

CIL: Contrastive Instance Learning Framework for Distantly Supervised Relation Extraction

Tao Chen^{1,2}, Haizhou Shi¹, Siliang Tang^{1,2*}, Zhigang Chen³,
Fei Wu¹ & Yueting Zhuang¹

¹Zhejiang University

²Alibaba-Zhejiang University Joint Research Institute of Frontier Technologies

³State Key Laboratory of Cognitive Intelligence, Hefei, China

{ttc, shihaizhou, siliang, yzhuang}@zju.edu.cn
zgchen@iflytek.com, wufei@cs.zju.edu.cn

Abstract

The journey of reducing noise from distant supervision (DS) generated training data has been started since the DS was first introduced into the relation extraction (RE) task. For the past decade, researchers apply the multi-instance learning (MIL) framework to find the most reliable feature from a bag of sentences. Although the pattern of MIL bags can greatly reduce DS noise, it fails to represent many other useful sentence features in the datasets. In many cases, these sentence features can only be acquired by extra sentence-level human annotation with heavy costs. Therefore, the performance of distantly supervised RE models is bounded. In this paper, we go beyond typical MIL framework and propose a novel Contrastive Instance Learning (CIL) framework. Specifically, we regard the initial MIL as the relational triple encoder and constraint positive pairs against negative pairs for each instance. Experiments demonstrate the effectiveness of our proposed framework, with significant improvements over the previous methods on NYT10, GDS and KBP.

1 Introduction

Relation extraction (RE) aims at predicting the relation between entities based on their context. Several studies have been carried out to handle this crucial and complicated task over decades as the extracted information can serve as a significant role for many downstream tasks. Since the amount of training data generally limits traditional supervised RE systems, current RE systems usually resort to distant supervision (DS) to fetch abundant training data by aligning knowledge bases (KBs) and texts. However, such a heuristic way inevitably introduces some noise to the generated data. Training a robust and unbiased RE system under DS data

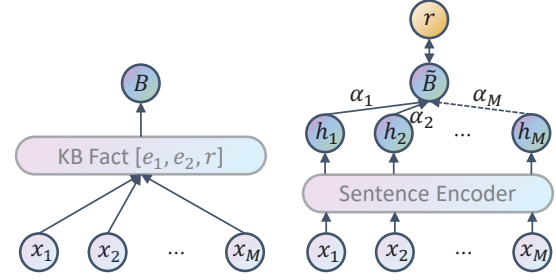


Figure 1: Classical MIL framework for DSRE. (Left) A set of instances (x_1, x_2, \dots, x_m) with the same KB fact $[e_1, e_2, r]$ form a bag B ; (Right) The MIL framework trains the DSRE model at bag level ($\hat{B} : \sum \alpha_i h_i$).

noise becomes the biggest challenge for distantly supervised relation extraction (DSRE).

With awareness of the existing DS noise, Zeng et al. (2015) introduces the multi-instance learning (MIL) framework to DSRE by dividing training instances into several bags and using bags as new data units. Regarding the strategy for selecting instances inside the bag, the soft attention mechanism proposed by Lin et al. (2016) is widely used for its better performance than the hard selection method. The ability to form accurate representations from noisy data makes the MIL framework soon become a paradigm of following-up works.

However, we argue that the MIL framework is effective to alleviate data noise for DSRE, but is not data-efficient indeed: As Figure 1 shows: The attention mechanism in the MIL can help select relatively informative instances (e.g. h_1, h_2) inside the bag, but may ignore the potential information of other abundant instances (e.g. h_m). In other words, no matter how many instances a bag contains, only the formed bag-level representation can be used for further training in the MIL, which is quite inefficient. Thus, our focus is on how to make the initial MIL framework *efficient enough to leverage all instances while maintaining the ability to obtain an accurate model under DS data noise*.

* Corresponding author

Here, we propose a contrastive-based method to help the MIL framework learn efficiently. In detail, we regard the initial MIL framework as the bag encoder, which provides relatively accurate representations for different relational triples. Then we develop contrastive instance learning (CIL) to utilize each instance in an unsupervised manner: In short, the goal of our CIL is that the instances sharing the same relational triples (*i.e.* positive pairs) ought to be close in the semantic space, while the representations of instances with different relational triples (*i.e.* negative pairs) should be far away.

Experiments on three public DSRE benchmarks — NYT10 (Riedel et al., 2010; Hoffmann et al., 2011), GDS (Jat et al., 2018) and KBP (Ling and Weld, 2012) demonstrate the effectiveness of our proposed framework CIL, with consistent improvements over several baseline models and far exceed the state-of-the-art (SOTA) systems. Furthermore, the ablation study shows the rationality of our proposed positive/negative pair construction strategy.

