A Distantly-Supervised Relation Extraction Method Based on Selective Gate and Noise Correction

Zhuowei Chen[1], Yujia Tian[1], Lianxi Wang✉[1,2], Shengyi Jiang[1,2]
[1]School of Information Science and Technology,
Guangdong University of Foreign Studies, Guangzhou, China, 510006
[2]Guangzhou Key Laboratory of Multilingual Intelligent Processing, Guangzhou, China, 510006
{20211003051,20211003065,wanglianxi}@gdufs.edu.cn
jiangshengyi@163.com

Abstract

Entity relation extraction, as a core task of information extraction, aims to predict the relation of entity pairs identified by text, and its research results are applied to various fields. To address the problem that current distantly supervised relation extraction (DSRE) methods based on large-scale corpus annotation generate a large amount of noisy data, a DSRE method that incorporates selective gate and noise correction framework is proposed. The selective gate is used to reasonably select the sentence features in the sentence bag, while the noise correction is used to correct the labels of small classes of samples that are misclassified into large classes during the model training process, to reduce the negative impact of noisy data on relation extraction. The results on the English datasets clearly demonstrate that our proposed method outperforms other baseline models. Moreover, the experimental results on the Chinese dataset indicate that our method surpasses other models, providing further evidence that our proposed method is both robust and effective.

1 Introduction

Entity Relation Extraction (RE) is a crucial task in information extraction that aims to identify the relation between entity pairs in text. The findings of RE have practical applications in several fields, such as the construction of knowledge graphs (KG), semantic web annotation, and the development and optimization of question-and-answer systems and search engines, which have a significant impact on daily life. However, the task of RE is challenging due to the limited availability of annotated data. To address this challenge, distant supervision has been proposed, which automatically annotates data, significantly increasing the number of annotated samples.

However, distant supervision suffers from a strong hypothesis, leading to a large number of noisy labels during data annotation. Training on a dataset with noisy labels can result in model overfitting to the noise, which adversely impacts the model's performance (Li et al., 2022b; Christou and Tsoumakas, 2021).

To mitigate these issues, this paper proposes a novel method for RE that incorporates selective gate and the end-to-end noise correction method. In our model, selective gate is utilized to rationally select sentence features in the sentence bag, while noise correction is used to correct the labels of small classes of samples that are misclassified into larger classes during model training. These techniques reduce the negative impact of noisy data on the distant DSRE model. Additionally, since common word embedding models, such as Word2Vec and Glove, produce static vectors that overlook contextual semantics and the flexible use of multiple-meaning words, this paper introduces a pre-trained language model (PLM) to encode and extract features from sentences. This approach provides richer sentence semantic features, effectively improving prediction accuracy and reducing training time. Experiment results demonstrate that this method significantly outperforms the baseline models, improving the DSRE model's performance.

The major contributions of this paper can be summarized as follows:

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 736-747, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

- We propose a DSRE method, named PLMG-Pencil, which combines PLM and selective gate and introduces an end-to-end noise correction training framework called pencil. Selective gate prevents the propagation of noisy representations and pencil corrects noise labels during the training process, reducing the impact of noise on the dataset and improving the performance of the DSRE model.

- We present a novel algorithm for DSRE that combines selective gate mechanism and pencil framework within a three-stage training process. This process involves training the backbone model, gradually correcting noisy labels, and subsequently fine-tuning our model using the corrected data. Empirical experiments demonstrate the robustness and effectiveness of our proposed method.

- Our experiments on three different Chinese and English datasets demonstrate that effective sentence-level feature construction methods and training methods, combined with noise correction, are crucial for improving the performance of models on DSRE tasks.

2 Related Work

2.1 Distantly Supervised Relation Extraction Models

Numerous RE models have been proposed, with deep learning-based models like convolutional neural networks (CNNs) being the current mainstream. CNNs can automatically extract features from sentences, making them a fundamental model for future research(Zhang and Wallace, 2015). However, the maximum pooling operation used in this model ignores important structural and valid information about the sentence.

Socher (2012) was the first to propose a recurrent neural network (RNN) to train relational extractors by encoding sentences. In addition, Zeng (2018) proposed a piecewise convolutional neural network (PCNN) that uses maximum pooling processing based on CNN to effectively preserve the information features of long texts while also reducing the time complexity. Zhou (2016) introduced an attention mechanism based on the long short-term memory network (LSTM) to form the classical BiLSTM-ATT model. The model can reasonably assign weights to features to obtain a better representation of the sentence. Riedel (2010) proposed a multi-instance learning (MIL) framework with a basic annotation unit of a sentence bag containing a common entity pair, rather than a single individual sentence. For sentence bag level labeled data, the model can be made to implicitly focus on correctly labeled sentences through an attention mechanism, thus learning from noisy data to become a stable and robust model. Subsequently, Ye and Ling (2019) proposed a DSRE method based on the intra- and inter-sentence bag, combining sentence-level and bag-level attention for noise correction. Alt(2019) introduced a transformer-based PLM for DSRE. Chen (2021) proposed a new contrastive instance learning method (CIL) to further improve the performance of DSRE models. Further, Li (2022a) introduces a hierarchical contrast framework (HiCLRE) on top of Chen's CIL method to enhance cross-layer semantic interaction and reduce the impact of noisy data. These methods are generally neural network driven and use neural network models that have strong generalization capabilities compared to traditional methods.

