Positive-Guided Knowledge Distillation for Document-Level Relation Extraction with Noisy Labeled Data

Daojian Zeng[1,2], Jianling Zhu[1], Lincheng Jiang[3], and Jianhua Dai[1(✉)]

[1] Hunan Provincial Key Laboratory of Intelligent Computing and Language Information Processing, Hunan Normal University, Changsha, China
{zengdj,zhujl,jhdai}@hunnu.edu.cn
[2] Institute of AI and Targeted International Communication, Hunan Normal University, Changsha, China
[3] College of Advanced Interdisciplinary Studies, National University of Defense Technology, Changsha, China
linchengjiang@nudt.edu.cn

Abstract. Since one entity may have multiple mentions and relations between entities may stretch across multiple sentences in a document, the annotation of document-level relation extraction datasets becomes a challenging task. Many studies have identified that these datasets contain a large number of noisy labels, hindering performance improvement for the document-level relation extraction task. The previous and most straightforward method is denoising noisy labeled data from a data annotation perspective. However, this time-consuming approach is not suitable for large-scale datasets. In this paper, we propose a novel **P**ositive-**G**uided **K**nowledge **D**istillation (PGKD) model to address the noisy labeled data problem for document-level relation extraction. We design a new teacher-student architecture. The teacher model trained with only positive samples can partially supervise the student model. The positive-guided knowledge distillation algorithm transfers the clean positive-class patterns from the teacher model to the student model. In this way, the student model trained with all samples can efficiently prevent the interference of false negative samples. Extensive experiments on MixDocRED demonstrate that PGKD achieves state-of-the-art effectiveness for document-level relation extraction with noisy labeled data. Moreover, PGKD also surpasses other baselines even on the well-annotated Re-DocRED.

Keywords: Positive-guided knowledge distillation · Noisy labeled data · False negative samples · Document-level relation extraction

1 Introduction

Relation extraction (RE) comprises a primary branch of information extraction. It plays a crucial role in extracting structured information from unstructured

F. Liu et al. (Eds.): NLPCC 2023, LNAI 14302, pp. 249–260, 2023.
https://doi.org/10.1007/978-3-031-44693-1_20

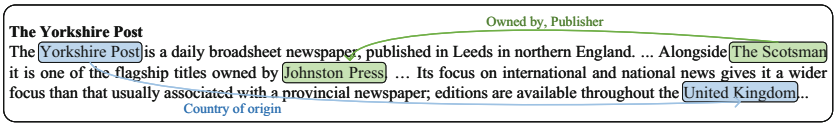

Fig. 1. A document in the DocRED dataset. New relations between entities are added by our revising. However, in the previous incomplete DocRED dataset, these triples are not in the ground truths.

textual data. Early studies focused primarily on the intra-sentence relation setting. Recently, there have been significant efforts on document-level RE, which deals with multiple mentions for each entity and complex inter-sentence relations. Consequently, the complexity of data annotation for document-level RE has increased dramatically. It is nearly impossible to rely completely on manual annotation to obtain large-scale, well-annotated datasets. To overcome this challenge, existing datasets (e.g., DocRED) employ a recommend-revise annotation scheme. This involves recommending relation triple candidates via distant supervision and then confirming their correctness through annotators. However, distant supervision relies on a sparse and incomplete knowledge base, which causes that human annotations fail to cover all ground-truth relation triples and suffer a major drawback - noisy labels[1]. For example, the relation between "Yorkshire Post" and "United Kingdom", i.e., "Country of origin", can be easily retrieved from the document in Fig. 1, whereas this triple is not included in DocRED. Current studies [5,8,13] have identified that the noisy labeled data is a key factor hindering the performance improvement of document-level RE. Nevertheless, how to resolve this issue has received limited attention.

Recent efforts in addressing noisy labeled data can be divided into two genres. The first and most direct solution is from the standpoint of data annotation. Huang et al. [8] pointed out the false negative problem[2] in DocRED, and assigned two expert annotators to relabel 96 documents from scratch. Subsequently, Tan et al. [13] adopted a human-in-the-loop approach to iteratively re-annotate 4,053 documents in DocRED. This approach involved reintroducing relation triples that were initially missed back into the original dataset. However, the complexity of the document-level RE task inevitably increases the difficulty and cost of producing high-quality benchmark datasets. Another economical solution is denoising from the model-based perspective. Reinforcement learning [1] and generative adversarial learning [5] successively solved the noise in the data. Compared with the data annotation solution, this model-based denoising method produces a smaller workload and higher reusability. Therefore, it is essential to study the problem of incompletely annotated data from a model perspective.

