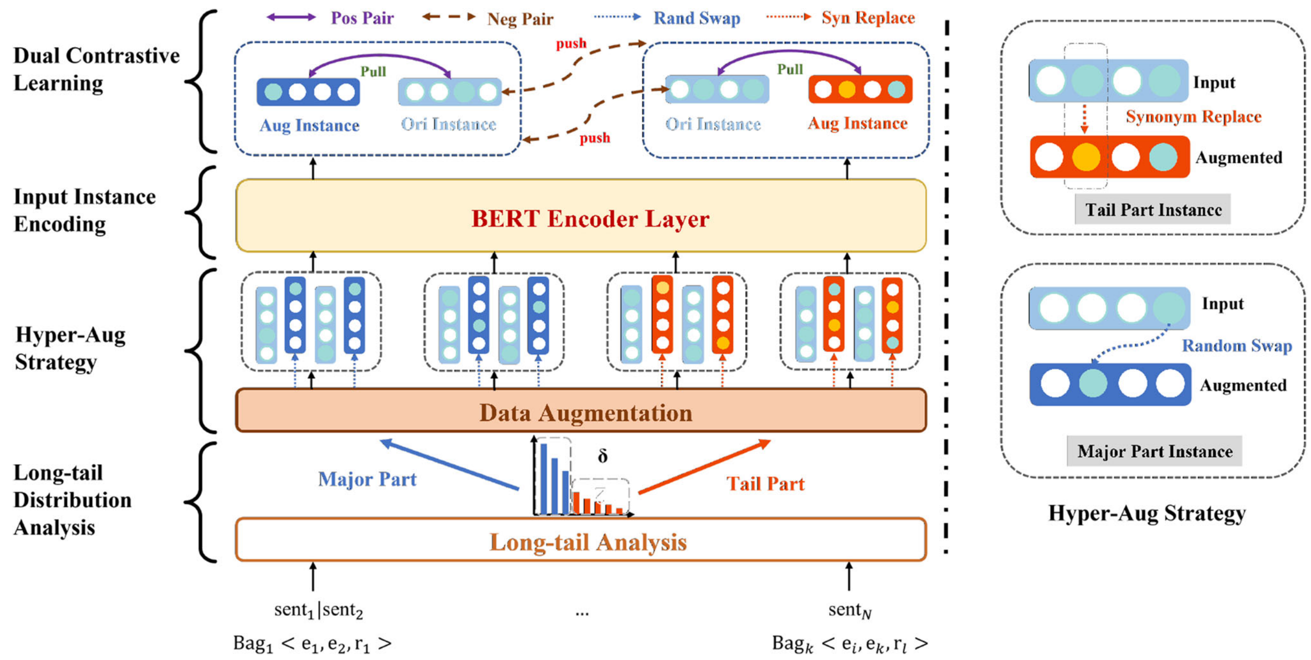


Fig. 3 Overall flowchart of LTACL. At first, the bag format input divided into major part and tail part according to a simple but effective analysis from the number of major samples. Then, different parts of the long-tail data lead to various levels augmentation strategies. Tokens and relations are encoded by BERT as distributed representations from multiple perspectives. Finally, a sentence-level positive pairs and a bag-level negative pairs are used to perform contrastive learning DS relation extraction

Hyper-augmentation strategy

In this section, we describe our long-tail awareness method. The purpose is direct and simple: given distribution of a long-tail dataset, the tail part requires hard augmentations for a better utilizing of instances while the major part just suits gentle operations.

Specifically, the dataset offers a collection of sentences $\{x_i\}_{i=1}^{m}$. In MIL learning paradigm, we choose the instances which has same relational triplet $\langle e_1, e_2, r_1\rangle$ to form bag B. Most of the textual data with relation labels present a long-tail distribution spontaneously, which means some of the relations own much fewer instances than others. To ensure that the tail parts of data not be ignored, we conduct our long-tail analysis module on input bags. According to the training set, we collect each relations number N_{ri}. For $\{r_m | m = \max(N_{r_1}, N_{r2}, ..., N_{ri})\}$, we employ r_m as the denominator to calculate the ratio to each relation:

$$\begin{cases} B_{\text{Major}} & \frac{r_i}{r_m} > \delta \\ B_{\text{Tail}} & \frac{r_i}{r_m} \le \delta \end{cases} \tag{1}$$

where $i = 1, ..., N_{ri}$, B_{Major} is the collection of the Major part of each bag, while B_{Tail} is the tail part. Note that we empirically choose $\delta = 0.03$ in the experiments, which means that we treat relations with instances less than 1000 as tail part relations.

We explore four data augmentation strategies of varying difficulties to generate views for contrastive learning, including synonym replacement (SR), random deletion (RD), term frequency-inverse document frequency (TF-IDF) and random swap (RS).

Synonym replacement is a widely used method for textual data augmentation. Randomly choose n words that are not stop words or entities. Then, replace chosen words with their synonyms in the word list.[1] Generally, SR is a robust augmentation in keeping instance representation.

Term frequency-inverse document frequency [21] adopts TF-IDF values to evaluate the importance of each word in the corpus. Based on that, the method replaces the unimportant words in the lexicon with other low TF-IDF score words to augment the original text.

Random deletion is a simple and direct strategy for randomly removing several words in the sentence with a fixed probability. To be noticed, we should not abandon the entity to avoid tampering the semantic of instances.

Random swap is a strategy which aims to randomly swap the order of chosen tokens in the input sentences. It is worth noting that we need to skip the entity in case of wrecking the whole representation of the instance. RS mainly affects the position encoding of the sentence.

[1] https://wordnet.princeton.edu/.

