Augmenting Document-level Relation Extraction with Efficient Multi-Supervision

Anonymous ACL submission

Abstract

Despite its popularity in sentence-level relation extraction, distantly supervised data is rarely utilized by existing work in document-level relation extraction due to its noisy nature and low information density. Among its current applications, distantly supervised data is mostly used as a whole for pertaining, which is of low time efficiency. To fill in the gap of efficient and robust utilization of distantly supervised training data, we propose Efficient Multi-Supervision for document-level relation extraction, in which we first select a subset of informative documents from the massive dataset by combining distant supervision with expert supervision, then train the model with Multi-Supervision Ranking Loss that integrates the knowledge from multiple sources of supervision to alleviate the effects of noise. The experiments demonstrate the effectiveness of our method in improving the model performance with higher time efficiency than existing baselines.

1 Introduction

Different from traditional sentence-level Relation Extraction (RE), document-level relation extraction (DocRE) aims to extract the relations between multiple entity pairs within a document. The input documents of DocRE typically contain many named entities and are involved in multiple relation facts. Compared with sentence-level RE, DocRE is a more challenging task with richer interactions between the entity mentions within the document. Previous work in DocRE generally learns in a fully supervised manner, using human-annotated datasets with ground-truth labels for training and evaluation. However, human annotations for DocRE are more expensive than that of sentence-level RE due to the complexity of the task. Therefore, the expansion of DocRE datasets is costly and slow, which limits the application of DocRE.

Distant supervision has already been used in RE to significantly augment the training data (Mintz et al., 2009; Riedel et al., 2010). In sentence-level RE, distant supervision automatically annotates the sentences by aligning the mentioned entity pairs with the relations in the existing knowledge bases, assuming that all the co-appearing entity pairs in a sentence express their existing relations in the knowledge base. Despite the potential risk of noisy instances (instances with wrong labels), Yao et al. (2019) introduces distant supervision into the construction of DocRED, the most widely used dataset in DocRE. The statistics of distantly supervised (DS) data and human-annotated data of DocRED are shown in Table 1. According to the statistics, the size of DS data is much larger (about 20 times) than the human-annotated data in DocRED, indicating that DS data holds great potential to improve the performance of DocRE. However, due to the noisy nature of distant supervision and the overly large size of DS data, the utilization of the DS dataset is rarely discussed in the area of DocRE.

Dataset	#Doc.	#Ins.	#Fact	#Ent.
Annotated	5k	63k	56k	132k
Distant	101k	1,508k	881k	2,558k

Table 1: The statistics of the human-annotated and distantly labeled datasets of DocRED (Yao et al., 2019). Doc., Ins., Fact and Ent. indicate the numbers of documents, relation instances, relation facts and entities respectively.

With the development of Pre-trained Language Models (PLMs), some of the recent work in DocRE proposes to utilize the DS data for pretraining PLMs and achieve considerable improvements (Tan et al., 2022; Ma et al., 2023; Li et al., 2023; Sun et al., 2023). However, existing methods typically use all of the DS data for pretraining, neglecting that the expansion of DS data is much faster and cheaper than that of human-annotated data. With the fast-growing size of DS data, utilizing all of it can make pretraining extremely expensive and lead to low time efficiency. Moreover, the wrong labeling problem of DS remains a major challenge and causes a lot of noise within the DS dataset.

To improve the efficiency of DS data utilization as well as reduce the effects of noise from distant supervision, we propose Efficient Multi-Supervision (EMS) which includes two steps: (1) Document Informativeness Ranking (DIR) for data augmentation with informative DS documents and (2) noise-resistant training using Multi-Supervision Ranking Loss (MSRL). In DIR, We describe the valid information in a DS document as the reliable labels it contains, and we define a scoring criterion to rank the documents in DS data according to their informativeness. Later, we use the top informative subset of DS documents to augment the training data. In the training step, we extend the adaptive ranking loss (Zhou et al., 2021) to a more robust and flexible form called MSRL to receive supervision from multiple sources. We consider three sources of supervision: distant supervision from automatically generated labels, expert supervision from a trained model and self supervision from the output of the training model. Distant supervision and expert supervision participate in determining the desired ranking of relation classes in the loss function. Self supervision is employed to dynamically adjust the fitting priority of the relation classes. We conduct experiments on the DocRED dataset to demonstrate our method's effectiveness. The results show that EMS can efficiently augment the training process of DocRE and the ablation study demonstrates that both DIR and MSRL play important roles in improving the performance. Our contributions are summarized below:

- The proposed Document Informativeness Ranking (DIR) is the first attempt to retrieve the most informative documents from the DS dataset. It augments the training data with higher efficiency and greatly saves the time cost of DS data utilization.