In this paper, we propose a novel **P**ositive-**G**uided **K**nowledge **D**istillation (**PGKD**) model for document-level RE. PGKD is based on a teacher-student architecture. The student model is partially supervised by the teacher model

[1] Noisy labels refer to incorrect or inaccurate annotations assigned to the samples.
[2] The relation triples are not in the ground truths of the dataset.

overlooking the NA (*no_relation*) instances. The proposed model works in a positive-guided manner. The distillation algorithm transfers the positive-class patterns to the student model. Although the student model is trained with all samples, it can avoid the pattern collapse [9] due to the supervision of the teacher model. Specifically, based on the soft label and prediction of the student model, we calculate a decoupled knowledge distillation loss. Moreover, the student model can benefit from a hard loss that enables it to learn the ground truths. Lastly, by incorporating both the hard loss and the knowledge distillation loss, the student model can learn from the teacher model's expertise while refining its own classification capabilities. In addition, we construct a new dataset called Mix-DocRED, consisting of both noisy labeled training data and well-labeled validation data. The evaluation of PGKD is performed on Mix-DocRED to validate our motivation. We can summarize our contributions as follows:

- We address the problem of the noisy labeled data in document-level RE from a model perspective, and design a new teacher-student architecture.
- We innovatively utilize positive samples to train the teacher model to ensure that the student model can avoid the pattern collapse and mimic the outputs of the teacher model on the positive classes.
- Experimental results on Mix-DocRED demostrate that PGKD achieves state-of-the-art performance for document-level relation extraction with noisy labeled data. Furthermore, PGKD outperforms existing competitive baselines even on the well-annotated Re-DocRED dataset.

2 Related Work

Document-Level Relation Extraction. There are two categories of approaches for document-level RE. On the one hand, researchers construct a delicately designed document graph [2]. Following this, many studies integrated similar structural dependencies to model documents. Otherwise, a special reasoning network was designed for relation inference [17]. On the other hand, there are some works [15,20] that attempt to use pre-trained language models directly for document-level RE without involving graph structure. Xu et al. [15] incorporated entity structure dependencies within the Transformer encoding part and throughout the overall system. Zhou et al. [20] introduced an adaptive-thresholding loss and a localized context pooling technique. These Transformer-based approaches are simple but very effective. However, most works were proposed under the assumption that datasets are completely annotated. Recently, Huang et al. [8] identified the false negative issue in DocRED and re-annotated 96 documents. Moreover, Tan et al. [13] adopted a more efficient semi-manual approach to re-annotate 4,053 documents. Despite their effectiveness, these annotation methods [8,13] are time-consuming and impractical for large-scale datasets. Therefore, we introduce the positive-guided knowledge distillation approach to address the problem of noisy labeled data in document-level RE.

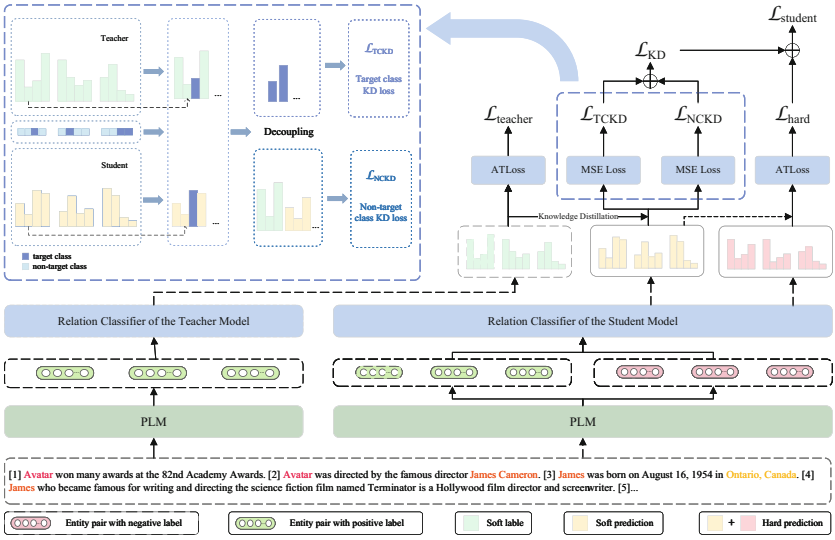

Fig. 2. The overall architecture of PGKD. The left part and the right part represent the teacher model and the student model, respectively. The part enclosed by the purple dotted box represents the process of obtaining the target and non-target knowledge distillation losses.