Since we have enough samples in the major part, we take RS as a further mining method for contrastive learning. Moreover, we utilize SR as a slight perturbation for the tail part to keep its primary distribution. Note that overmuch data augmentation will hurt the performance, hence we apply the configuration of EDA [34] to initialize it. Besides, we give extra analysis for testing the robustness of four augmentation strategies mentioned before in "Ablation study".

Input instance encoding

Sentence level encoder

Following the hyper-augmentation strategy module, we use BERT Encoder for our sentence-level embedding, of which details can be found from [35, 36]. Given an input of sentence S_i, we can obtain $[w_{i1}, e_{i1}, w_{i2}, \ldots, e_{i2}, w_{iN_w-1}, w_{iN_w}]$ by BERT Tokenizer, where N_w is the number of whole words, i is the i-th sentences of input instances, e_{i1} and e_{i2} are the corresponding subject and object entities respectively. Following the setting of encoder-only transformer, we add special tokens [18] to mark the beginning ([CLS]) and the end ([SEP]) of each sentence. After that, we transmit the token embedding and the position embedding to hidden information $[t_{[\mathrm{CLS}]}, t_{i1}, t_{ie_1}, t_{i2}, \ldots, t_{ie_2}, t_{iN_t-1}, t_{iN_t}, t_{[\mathrm{SEP}]}]$, where N_t is the number of all tokens. We denote the concatenation of two entity hidden vectors as the entity-aware sentence representation $\mathcal{H}(S_i) = h_i = [t_{e_{i1}} : t_{e_{i2}}]$.

Bag level encoder

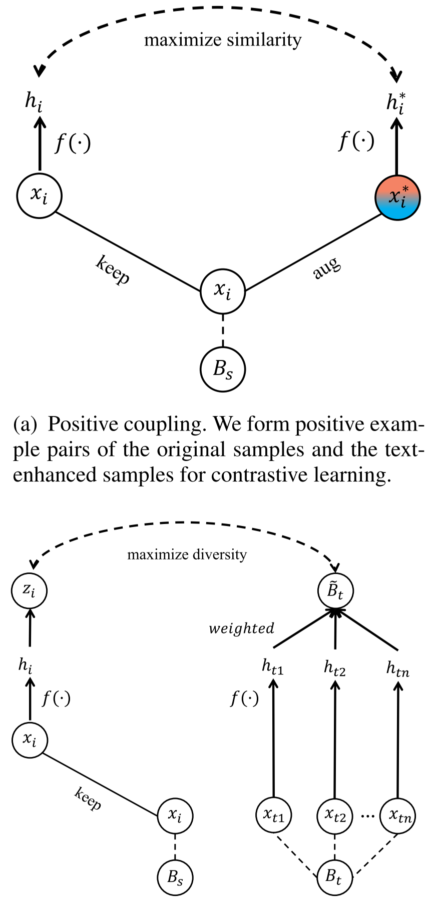
Sticking with the DS assumption is hard to avoid noise. Hence, we follow MIL framework and adopt selective attention [15] to highlight sentences that better express the current relation of each bag. Given an encoder $\mathcal{F}$ for bag B, B is a group of sentences that under the same fact $\langle e_1, e_2, r_c\rangle$, r_c is the correct relation. $\widetilde{B}$ is the weighted sum of all instances representations:

$$\widetilde{B} = \mathcal{F}(\langle e_1, e_2, r_c\rangle) = \sum_{i=1}^{N_b} \alpha_{ki} h_i \tag{2}$$

where N_b is the bag size, h_i follows the definition in sentence encoder. α_{ki} is soft-attention score between the kth relation and the ith instance, $\alpha_{ki} = \exp(a_{ki})/\sum_{j=1}^{N_b} \exp(a_{kj})$, where a_{ki} is the matching degree between the kth relation query and the ith sentence in a bag.

After that, we utilize a fully connected layer with Softmax to obtain the probability distribution over the relations:

$$p(r|\widetilde{B}, \theta) = \frac{e^{o_r}}{\sum_{i=1}^{N_r} e^{o_i}} \tag{3}$$



(a) Positive coupling. We form positive example pairs of the original samples and the text-enhanced samples for contrastive learning.

(b) Negative coupling. We form negative example pairs of the original sample and the other samples in the whole batch for contrastive learning.

Fig. 4 The framework of our dual-contrastive learning

where θ is the learnable parameter, $o = M\widetilde{B} + b$ represents the relative score to all relation labels, N_r is the total number of relations, M is a relation weight matrix, and b is bias term.

In bag encoder, our model takes cross-entropy loss as DSRE task loss $\mathcal{L}_{\mathcal{B}}(\theta)$, where N_s is the set of all training samples:

$$\mathcal{L}_{\mathcal{B}}(\theta) = -\sum_{i=1}^{N_s} \log p(r_i|\widetilde{B}_i, \theta) \tag{4}$$

Dual contrastive learning

The secret of contrastive learning is constructing positive and negative sample pairs to extract information by itself. Following previous work [37], we aim to maximize the mutual information between the positive pairs and push apart the cluster of negative samples.

Positive pair construction

The critical question is how to construct (x_i, x_i^+) pairs. As Fig. 4a illustrates, given the major instances x_m and the tail instance x_t in a batch, where $\{x_i = x_m + x_t | x_m \in B_{\text{Major}}, x_t \in B_{\text{Tail}}\}$. We denote the mixed view $x^* = z_{\text{SR}}(x_t) + z_{\text{RS}}(x_m)$, where $z(\cdot)$ is the selective augmentation. Finally, we adopt a bert-based encoder $f(\cdot)$ to extract the representation.