2.2 Noise Correction Methods

There are three categories of noise correction methods for DSRE: rule-based statistical methods, multi-instance learning-based methods, and adversarial and reinforcement learning-based methods. Multi-instance learning-based approaches have received the most attention from scholars, due to their effectiveness in correcting noise labels as demonstrated by Yao (2018).

In deep neural networks, designing robust loss functions has also proven effective in coping with noise by making models robust during training. Several studies have examined the robustness of different loss functions such as mean square loss and cross-entropy loss. Zhang (2018) combined the advantages of mean absolute loss and cross-entropy loss to obtain a better function. Li (2020a) proposed DivideMix framework that separates noisy samples using a Gaussian mixture model before training the model. Tanaka (2018) proposed an optimization strategy while Jiang(2018) introduced MentorNet technique for regularizing deep CNNs on test data with weakly supervised labels.

Moreover, Wu(2017) and Shi(2018) have investigated adversarial training based approaches where simulated noise is mixed with real samples during training in order to improve model's robustness against noisy datasets by distinguishing between real versus noisy samples. Although this type of approaches improves corpus quality up to some extent, it requires simultaneous training of two models which can lead to instability and difficulty when applied directly into production systems at scale.

3 Methodology

To mitigate the impact of noise on the DSRE model, this paper proposes a two-pronged approach, PLM-based selective gate pencil (PLMG-Pencil) method. As shown in Figure 1, first, we encode the text using PLM and employ the selective gate mechanism to select sentence-level features that contribute to the bag-level feature. Second, we replace all labels with soft labels and train the model in the pencil framework. This framework uses soft labels that are updated during training and can be corrected for noisy data. This approach reduces the chances of noise being selected in the selective gate, even if it cannot be corrected in the pencil framework. These two methods complement each other, reducing the degree of noise interference and improving the model's RE performance. In this section, we will describe our approach from the backbone model architecture, noise correction framework, and the RE algorithm.