Knowledge Distillation. Hinton et al. [7] first introduced the concept of knowledge distillation. Its core idea is to transfer "dark knowledge" of the teacher model to the student model. Afterwards, Heo et al. [6] skipped redundant information that adversely affects the compression of the student model. Mirzadeh et al. [11] introduced a teacher assistant as an intermediary between teachers and students. Differing from the above methods, Zhao et al. [19] proposed the new concept of decoupled knowledge distillation (DKD) which comprises two components: target class knowledge distillation (TCKD) and non-target class knowledge distillation (NCKD). TCKD focuses on transferring knowledge related to the "difficulty" of training samples, whereas NCKD plays a crucial role in the effectiveness of logit distillation. Recently, Tan et al. [12] attempted to address the disparity between human-annotated data and distantly supervised data by knowledge distillation.

3 Problem Formulation

Given a set of n entities $\{e_1, \ldots, e_n\}$ in a document D, the goal of the document-level RE task is to predict all relation types $r \in R \cup \{no\_relation\}$ for each entity pair (e_h, e_t). e_h and e_t is the head and the tail entities, respectively. R stands for a collection of predefined relation classes. The setting of this work is to employ the incompletely labeled training set to train a document-level RE model and then evalutae this model with a well-annotated test set.

4 Model Architecture

As shown in Fig. 2, the proposed PGKD model is based on a teacher-student architecture. Specifically, we first follow Zhou et al. [20] to get the entity pair embedding with the local contextual representations. Next, we only apply positive samples to train the teacher model, which allows it to learn complete and clean positive-class patterns without interference from false negative samples. For the student model, we utilize all samples as the training data to generate the predicted results under the assistance of the teacher model and the guidance of the ground truths.

4.1 Teacher Model

Specifically, we first follow Zhou et al. [20] to obtain an entity pair representation that incorporates the localized context. The entity pair encoding method is equally applicable to the teacher model and the student model. Given a document $D = [h_l]_{l=1}^{L}$ containing L words, we utilize a special token "*" which is inserted before and after each mention. This approach allows us to easily identify the specific locations of mentions within the document. The document D is subsequently encoded using a pre-trained language model (PLM) to obtain its contextual embedding $H = [h_1, \ldots, h_L]$, $h_l \in \mathbb{R}^d$ of each token and cross token attention A. We adopts the vector representation of the marker "*" before the mention as the embedding of the mention m_j^i, where m_j^i represents the j^{th} mention of the i^{th} entity. All mentions to the same entity are adopted the logsumexp pooling to get the entity embedding e_i. The local contextual embedding $c_{h,t}$ of the entity pair is computed as follows:

$$c_{h,t} = H^T \frac{A_h \circ A_t}{A_h^T A_t}, \tag{1}$$

where A_h, A_t denote the attentions of the head and the tail entities, A_h, $A_t \in \mathbb{R}^L$, respectively. Then we compute the entity pair embedding $g^{(h,t)} \in \mathbb{R}^d$ as follows:

$$\begin{aligned} z_h &= \tanh\left(W_h e_h + W_{c_h} c_{h,t}\right), \\ z_t &= \tanh\left(W_t e_t + W_{c_t} c_{h,t}\right), \\ g_i^{(h,t)} &= \sum_{j=1}^{k} \left(z_h^j W_{g_i}^j z_t^j\right) + b_i, \\ g^{(h,t)} &= \left[g_1^{(h,t)}, g_2^{(h,t)}, \ldots, g_d^{(h,t)}\right] \end{aligned} \tag{2}$$

where $W_h, W_{c_h}, W_t, W_{c_t}, W_{g_i}^j$ and b_i are model parameters, $W_h, W_{c_h}, W_t, W_{c_t} \in \mathbb{R}^d, W_{g_i}^j \in \mathbb{R}^{d/k \times d/k}$. z_h and z_t indicate the embeddings of the head entity and the tail entity, $z_h, z_t \in \mathbb{R}^d$, respectively.