Negative pair construction

As the Fig. 4b exhibits, through out the bag encoder, we get a batch of bags $(B_1, B_2, ..., B_G)$ with their weighted representation which is denoted as $(\widetilde{B}_1, \widetilde{B}_2, ..., \widetilde{B}_G)$, where G is batch size. To avoid the "collapse", vast quantities of negative samples are needed. Following the negative pairs construction [21], we use another weighted bag representation $\widetilde{B}_t$ in the same batch serve as x_i^-, where $x_i \in B_s$ and $t \neq s$.

Training objective

To employ bag-level and sentence-level mutual information, we define a two-part training objective:

$$\mathcal{L}(\theta) = \sum_{B} \sum_{x \in B} \mathcal{L}_c(x; \theta) + \mathcal{L}_B(\theta) \tag{5}$$

where θ is learnable parameters, and B is the bag set which consists of corresponding instances.

The main objective for sentence-level relation extraction is defined to minimize a contrastive loss L_c. We choose InfoNCE Loss [38] to be our loss function:

$$\mathcal{L}_c(x_s; \theta) = -\log \frac{e^{\text{sim}(h_s, h_s^*)}}{e^{\text{sim}(h_s, h_s^*)} + \sum_t e^{\text{sim}(h_s, \widetilde{B}_t)}} \tag{6}$$

where h_s and h_s^* are the representation of positive instances pairs, $\text{sim}(h_s, h_s^*)$ denotes the similarity between two vectors, and $t \neq s$.

As discussed in "Input instance encoding", the bag-level objective can be expressed by Eq. (4). According to [39], introducing an randomly masks on some tokens to avoid catastrophic forgetting. We denote $L_{\mathcal{M}}(\theta)$ as an auxiliary objective to improve generalization.

Following the CIL framework, we take the bag encoder loss as domination at the beginning of training, and finally increases the CL loss equally. This helps smooth bag-level and sentence-level training.

Therefore, our final objective combines Eq. (5) and $L_{\mathcal{M}}(\theta)$ as

$$\mathcal{L}(\theta) = \frac{\lambda(t)}{N_B} \sum_{B} \sum_{x \in B} \mathcal{L}_C(x; \theta) + \mathcal{L}_B(\theta) + \lambda_M \mathcal{L}_M(\theta) \tag{7}$$

where $\lambda(t) = \frac{2}{1+e^{-t}} - 1$ denotes a balance coefficient, t is relative training steps, N_B is the number of whole instances in the batch, and λ_M is the weight of language model objective $L_{\mathcal{M}}$.

Experiments

Experimental setup

Datasets

We evaluate our method on three popular datasets including NYT10, NYT10m and Wiki20. NYT10 [40] has been broadly used as a benchmark dataset. The dataset is constructed with New York Times corpus from 2005 to 2007 and aligned with Freebase knowledge base [40]. The dataset has 52 common relations and a special relation NA indicated an unknown relation between entity pair. The training data contains 522,611 sentences, 281,270 entity pairs and 18,252 relational facts. NYT10m [41] is built by manually annotated test sets of classical NYT10 to make DSRE evaluation more credible. Specifically, the NYT10m dataset knockouts redundant samples and compresses relation in the new dataset. There are 25 relations on the dataset. The training set contains 417,893 sentences and 17,137 facts with 80% NA relations. With efforts of human labeling, there are 157,859 sentences and 1940 facts with 96% N/A relations in test set, which are an important part of DS evaluation. Wiki20m [41] is re-organized by searching the same relation ontology in Wiki80 [42] and re-splitting the train/validation/test sets, which is constructing by distantly supervised label from Wikipedia and Wikidata [43]. There are 81 relation classes on the dataset, and we remove the sentences which have only one entity. The training set contains 698,721 sentences and 157,740 facts with 59% NA relations. After manual labeling, there are 137,986 sentences and 56,000 facts with 25% N/A relations in test set, which is an high quality benchmark on DS problem. It is worth noting that we did not conduct our method on GIDS [44] because this dataset is well designed on the relation with balanced instances, which means it does not fit the long-tail distribution. For exploring the ability of our proposed method on different data distributions and diverse relation types, we crop the dataset with a long-tailed way based on these types, each of which is carried out under adoption construct the dataset.

Metrices

Following previous work [19, 28], we conduct a held-out evaluation on NYT10. On the human-annotated dataset NYT10m, we further test our model. We adopt standard Precision–Recall curves (PR-curve), the Area Under Curve

(AUC) and the Precision at N (P@N) to evaluate the model. More concretely, the Marco-F1 score and the accuracy of Hits@K are also reported to consider the long-tail metrics for different cutoffs.

Compared baselines

We compare our LTACL to several baseline models for a comprehensive comparison to demonstrate the superiority of our LTACL. Our comparison mainly focuses on two groups of models: the models for traditional DSRE and the models designed for the long-tail problem. And some baselines are used on both sides.

Traditional DSRE models Mintz [13] is a multi-class logistic regression RE model under DS setting. PCNN-ATT [14] is a piece-wise CNN model with selective attention over instances. RESIDE [28] is a NN model that makes use of relevant side information (entity types and relational phrases) and employs Graph CNN to capture syntactic information of instances. REDSandT [20] is a transformer-based DSRE method that manages to capture highly informative instances and label embeddings by exploiting BERT pretrained model. CIL [21] integrates contrastive learning framework with multi-instance learning, and takes full advantage of sentence information. HiCLRE [24] introduces a multi-level and multi-Granularity contrastive learning with recontextualization module to reduce the influence of noisy data in DSRE. Bert-ATT [45] replaces the piece-wise CNN model with BERT in sentence-level learning and adopts ONE aggregator for bag-level training. RH-Net [32] employs the tree search for relation extraction and fitters noisy instances based on reinforcement learning method, solving the noisy labeling and long-tail problem simultaneously for DSRE.