- We extend the previous ranking-based loss of DocRE as Multi-Supervision Ranking-based Loss (MSRL) which enables the model to combine multiple sources of supervision in the calculation of training loss. Compared with the original ranking-based loss, MSRL is more robust against the noise from incorrect labels and is flexible in handling supervision from multiple sources.

- We provide detailed experiments and efficiency analysis for EMS. The experiments and analysis show that EMS can improve the training of DocRE with high efficiency.

2 Related Work

Relation Extraction(RE) has been a long-discussed topic in information extraction. Traditional RE mostly extracts relations between an entity pair within a sentence (Zeng et al., 2014; Wang et al., 2016; Zhang et al., 2017). However, it is shown by prior works that a large number of relation facts can only be extracted from multiple sentences (Verga et al., 2018; Yao et al., 2019). Therefore, various methods have been proposed to explore document-level relation extraction (DocRE) recently. Early methods in DocRE are mostly based on Graph Neural Networks (Scarselli et al., 2008). Quirk and Poon (2017) first introduces document-level graphs, in which they use words as nodes and dependency information as edges. Later graph-based methods (Peng et al., 2017; Song et al., 2018; Jia et al., 2019; Christopoulou et al., 2019; Nan et al., 2020; Zeng et al., 2021) typically extends the GNN architectures to learn better representations for the entity mentions. Recently, transformer-based methods, especially those with pretrained language models, have become popular since they can automatically learn the dependency information (Verga et al., 2018; Wang et al., 2019; Tang et al., 2020; Ye et al., 2020). Particularly, Zhou et al. (2021) proposes the adaptive thresholding loss to make the classification threshold adjustable to different entity pairs. Tan et al. (2022) adopts knowledge distillation to utilize the large but noisy distantly supervised data. Some recent work also leverages the DS data for better performance (Ma et al., 2023; Li et al., 2023; Sun et al., 2023).

However, previous methods typically use all the DS data for pertaining, which is of low efficiency. Therefore, we seek to utilize only the most informative part of the DS data to improve the model performance with higher efficiency. Moreover, we modify the widely used adaptive thresholding loss (Zhou et al., 2021) to a generalized form integrating multiple sources of supervision to mitigate the noisy instance problem in DS data.

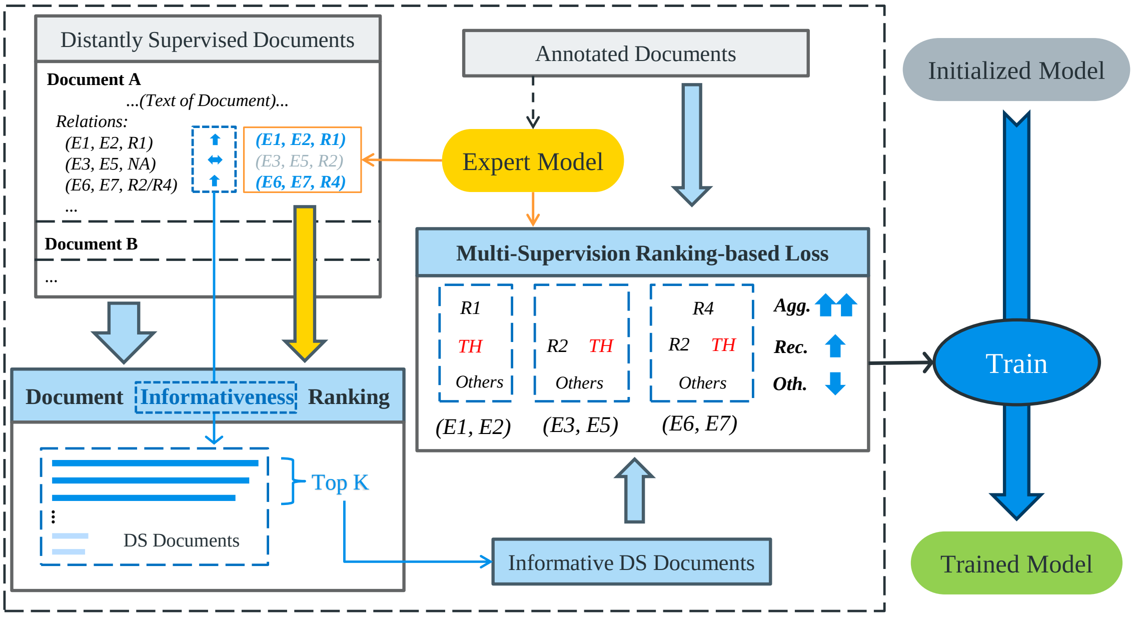


Figure 1: The illustration of EMS, contains two main components: DIR and MSRL. In MSRL, **Agg.** represents *aggreements*, **Rec.** represents *recommendations* and **Oth.** represents *others*.