Accordingly, the major contributions of this paper are summarized as follows:

- We discuss the long-standing MIL framework and point out that it can not effectively utilize abundant instances inside MIL bags.
- We propose a novel contrastive instance learning method to boost the DSRE model performances under the MIL framework.
- Evaluation on held-out and human-annotated sets shows that CIL leads to significant improvements over the previous SOTA models.

2 Methodology

In this paper, we argue that the MIL framework is effective to denoise but is not efficient enough, as the initial MIL framework only leverages the formed bag-level representations to train models and sacrifices the potential information of numerous instances inside bags. Here, we go beyond the typical MIL framework and develop a novel contrastive instance learning framework to solve the above issue, which can prompt DSRE models to utilize each instance. A formal description of our proposed CIL framework is illustrated as follows.

2.1 Input Embeddings

Token Embedding For input sentence/instance x , we utilize BERT Tokenizer to split it into several

tokens: $(t_1, t_2, \dots, e_1 \dots e_2 \dots t_L)$, where e_1, e_2 are the tokens corresponding to the two entities, and L is the max length of all input sequences. Following standard practices (Devlin et al., 2019), we add two special tokens to mark the beginning ([CLS]) and the end ([SEP]) of sentences.

In BERT, token [CLS] typically acts as a pooling token representing the whole sequence for downstream tasks. However, this pooling representation considers entity tokens e_1 and e_2 as equivalent to other common word tokens t_i , which has been proven (Baldini Soares et al., 2019) to be unsuitable for RE tasks. To encode the sentence in an entity-aware manner, we add four extra special tokens ([H-CLS], [H-SEP]) and ([T-CLS], [T-SEP]) to mark the beginning and the end of two entities.

Position Embedding In the Transformer attention mechanism (Vaswani et al., 2017), positional encodings are injected to make use of the order of the sequence. Precisely, the learned position embedding has the same dimension as the token embedding so that the two can be summed.

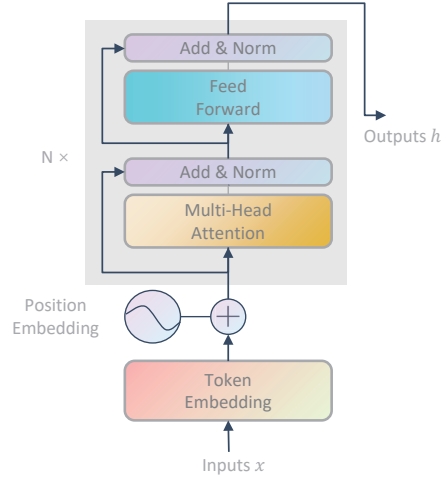


Figure 2: BERT Encoder: $N \times$ Transformer Blocks.

2.2 Sentence Encoder

BERT Encoder (Transformer Blocks, see Figure 2) transforms the above embedding inputs (token embedding & position embedding) into hidden feature vectors: $(h_1, h_2, \dots, h_{e_1} \dots h_{e_2} \dots h_L)$, where h_{e_1} and h_{e_2} are the feature vectors corresponding to the entities e_1 and e_2 . By concatenating the two entity hidden vectors, we can obtain the entity-aware sentence representation $h = [h_{e_1}; h_{e_2}]$ for the input sequence x . We denote the sentence encoder \mathcal{H} as:

$$\mathcal{H}(x) = [h_{e_1}; h_{e_2}] = h$$

2.3 Bag Encoder

Under the MIL framework, a couple of instances x with the same relational triple $[e_1, e_2, r]$ form a bag B . We aim to design a bag encoder \mathcal{F} to obtain representation \tilde{B} for bag B , and the obtained bag representation is also a representative of the current relational triple $[e_1, e_2, r]$, which is defined as:

$$\mathcal{F}(B) = \mathcal{F}([e_1, e_2, r]) = \tilde{B}$$

With the help of the sentence encoder described in section 2.2, each instance x_i in bag B can be first encoded to its entity-aware sentence representation $h_i = \mathcal{H}(x_i)$. Then the bag representation \tilde{B} can be regarded as an aggregation of all instances' representations, which is further defined as:

$$\mathcal{F}([e_1, e_2, r]) = \tilde{B} = \sum_{i=1}^K \alpha_i h_i$$

where K is the bag size. As for the choice of weight α_i , we follow the soft attention mechanism used in (Lin et al., 2016), where α_i is the normalized attention score calculated by a query-based function f_i that measures how well the sentence representation h_i and the predict relation r matches:

$$\alpha_i = \frac{e^{f_i}}{\sum_j e^{f_j}}$$

where $f_i = h_i A q_r$, A is a weighted diagonal matrix and q_r is the query vector which indicates the representation of relation r (randomly initialized).

Then, to train such a bag encoder parameterized by θ , a simple fully-connected layer with activation function *softmax* is added to map the hidden feature vector \tilde{B} to a conditional probability distribution $p(r|\tilde{B}, \theta)$, and this can be defined as:

$$p(r|\tilde{B}, \theta) = \frac{e^{o_r}}{\sum_{i=1}^{n_r} e^{o_i}}$$

where $o = M\tilde{B} + b$ is the score associated to all relation types, n_r is the total number of relations, M is a projection matrix, and b is the bias term.