Long-tail awareness models ToHRE [46] formulates the DSRE as a hierarchical classification task and proposes a top-down classification strategy with a hierarchical bag presentation. PCNN+HATT [47] leverages relation hierarchies information and gain a coarse-to-fine grained attention for DSRE. PCNN+KATT [31] is an attention-based method and utilizes knowledge base embedding with graph neural network to represent the hierarchical relational label.

Implementation details

During our experiments, the nlpaug[2] is used to perform our selective augmentation. In random swap, we set aug_ min $=$ 1, aug_ max $=$ 10 and aug_p $=$ 0.3, while in synonym replacement, we set aug_ min $=$ 0, aug_ max $=$ 10 and aug_p $=$ 0.2. Following CIL, we reuse pretrained checkpoints of BERT [18] (uncased) as sentence encoder. While training, due to limitation of equipment, we set our batch size=16, after 3 epochs we test our model. In experiments, we use the grid search to find the learning rate for the learning rate within [1×10e−6,5×10e−4] and find the optimal lr = $3e-5$. Experiments conducted on a NVIDIA's GeForce RTX 3090.

Computational burden and complexity

In detail, we deployed our experimental models and training/testing codes on an Ubuntu 20.04 environment. The GPU used was an RTX3090 with 24 GB of memory, and the CPU used was an Intel(R) i7-10700 2.90 GHz with 40 GB of RAM. Our deep neural network development platform was PyTorch 1.10.1, and we utilized CUDA version 11.1 and cuDNN version 8.0.5.35 for neural network acceleration. Following the metrics proposed in [8] for analyzing model complexity, we assessed the complexity of our proposed LTACL model based on four aspects, including the number of neural network parameters (Params), floating-point operations (FLOPs), single-model inference time, and the computational resources (GPU memory) required for model deployment. This analysis allows us to measure the computational burden imposed by our proposed method. (1) Params. We measure the spatial complexity of our proposed method by calculating the number of neural network parameters of the LTACL model. The total number of parameters of the model is 112 M (112002701). (2) FLOPs. We calculate the FLOPs generated by our model with the assistant of the DeepSpeed library.[3] Our total FLOPs are up to 913.27 G times. (3) Inference Time. When training our model in 30 epoches, it almost takes 12660.84 s, while for the inference phase, we calculated the average inference speed per sample, and our model achieves a relation extraction speed of 16 samples/s on the current evaluation. (4) GPU memory. After five times repeat test, our model can to reach a peak GPU memory usage of 19.2 GBs when batchsize = 16 during training.

Overall results

General experimental results

In this section, we present the performance of different models on two datasets. For the NTY10 dataset, Table 1 shows more numerical comparisons among several competitive methods. It can be observed that our model achieves state-of-the-art performance on DSRE by obtaining 51.41% AUC. Our LTACL outperformer the HiCLRE by 6.11% at AUC score. Compared to the latest CIL method that relies on single view of augmentation, our approach gets absolute AUC improvements of 2.61% on DSRE. We find

[2] https://github.com/makcedward/nlpaug.

[3] https://github.com/microsoft/DeepSpeed.

Table 1 AUC and P@N evaluation results on NYT10

RE methods	AUC	P@100	P@200	P@300	P@500	P@M
Mintz[a]	10.70	52.30	50.20	45.00	39.70	46.80
PCNN-ATT[a]	32.81	73.27	68.66	61.79	59.88	65.90
RH-Net[a]	37.81	81.00	84.50	80.67	72.00	79.54
RESIDE[a]	41.50	81.80	75.40	74.30	69.70	77.20
REDSandT	42.40	78.00	74.00	73.00	67.60	73.15
HiCLRE[a]	45.3	82.0	78.5	74.0	–	78.2
CIL (bs = 16)[b]	48.80	77.20	76.10	76.10	73.30	76.50
LTACL	**51.41**	**87.13**	80.10	78.74	**76.45**	**80.60**
	2.61%	6.13%	−4.4%	−1.93%	3.15%	1.06%

P@N represents precision calculated for the top-N rated relation instances

[a] Results from the corresponding official codes

[b] Baseline results given by our implementation. The Bold in the table indicates our LTACL achieves the best performance compare to all baselines, while the Underline signifies the best performance among the compared baselines

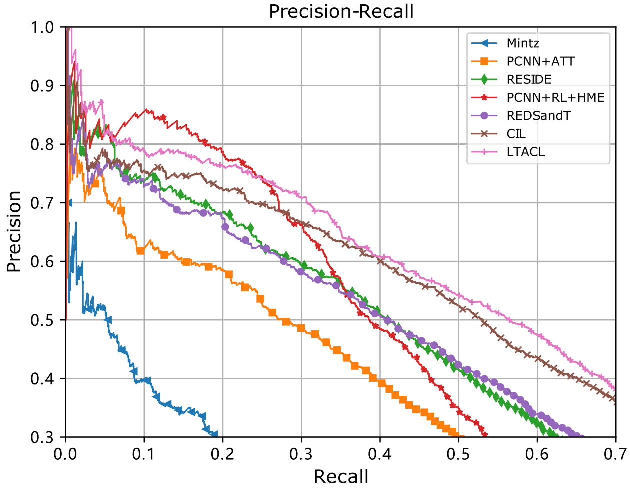


Fig. 5 Precision–Recall curves on NYT10 dataset

even stronger performance increases with respect to DSRE (+9.01%) when compared to the REDSandT baseline, which uses transformer-based sentence encoder for tokens with their types and applies attention mechanism on relations. For P@N metric, our method underperforms in P@200 (−4.4%) and P@300 (−1.9%) against RH-Net, which preprocesses the relations with an elaborate hierarchical search. However, Table 1 shows that our model significantly outperforms previous methods in most cases for averages of P@100 to P@500, indicating that our model has a consistent performance.