3 Methodology

The overall framework of EMS, our proposed method, is illustrated in Figure 1. In both DIR and MSRL, we adopt a pretrained expert model to provide an extra source of supervision by making predictions on the DS data. First, we retrieve a set of the most informative documents from the DS dataset using Document Informativeness Ranking (DIR) to augment the training data of DocRE. Then, the model is trained using the augmented training data with the help of Multi-Supervision Ranking-based Loss (MSRL). MSRL enhances the model's training by ensuring the relation classes adhere to specific rankings based on their logits. It also employs self supervision to dynamically adjust the learning of relation classes.

3.1 Preliminary

The task of document-level relation extraction is to predict the relation classes between pairs of entities $(e_s, e_o)_{s,o=1...n;s\neq o}$ given a document D containing the entity set $\{e_i\}_{i=1}^{n}$. where e_s and e_o represent the subject and object respectively. The set of predefined relation classes is $R \cup \{NA\}$, where NA stands for *no relation* between the entity pair. During the testing process, the relations between all the possible entity pairs $(e_s, e_o)_{s,o=1...n;s\neq o}$ are predicted and there may be multiple relation classes between e_s and e_o. Each of the entity pairs is called an instance in the following parts.

The annotation process of distantly supervised data is based on the assumption that if two entities participate in a relation, any document that contains those two entities expresses that relation (Mintz et al., 2009; Yao et al., 2019). This assumption is too strong and causes the unreliability of DS labels. In order to provide an extra source of automatic annotation, we utilize a trained expert model to make predictions on the distantly supervised dataset. The relation triples provided by distant supervision are denoted as (e_s, e_o, r_{DS}) and the expert predictions are denoted as (e_s, e_o, r_{EX}), where both r_{DS} and r_{EX} are within $R \cup \{NA\}$ and may indicate multiple relation classes.

3.2 Document Informativeness Ranking

The information within distantly labeled documents is hard to obtain due to its noisy nature. Still, it holds the potential to improve the performance of DocRE models and most of the relevant methods use all the DS data for pretraining before fine-tuning on the annotated set (Tan et al., 2022; Ma et al., 2023; Sun et al., 2023). However, using all of the DS data is often of low efficiency. As shown in Table 1, the size of existing DS data is far larger than the annotated dataset. Moreover, the automatic annotation process of DS data enables it to expand much faster and cheaper than the human-annotated dataset. Pretraining using the DS dataset before refining the model with the human-annotated dataset can help with performance gains, but this approach makes the whole process too expensive in time cost for realistic use.

To overcome this challenge and efficiently utilize the DS dataset for the DocRE task, we propose the Document Informativeness Ranking (DIR) to retrieve the most informative subset of the DS dataset to augment the training data of DocRE. We argue that relying solely on the DS label is inevitably biased and at least one extra source of reference labels should be introduced. Therefore, we employ an expert model to automatically generate predictions on the DS data, which can be pretrained on an annotated dataset to increase efficiency. For a fair comparison in the experiments, the expert model and the training model share identical network architectures.

Based on the consistency between the DS labels and the expert predictions, we divide the relation classes of each instance into three groups:

- Agreements (Agg.): relation classes indicated by both the DS label and the expert prediction.
- Recommendations (Rec.): relation classes indicated by either the DS label or the expert prediction, but not both.
- Others (Oth.): the rest of the relation classes (not indicated by either the DS label or the expert prediction).

Distant supervision typically aligns the entities with existing pairs in the knowledge base, while the expert makes predictions based on the knowledge it has learned (from the human-annotated dataset). In other words, distant supervision is a source of prior knowledge while the expert is usually context-based. Thus, *agreements* can be seen as relatively reliable labels as they are supported by both prior knowledge and contextual information. *Recommendations* have discrepancies between prior knowledge and the context, which could be attributed to either the incompleteness of existing knowledge or potential biases held by the expert. In either case, the document may still express the relation classes in the *Recommendations* group. However, the relation classes in the *Others* group lack grounding in prior knowledge or context, making their presence in the document less probable.