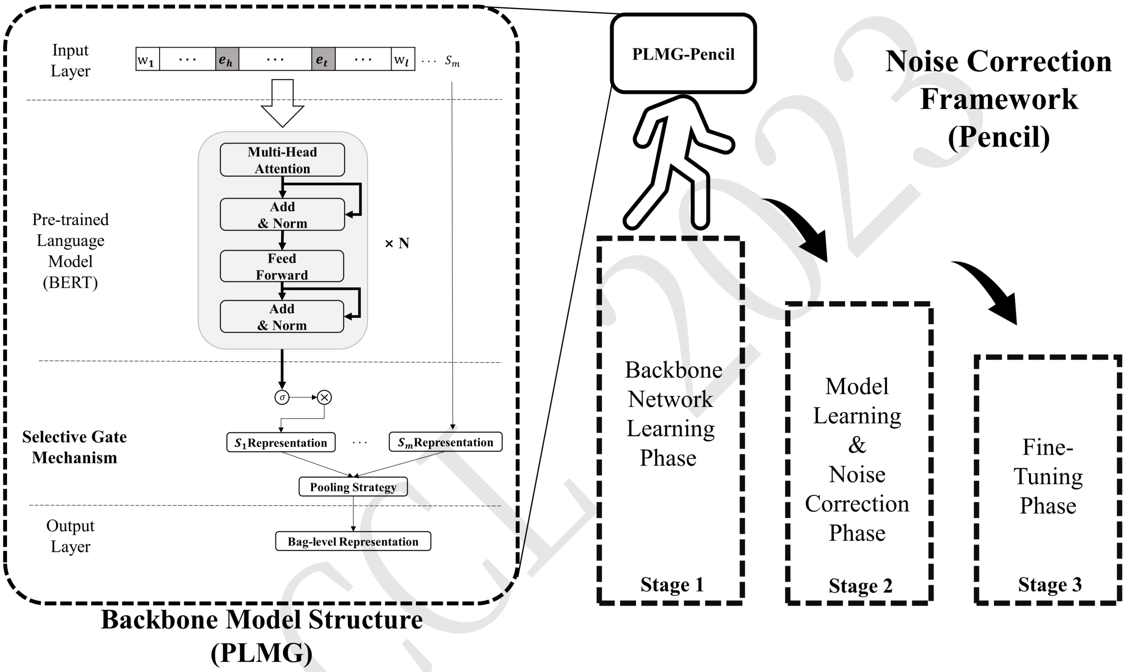

Figure 1: Overview of PLMG-Pencil

3.1 Backbone Model

This paper proposes the PLM-based selective gate as the backbone model, inspired by the Entity-aware Self-attention Enhanced selective gate (SeG) framework proposed by Li (2020b). The primary architecture of our model is presented in the backbone model structure part in Figure 1, and it comprises two main components: (1) **PLM**, structured to encode sentence, entity, and location features for semantic enhancement. (2) **Selective gate**, which enhances the representation of bag-level features by assigning weights to different sentences in the bag. The selective gate mechanism reduces the impact of noise on the model by weighing the contribution of each sentence in the bag.

3.1.1 Input Embeddings

To convert a sentence into a sequence of tokens, we use the BERT tokenizer, which results in a token sequence $S = \{t_1, t_2, \ldots, e_1, \ldots, e_2, \ldots, t_L\}$, where t_n denotes tokens, e_1 and e_2 denote the head and tail entities, respectively, and L represents the maximum length of the input sentence. We add two special tokens, [CLS] and [SEP], to signify the beginning and end of the sentence, respectively.

However, the [CLS] token is not ideal for RE tasks as it only serves as a pooling token to represent the entire sentence. Therefore, to incorporate entity information into the input, we introduce four tokens: [unused1], [unused2], [unused3], and [unused4], which mark the start and end of each entity.

3.1.2 Selective Gate Enhanced Bag Representation

To obtain an effective bag representation, we introduce the selective gate mechanism, which dynamically calculates the weight of each sentence in the bag. We first represent each sentence using a PLM, such as BERT, which accepts structured sequences of tokens S that integrate entity information e_1 and e_2. The PLM's sentence encoder then sums the embeddings, including tokens, entities, and position, to generate context-aware sentence representations $H = \{h_{t_1}, h_{t_2}, \ldots, h_{e_1}, \ldots, h_{e_2}, \ldots, h_{t_L}\}$:

$$H = \mathrm{PLM}(S) \tag{1}$$

where h_{t_n} denotes the hidden features of the token t_n and PLM represents a pre-trained language model, such as BERT, that serves as the sentence encoder. We use special tokens to encode sentences to generate structural representations of sentences for RE task, including [CLS] for sentence-level pooling, its hidden features denoted as $h_{[CLS]}$. [unused1] and [unused2] mark the start and end of the head entity, [unused3] and [unused4] for the tail entity.

$$h_{e_h} = \mathrm{mean}(h_{t_{[unused1]}}, h_{t_{[unused2]}}) \tag{2}$$

$$h_{e_t} = \mathrm{mean}(h_{t_{[unused3]}}, h_{t_{[unused4]}}) \tag{3}$$

Representations of two entities, h_{e_h} and h_{e_t}, are generated by Equation (2) and Equation (3). The hidden features of these special tokens are denoted as $h_{t_{[unused1]}}$, $h_{t_{[unused2]}}$, $h_{t_{[unused3]}}$ and $h_{t_{[unused4]}}$. The sentence representations are generated using the following formulas:

$$h_{S_i} = \sigma([h_{e_h} \mid\mid h_{e_t} \mid\mid h_{[CLS]}] \cdot W_S) + b_S \tag{4}$$

where $||$ represents the concatenation operation, σ is the activation function, W_S is a weight matrix, and b_S is the bias.

Bag Representation The use of PLMs allows us to obtain sentence representations S_n, which can be stacked to form the initial bag representation $B = \{S_1, S_2, ..., S_n\}$. While selective attention modules are commonly used to aggregate sentence-level representations into bag-level representations, our proposed model leverages SeG's novel selective gate mechanism for this purpose. Specifically, when dealing with noisy data, the selective attention mechanism may be inefficient or ineffective if there is only one sentence in the bag, or if that sentence is mislabeled. Given that approximately 80% of the RE benchmark datasets contain single-sentence bags with mislabeled instances, our selective gate mechanism offers a more effective solution by dynamically reducing the alignment of gating values with instances of mislabeling, thereby preventing the propagation of noisy representations.