We also compare the Precision–Recall curves of our model with several major baselines to further evaluate the overall performance in Fig. 5. Consistent with the previous results, RH-Net performs better at first but drops rapidly in precision after a recall-level of approximately 0.25. Notably, our method can achieve higher precision over most part of the entire range of recall.

Table 2 AUC and Micro-F1 scores evaluation on NYT10m

RE methods	AUC	Micro-F1
PCNN-ATT[a]	56.80	56.50
RESIDE[a]	35.80	43.30
Bert-ATT[a]	51.20	54.10
HiCLRE[a]	61.4	36.9
CIL[b]	57.66	60.91
LTACL[b]	**59.42**	**62.38**
	−1.98%	1.47%

[a] Baseline results reported in [41]

[b] Baseline results given by our implementation. The Bold in the table indicates our LTACL achieves the best performance compare to all baselines, while the Underline signifies the best performance among the compared baselines

Table 3 AUC and Micro-F1 scores evaluation on Wiki20m

RE methods	AUC	Micro-F1
PCNN-ATT[a]	77.50	71.20
Bert-ATT[a]	70.90	66.80
CIL[b]	77.00	70.70
LTACL[b]	**79.19**	**71.22**
	2.19%	0.52%

[a] Baseline results reported in [41]

[b] Baseline results given by our implementation. The Bold in the table indicates our LTACL achieves the best performance compare to all baselines, while the Underline signifies the best performance among the compared baselines

To avoid the influence of noisy labeling by held-out, we take NYT10m, a human-labeled test data, for bag-level evaluation on our model. Table 2 shows that while PCNN-ATT achieves a surprising result against graph-based and Bert-based method, the contrastive learning framework is a notch above the rest. Our model surpasses the CIL in AUC

Table 4 AUC and Micro-F1 scores evaluation on long-tailed GIDS

RE methods	AUC	F1	P@500	P@1000	P@2000	P@M
CIL‡	88.73	85.19	93.41	90.61	83.26	89.09
LTACL	**89.96**	**86.42**	**95.21**	**93.31**	**88.26**	**92.26**
	1.23%	1.23%	1.8%	2.7%	5%	3.17%

We collect the scores for five variant versions of the GIDS dataset and retain their average performance. The old in the table indicates our LTACL achieves the best performance compare to the CIL

Table 5 Accuracy (%) of Hits@K(Macro) on relations with training instances less than 100/200 on NYT10

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
RH-Net	36.6	64.1	68.9	44.5	62.3	71.7
ToHRE	62.9	75.9	81.4	69.7	80.3	84.8
CIL	71.2	74.87	85.86	75.77	78.85	88.11
LTACL	**91.91**	**91.91**	**97.69**	75.57	**81.94**	**89.43**
	20.7%	16.52%	11.83%	−0.2%	1.64%	1.32 %

The Bold in the table indicates our LTACL achieves the best performance compare to all baselines, while the Underline signifies the best performance among the compared baselines

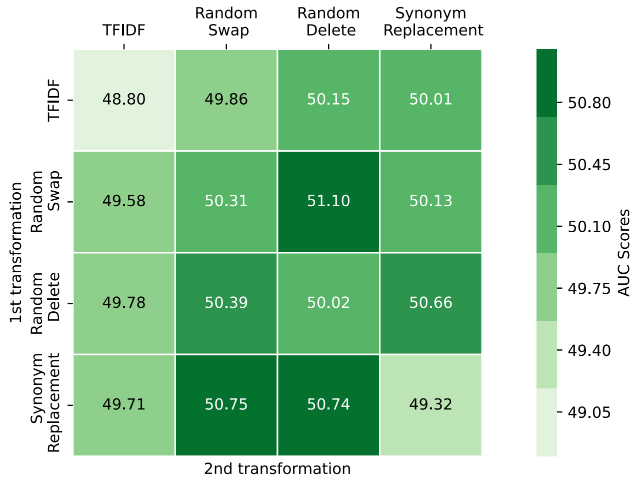


Fig. 6 The performance visualization (AUC score) under individual or composition of different augmentations. All these experiments follow our contrastive learning setting on NYT10 dataset. Columns and diagonal entries correspond to single transformation, and off-diagonals correspond to composition of two transformations applied sequentially (the row indicates data augmentation strategy for the 1st subset of data)

score (+1.76%). The above findings indicate that our model improves the generalization against the noise by various view with contrastive learning. However, our method reduces the AUC by 1.98% relative to HiCLRE, which implies that HiCLRE has a really strong ability in the prediction of positive examples. However, on the more balanced Micro-F1 metric, HiCLRE dramatically decreases to 36.9%, which also proves that the long-tail problem has a great impact on HiCLRE in DSRE. The performance on metric Micro-F1 further verifies that the proposed model outperforms CIL (+1.47%). Hence, through the results, we find out that LTACL is able to tackle the noise issue for relation extraction.

Moreover, we adopt Wiki20m dataset, which is a non-NYT source data for an additional evaluation of our model. As Table 3 presents, LTACL achieves a 79.19% AUC score and exceeds the CIL by 2.19%. Though our model obtains a minor discrepancy (+0.02%) on Micro-F1 when compared with PCNN-ATT, while CIL even fall behind (−0.5%). As we know the Wiki20m has larger number of instances with less long-tail relations, the well labeled on distantly supervised instances may weaken our advantages.