The informativeness of a document can be described as the amount of **reliable** and **valuable** information it contains. In our work, each DS instance (e_s, e_o, r_{DS}) is considered an individual information contributor, and the amount of information it contributes is determined by the reliability and scarcity of its label. Considering the presence of two labeling sources (distant supervision and expert prediction), we propose the equation below as an attempt to quantify a document's informativeness:

$$I(D) = \sum_{i=1}^{N_D} \sum_{j=1}^{N_r} V_{r_j} (y_{DS}^{ij} \cdot y_{EX}^{ij}) P_{EX}^{ij} \qquad (1)$$

where y_{DS}^{ij} and y_{EX}^{ij} are the one-hot labels from distant supervision and expert prediction respectively. P_{EX}^{ij} is the softmaxed output distribution from the expert. N_D is the number of entity pairs in the document. N_r is the number of predefined relation classes R, note that NA is not considered in DIR. V_r is a class-weighting vector to encourage the retrieval of the instances expressing rare relation classes. In the experiments, we directly apply the class weight function of scikit-learn [1] toolkit based on the distribution of classes in the human-annotated dataset. After computing informativeness, we retrieve a subset of the documents with the largest informativeness $I(D)$ in the DS dataset, forming an augmentation set S_{aug} where $S_{aug} \subset S_{DS}$. During training, the augmentation set S_{aug} is mixed with the human-annotated dataset S_{ann}. The DS documents in S_{aug} are considered of relatively high quality, but they still contain many wrongly labeled instances to be addressed by MSRL in training.

3.3 Multi-Supervision Ranking-based Loss

Ranking-based loss functions like the Adaptive-Thresholding Loss (ATL) (Zhou et al., 2021) are widely used by previous DocRE methods. In ATL, an adaptive threshold TH is introduced to separate the relation classes expressed by the instance (positive) and those unexpressed relation classes (negative). The goal of training is to push positive classes above the threshold and keep negative ones below the threshold, adhering to the $positive \rightarrow TH \rightarrow negative$ ranking order. However, the labels from distant supervision can be misleading and may cause a lot of false positive or false negative instances. To alleviate this issue, we extend ATL to the Multi-Supervision Ranking-based

[1] https://scikit-learn.org/

Loss (MSRL) which combines multiple sources of supervision to alleviate the effects of noisy instances.

Different from ATL, MSRL receives two sources of labels: distant supervision and expert prediction. As stated previously, the relation classes for each instance can be divided into *agreements*, *recommendations* and *others*. Intuitively, we hope to push *agreements* above threshold TH and keep *others* below TH. As for *recommendations*, we also keep them above *others* without additional ranking restrictions. The idea is to fit the *recommendations* in a self-paced manner, hoping that reliable *recommendations* can rise above the threshold TH while unreliable ones stay below.

Similar to ATL (Zhou et al., 2021), the logit vector is broken down into two parts to compute the probability vectors:

$$\begin{aligned} P_r^a &= \frac{exp(O_r)\mathbb{1}(r \in R_{agg.})}{\sum_{r' \in R_{agg.} \cup \{TH\}} exp(O_{r'})} \\ P_r^b &= \frac{exp(O_r)\mathbb{1}(r \in R_{rec.} \cup \{TH\})}{\sum\limits_{r' \in R_{rec.} \cup R_{oth.} \cup \{TH\}} exp(O_{r'})}) \end{aligned} \tag{2}$$

where P_r^a only involves *aggreements* and the TH class, with an indicator function $\mathbb{1}$ filtering the relation classes except *aggreements*. P_r^b involves *recommendations*, *others* and the TH class.

Since *recommendations* potentially contain wrong labels reflecting incomplete knowledge or biases, we hope to carefully adjust their fitting priorities during training. Intuitively, the *recommendations* confirmed by the current predictions y of the training model and with higher probabilities P_r^b are less likely to be noisy. Thus, we design an extra class weighting mechanism based on self supervision to mitigate noisy *recommendations*. On the other hand, we hope to encourage the model to focus more on the under-fitted *aggreements* to effectively learn reliable knowledge. Therefore, the class weighting mechanism within MSRL is also divided into two parts:

$$\begin{aligned} w_r^a &= \gamma_a + (1 - y_r)(1 - P_r^a) \\ w_r^b &= \gamma_b + y_r P_r^b \end{aligned} \tag{3}$$

where γ_a and γ_b are the offsets of class weights which are based on the need for normalization. When y_r is negative (equals 0) and P_r^a is small, it indicates that the class belonging to *agreements* is under-fitted. In this case, a larger w_r^a can prompt the model towards better fitting of the *agreements*. In contrast, w_r^b only rewards those reliable *recommendations* with positive predictions $y_r = 1$ and large probability values P_r^b.

Finally, MSRL is defined in the following form:

$$L = -\sum_r \log(w_r^a P_r^a) + \log(w_r^b P_r^b) \tag{4}$$

where the first term pushes *agreements* above TH and the second term keeps *recommendations* and TH above *others*. Both distant and expert supervisions are involved in dividing the relation classes into *agreements*, *recommendations* and *others*, while self supervision dynamically adjusts the learning priorities within the groups. In summary, the idea of multi-supervision not only allows MSRL to divide the relation classes in a more fine-grained manner but also enables flexibility in handling uncertainty. When using MSRL to train on human-annotated data, there are only expert supervision (human annotations) and self supervision available. In this case, there is no *recommendations*, and MSRL is equivalent to a ranking-based loss with adaptive thresholds and class weights w_r^a accelerating the learning of under-fitted positive relation classes.