For the teacher model, the entity pair embedding $g_t^{(h,t)}$ of a positive sample is fed into a feed-forward linear layer to get the predicted result $P_t\left(r|e_h, e_t\right)$:

$$P_t\left(r|e_h, e_t\right) = \sigma\left(W_t g_t^{(h,t)} + b_t\right), \tag{3}$$

where W_t and b_t are the learnable parameters, $W_t, b_t \in \mathbb{R}^d$. σ is an activation function (e.g., *Sigmoid*). Because different entity pairs or classes have distinct interpretations for the same predicted score, a global threshold is inadequate. Therefore, we employ the adaptive-thresholding loss function [20] for the teacher model. A TH class is introduced to distinguish between positive and negative classes. Positive classes P_T are expected to have probabilities higher than TH, whereas negative classes N_T should have probabilities lower than TH. The loss function of the teacher model is as shown below:

$$\begin{aligned}\mathcal{L}_{teacher} = &- \sum_{r \in P_T} \log \left(\frac{\exp\left(logit_r\right)}{\sum_{r' \in P_T \bigcup \{TH\}} \exp\left(logit_{r'}\right)} \right) \\ &- \log \left(\frac{\exp\left(logit_{TH}\right)}{\sum_{r' \in N_T \bigcup \{TH\}} \exp\left(logit_{r'}\right)} \right),\end{aligned} \tag{4}$$

where *logit* is the hidden representation in the last layer before *Sigmoid*.

4.2 Student Model

The student model utilizes both positive and negative samples as training data. It generates a hard loss and a knowledge distillation loss, supervised by the true labels and the teacher model, respectively. Similarly to the teacher model, the student model first obtains the entity pair embedding $g_s^{(h,t)}$. Then the probability score $P_s\left(r|e_h, e_t\right)$ is obtained by inputting $g_s^{(h,t)}$ into a linear layer:

$$P_s\left(r|e_h, e_t\right) = \sigma\left(W_s g_s^{(h,t)} + b_s\right), \tag{5}$$

where W_s and b_s are the learnable parameters, $W_s, b_s \in \mathbb{R}^d$. Subsequently, we regard the ground-truth label of this input sample as the hard label and adopt the adaptive-thresholding loss function to compute the hard loss $\mathcal{L}_{hard}$, which can optimize its performance on individual sample classification.

Additionally, a knowledge distillation (KD) loss is introduced to obtain the supervised knowledge of the teacher on the student. We assign separate weights to the TCKD and NCKD losses. The true labels of the entity pair and the rest of the 97 relation types are considered as the target and the non-target classes, respectively. Specifically, we feed a positive sample to the trained teacher model that generates a soft label $P_t\left(r|e_h, e_t\right)$, while the student model produces a corresponding soft prediction $P_s\left(r|e_h, e_t\right)$. We then calculate the target class mask M_{TC} and non-target class mask M_{NC} based on the ground truth L_{PT}. The target class soft label L_{TC}, non-target class soft label L_{NC}, target class soft prediction P_{TC} and non-target class soft prediction P_{NC} can be available as follows:

$$\begin{aligned}L_{TC}\left(r|e_h, e_t\right) &= M_{TC} \cdot P_t\left(r|e_h, e_t\right), \\ L_{NC}\left(r|e_h, e_t\right) &= M_{NC} \cdot P_t\left(r|e_h, e_t\right), \\ P_{TC}\left(r|e_h, e_t\right) &= M_{TC} \cdot P_s\left(r|e_h, e_t\right), \\ P_{NC}\left(r|e_h, e_t\right) &= M_{NC} \cdot P_s\left(r|e_h, e_t\right)\end{aligned} \tag{6}$$

Table 1. Statistics of Re-DocRED and Mix-DocRED.