And we define the objective of bag encoder using cross-entropy function as follows:

$$\mathcal{L}_B(\theta) = - \sum_{i=1} \log p(r_i|\tilde{B}_i, \theta)$$

2.4 Contrastive Instance Learning

As illustrated in section 1, the goal of our framework CIL is that the instances containing the same relational triples (*i.e.* positive pairs) should be as close (*i.e.* \sim) as possible in the hidden semantic space, and the instances containing different relational triples (*i.e.* negative pairs) should be as far (*i.e.* ∞) away as possible in the space. A formal description is as follows.

Assume there is a batch bag input (with a batch size G): (B_1, B_2, \dots, B_G) , the relational triples of all bags are different from each other. Each bag B in the batch is constructed by a certain relational triple $[e_1, e_2, r]$, and all instances x inside the bag satisfy this triple. The representation of the triple can be obtained by bag encoder as \tilde{B} .

We pick any two bags B_s and $B_{t:t \neq s}$ in the batch to further illustrate the process of contrastive instance learning. B_s is defined as the source bag constructed with relational triple $[e_{s1}, e_{s2}, r_s]$ while B_t is the target bag constructed with triple $[e_{t1}, e_{t2}, r_t]$. And we discuss the positive pair instance and negative pair instances for any instance x_s in bag B_s .

It is worth noting that all bags are constructed automatically by the distantly supervised method, which extracts relational triples from instances in a heuristic manner and may introduce true/false positive label noise to the generated data. In other words, though the instance x is included in the bag with relational triple $[e_1, e_2, r]$, it may be noisy and fail to express the relation r .

2.4.1 Positive Pair Construction

Instance $x_s \sim$ Random Instance $x_{s'}$ One intuitive choice of selecting positive pair instance for instance x_s is just picking another instance $x_{s'} \neq x_s$ from the bag B randomly. However, both of the instances x_s and $x_{s'}$ may suffer from data noise, and they are hard to express the same relational triple simultaneously. Thus, taking instance x_s and randomly selected instance $x_{s'}$ as a positive pair is not an optimal option.

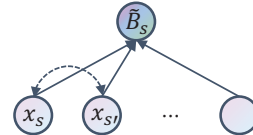


Figure 3: Instance $x_s \sim$ Random Instance $x_{s'}$

Instance $x_s \sim$ Relational Triple \tilde{B}_s Another positive pair instance candidate for instance x_s is

the relational triple representation \tilde{B}_s of current bag B . Though B_s can be regarded as a de-noised representation, x_s may be still noisy and express other relation $r \neq r_s$. Besides, the quality of constructed positive pairs heavily relies on the model performance of the bag encoder.

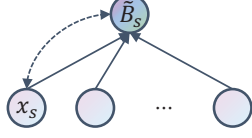


Figure 4: Instance $x_s \sim$ Relational Triple \tilde{B}_s

Instance $x_s \sim$ Augmented Instance x_s^* From the above analysis, we can see that the general positive pair construction methods often encounter the challenge of DS noise. Here, we propose a noise-free positive pair construction method based on **TF-IDF data augmentation**: If we only make small and controllable data augmentation to the original instance x_s , the augmented instance x_s^* should satisfy the same relational triple with instance x_s .

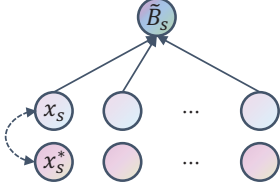


Figure 5: Instance $x_s \sim$ Augmented Instance x_s^*

In detail: (1) We first view each instance as a document and view each word in the instances as a term, then we train a *TF-IDF* model on the total training corpus. (2) Based on the trained *TF-IDF* model, we insert/substitute some unimportant (low *TF-IDF* score, see Figure 6) words to/in instance x_s with a specific ratio, and can obtain its augmented instance x_s^* . Particularly, special masks are added to entity words to avoid them being substituted.



Figure 6: An example of word substitution: The low-scoring word *of* is replaced with word *goehr*, and entity words *stephen king* and *maine* are protected.

2.4.2 Negative Pair Construction

Instance $x_s \sim$ Random Instance x_t Similarly, for instance x_s in bag B_s , we can randomly select an instance x_t from another different bag B_t as its

negative pair instance. Under this strategy, x_s is far away from the average representation $\sum_{i=1}^K \alpha_i h_i$ of the bag B_t , where all $\alpha_i = \frac{1}{K}$ approximately. And the randomly selected instance x_t may be too noisy to represent the relational triple of bag B_t , so that the model performance may be influenced.