Different from knowledge distillation, which uses soft labels (logits) from the teacher model as an extra source of knowledge. EMS uses one-hot labels from the expert model in both DIR and MSRL. Intuitively, soft labels contain more information than one-hot labels. However, soft labels may not be accessible in some cases, for example, when using text-to-text language models or human annotators. Therefore, employing one-hot labels enables more flexible choices for the expert in real applications.

4 Experiments

In this section, we first introduce the dataset and experimental settings used in our experiments. Then, we provide our main experiment results and compare EMS with several strong baselines. Finally, we discuss the effects of DIR and MSRL through ablation study.

4.1 Datasets and Settings

We employ the DocRED (Yao et al., 2019) dataset in our experiments. DocRED is a large-scale DocRE dataset constructed from Wikipedia and Wikidata. It is the most widely used dataset for DocRE so far and has the largest available

Hyperparameter	Value
Batch size	4
Number of epochs	30
Number of relation classes N_r	96
Class Weight Offsets γ_a/γ_b	1.0 / 0.9

Table 2: The details of experimental settings.

DS dataset. The statistics of DocRED are already displayed in Table 1. The human-annotated dataset is divided into train/dev/test sets, with 3053/1000/1000 documents respectively. We use the dev set for evaluation and choose the best model for testing.

The base model of our experiment is ATLOP (Zhou et al., 2021), which is a popular benchmark in DocRED. We use the same ATLOP architecture for the expert model and the training model for fair comparisons. The encoder is initialized using *bert-base-cased* checkpoint(Devlin et al., 2018). Due to the limitation of infrastructure, the experiments are run using smaller batch size settings. Our model is optimized with AdamW (Loshchilov and Hutter, 2017) using a 5e-5 learning rate for the encoder and 1e-4 for the classifier, with the first 6% steps as warmup steps. Other details of hyperparameters are shown in Table 2.

The evaluation metrics are F_1 and Ign F_1. The Ign F_1 represents the F_1 score excluding the relation triples shared by the human-annotated training set.

4.2 Compared Baselines

We compare our EMS with several strong baselines, some of which also utilize DS data in their frameworks. ATLOP (Zhou et al., 2021) proposes a localized context pooling layer to aggregate related context for entity pairs to get better entity representations and utilizes an adaptive thresholding loss function to replace the global threshold with an entity-pair-dependent threshold. ATLOP is also the expert model adopted in our experiments. SSAN(Xu et al., 2021) utilizes co-occurrence information between entity mentions and extends the standard self-attention mechanism with structural guidance. SIRE(Zeng et al., 2021) employs a sentence-level encoder to extract intra-sentence relations and a document encoder to extract inter-sentence relations respectively to represent two types of relations in different ways. DocuNet (Zhang et al., 2021) regards the DocRE task as a semantic segmentation task, attempting to capture both local context information and global interdependency among triples. NCRL (Zhou and Lee, 2022) proposes a multi-label loss that prefers large label margins between the NA class and the predefined relation classes. KD-DocRE (Tan et al., 2022) proposes an adaptive focal loss to alleviate the long-tailed problem and uses knowledge distillation to utilize the DS dataset. The compared methods all use the BERT (Devlin et al., 2018) encoder for fair comparisons.

As for methods concerning DS data, we choose KD-DocRE for comparison, which uses all the DS data in pretraining. KD-DocRE also shares a similar network architecture with ATLOP, which makes it a good fit for comparison. We also present the result of ATLOP pretrained by DS data and fine-tuned by human-annotated data to compare with the performance and efficiency of EMS.

4.3 Main Results

Table 3 shows the experimental results on the DocRED dataset. According to the results, DS data greatly improves the performance of DocRE models. With only 3% of DS data, the performance of ATLOP+EMS almost surpasses the state-of-the-art non-DS methods. However, DS data significantly increases the time costs of the models due to its massive size. Taking ATLOP as an example, with DS data, the performance increases by 2.63 on test F_1 and 2.72 on test Ign F_1. However, the time cost dramatically increases to more than 30 times due to the use of all DS data in pretraining. KD-DocRED, the state-of-the-art method, also requires a substantial cost of time to achieve good performance. By retrieving informative instances and denoised training with MSRL, EMS can improve performance with higher efficiency than the baselines. Using only 3% of DS data, ATLOP+EMS achieves 1.2 and 1.17 improvements on test F_1 and test Ign F_1 respectively, only increasing the time cost to 4 times. Using 30% of DS data, ATLOP+EMS even surpasses ATLOP with DS pretraining by 0.41 test F_1 and 0.55 test Ign F_1. It also achieves a comparable performance to the state-of-the-art method with 13 times the cost of the original ATLOP, which is significantly smaller than KD-DocRE.