	Re-DocRED	Mix-DocRED		
data sets	Train	Train	Dev	Test
#Documents	3053	3053	500	500
Avg. #Entities	19.4	19.5	19.4	19.6
Avg. #Triples	28.1	12.5	34.6	34.9
Avg. #Sentences	7.99	7.9	8.2	7.9
#Positive samples	67,808	35,615	13,362	13,672
NA rate	93.5%		96.0%	

Ultimately, we employ the mean squared error loss function to calculate the target class KD loss $\mathcal{L}_{TCKD}$ and non-target class KD loss $\mathcal{L}_{NCKD}$. The student model can adequately learn the logit distribution of the teacher model on the positive classes. The positive-class pattern collapse of the student model can be avoided in this positive-guided manner.

$$\begin{aligned}\mathcal{L}_{TCKD} &= \frac{1}{|\mathcal{E}^t|}\sum_{(e_h,e_t)\in\mathcal{E}^t}\sum_{r\in\mathcal{R}}(L_{TC}\left(r|e_h,e_t\right)-P_{TC}\left(r|e_h,e_t\right))^2,\\ \mathcal{L}_{NCKD} &= \frac{1}{|\mathcal{E}^t|}\sum_{(e_h,e_t)\in\mathcal{E}^t}\sum_{r\in\mathcal{R}}(L_{NC}\left(r|e_h,e_t\right)-P_{NC}\left(r|e_h,e_t\right))^2,\end{aligned} \tag{7}$$

where $\mathcal{E}^t$ means the number of positive samples. $\mathcal{R}$ represents all predicted relation types. The KD loss is reformulated into a weighted sum of $\mathcal{L}_{TCKD}$ and $\mathcal{L}_{NCKD}$ as follows:

$$\mathcal{L}_{KD} = \alpha * \mathcal{L}_{TCKD} + \beta * \mathcal{L}_{NCKD}, \tag{8}$$

where the hyper-parameters α and β are utilized to disentangle the classical knowledge distillation process. By incorporating both the hard loss $\mathcal{L}_{hard}$ and the knowledge distillation loss $\mathcal{L}_{KD}$, the student model aims to learn from the teacher model's expertise while refining its own classification capabilities. Its final loss function is shown as follows:

$$\mathcal{L}_{student} = \gamma * \mathcal{L}_{hard} + \delta * \mathcal{L}_{KD}, \tag{9}$$

where γ and δ are the hyper-parameters to make trade-offs.

5 Experiments

5.1 Datasets

DocRED[3] [16] is a well-known benchmark dataset for document-level RE, but it is plagued by a high rate of false negative samples. To overcome this limitation,

[3] https://github.com/thunlp/DocRED.

Tan et al. [13] performed a re-annotation of the 4,053 documents in DocRED, creating a new dataset called Re-DocRED[4]. However, there is currently no reliable benchmark dataset for document-level denoising RE. So we fuse DocRED and Re-DocRED to construct a new dataset named Mix-DocRED, which comprises the training set of DocRED, as well as the dev and testing sets of Re-DocRED. Table 1 provides the statistics of Re-DocRED and Mix-DocRED.

5.2 Implementation Details

In this work, we employed BERT_{Base} [3] and RoBERTa_{Large} [10] as document encoders. AdamW was used as the optimizer of our model. We performed warm-up [4] on the initial 6% steps during training and set the learning rates to 5e-5 and 3e-5 for BERT_{Base} and RoBERTa_{Large}, respectively. We performed the grid search on the development set to optimize the hyper-parameters, which include α, β, γ, and δ. The values for these hyper-parameters were 2, 1, 0.7, and 1, respectively. We reported the mean results with three different seeds. A single NVIDIA RTX 3090 GPU was used for all experiments. Precision, recall, Ign F1 and F1 scores served as the primary evaluation metrics.

5.3 Baselines

We conducted two sets of comparisons to evaluate the proposed PGKD model. Firstly, we compared PGKD against some existing competitive baselines on the DocRED leaderboard[5]. These baselines were developed under the assumption that the dataset is well-annotated, including BiLSTM [16], GAIN [17], ATLOP [20], and DocuNet [18]. Secondly, we compared PGKD with SSR-PU [14], the current state-of-the-art framework for document-level denoising RE. SSR-PU was a unified positive-unlabeled learning framework that effectively solved the incomplete labeling problem.