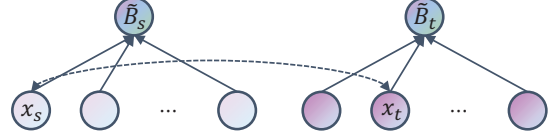


Figure 7: Instance $x_s \sim$ Random Instance x_t

Instance $x_s \sim$ Relational Triple \tilde{B}_t Compared to the random selection strategy, using relational triple representation \tilde{B}_t as the negative pair instance for x_s is a better choice to reduce the impact of data noise. As the instance x_i can be seen as be far away from a weighted representation $\sum_{i=1}^K \alpha_i h_i$ of the bag B_t , where all α_i are learnable. Though the instance x_s may still be noisy, x_s and \tilde{B}_t can not belong to the same relational triple.

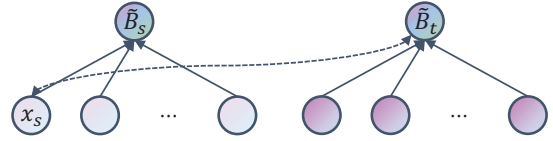


Figure 8: Instance $x_s \sim$ Relational Triple \tilde{B}_t

2.5 Training Objective

As discussed above, for any instance x_s in the source bag B_s : (1) The instance x_s^* after controllable data augmentation based on x_s is its positive pair instance. (2) The relational triple representations \tilde{B}_t of other different ($t \neq s$) bags in the batch are its negative pair instances. The overall schematic diagram of CIL is shown in Figure 9.

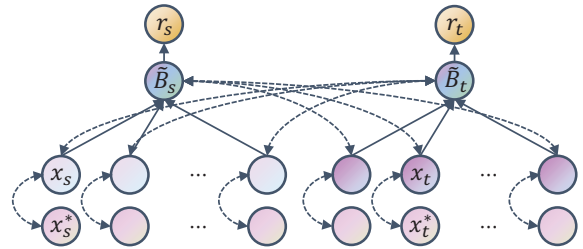


Figure 9: Contrastive Instance Learning

And we define the objective for instance x_s in

bag B_s using InfoNCE (Oord et al., 2018) loss:

$$\mathcal{L}_C(x_s; \theta) = -\log \frac{e^{\text{sim}(h_s, h_s^*)}}{e^{\text{sim}(h_s, h_s^*)} + \sum_{t:t \neq s} e^{\text{sim}(h_s, \tilde{B}_t)}}$$

where $\text{sim}(a, b)$ is the function to measure the similarity between two representation vectors a, b , and $h_s = \mathcal{H}(x_s)$, $h_s^* = \mathcal{H}(x_s^*)$ are the sentence representations of instances x_s, x_s^* .

Besides, to inherit the ability of language understanding from BERT and avoid catastrophic forgetting (McCloskey and Cohen, 1989), we also add the masked language modeling (MLM) objective to our framework. Pre-text task MLM randomly masks some tokens in the inputs and allows the model to predict the masked tokens, which prompts the model to capture rich semantic information in the contexts. And we denote this objective as $\mathcal{L}_M(\theta)$.

Accordingly, the total training objective of our contrastive instance learning framework is:

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

where $N = KG$ is the total number of instances in the batch, λ_M is the weight of language model objective \mathcal{L}_M , and $\lambda(t) \in [0, 1]$ is an increasing function related to the relative training steps t :

$$\lambda(t) = \frac{2}{1 + e^{-t}} - 1$$

At the beginning of our training, the value of $\lambda(t)$ is relatively small, and our framework CIL focuses on obtaining an accurate bag encoder (\mathcal{L}_B). The value of $\lambda(t)$ gradually increases to 1 as the relative training steps t increases, and more attention is paid to the contrastive instance learning (\mathcal{L}_C).

3 Experiments

Our experiments are designed to verify the effectiveness of our proposed framework CIL.

3.1 Benchmarks

We evaluate our method on three popular DSRE benchmarks — NYT10, GDS and KBP, and the dataset statistics are listed in Table 1.

NYT10 (Riedel et al., 2010) aligns Freebase entity relations with New York Times corpus, and it has two test set versions: (1) NYT10-D employs held-out KB facts as the test set and is still under distantly supervised. (2) NYT10-H is constructed manually by (Hoffmann et al., 2011), which contains 395 sentences with human annotations.

Dataset	# Rel.	# Ins.	# Test Ins.	# Test Set
NYT10-D	53	694,491	172,448	DS
NYT10-H	25	362,691	3,777	MA
GDS	5	18,824	5,663	Partly MA
KBP	7	148,666	1,940	MA

Table 1: Statistics of various used datasets. Rel.: relation, Ins.: instance and MA: manually annotated.

GDS (Jat et al., 2018) is created by extending the Google RE corpus with additional instances for each entity pair, and this dataset assures that the at-least-one assumption of MIL always holds.