To further evaluate the performance of our method on datasets from different domains under a long-tailed distribution, we manually design a new DSRE dataset based on the GIDS dataset. In this new dataset, we down sample the data from each of the five relationship categories in GIDS and select one relationship category to reduce to only 300 samples, creating a long-tailed dataset. We construct five different versions of this dataset and compare the performance of our LTACL method with the CIL method. We average the results across all versions of the dataset. Table 4 shows the comparison results between our LTACL method and the CIL method in terms of F1 score. Our LTACL method outperforms the CIL method by 1.23% on average. These experiments demonstrate that our method is capable of achieving non trivial performance in solving the DSRE problem on datasets from different domains, even under a long-tailed distribution.

Table 6 Ablations on the NYT10 dataset

Model	AUC	Micro-F1	Marco-F1	P@M
LTACL$_{0.03}$	**51.41**	**53.36**	**39.22**	**80.60**
Reverse	50.15 (−1.26)	53.36 (−0.00)	38.34 (−0.88)	78.07 (−2.53)
δ : 0.0 (RS)	50.31 (−1.10)	53.05 (−0.31)	37.73 (−1.49)	80.49 (−0.11)
δ : 0.1	49.61 (−1.80)	52.17 (−1.19)	38.26 (−0.08)	79.24 (−1.36)
δ : 1.0 (SR)	49.32 (−2.09)	52.24 (−1.12)	38.24 (−0.98)	77.74 (−2.86)

The threshold value δ is set 0.03 empirically to divide the whole data into two parts. Specifically, RS is adopted for the long part while SR is utilized for tail part.The Bold in the table indicates the performance of our proposed LTACL with complete modules

Table 7 A case study selected from the subset of testing where entities are marked in bold

Sentence	CIL predicted	Our predicted	Ground truth
Steven Spielberg of **DreamWorks** are the hosts with a private dinner afterward at Mr. Geffen 's.(S1)	Founders ✓	Founders ✓	founders
The **Cajun** rocker **Zachary Richard** had a song that vowed...(S2)	Place_of _birth ✗	Ethnicity ✓	Ethnicity
He had grown up in **Louisiana** and learned to ride on the region's fabled **Cajun** bush tracks.(S3)	Place _lived ✗	Place _lived ✗	Geographic _distribution

Evaluation on long-tail relations

To further demonstrate the improvements on uneven relations, we evaluate our model on a subset of test dataset in which all the relations have fewer than 100 or 200 instances. The macro-average Hits@K accuracy is introduced for a fair comparison with methods which are dedicated to long-tail problem. Following previous work [46], we select K from $\{10, 15, 20\}$. Table 5 shows that our method obtains a surprising performance on these benchmarks. In the aspect of Hits@10 which are less than 200 training instances, even if our results are slightly less than CIL (−0.2%), our model achieves significantly performance on Hits@15 (+3.09%) and Hits@20 (+1.32%) settings. Note that our LTACL evaluates on the less than 100 training instances, has largely improved the performance in metric Hits@10 (+20.71%), Hits@15 (+16.01%) and Hits@20 (+11.83%) respectively. This demonstrates that our different level augmentation strategy for long-tail parts can better leverage different relation samples without being trapped in the majority class.

Ablation study

In this section, we conduct extensive ablation studies on the effect of mixed views and long-tail data processing.

Effect of mixed data augmentation To verify the impact of mixed views, we select four options for individual and composition of transformations including term frequency-inverse document frequency (TF-IDF), Random Swap (RS), Random Delete (RD) and Synonym Replacement (SR), according to the configuration of EDA [34]. Besides, we remove the long-tail discrimination by randomly dividing the NYT10 dataset into two parts equally, and adopt different operations on it in sequence. In Fig. 6, we observe that:

1. There is hardly a single transformation which suffices to learn better representations than the mixed views under the positive pairs construction. This confirms that the contrastive learning with mixed augmentations benefits DSRE task.
2. Even if we select the same augmentation combinations of RS and SR, the AUC scores of RS-SR (50.13%) and SR-RS (50.75%) are less than the score of long-tail awareness (51.41%). It reveals that even if the model adopts the mixed augmentation, ignoring long-tail distribution problem would affect the DSRE performance.

Ablations of long-tail partition processing In Table 6, we conduct extensive ablations on the hyperparameters of the δ and the different level augmentations to investigate the effectiveness of our long-tail awareness architecture. According to Eq. (1), we select δ from 0.0 (consider all data as major part), 1.0 (consider all data as tail part) and 0.1 (the tail part has 47 instances of classes). For a fair comparison, we also take a reverse strategy at δ is 0.03. We find that:

1. The inappropriate augmentation for major part and the tail degrades the performance of our model. When adopting a reverse strategy (RS for tail part and SR for major part), the AUC score decreases by 1.26% while the P@M decreases by 2.53%.
2. Incorrect division of long-tail parts lead to inefficiently learning of overall representation. When $\delta = 0.1$, more major classes have been brought into the tail part. The AUC drops by 1.80% while the Micro-F1 drops by 1.19%.
3. The single augmentation strategy neither suffices to learn better representations nor alleviates the long-tail distribution problem. When $\delta = 1.0$, it denotes that we only adopt RS for the whole data. The Micro-F1 score decreases by 0.31% while the Macro F1 decreases by 1.49%, indicating that the model has been trapped in long-tail bias. When $\delta = 0.0$, it denotes that we only adopt SR for the whole data. The results of Micro F1 score 52.24%(−1.12%) and Macro F1 38.24%(−0.98%) further prove our viewpoint.