5.4 Main Results

We report the mean and standard deviation of PGKD on the Mix-DocRED test set compared to other strong baselines. As seen in Table 2, PGKD outperforms the competitive models, achieving the highest F1 score of 56.50 and 59.92, respectively. To facilitate a more direct comparison with the latest denoising framework SSR-PU, we adopt ATLOP as the backbone. Our PGKD outperforms SSR-PU by 0.36 and 0.42 F1 points, respectively. This implies that PGKD is superior to SSR-PU in the RE ability from incomplete annotated data. Furthermore, It is worth noting that BiLSTM, GAIN, ATLOP, and DocuNet experience significant drops in F1 score when facing with incomplete labeling scenarios. For instance, DocuNet shows a decrease of 10.51 and 10.18 F1 points compared to PGKD. The conspicuous gap between these baselines and PGKD is mainly due

[4] https://github.com/tonytan48/Re-DocRED.
[5] https://competitions.codalab.org/competitions/20717.

Table 2. Experimental results (%) on the Mix-DocRED test set. Results with † are reported from [14]. Results with ∗ are based on our implementations.

Model	Ign F1	F1	Precision	Recall
BiLSTM†	32.57 ± 0.22	32.86 ± 0.22	77.04 ± 1.01	20.89 ± 0.17
GAIN+BERT$^{\dagger}_{Base}$ [17]	45.57 ± 1.36	45.82 ± 1.38	88.11 ± 1.07	30.98 ± 1.36
ATLOP+BERT$^{*}_{Base}$ [12]	45.18 ± 0.23	45.48 ± 0.25	85.66 ± 0.30	30.96 ± 0.28
DocuNet+BERT$^{\dagger}_{Base}$ [18]	45.88 ± 0.33	45.99 ± 0.33	**94.16 ± 0.32**	30.42 ± 0.29
SSR-PU+ATLOP+BERT$^{\dagger}_{Base}$ [14]	55.21 ± 0.12	56.14 ± 0.12	70.42 ± 0.18	46.67 ± 0.14
PGKD+BERT$^{*}_{Base}$	**55.45 ± 0.20**	**56.50 ± 0.21**	65.85 ± 0.21	**49.50 ± 0.22**
GAIN+RoBERTa$^{*}_{Large}$ [17]	48.65 ± 0.24	48.76 ± 0.25	88.60 ± 0.25	33.64 ± 0.26
ATLOP+RoBERTa$^{*}_{Large}$ [12]	48.70 ± 0.30	48.91 ± 0.30	89.68 ± 0.32	33.63 ± 0.35
DocuNet+RoBERTa$^{*}_{Large}$ [18]	49.54 ± 0.27	49.74 ± 0.25	**94.81 ± 0.26**	34.27 ± 0.27
SSR-PU+ATLOP+RoBERTa$^{\dagger}_{Large}$ [14]	58.68 ± 0.43	59.50 ± 0.45	74.21 ± 0.53	49.67 ± 0.77
PGKD+RoBERTa$^{*}_{Large}$	**58.87 ± 0.24**	**59.92 ± 0.25**	67.61 ± 0.25	**53.79 ± 0.23**

Table 3. Error distributions of ATLOP and PGKD on the dev set of Mix-DocRED. In each cell, the data on the left (or right)

Predictions	Ground Truth				
	$r \in R$		NR		
$r \in R$	C 3,523 (25.49%)	6,203 (35.61%)	MR 457 (3.31%)	4,055 (23.28%)	
	W 2,021 (14.67%)	2,928 (16.81%)			
NR	MS 7,818 (56.57%)	4,231 (24.29%)	CN 179,413	175,815	

to that the former prioritize precision over recall, at the cost of sacrificing overall performance. Without the ability to systematically identify relation triples that are overlooked in the dataset, these baselines simply treat unlabeled data as negative samples. Fortunately, our PGKD is able to learn clean positive-class patterns that aid in better distinguishing between positive and negative samples. PGKD overcomes the challenge posed by noisy labeled data and appropriately increases the recall score, leading to an overall improvement in performance.

Additionally, while the decreased precision of PGKD is indeed a concern, it can be attributed to that the teacher model monitors the student model. Since the teacher model is trained only on positive samples, the student model tends to be biased towards predicting more positive samples under its guidance. Nevertheless, despite the decrease in precision, our model (PGKD) exhibits significant improvement over the baselines according to recall and F1 scores. Thus, PGKD outperforms the competitive baselines by effectively balancing the trade-off between precision and recall scores.