KBP (Ling and Weld, 2012) uses Wikipedia articles annotated with Freebase entries as the training set, and employs manually-annotated sentences from 2013 KBP slot filling assessment results (Ellis et al., 2012) as the extra test set.

3.2 Evaluation Metrics

Following previous literature (Lin et al., 2016; Vashishth et al., 2018; Alt et al., 2019), we first conduct a held-out evaluation to measure model performances approximately on NYT10-D and GDS. Besides, we also conduct an evaluation on two human-annotated datasets (NYT10-H & KBP) to further support our claims. Specifically, Precision-Recall curves (PR-curve) are drawn to show the trade-off between model precision and recall, the Area Under Curve (AUC) metric is used to evaluate overall model performances, and the Precision at N (P@N) metric is also reported to consider the accuracy value for different cut-offs.

3.3 Baseline Models

We choose six recent methods as baseline models.

Mintz (Mintz et al., 2009) A multi-class logistic regression RE model under DS setting.

PCNN-ATT (Lin et al., 2016) A piece-wise CNN model with selective attention over instances.

MTB-MIL (Baldini Soares et al., 2019) A relation learning method based on distributional similarity, achieves amazing results for supervised RE¹.

RESIDE (Vashishth et al., 2018) 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.

¹For MTB-MIL, we firstly conduct MTB pre-training to learn relation representations on the entire training corpus and continually fine-tune the model by the MIL framework.

Method	NYT10-D					GDS				
	AUC	P@100	P@200	P@300	P@M	AUC	P@500	P@1000	P@2000	P@M
Mintz [†]	10.7	52.3	50.2	45.0	49.2	-	-	-	-	-
PCNN-ATT [‡]	34.1	73.0	68.0	67.3	69.4	79.9	90.6	87.6	75.2	84.5
MTB-MIL	40.8	76.2	71.1	69.4	72.2	88.5	94.8	92.2	87.0	91.3
RESIDE [‡]	41.5	<u>81.8</u>	75.4	<u>74.3</u>	<u>77.2</u>	89.1	94.8	91.1	82.7	89.5
REDSandT [‡]	<u>42.4</u>	78.0	75.0	73.0	75.3	86.1	95.6	92.6	84.6	91.0
DISTRE [†]	42.2	68.0	67.0	65.3	66.8	<u>89.9</u>	<u>97.0</u>	<u>93.8</u>	<u>87.6</u>	<u>92.8</u>
CIL [*]	50.8	90.1	86.1	81.8	86.0	91.6	98.4	95.3	88.7	94.1

Table 2: Model performances on NYT10-D and GDS. (†)/(‡) marks the results on (NYT10-D column)/(both columns) are reported in the previous papers. Bold and underline indicate the best and the second best scores, and * indicates that our model shows significant gains ($p > 0.05$) over the second-best model based on Student’s t-test.

REDSandT (Christou and Tsoumakas, 2021) A transformer-based DSRE method that manages to capture highly informative instance and label embeddings by exploiting BERT pre-trained model.

DISTRE (Alt et al., 2019) A transformer-based model, GPT fine-tuned for DSRE under the MIL.

3.4 Evaluation on Distantly Supervised Set

We summarize the model performances of our method and above-mentioned baseline models in Table 2. From the results, we can observe that: (1) On both two datasets, our proposed framework CIL achieves the best performance in all metrics. (2) On NYT10-D, compared with the previous SOTA model DISTRE, CIL improves the metric AUC (42.2→50.8) by 20.4% and the metric P@Mean (66.8→86.0) by 28.7%. (3) On GDS, though the metric of previous models is already high (≈ 90.0), our model still improves it by nearly 2 percentage points. (89.9→91.6 & 92.8→94.1).

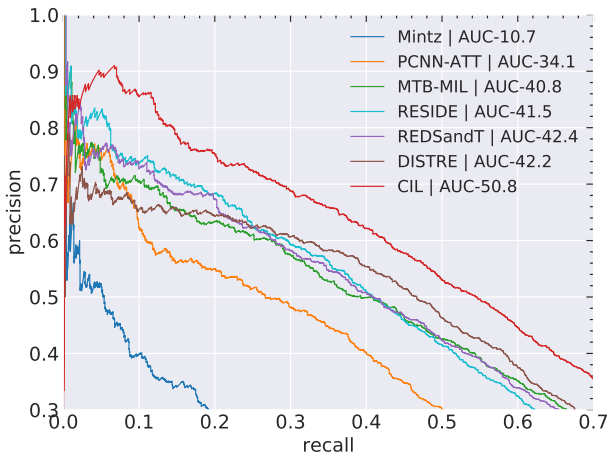


Figure 10: PR-curve on NYT10-D.