In practice, the size of DS data grows faster than the human-annotated dataset because DS labels are much cheaper and faster to obtain. Therefore, EMS can save even more time in the real application of distant supervision.

Model	Dev		Test		Relative Time Cost
	F_1	Ign F_1	F_1	Ign F_1	
Without distantly supervised data					
ATLOP* (Zhou et al., 2021)	61.05	59.18	60.85	58.71	1x
SSAN(Xu et al., 2021)	59.19	57.03	58.16	55.84	-
SIRE(Zeng et al., 2021)	61.60	59.82	62.05	60.18	-
DocuNet(Zhang et al., 2021)	61.83	59.86	61.86	59.93	-
NCRL*(Zhou and Lee, 2022)	61.10	59.22	60.91	58.77	-
KD-DocRE (Tan et al., 2022)	62.03	60.08	62.08	60.04	-
With all of distantly supervised data					
ATLOP with DS	63.42	61.57	63.48	61.43	34x
KD-DocRE with DS	**64.81**	**62.62**	**64.76**	**62.56**	111x
With EMS					
ATLOP+EMS (3% DS data)*	62.39	60.56	62.05	59.88	4x
ATLOP+EMS (30% DS data)*	<u>64.08</u>	<u>62.11</u>	<u>63.89</u>	<u>61.98</u>	13x

Table 3: Results of EMS and baselines on DocRED. Models marked with * are reproduced or implemented by us, others are from the papers. The relative time costs are estimated using the method in Appendix A. **Bold** indicates the best results among the compared methods, the second best results are <u>underlined</u>.

4.4 Ablation Study

Model	Dev	
	F_1	Ign F_1
ATLOP+EMS	**62.39**	**60.56**
- Self Sup.	62.19	60.23
- Expert Sup.	61.33	59.21
- Distant Sup.	61.59	59.70
Rand+MSRL	61.72	59.84
ATLOP	61.05	59.18

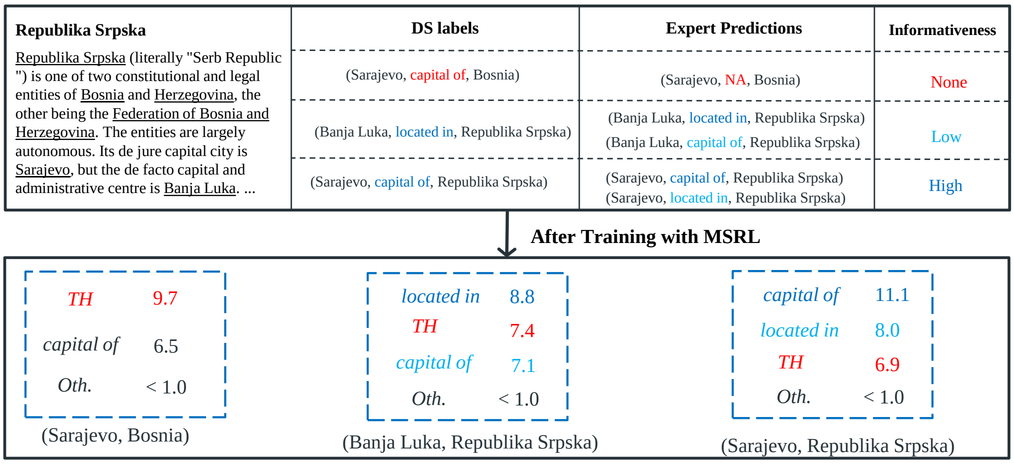
Table 4: Ablation study of our method using top 3% of the DS data. **Sup.** is the abbreviation of **supervision**.

Table 4 shows the results of the ablation study. We conduct this part of experiments on the dev set of DocRED using the top 3% of the DS data. Removing self supervision means removing the class weights w_r^a and w_r^b defined in Equation 3. This slightly affects the performance because class weights not only accelerate the learning of underfitted *agreements* but also reduce the effects of noisy *recommendations*. If the size of the augmentation set increases, the effects of removing self supervision will be more significant because more noisy instances are introduced into the training process.

Since MSRL distinguishes *agreements*, *recommendations* and *Others* based on the consistency between distant supervision and expert supervision, removing either of them essentially disables MSRL. Removing distant supervision leads to sole dependence on an expert model trained on a smaller dataset, which can lead to inaccurate predictions due to unseen patterns. Removing expert supervision, on the other hand, leaves a large number of noisy instances unaddressed. Thus, both distant supervision and expert supervision are crucial for MSRL. According to the results in the third and fourth row of Table 4, removing either distant supervision or expert supervision leads to a significant performance decline.