To generate gate values for each S_j, we employ a two-layer feed-forward network with the following formula:

$$g_j = \sigma(W^{(g_1)}\sigma(W^{(g_2)}S_j + b^{(g_2)}) + b^{(g_1)}), \forall j = 1, ..., m \tag{5}$$

We have $W^{(g2)} \in R^{3d_c \times d_h}$ and $W^{(g1)} \in R^{d_h \times d_h}$, $\sigma(\cdot)$denotes the activation function and $g_i \in (0, 1)$, after that, values of the gates are calculated and the mean pooling aggregation is performed in the bag to generate bag-level representation thus the further relation classification can be performed. The formula of this process is as follows, and m denotes the size of the sentence bag.

$$Q = \frac{1}{m}\sum\nolimits_{j=1}^{m} S_j g_j \tag{6}$$

3.1.3 Classifier

We feed Q into a multi-layer perception (MLP) and apply the $|c|$-way softmax function to determine the relation between the head and tail entities, where $|c|$ represents the number of distinct relation classes. The formula for this process is as follows:

$$p = \text{Softmax}(\text{MLP}(Q)) \in R^{|c|} \tag{7}$$

3.1.4 Model Learning

To train the model, we minimize the negative log-likelihood loss plus an L2 regularization penalty, which is expressed by the following formula:

$$L_{NLL} = -\frac{1}{|D|}\sum_{k=1}^{|D|} \log p^k + \beta||\theta||_2^2 \tag{8}$$

where p^k represents the predicted distribution of the k-th example in the dataset D from Equation (8). The term $\beta||\theta||_2^2$ is the L2 regularization penalty, where θ is the set of model parameters, and β controls the strength of the regularization.

By minimizing this loss function using an optimization algorithm such as stochastic gradient descent, the model learns to predict the correct relation between the head and tail entities.

3.2 Noise Correction Framework

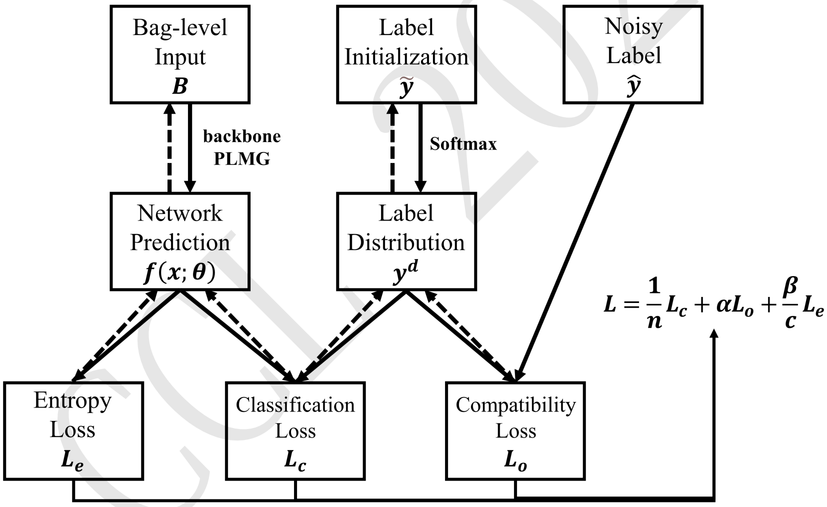
Bag-level Input B
Label Initialization $\tilde{y}$
Noisy Label $\hat{y}$
backbone PLMG
Softmax
Network Prediction $f(x;\theta)$
Label Distribution y^d
$L = \frac{1}{n}L_c + \alpha L_o + \frac{\beta}{c}L_e$
Entropy Loss L_e
Classification Loss L_c
Compatibility Loss L_o

Figure 2: Pencil Framework

In this section, we introduce pencil, a noise correction framework based on the end-to-end noise-labeled learning correction framework proposed by Yi and Wu (2019). The framework is illustrated in Figure 2, with solid arrows representing forward computation and dashed arrows indicating backward propagation.

The pencil framework is designed to update both the network parameters and the data labels simultaneously using gradient descent and backpropagation. To accomplish this, the model generates a vector $\tilde{y}$ to construct soft labels.

$$y^d = \text{Softmax}(\tilde{y}) \tag{9}$$

With Equation (9), $\tilde{y}$ can be updated by gradient descent and backpropagation. The following equation shows the initialized representation of the label with noise in the initial value.

$$\widetilde{y} = K\hat{y} \tag{10}$$

where $\hat{y}$ is the original label with noise, and K is a large constant which ensures y^d and $\hat{y}$ has the most similar distribution in Equation (9), i.e., $y^d \approx \hat{y}$.

An intricately devised loss function is employed to correct the noise labels during the model training procedure, with L_e and L_o as penalty terms and L_c as the classification loss. This loss function incorporates two hyperparameters, denoted as α and β, which can be flexibly adjusted to accommodate diverse datasets with varying proportions of noisy data. Specifically, increasing the value of α and reducing the value of β will yield a diminished degree of label correction. In a c-class classification problem, the loss function is presented as follows.

$$L = \frac{1}{c}L_c + \alpha L_o + \frac{\beta}{c}L_e \tag{11}$$

where c denotes the number of classes.