Table 4. Results (%) on the revised Re-DocRED test set. Results with † are reported from [13]. Results with ∗ are based on our implementations.

Model	Ign F1	F1
ATLOP+BERT$^{*}_{Base}$ [12]	73.14	73.86
DocuNet+BERT$^{*}_{Base}$ [18]	73.94	74.03
PGKD+BERT$^{*}_{Base}$	**74.28**	**74.35**
ATLOP+RoBERTa$^{\dagger}_{Large}$ [12]	76.94	77.73
DocuNet+RoBERTa$^{\dagger}_{Large}$ [18]	77.27	77.92
KD-DocRE+RoBERTa$^{\dagger}_{Large}$ [12]	77.63	78.35
PGKD+RoBERTa$^{*}_{Large}$	**77.67**	**78.38**

5.5 Error Analysis

In this section, we follow Tan et al. [12] and provide a detailed error analysis to specify ATLOP and PGKD. We compare the predictions with the ground truths to form five categories, including: 1) **C**: where all predicted relations are correct. 2) **W**: the entity pair is correctly identified, but there are certain predicted relations that are incorrect. 3) **MS**: where the model fails to identify the entity pair in the ground truth. 4) **MR**: where the model generates a relation label for a negative sample. 5) **CN**: where both the head entity and tail entity are not included in the ground truth and the predicted relation does not correspond to any existing relation. Afterwards, we design a confusion matrix in Table 3 to present each predicted category's number and score. Given that the final evaluation score is assessed based on $r \in R$ triples, the **CN** category is ignored when calculating the final score. Based on the above matrix, we can draw conclusions. Table 3 indicates that the sum of error scores under the **MR** and **MS** categories for ATLOP is 59.88%, exceeds PGKD's sum by 12.31%. This comparison proves that PGKD outperforms ATLOP when determining the relation between head and tail entities. Moreover, PGKD's score under the **C** category is 35.61% (6,203), while ATLOP's score is 25.49% (3,523), further indicating that PGKD has a higher recall score than ATLOP. These strong contrasts demonstrate the effectiveness of positive-guided knowledge distillation in addressing the problem of noisy labeled data.

5.6 Experiment on the Well-Annotated Re-DocRED

In this section, we assess the performance of PGKD in comparison with various state-of-the-art baselines, namely ATLOP, DocuNet, and KD-DocRE [12], utilizing the well-annotated Re-DocRED for both training and testing. As shown in Table 4, PGKD achieves a superior F1 score of 78.38%, outperforming the other baselines. Despite being designed to handle noisy labels, PGKD shows relative improvement over the baselines even in well-annotated scenarios. Additionally, these results can be regarded as a maximum value for document-level RE with incompletely annotated data.

Table 5. Experimental results (%) of positive relation classification on the Mix-DocRED and Re-DocRED test sets. Results with $*$ are based on our implementation.

Model	Mix-DocRED		Re-DocRED	
	Ign F1	F1	Ign F1	F1
Teacher+BERT$^*_{Base}$	78.45	78.56	90.97	91.36
Teacher+RoBERTa$^*_{Large}$	79.85	80.19	92.62	92.95

5.7 Analysis and Discussion on Positive Relation Classification

We conduct an experiment to explore the eligibility of a model trained solely with positive samples to be the teacher model. The experiment focuses on classifying relation types for positive samples, referred to as positive relation classification (PRC). The number of positive samples for each set on Mix-DocRED and Re-DocRED is summarized in Table 1. The results in Tables 2 and 5 show that the teacher model on Mix-DocRED exceeds PGKD by 22.06 (78.56 vs. 56.50) and 20.27 (80.19 vs. 59.92) F1 scores, respectively. This strongly proves that the teacher model has effectively learned patterns of positive samples, making it a valuable source of knowledge for the student model. The PRC metric, which excludes all NA samples, provides an upper bound of the RE performance for a given dataset. The performance gap between PRC and RE should not be significant if the dataset has good annotation quality. However, we observe that the performance on the standard RE task is inferior to counterpartrpar on the PRC task, suggesting that the annotations for positive samples are of higher quality compared to the entire dataset. Therefore, it is reasonable to use a model trained only with positive samples as the teacher model.