Rand+MSRL is a variation that selects augmentation set randomly instead of on the basis of informativeness. Other settings are identical to ATLOP+EMS. The presented result is from the best model among five runs using five different random seeds. The performance decreases by 0.67 for F_1 and 0.72 for Ign F_1 compared with using DIR. The difference in performance demonstrates that DIR is effective in retrieving informative documents from the DS dataset.

From the above results and discussions, we can conclude that distant supervision, expert supervision and self supervision all proved useful in the EMS framework. Also, it is clear that the two main components of EMS, DIR and MSRL, are both effective in improving the performance of DocRE.



Republika Srpska	DS labels	Expert Predictions	Informativeness
Republika Srpska (literally "Serb Republic ") is one of two constitutional and legal entities of Bosnia and Herzegovina, the other being the Federation of Bosnia and Herzegovina. The entities are largely autonomous. Its de jure capital city is Sarajevo, but the de facto capital and administrative centre is Banja Luka. ...	(Sarajevo, capital of, Bosnia)	(Sarajevo, NA, Bosnia)	None
	(Banja Luka, located in, Republika Srpska)	(Banja Luka, located in, Republika Srpska) (Banja Luka, capital of, Republika Srpska)	Low
	(Sarajevo, capital of, Republika Srpska)	(Sarajevo, capital of, Republika Srpska) (Sarajevo, located in, Republika Srpska)	High

Figure 2: A retrieved document with some representative instances. The numbers are the logit values of the relation classes after training, and "located in" is the abbreviation of relation class "located in the administrative territorial entity".

5 Case Study

In order to illustrate the idea of multi-supervision, we choose an example document retrieved by DIR from the DS dataset and present it in Figure 2. At the upper part of the figure, DIR estimates the informativeness of each instance. Since *capital of* is a rare relation class with only dozens of instances in the human-annotated set while *located in* is a more common one, the informativeness of the third instance is higher than the second one. After training on the augmented dataset $\{S_{ann} \cup S_{aug}\}$ with MSRL, the logit values of the instances are shown at the bottom part of Figure 2. For the pair (Sarajevo, Bosnia), *capital of* is not the gold label according to the context of the document, but distant supervision indicates that *capital of* could be applicable to this entity pair in other contexts. Therefore, it is acceptable and reasonable that *capital of*, which is a *recommendation* from distant supervision, has a higher logit value than the classes from *Others*. Regarding entity pairs (Banja Luka, Republika Srpska) and (Sarajevo, Republika Srpska), the *agreements* are both far above the TH threshold. The *recommendation* for (Banja Luka, Republika Srpska) is ambiguous because it is undefined whether de jure capital indicates the *capital of* relation, so the logit value is near the *TH* threshold. The *recommendation* for (Sarajevo, Republika Srpska), *located in*, is a missing gold label due to the incompleteness of distant supervision. Therefore, the logit value of *located in* tends to rise above the threshold after learning from the augmented dataset. This case study illustrates the process of DIR and the outcome of MSRL and shows that integrating multiple sources of supervision enables the model to learn from DS instances with better robustness and flexibility.

6 Conclusions

In this paper, we introduce EMS, an efficient and effective approach leveraging DS data to enhance DocRE models. EMS comprises two key components: DIR and MSRL. Unlike traditional methods that costly pretrain on the entire DS dataset, DIR retrieves the most informative documents from DS to create an augmentation set. Subsequently, the model undergoes training with MSRL, which flexibly mitigates noisy DS labels by integrating multiple sources of supervision. Our experiments demonstrate that EMS can significantly boost the DocRE model with higher time efficiency than existing baselines.

7 Limitations

Our work still has some limitations. Firstly, EMS depends on an expert model to provide an extra source of supervision, meaning that the capability of the expert is crucial to the effectiveness of EMS. Secondly, the useful information within the informative documents is still very sparse due to the highly noisy nature of distant supervision, which makes the learning on the augmentation set inefficient compared with that on the annotated set. Thirdly, though the network architecture is not likely to affect the efficacy of EMS, there is still a lack of combinations between EMS and all kinds of DocRE models in our experiments.

References

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. 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 4925–4936, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Robin Jia, Cliff Wong, and Hoifung Poon. 2019. Document-level n-ary relation extraction with multiscale representation learning. 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 3693–3704, Minneapolis, Minnesota. Association for Computational Linguistics.

Junpeng Li, Zixia Jia, and Zilong Zheng. 2023. Semi-automatic data enhancement for document-level relation extraction with distant supervision from large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5495–5505, Singapore. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Youmi Ma, An Wang, and Naoaki Okazaki. 2023. DREEAM: Guiding attention with evidence for improving document-level relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1971–1983, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu. 2020. Reasoning with latent structure refinement for document-level relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1546–1557, Online. Association for Computational Linguistics.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence n-ary relation extraction with graph lstms. *Transactions of the Association for Computational Linguistics*, 5:101–115.