The classification loss, which works as the main loss of the model guiding the model to learn, is measured using the dual form of the KL divergence between the predicted distribution and the soft labels. The formula for the classification loss L_c is given by:

$$L_c = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c} f_j(x_j;\theta)\log\left(\frac{f_j(x_i;\theta)}{y_{ij}^d}\right) \tag{12}$$

where n denotes the batch size and y_{ij}^d denotes the corresponding soft label. In this equation, KL divergence is used in a symmetric form, which has been shown to perform better than using it directly in this framework in previous studies (Wu et al., 2017). Based on the gradient of the loss function L_c, it can be observed that a larger gap between the predicted value and the true label tends to correspond to a larger gradient. In this framework, the model parameter and noise labels can be updated together, which effectively serves to balance the disparity between the prediction and the true label, facilitating the gradual correction of noisy labels.

To avoid falling into a local optimum, the model sets the entropy loss L_e, using the predicted values of the network and its calculation of the cross-entropy loss. The formula is as follows.

$$L_e = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c} f_j(x_i;\theta)\log f_j(x_i;\theta) \tag{13}$$

The compatibility loss function L_o is formulated as follows, which uses noise labels and soft labels to calculate the cross-entropy loss so as to avoid large deviations between the corrected label and the original noise label.

$$L_o = -\frac{1}{n}\sum\nolimits_{i=1}^{n}\sum\nolimits_{j=1}^{c}\hat{y}_{ij}\log y_{ij}^d \tag{14}$$

3.3 PLMG-Pencil Relation Extraction Method

This paper presents a RE algorithm that utilizes selective gate and noise correction, as shown in Algorithm 1. The complete training process of the algorithm is described below.

- **Stage 1 - Backbone Network Learning Phase**: Initially, the PLMG-Pencil network is trained from scratch with a larger fixed learning rate. The noise in the data is not processed in this stage, and the loss calculation formula only utilizes the classification loss. The network parameters obtained at this stage serve as the initialized network parameters for the next training step.

- **Stage 2 - Model Learning and Noise Correction Phase**: In this stage, the network parameters and label distributions are updated together using the model, thus, noisy labels can be corrected. To avoid overfitting the label noise, the label distribution is corrected for the noise in the original labels. We obtain a vector of label distributions for each sentence bag at the end of this stage. Due to the dissimilarity of the learning rate used for soft labels update and the global model parameters update, a hyperparameter λ is set to adjust it.

- **Stage 3 - Final Fine-Tuning Phase**: The label distribution learned by the model in the previous stages are utilized to fine-tune the network in this stage. Sample labels in the training set are not updated, and the network parameters are updated using the classification loss as the loss function of the model. There are no additional adjustments to the learning rate, and the same decay rules are followed for general neural network training.

Algorithm 1: PLMG-Pencil Distantly Supervised Relation Extraction Algorithm

```
Input: Dataset $D = x_i, \widetilde{y}_i (1 < i < n)$, epoch of stages $T_1$, $T_2$.

Stage 1:
Initialization: $t \leftarrow 1$.
while $t \leq T_1$ do
    Train and update the model parameters $\theta$, while calculating the loss in equation (14) with
      $\alpha = 0$ and $\beta = 0$. Hold off on using $\widetilde{y}_i$;
    $t \leftarrow t + 1$;

Stage 2:
Initialization: $\widetilde{y}_i = K\hat{y}_i$.
while $T_1 \leq t \leq T_2$ do
    Train and update the model parameters $\theta$ and $y_i^d$;
    $t \leftarrow t + 1$;

Stage 3:
while $T_2 \leq t$ do
    Train and update the model parameters $\theta$ and $y_i^d$;
    Train and update the model parameters $\theta$, while calculating the loss in equation (14) with
      $\alpha = 0$ and $\beta = 0$. Do not update sample labels.
    $t \leftarrow t + 1$;
Output: $\theta$, noise-corrected labels.
```

4 Experiments

4.1 Datasets

We evaluate our proposed model on three different datasets: the New York Times (NYT10) dataset and the GDS dataset in English, the SanWen dataset in Chinese. Datasets statistics are shown in Appendix A.

NYT10 (Riedel et al., 2010): This dataset is widely used in models based on DSRE, which is annotated with 58 different relations and the NA relation account for over 80% of the total. It has 522K and 172K sentence sets in the training and test sets respectively.

GDS (Jat et al., 2018): This dataset is created from the Google RE corpus, which contains 5 relations. It has 18K and 5K instances in the training and test sets, respectively.

SanWen (Xu et al., 2017): This dataset contains 9 relations from 837 Chinese documents. It has 10K, 1.1K, and 1.3K sentences in the training set, test set, and validation set respectively.

4.2 Baselines

To validate the effectiveness of the RE model proposed in this paper, we compare it with mainstream RE methods on the three datasets mentioned above. The following baseline methods are used.

Mintz (Mintz et al., 2009): It concatenates various features of sentences to train a multi-class logistic regression classifier.