Consequently, our model gives the appropriate partition for each part of the long-tail distribution. These metrics prove that LTACL selects right augmentations for capturing varied representations and achieves a competitive overall performance in DSRE.

Case study

Table 7 shows the predicted relation under the contrastive learning framework in bag-level testing. Note that we select three relational triplets from the long-tail class. S1 is a easy sample that both methods give the correct prediction. S2 is a hard sample and may confuse the model. The CIL gives the wrong relation prediction place_of_birth, which contains more than 8000 training instances. The different parts of the long-tail data are taken into account by our model thus we make the right choice. Our method makes the wrong prediction geographic_distribution on the last sentence. We analyze that LTACL may not capture the key information because the word related to geographical location has been replaced by other inappropriate synonyms.

Conclusion

In this paper, we propose a novel long-tail awareness contrastive learning (LTACL) model for distantly supervised relation extraction. We first show that long-tail effect in previous contrastive learning framework is largely underestimated. We then identify major and tail parts on ubiquitous long-tail distribution dataset. Our model LTACL selects varying degrees of data enhancements on different parts. And it is conceptual simple since it requires no redundant operation to construct positive pairs. The experimental results show that our approach provides nontrivial performance on various distantly supervised benchmarks and is effective in handling long-tail relation bias.

In the future, we will try to explore other contrastive learning methods to get rid of complicated object function terms, and experiment with long-tail selective on other relation extraction tasks to further prove its effectiveness.

Acknowledgements This work was supported as part of A Multi-Scale and High-Efficiency Computing Platform for Advanced Functional Materials, a scientific research program funded by Haihe Laboratory in Tianjin, Grants no. 22HHXCJC00007.

References

1. Lou J, Lu Y, Dai D, Jia W, Lin H, Han X, Sun L, Wu H (2023) Universal information extraction as unified semantic matching. arXiv preprint arXiv:2301.03282
2. Lu Y, Liu Q, Dai D, Xiao X, Lin H, Han X, Sun L, Wu H (2022) Unified structure generation for universal information extraction. arXiv preprint arXiv:2203.12277
3. Li Q, Peng H, Li J, Wu J, Ning Y, Wang L, Philip SY, Wang Z (2021) Reinforcement learning-based dialogue guided event extraction to exploit argument relations. IEEE/ACM Trans Audio Speech Lang Process 30:520–533
4. Liu F, Lin H, Han X, Cao B, Sun L (2022) Pre-training to match for unified low-shot relation extraction. arXiv preprint arXiv:2203.12274
5. Li Q, Li J, Wang L, Ji C, Hei Y, Sheng J, Sun Q, Xue S, Xie P (2023) Type information utilized event detection via multi-channel gnns in electrical power systems. ACM Trans Web 17(3):1–26
6. Yuan L, Cai Y, Wang J, Li Q (2023) Joint multimodal entity-relation extraction based on edge-enhanced graph alignment network and word-pair relation tagging. Proceedings of the AAAI conference on artificial intelligence 37:11051–11059
7. Wang X, Cai J, Jiang Y, Xie P, Tu K, Lu W (2022) Named entity and relation extraction with multi-modal retrieval. arXiv preprint arXiv:2212.01612
8. Chen Y, Harbecke D, Hennig L (2022) Multilingual relation classification via efficient and effective prompting. arXiv preprint arXiv:2210.13838