Chris Quirk and Hoifung Poon. 2017. Distant supervision for relation extraction beyond the sentence boundary. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1171–1182, Valencia, Spain. Association for Computational Linguistics.

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.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80.

Linfeng Song, Yue Zhang, Zhiguo Wang, and Daniel Gildea. 2018. N-ary relation extraction using graph-state LSTM. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2226–2235, Brussels, Belgium. Association for Computational Linguistics.

Qi Sun, Kun Huang, Xiaocui Yang, Pengfei Hong, Kun Zhang, and Soujanya Poria. 2023. Uncertainty guided label denoising for document-level distant relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15960–15973, Toronto, Canada. Association for Computational Linguistics.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1672–1681, Dublin, Ireland. Association for Computational Linguistics.

Hengzhu Tang, Yanan Cao, Zhenyu Zhang, Jiangxia Cao, Fang Fang, Shi Wang, and Pengfei Yin. 2020. Hin: Hierarchical inference network for document-level relation extraction. In *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part I 24*, pages 197–209. Springer.

Patrick Verga, Emma Strubell, and Andrew McCallum. 2018. Simultaneously self-attending to all mentions for full-abstract biological relation extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 872–884, New Orleans, Louisiana. Association for Computational Linguistics.

Hong Wang, Christfried Focke, Rob Sylvester, Nilesh Mishra, and William Wang. 2019. Fine-tune bert for docred with two-step process. *arXiv preprint arXiv:1909.11898*.

Linlin Wang, Zhu Cao, Gerard De Melo, and Zhiyuan Liu. 2016. Relation classification via multi-level attention cnns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1298–1307.

Benfeng Xu, Quan Wang, Yajuan Lyu, Yong Zhu, and Zhendong Mao. 2021. Entity structure within and throughout: Modeling mention dependencies for document-level relation extraction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14149–14157.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Peng Li, Maosong Sun, and Zhiyuan Liu. 2020. Coreferential Reasoning Learning for Language Representation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7170–7186, Online. Association for Computational Linguistics.

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

Shuang Zeng, Yuting Wu, and Baobao Chang. 2021. SIRE: Separate intra- and inter-sentential reasoning for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 524–534, Online. Association for Computational Linguistics.

Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. 2021. Document-level relation extraction as semantic segmentation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3999–4006. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14612–14620.

Yang Zhou and Wee Sun Lee. 2022. None class ranking loss for document-level relation extraction. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4538–4544. International Joint Conferences on Artificial Intelligence Organization. Main Track.

A Time Efficiency

Previous methods concerning DS data mostly involve pertaining using the whole DS data. Taking KD-DocRE(Tan et al., 2022) as an example, it first pretrains the teacher model on DS data, then inference logits for the DS data. It also needs to pretrain the student model on DS data before fine-tuning it on the human-annotated dataset. In contrast, DIR pretrains the expert on human-annotated dataset, inferences using the DS data for informativeness ranking, and trains the model with the augmented dataset. Since DIR does not need to repeatedly train on the large DS dataset, it is much more efficient in the cost of time compared with previous baselines. For better comparisons, we give a rough estimation to support our idea based on the number of processing steps.

For convenience of estimation, we assume the processing time needed for inference or training on the same set of data is similar. Under this assumption, we further assume the time needed for one processing step in inference or training as t, which is the minimal unit of time in our analysis. We represent the sizes of DS data S_{DS} and human-annotated data S_{ann} as M and m respectively. With the above notations, we are able to represent the estimated time costs of DocRE methods. For example, the time cost of training the original ATLOP model for 30 epochs can be estimated as $30mt$.

For EMS, we assume that each training round includes $\{k_n, n = 1, 2\}$ epochs. Then, the time cost of ATLOP+EMS can be estimated as $((k_1 + k_2)m + k_2 m_A + M)t$ with m_A being the size of the augmentation set S_{aug}. By taking $\{k_n, n = 1, 2\}$ as $\{30, 30\}$ respectively, $\frac{M}{m} \approx 33$ in DocRED, and $\frac{m_A}{m} \approx 10$ in our setting, the estimated time cost is $393mt$. We adopt the time cost of the original ATLOP ($30mt$) as the standard time cost for ease of comparison, and the relative time cost of KD-DocRE is $\frac{393mt}{30mt} \approx 13$. We estimate the relative time cost of DS-related methods using the same idea and present the results in Table 3.

Notably, ATLOP has the simplest architecture among the analyzed methods and intuitively has the shortest processing time in each training step.

Therefore, the relative time costs of KD-DocRED are likely to be underestimated.