PCNN+ATT (Lin et al., 2016): It uses selective attention to multiple instances to alleviate the problem of mislabelling.
RESIDE (Vashishth et al., 2018): It exploits the information of entity type and relation alias to add a soft limitation for relation classification.
MTB-MIL (Baldini Soares et al., 2019): It proposes a method for matching gaps and learning sentence representations through entity-linked text.
DISTRE (Alt et al., 2019): It combines the selective attention with its PLM.
SeG (Li et al., 2020b): It uses an entity-aware embedding-based self-attentive enhancement selective gate based on PCNN+ATT to rationally select sentence features within sentence bags to reduce the interference of noise.
CIL (Chen et al., 2021): It proposes a comparative instance learning method in the MIL framework.
HiCLRE (Li et al., 2022a): It incorporates global structural information and local fine-grained interactions to reduce sentence noise.

4.3 Parameter Settings

Table 1 presents the hyperparameter settings used in our experiments. The English datasets are trained on the bert-base-uncased model from the Huggingface platform, while the Chinese dataset uses the bert-base-chinese model. To effectively train our model, we use the parameter settings from Yi and Wu (2019) as initialization settings for our experiments. The model's dropout rate, learning rate, α, β, batch size, and epoch settings are shown in the table.

Params	Dropout	LR	α	β	BatchSize	Epoch 1	Epoch 2
Value	0.5	0.035	0.1	0.4	64	15	20

Table 1: Parameter Settings. Epoch 1 and Epoch 2 mark the end of Stage 1 and Stage 2, respectively, and LR stands for the learning rate.

It is important to note that the optimal values for α and β may vary based on the level of noise in different datasets. Therefore, these values should be adjusted accordingly to improve the loss calculation and enhance the overall performance of the model.

4.4 Results

To evaluate the performance of our model in DSRE tasks, we use AUC and P@N values as evaluation metrics. AUC measures the area under the ROC curve, while P@N indicates the average accuracy of top N instances. Finally, P@M represents the average of these three P@N results.

4.4.1 Evaluation on English Dataset

Table 2 and Table 3 present a comparison of our proposed model with baseline models on dataset GDS and NYT10, respectively. Our model achieves promising results, as shown by the following observations: (1) Our proposed model shows competitive performance in terms of AUC values on both datasets. As shown in Table 2, on the GDS dataset, the AUC values of our model reach comparable levels with CIL and HiCLRE. Furthermore, as shown in Table 3, on the NYT10 dataset, our model outperforms CIL and DISTRE by 4.1% and 5.2% in AUC values respectively. (2) Our model demonstrates a clear advantage in terms of P@N values. On the NYT10 dataset, the P@100 value is 2.5% higher than CIL, which uses a contrast learning framework. The maximum difference in P@N values appears on the P@300 value, of which our method is 5.9% higher. In comparison to the DISTRE model, which also uses the PLM and MIL framework, our model outperforms it by 16%, 13.5%, and 12.7% on P@100, P@200, and P@300 values respectively.

We further conduct ablation experiments to highlight the benefits of the pencil framework. Specifically, we train our model using a conventional MIL training framework. When comparing the results of the PLMG model with the PLMG-Pencil model on the GDS dataset, we observe a 0.2% decrease in the AUC value and a 0.1% decrease in the P@1K value for the PLMG model. These findings provide compelling evidence for the effectiveness of the pencil framework and our proposed algorithm. On the dataset NYT10, the proposed model shows a significant improvement compared to the model without pencil framework. Precisely, we observe a 6%, 2.5% and 2% improvement in P@100, P@200 and P@300 values respectively.

Dataset	Models	AUC	P@500	P@1K	P@2K	P@M
GDS	Mintz† (Mintz et al., 2009)	-	-	-	-	-
	PCNN-ATT† (Lin et al., 2016)	79.9	90.6	87.6	75.2	84.5
	MTB-MIL† (Baldini Soares et al., 2019)	88.5	94.8	92.2	87.0	91.3
	RESIDE† (Alt et al., 2019)	89.1	94.8	91.1	82.7	89.5
	REDSandT† (Christou and Tsoumakas, 2021)	86.1	95.6	92.6	84.6	91.0
	DISTRE† (Alt et al., 2019)	89.9	97.0	93.8	87.6	92.8
	CIL† (Chen et al., 2021)	90.8	**97.1**	94.0	87.8	93.0
	HiCLRE(Li et al., 2022a)	90.8	96.6	93.8	88.8	**93.1**
	PLMG-Pencil	**91.0**	95.4	**94.1**	88.8	92.8
	-without pencil (PLMG)	90.8	95.4	94.0	**89.0**	92.8

Table 2: Model Performances on GDS. (†) marks the results are reported in the previous research.