The overall PR-curve on NYT10-D is visualized

in Figure 10. From the curve, we can observe that: (1) Compared to PR-curves of other baseline models, our method shifts up the curve a lot. (2) Previous SOTA model DISTRE performs worse than model RESIDE at the beginning of the curve and yields a better performance after a recall-level of approximately 0.25, and our method CIL surpasses previous two SOTA models in all ranges along the curve, and it is more balanced between precision and recall. (3) Furthermore, as a SOTA scheme of relation learning, MTB fails to achieve competitive results for DSRE. This is because MTB relies on label information for pre-training, and noisy labels in DSRE may influence its model performance.

3.5 Evaluation on Manually Annotated Set

The automated held-out evaluation may not reflect the actual performance of DSRE models, as it gives false positive/negative labels and incomplete KB information. Thus, to further support our claims, we also evaluate our method on two human-annotated datasets, and the results² are listed in Table 3.

Method	NYT10-H			KBP		
	AUC	F1	P@M	AUC	F1	P@M
PCNN-A	38.9	47.0	58.6	15.4	31.5	32.8
DISTRE	37.8	50.9	54.1	<u>22.1</u>	<u>37.5</u>	<u>46.4</u>
CIL	46.0	55.5	63.0	30.1	44.0	48.2

Table 3: Model performances on NTY10-H and KBP. PCNN-A denotes PCNN-ATT. F1 refers to Micro-F1.

From the above result table, we can see that: (1) Our proposed framework CIL can still perform well under accurate human evaluation, with averagely 21.7% AUC improvement on NYT10-H and 36.2% on KBP, which means our method can generalize

²Manual evaluation is performed for each test sentence.

to real scenarios well. (2) On NYT10-H, DISTRE fails to surpass PCNN-ATT in metric P@Mean. This indicates that DISTRE gives a high recall but a low precision, but our method CIL can boost the model precision (54.1→63.0) while continuously improving the model recall (37.8→46.0). And the human evaluation results further confirm the observations in the held-out evaluation described above.

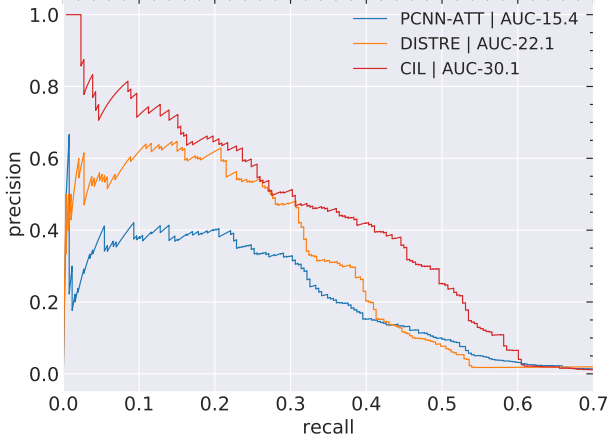


Figure 11: PR-Curve on KBP.

We also present the PR-curve on KBP in Figure 11. Under accurate sentence-level evaluation on KBP, the advantage of our model is more obvious with averagely 36.2% improvement on AUC, 17.3% on F1 and 3.9% on P@Mean, respectively.

3.6 Ablation Study

To further understand our proposed framework CIL, we also conduct ablation studies.

We firstly conduct an ablation experiment to verify that CIL has utilized abundant instances inside bags: (1) By removing our proposed contrastive instance learning, the framework degenerates into vanilla MIL framework, and we train the MIL on regular bags (MIL_{bag}). (2) To prove the MIL can not make full use of sentences, we also train the MIL on sentence bags (MIL_{sent}), which repeats each sentence in the training corpus to form a bag³.

Method	AUC	F1	P@M
CIL	50.8	52.2	86.0
MIL_{bag}	40.3(-10.5)	47.1(-5.1)	70.0(-16.0)
MIL_{sent}	36.0(-14.8)	43.5(-8.7)	63.3(-22.7)

Table 4: Model performances of three training patterns.

³ All the results in Table 4 are obtained under the same test setting that uses MIL bags (*i.e.* BERT+ATT) as test units.

From Table 4 we can see that: (1) MIL_{bag} only resorts to the accurate bag-level representations to train the model and fails to play the role of each instance inside bags; thus, it performs worse than our method CIL (50.8→40.3). (2) Though MIL_{sent} can access all training sentences, it loses the advantages of noise reduction in MIL_{bag} (40.3→30.6). The noisy label supervision may wrongly guide model training, and its model performance heavily suffers from DS data noise (86.0→63.3). (3) Our framework CIL succeeds in leveraging abundant instances while retaining the ability to denoise.

To validate the rationality of our proposed positive/negative pair construction strategy, we also conduct an ablation study on three variants of our framework CIL. We denote these variants as:

CIL_{randpos}: Randomly select an instance $x_{s'}$ also from bag B_s as the positive pair instance for x_s .

CIL_{bagpos}: Just take the relational triple representation B_s as the positive pair instance for x_s .

CIL_{randneg}: Randomly select an instance x_t from another bag B_t as the negative pair instance for x_s .