9. Hennig L, Thomas P, Möller S (2023) Multitacred: a multilingual version of the tac relation extraction dataset. arXiv preprint arXiv:2305.04582
10. Wadhwa S, Amir S, Wallace BC (2023) Revisiting relation extraction in the era of large language models. arXiv preprint arXiv:2305.05003
11. Tran V-H, Phi V-T, Shindo H, Matsumoto Y(2019) Relation classification using segment-level attention based CNN and dependency-based RNN. In: Proceedings of the 2019 Conference conference of the North American Chapter of the association for computational linguistics: human language technologies, Minneapolis, MN, USA, June 2-7, vol 1 (long and short papers). pp 2793–2798
12. Tian Y, Chen G, Song Y, Wan X (2021) Dependency-driven relation extraction with attentive graph convolutional networks. In: Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing, vol 1 (long papers). pp 4458–4471
13. Mintz M, Bills S, Snow R, Jurafsky D (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, Singapore, 2-7 August, pp 1003–1011
14. Zeng D, Liu K, Chen, Y, Zhao J (2015) Distant supervision for relation extraction via piecewise convolutional neural networks. In: Proceedings of the 2015 conference on empirical methods in natural language processing, Lisbon, Portugal, 17-21 September, pp 1753–1762
15. Lin Y, Shen S, Liu Z, Luan H, Sun M (2016) Neural relation extraction with selective attention over instances. In: Proceedings of the 54th Annual Meeting annual meeting of the association for computational linguistics, Berlin, Germany, 7-12 August, vol 1 (long papers). pp 2124–2133
16. Du J, Han J, Way A, Wan D (2018) Multi-level structured self-attentions for distantly supervised relation extraction. arXiv preprint arXiv:1809.00699
17. Ye Z-X, Ling Z-H (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, vol 1 (long and short papers). pp 2810–2819
18. Devlin J, Chang M-W, Lee K, Toutanova K (2019) Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the north american chapter of the association for computational linguistics: human language technologies, Minneapolis, MN, USA, 2-7 June, Volume 1, pp 4171–4186
19. Alt C, Hübner M, Hennig L (2019) Fine-tuning pre-trained transformer language models to distantly supervised relation extraction. arXiv preprint arXiv:1906.08646
20. Christou D, Tsoumakas G (2021) Improving distantly-supervised relation extraction through bert-based label and instance embeddings. IEEE Access 9:62574–62582
21. Chen T, Shi H, Tang S, Chen Z, Wu F, Zhuang Y (2021) Cil: contrastive instance learning framework for distantly supervised relation extraction. arXiv preprint arXiv:2106.10855
22. Chen T, Kornblith S, Norouzi M, Hinton G (2020) A simple framework for contrastive learning of visual representations. In: International conference on machine learning. PMLR, pp 1597–1607
23. Chen X, Fan H, Girshick R, He K (2020) Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297
24. Li D, Zhang T, Hu N, Wang C, He X (2022) Hiclre: a hierarchical contrastive learning framework for distantly supervised relation extraction. arXiv preprint arXiv:2202.13352
25. Peng T, Han R, Cui H, Yue L, Han J, Liu L (2022) Distantly supervised relation extraction using global hierarchy embeddings and local probability constraints. Knowl Based Syst 235:107637
26. Shang Y-M, Huang H, Sun X, Wei W, Mao X-L (2022) Learning relation ties with a force-directed graph in distant supervised relation extraction. ACM Trans Inf Syst (TOIS) 41:1–23
27. Qin P, Xu W, Wang WY(2018) Robust distant supervision relation extraction via deep reinforcement learning. In: Proceedings of the 56th annual meeting of the association for computational linguistics, Melbourne, Australia, 15-20 July, vol 1 (long papers). pp 2137–2147
28. Vashishth S, Joshi R, Prayaga SS, Bhattacharyya C, Talukdar P (2018) Reside: improving distantlysupervised neural relation extraction using side information. In: Proceedings of EMNLP, Brussels, Belgium, October 31 - November 4, pp 1257–1266
29. Sui D, Chen Y, Liu K, Zhao J (2020) Distantly supervised relation extraction in federated settings. arXiv preprint arXiv:2008.05049
30. Gui Y, LiuQ, Zhu M, Gao Z (2016) Exploring long tail data in distantly supervised relation extraction. In: Lin, C., Xue, N., Zhao, D., Huang, X., Feng, Y. (eds.) Natural language understanding and intelligent applications, Springer, Kunming, China, 2-6 December, pp . 514–522
31. Zhang N, Deng S, Sun Z, Wang G, Chen X, Zhang W, Chen H (2019) Long-tail relation extraction via knowledge graph embeddings and graph convolution networks. In: Proceedings of NAACL-HLT, Minneapolis, MN, USA, 2-7 June, pp 3016–3025
32. Wang J (2020) Rh-net: improving neural relation extraction via reinforcement learning and hierarchical relational searching. arXiv preprint arXiv:2010.14255
33. Yan Y, Li R, Wang S, Zhang F, Wu W, Xu W (2021) Consert: a contrastive framework for self-supervised sentence representation transfer. arXiv preprint arXiv:2105.11741
34. Wei J, Zou K (2019) Eda: easy data augmentation techniques for boosting performance on text classification tasks. arXiv preprint arXiv:1901.11196
35. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems. pp 5998–6008
36. Soares LB, FitzGerald N, Ling J, Kwiatkowski T (2019) Matching the blanks: Distributional similarity for relation learning. arXiv preprint arXiv:1906.03158
37. Wang T, Isola P (2020) Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International conference on machine learning. PMLR, pp 9929–9939
38. Oord A, Li Y, Vinyals O (2018) Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748
39. McCloskey M, Cohen NJ (1989) Catastrophic interference in connectionist networks: the sequential learning problem. In: Psychology of learning and motivation, vol 24. pp 109–165
40. Riedel S, Yao L, McCallum A (2010) Modeling relations and their mentions without labeled text. In: Joint European conference on machine learning and knowledge discovery in databases. Springer, pp 148–163
41. Gao T, Han X, Qiu K, Bai Y, Xie Z, Lin Y, Liu Z, Li P, Sun M, Zhou J (2021) Manual evaluation matters: Reviewing test protocols of distantly supervised relation extraction. arXiv preprint arXiv:2105.09543
42. Han X, Gao T, Yao Y, Ye D, Liu Z, Sun M (2019) Opennre: an open and extensible toolkit for neural relation extraction. In: Proceedings of EMNLP, Hong Kong, China, 3-7 November, pp 169–174
43. Han X, Gao T, Lin Y, Peng H, Yang Y, Xiao C, Liu Z, Li P, Sun M, Zhou J (2020) More data, more relations, more context and more openness: a review and outlook for relation extraction. arXiv preprint arXiv:2004.03186

44. Jat S, Khandelwal S, Talukdar P (2018) Improving distantly supervised relation extraction using word and entity based attention. arXiv preprint arXiv:1804.06987
45. Amin S, Dunfield KA, Vechkaeva A, NeumannG (2020) A data-driven approach for noise reduction in distantly supervised biomedical relation extraction. In: SIGBioMed workshop on biomedical language processing, Online, 9 July, pp 187–194
46. Yu E, Han W, Tian Y, Chang Y(2020) Tohre: a top-down classification strategy with hierarchical bag representation for distantly supervised relation extraction. In: Proceedings of COLING, Barcelona, Spain (Online), 8-13 December, pp 1665–1676
47. Han X, Yu P, Liu Z, Sun M, Li P (2018) Hierarchical relation extraction with coarse-to-fine grained attention. In: Proceedings of EMNLP, Brussels, Belgium, October 31 - November 4, pp 2236–2245