Dataset	Models	AUC	P@100	P@200	P@300	P@M
NYT10	Mintz† (Mintz et al., 2009)	10.7	52.3	50.2	45.0	49.2
	PCNN-ATT† (Lin et al., 2016)	34.1	73.0	68.0	67.3	69.4
	MTB-MIL† (Baldini Soares et al., 2019)	40.8	76.2	71.1	69.4	72.2
	RESIDE† (Alt et al., 2019)	41.5	81.8	75.4	74.3	77.2
	REDSandT† (Christou and Tsoumakas, 2021)	42.4	78.8	75.0	73.0	75.3
	DISTRE† (Alt et al., 2019)	42.2	68.0	67.0	65.3	66.8
	CIL† (Chen et al., 2021)	43.1	81.5	75.5	72.1	76.9
	HiCLRE(Li et al., 2022a)	45.3	82.0	78.5	74.0	78.2
	PLMG-Pencil	**47.0**	**84.0**	**80.5**	**78.0**	**80.8**
	-without pencil (PLMG)	47.0	78.0	78.0	76.0	77.3

Table 3: Model Performances on NYT10. (†) marks the results are reported in the previous research.

Figure 3 shows the PR curves for our proposed model and the baseline model. Our model clearly outperforms the baselines, particularly compared to the DISTRE model, which also uses PLM and MIL. Based on the ablation experiments conducted on the NYT10 dataset, it can be observed that the PLMG-Pencil method demonstrates a notable superiority in terms of precision at N (P@N) values. These results suggest that the selective gate has a positive impact on constructing sentence bag features and improving model performance. Furthermore, the pencil framework effectively corrects for noisy samples during training, leading to improved performance.

4.4.2 Evaluation on Chinese Dataset

We conduct additional experiments on the SanWen dataset to further validate the effectiveness of the pencil framework and selective gate mechanism. Figure 4 presents the model performances on this dataset.

Our model exhibits superior performance compared to HiCLRE, which utilizes the contrast learning framework, with a notable increase of 4.4% in AUC values. Furthermore, when compared to the SeG model that employs the selective gate mechanism, our PLMG-Pencil model, which incorporates the pencil approach, demonstrates a significant enhancement in AUC values. The ablation experiment further

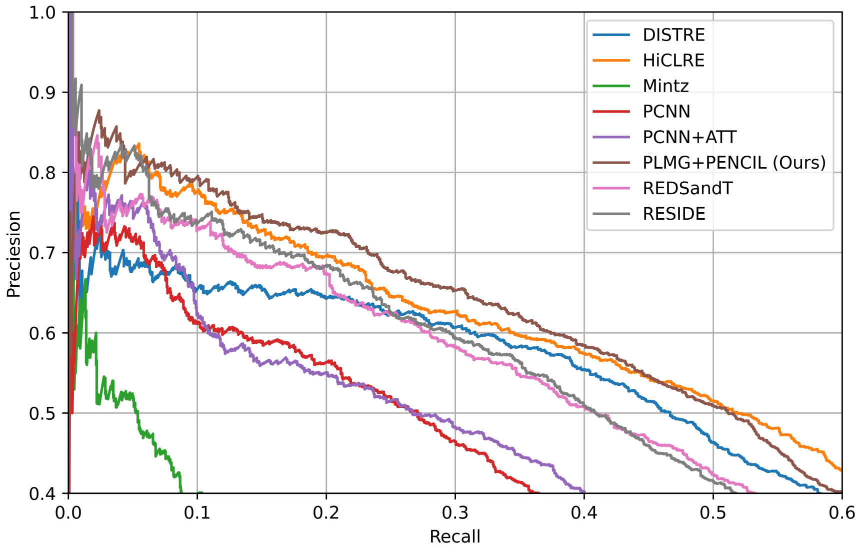

Figure 3: PR-Curve on NYT10

validates the effectiveness and robustness of our method. These results highlights the positive influence of the PLM and noise correction framework on the RE task.