6 Conclusion

In this work, we propose a novel PGKD model to address the noisy labeled data problem for document-level RE. Our model is based on the teacher-student architecture. PGKD innovatively only utilizes positive samples to train the teacher model. The student model is partially supervised by the teacher model to avoid positive class pattern collapse and interference of noisy labeled data. We conducted experiments on two distinct datasets, Mix-DocRED and Re-DocRED. Extensive experimental results demostrate that the proposed PGKD exhibits SOTA effectiveness in denoising noisy labeled data, outperforming competitive baselines.

Acknowledgements. This work is supported by the National Natural Science Foundation of China (Grant No. 62276095 and 72204261), the National Social Science Foundation of China (Grant No. 20&ZD047), and the Hunan Provincial Natural Science Foundation of China (Grant No. 2021JJ40681).

References

1. Chen, J., Fu, T., Lee, C., Ma, W.: H-FND: hierarchical false-negative denoising for distant supervision relation extraction. In: ACL/IJCNLP, pp. 2579–2593 (2021)
2. Christopoulou, F., Miwa, M., Ananiadou, S.: Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. In: EMNLP-IJCNLP, pp. 4924–4935 (2019)
3. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT, pp. 4171–4186 (2019)
4. Goyal, P., et al.: Accurate, large minibatch SGD: training imagenet in 1 hour. arXiv preprint arXiv:1706.02677 (2017)
5. Hao, K., Yu, B., Hu, W.: Knowing false negatives: an adversarial training method for distantly supervised relation extraction. In: EMNLP, pp. 9661–9672 (2021)
6. Heo, B., Kim, J., Yun, S., Park, H., Kwak, N., Choi, J.Y.: A comprehensive overhaul of feature distillation. In: IEEE/CVF, pp. 1921–1930 (2019)
7. Hinton, G.E., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
8. Huang, Q., Hao, S., Ye, Y., Zhu, S., Feng, Y., Zhao, D.: Does recommend-revise produce reliable annotations? an analysis on missing instances in docred. In: ACL, pp. 6241–6252 (2022)
9. Li, T., Hu, Y., Ju, A., Hu, Z.: Adversarial active learning for named entity recognition in cybersecurity. Comput. Mater. Continua **66**(1) (2021)
10. Liu, Y., et al.: Roberta: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
11. Mirzadeh, S., Farajtabar, M., Li, A., Levine, N., Matsukawa, A., Ghasemzadeh, H.: Improved knowledge distillation via teacher assistant. In: AAAI, pp. 5191–5198 (2020)
12. Tan, Q., He, R., Bing, L., Ng, H.T.: Document-level relation extraction with adaptive focal loss and knowledge distillation. In: Findings of ACL, pp. 1672–1681 (2022)
13. Tan, Q., Xu, L., Bing, L., Ng, H.T.: Revisiting docred - addressing the overlooked false negative problem in relation extraction. arXiv preprint arXiv:2205.12696 (2022)
14. Wang, Y., Liu, X., Hu, W., Zhang, T.: A unified positive-unlabeled learning framework for document-level relation extraction with different levels of labeling. In: EMNLP, pp. 4123–4135 (2022)
15. Xu, B., Wang, Q., Lyu, Y., Zhu, Y., Mao, Z.: Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction. In: IAAI, EAAI, pp. 14149–14157 (2021)
16. Yao, Y., et al.: Docred: a large-scale document-level relation extraction dataset. In: ACL, pp. 764–777 (2019)
17. Zeng, S., Xu, R., Chang, B., Li, L.: Double graph based reasoning for document-level relation extraction. In: EMNLP, pp. 1630–1640 (2020)
18. Zhang, N., et al.: Document-level relation extraction as semantic segmentation. In: IJCAI, pp. 3999–4006 (2021)
19. Zhao, B., Cui, Q., Song, R., Qiu, Y., Liang, J.: Decoupled knowledge distillation. In: IEEE/CVF, pp. 11943–11952 (2022)
20. Zhou, W., Huang, K., Ma, T., Huang, J.: Document-level relation extraction with adaptive thresholding and localized context pooling. In: AAAI, pp. 14612–14620 (2021)