And we summarize the model performances of our CIL and other three variants in Table 5.

Method	AUC	F1	P@M
CIL	50.8	52.2	86.0
CIL _{randpos}	49.2(-1.6)	50.9(-1.3)	83.8(-2.2)
CIL _{bagpos}	47.8(-3.0)	50.5(-1.7)	79.2(-6.8)
CIL _{randneg}	48.4(-2.4)	50.6(-1.6)	78.2(-7.8)

Table 5: Model performances of our proposed framework CIL and its three variants.

As the previous analysis in section 2.4, the three variants of our CIL framework may suffer from DS noise: (1) Both variants CIL_{randpos} and CIL_{bagpos} may construct noisy positive pairs; therefore, their model performances have a little drop (50.8→49.2, 50.8→47.8). Besides, the variant CIL_{bagpos} also relies on the bag encoder, for which it performs worse than the variant CIL_{randpos} (49.2→47.8). (2) Though the constructed negative pairs need not be as accurate as positive pairs, the variant CIL_{randneg} treats all instances equally, which gives up the advantage of formed accurate representations. Thus, its model performance also declines (50.8→48.4).

3.7 Case Study

We select a typical bag (see Table 6) from the training set to better illustrate the difference between MIL_{sent} , MIL_{bag} and our framework CIL.

Sentence	Predicted Relation
<i>john mcgahern</i> , the eldest of seven children, was born on nov.12, 1934, in <i>dublin</i> .	S: <i>/place_borned</i> ✓ B: <i>/place_borned</i> ✓ C: <i>/place_borned</i> ✓
<i>john mcgahern</i> , whose stark ..., died yesterday in <i>dublin</i> .	S: <i>/place_borned</i> ✗ B: <i>/place_borned</i> ✗ C: <i>/place_deaded</i> ✓

Table 6: A typical bag selected from the training set: The bag is constructed with relational triple (*john mcgahern*, */place_borned*, *dublin*), and the first sentence (S1) is clean to express relation */place_borned* while the second instance (S2) are noisy with true relation */place_deaded*. S: MIL_{sent} , B: MIL_{bag} and C: CIL.

Under MIL_{sent} pattern, both S1, S2 are used for model training, and the noisy sentence S2 may confuse the model. As for MIL_{bag} pattern, S1 is assigned with a high attention score while S2 has a low attention score. However, MIL_{bag} only relies on the bag-level representations, and sentences like S2 can not be used efficiently. Our framework CIL makes full use of all instances (S1, S2) and avoids the negative effect of DS data noise from S2.

4 Related Work

Our work is related to DSRE, pre-trained language models, and recent contrastive learning methods.

DSRE Traditional supervised RE systems heavily rely on the large-scale human-annotated dataset, which is quite expensive and time-consuming. Distant supervision is then introduced to the RE field, and it aligns training corpus with KB facts to generate data automatically. However, such a heuristic process results in data noise and causes classical supervised RE models hard to train. To solve this issue, Lin et al. (2016) applied the multi-instance learning framework with selective attention mechanism over all instances, and it helps RE models learn under DS data noise. Following the MIL framework, recent works improve DSRE models from many different aspects: (1) Yuan et al. (2019) adopted relation-aware attention and constructed super bags to alleviate the problem of bag label error. (2) Ye et al. (2019) analyzed the label distribution of dataset and found the shifted label problem that significantly influences the performance of DSRE models. (3) Vashishth et al. (2018) employed Graph Convolution Networks (Defferrard et al., 2016) to encode syntactic information from the text and improves DSRE models with additional side information from KBs. (4) Alt et al.

(2019) extended the GPT to the DSRE, and fine-tuned it to achieve SOTA model performance.

Pre-trained LM Recently pre-trained language models achieved great success in the NLP field. Vaswani et al. (2017) proposed a self-attention based architecture — Transformer, and it soon becomes the backbone of many following LMs. By pre-training on a large-scale corpus, BERT (Devlin et al., 2019) obtains the ability to capture a notable amount of “common-sense” knowledge and gains significant improvements on many tasks following the fine-tune scheme. At the same time, GPT (Radford et al., 2018), XL-Net (Yang et al., 2019) and GPT2 (Radford et al., 2019) are also well-known pre-trained representatives with excellent transfer learning ability. Moreover, some works (Radford et al., 2019) found that considerably increasing the size of LM results in even better generalization to downstream tasks.