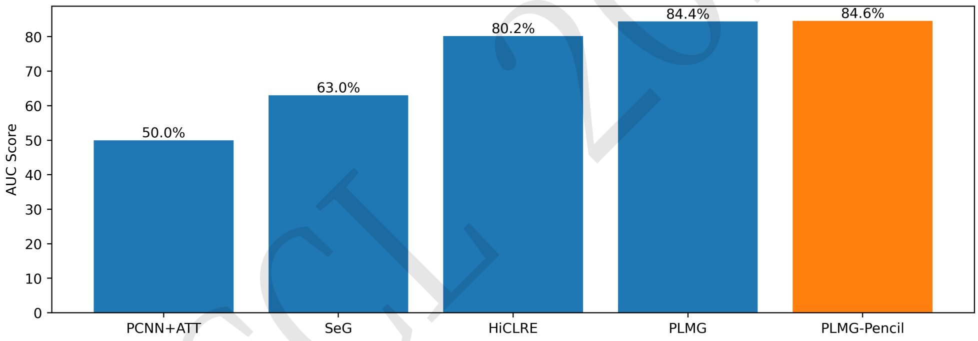

Figure 4: AUC Values of Models on SanWen

Based on the experimental results and the analysis of the dataset features described in Section 4.1, our model tends to perform better on datasets with more relations, such as NYT and SanWen. Compared with baselines, our model can achieve greater advantages on such datasets. In addition, the experimental results on the NYT10 dataset reveal that the pencil framework generates more significant performance enhancements compared to those obtained through experiments performed on the GDS dataset. The GDS dataset employs various methods to mitigate noise interferences and thus has higher quality annotations (Jat et al., 2018). Moreover, the pencil framework is designed to conduct a noise correction process for optimizing model performance, thus, it tends to bring larger improvements on datasets with greater amounts of noisy data.

5 Conclusion

In this paper, we propose the PLMG-Pencil method for DSRE. Our approach automatically learns the weights of different sentences in a sentence bag and selects the features that best represent the sentence bag through a gate mechanism. Additionally, we introduce a noise correction framework based on end-to-end probability with noise label learning for improved performance in RE. The experimental results clearly demonstrate that our proposed model outperforms baselines and achieves significant improvement in the RE task. Our approach shows great potential for practical application in the field of information extraction.

Acknowledgements

This work was supported by the National Social Science Fund of China (No. 22BTQ045).

A Datasets statistics

Dataset	#Relation	#Train	#Dev	#Test	Language
NYT	58	520K	-	172K	English
GDS	5	18K	-	5K	English
SanWen	9	10K	1.1K	1.3K	Chinese

Table 4: Datasets statistic.

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, July. 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, July. Association for Computational Linguistics.

Tao Chen, Haizhou Shi, Siliang Tang, Zhigang Chen, Fei Wu, and Yueting Zhuang. 2021. CIL: Contrastive instance learning framework for distantly supervised relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6191–6200, Online, August. Association for Computational Linguistics.

Despina Christou and Grigorios Tsoumakas. 2021. Improving distantly-supervised relation extraction through bert-based label and instance embeddings. *IEEE Access*, 9:62574–62582.

Sharmistha Jat, Siddhesh Khandelwal, and Partha Talukdar. 2018. Improving distantly supervised relation extraction using word and entity based attention. *arXiv preprint arXiv:1804.06987*.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. 2018. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International conference on machine learning*, pages 2304–2313. PMLR.

Junnan Li, Richard Socher, and Steven C. H. Hoi. 2020a. Dividemix: Learning with noisy labels as semi-supervised learning. *ArXiv*, abs/2002.07394.

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

Dongyang Li, Taolin Zhang, Nan Hu, Chengyu Wang, and Xiaofeng He. 2022a. HiCLRE: A hierarchical contrastive learning framework for distantly supervised relation extraction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2567–2578, Dublin, Ireland, May. Association for Computational Linguistics.

Rui Li, Cheng Yang, Tingwei Li, and Sen Su. 2022b. Midtd: A simple and effective distillation framework for distantly supervised relation extraction. *ACM Transactions on Information Systems (TOIS)*, 40(4):1–32.

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.

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

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2010, Barcelona, Spain, September 20-24, 2010, Proceedings, Part III 21*, pages 148–163. Springer.

Ge Shi, Chong Feng, Lifu Huang, Boliang Zhang, Heng Ji, Lejian Liao, and He-Yan Huang. 2018. Genre separation network with adversarial training for cross-genre relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1018–1023.

Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 1201–1211.

Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2018. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5552–5560.

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, October-November. Association for Computational Linguistics.

Yi Wu, David Bamman, and Stuart Russell. 2017. Adversarial training for relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1778–1783.

Jingjing Xu, Ji Wen, Xu Sun, and Qi Su. 2017. A discourse-level named entity recognition and relation extraction dataset for chinese literature text. *arXiv preprint arXiv:1711.07010*.

Jiangchao Yao, Jiajie Wang, Ivor W Tsang, Ya Zhang, Jun Sun, Chengqi Zhang, and Rui Zhang. 2018. Deep learning from noisy image labels with quality embedding. *IEEE Transactions on Image Processing*, 28(4):1909–1922.

Zhi-Xiu Ye and Zhen-Hua Ling. 2019. Distant supervision relation extraction with intra-bag and inter-bag attentions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2810–2819, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Kun Yi and Jianxin Wu. 2019. Probabilistic end-to-end noise correction for learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7017–7025.

Daojian Zeng, Yuan Dai, Feng Li, R Simon Sherratt, and Jin Wang. 2018. Adversarial learning for distant supervised relation extraction. *Computers, Materials & Continua*, 55(1):121–136.

Zhilu Zhang and Mert Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31.

Ye Zhang and Byron Wallace. 2015. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 207–212.