Contrastive Learning As a popular unsupervised method, contrastive learning aims to learn representations by contrasting positive pairs against negative pairs (Hadsell et al., 2006; Oord et al., 2018; Chen et al., 2020; He et al., 2020). Wu et al. (2018) proposed to use the non-parametric instance-level discrimination to leverage more information in the data samples. Our approach, however, achieves the goal of data-efficiency in a more complicated MIL setting: instead of contrasting the instance-level information during training, we find that instance-bag negative pair is the most effective method, which constitutes one of our main contributions. In the NLP field, Dai and Lin (2017) proposed to use contrastive learning for image caption, and Clark et al. (2020) trained a discriminative model for language representation learning. Recent literature (Peng et al., 2020) has also attempted to relate the contrastive pre-training scheme to classical supervised RE task. Different from our work, Peng et al. (2020) aims to utilize abundant DS data and help classical supervised RE models learn a better relation representation, while our CIL focuses on learning an effective and efficient DSRE model under DS data noise.

5 Conclusion

In this work, we discuss the long-standing DSRE framework (*i.e.* MIL) and argue the MIL is not efficient enough, as it aims to form accurate bag-level representations but sacrifices the potential informa-

tion of abundant instances inside MIL bags. Thus, we propose a contrastive instance learning method CIL to boost the MIL model performances. Experiments have shown the effectiveness of our CIL with stable and significant improvements over several baseline models, including current SOTA systems.

Acknowledgments

This work has been supported in part by National Key Research and Development Program of China (2018AAA0101900), Zhejiang NSF (LR21F020004), Key Technologies and Systems of Humanoid Intelligence based on Big Data (Phase ii) (2018YFB1005100), Zhejiang University iFLYTEK Joint Research Center, Funds from City Cloud Technology (China) Co. Ltd., Zhejiang University-Tongdun Technology Joint Laboratory of Artificial Intelligence, Chinese Knowledge Center of Engineering Science and Technology (CKCEST).

References

- Christoph Alt, Marc Hübner, and Leonhard Hennig. 2019. [Fine-tuning pre-trained transformer language models to distantly supervised relation extraction](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1388–1398, Florence, Italy. Association for Computational Linguistics.
- Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. [Matching the blanks: Distributional similarity for relation learning](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. [A simple framework for contrastive learning of visual representations](#). In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR.
- Despina Christou and Grigorios Tsoumakas. 2021. [Improving distantly-supervised relation extraction through bert-based label & instance embeddings](#). *CoRR*, abs/2102.01156.
- Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. [ELECTRA: pre-training text encoders as discriminators rather than generators](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Bo Dai and Dahua Lin. 2017. [Contrastive learning for image captioning](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 898–907.
- Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. 2016. [Convolutional neural networks on graphs with fast localized spectral filtering](#). In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3837–3845.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Joe Ellis, Xuansong Li, Kira Griffitt, Stephanie M Strassel, and Jonathan Wright. 2012. Linguistic resources for 2013 knowledge base population evaluations. In *TAC*.
- Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*, volume 2, pages 1735–1742. IEEE.
- Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. [Momentum contrast for unsupervised visual representation learning](#). In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9726–9735. IEEE.
- Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. [Knowledge-based weak supervision for information extraction of overlapping relations](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 541–550, Portland, Oregon, USA. Association for Computational Linguistics.
- Sharmistha Jat, Siddhesh Khandelwal, and Partha Talukdar. 2018. Improving distantly supervised relation extraction using word and entity based attention. *arXiv preprint arXiv:1804.06987*.
- Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. [Neural relation extraction with selective attention over instances](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2124–2133, Berlin, Germany. Association for Computational Linguistics.

- Xiao Ling and Daniel S. Weld. 2012. [Fine-grained entity recognition](#). In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, July 22–26, 2012, Toronto, Ontario, Canada*. AAAI Press.
- Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.
- Mike Mintz, Steven Bills, Rion Snow, and Daniel 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*, pages 1003–1011, Suntec, Singapore. Association for Computational Linguistics.
- Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.
- Hao Peng, Tianyu Gao, Xu Han, Yankai Lin, Peng Li, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2020. [Learning from Context or Names? An Empirical Study on Neural Relation Extraction](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3661–3672, Online. Association for Computational Linguistics.
- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- 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*, pages 148–163. Springer.
- Shikhar Vashishth, Rishabh Joshi, Sai Suman Prayaga, Chiranjib Bhattacharyya, and Partha Talukdar. 2018. [RESIDE: Improving distantly-supervised neural relation extraction using side information](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1257–1266, Brussels, Belgium. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, Long Beach, CA, USA*, pages 5998–6008.
- Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. 2018. [Unsupervised feature learning via non-parametric instance discrimination](#). In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018*, pages 3733–3742. IEEE Computer Society.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. [Xlnet: Generalized autoregressive pretraining for language understanding](#). In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, Canada*, pages 5754–5764.
- Qinyuan Ye, Liyuan Liu, Maosen Zhang, and Xiang Ren. 2019. [Looking beyond label noise: Shifted label distribution matters in distantly supervised relation extraction](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3841–3850, Hong Kong, China. Association for Computational Linguistics.
- 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*, volume 33, pages 419–426.
- Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. [Distant supervision for relation extraction via piecewise convolutional neural networks](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762, Lisbon, Portugal. Association for Computational Linguistics.