Towards Integration of Discriminability and Robustness for Document-Level Relation Extraction

Jia Guo[*1,2], **Stanley Kok**[1], **Lidong Bing**[2]
[1]School of Computing, National University of Singapore
[2]DAMO Academy, Alibaba Group
guojia@u.nus.edu, skok@comp.nus.edu.sg
l.bing@alibaba-inc.com

Abstract

Document-level relation extraction (DocRE) predicts relations for entity pairs that rely on long-range context-dependent reasoning in a document. As a typical multi-label classification problem, DocRE faces the challenge of effectively distinguishing a small set of positive relations from the majority of negative ones. This challenge becomes even more difficult to overcome when there exists a significant number of annotation errors in the dataset. In this work, we aim to achieve better integration of both the discriminability and robustness for the DocRE problem. Specifically, we first design an effective loss function to endow high discriminability to both probabilistic outputs and internal representations. We innovatively customize entropy minimization and supervised contrastive learning for the challenging multi-label and long-tailed learning problems. To ameliorate the impact of label errors, we equipped our method with a novel negative label sampling strategy to strengthen the model robustness. In addition, we introduce two new data regimes to mimic more realistic scenarios with annotation errors and evaluate our sampling strategy. Experimental results verify the effectiveness of each component and show that our method achieves new state-of-the-art results on the DocRED dataset, its recently cleaned version, Re-DocRED, and the proposed data regimes. [1]

1 Introduction

The problem of document-level relation extraction (DocRE) has garnered increasing attention from the research community (Quirk and Poon, 2017; Peng et al., 2017; Yao et al., 2019) due to its importance to real-world applications. DocRE is inherently a multi-label problem, in which we have to predict a set of relations from the pre-defined label set for every entity pair in a document. Thus, it is crucial for DocRE models to adopt an effective learning objective that can clearly distinguish massive semantically close relations.

Recently, several works have proposed new loss functions to learn an adaptive threshold for better separating positive and negative relations. However, these approaches (Zhou et al., 2021; Tan et al., 2022a) either enforce learning a *total order* among all relations that leads to superfluous comparisons and diminishing differences among them or improperly penalize all pre-defined labels of positive entity pairs if their average margins are lower than the threshold when addressing the label imbalance problem (Zhou and Lee, 2022). In contrast, we propose an approach that learns a *partial order*, ranking all positive relations above a threshold individually, which is in turn ranked above all negative relations. Our approach does not waste precious data and probability mass in modeling the ordering among positive relations (likewise for negative relations). We further sharpen the distinction in each distribution of a relation and the threshold through the principled use of entropy minimization.

Besides, none of the above methods take the discriminability of internal representations into account, as well as the model robustness against annotation errors. To solve these issues, we introduce novel modifications to the supervised contrastive learning (Khosla et al., 2020) to accentuate the differences among the embeddings of entity pairs from different classes and the similarities of that from the same class. Our method can better accommodate the multi-label setting and the long-tail phenomenon that is typically present in DocRE datasets. To combat the annotation error problem stated in Tan et al. (2022b), we design two new data regimes and a novel negative label sampling strategy that gives consistently strong performance even with incomplete annotations. In sum, our

*This work was partially done when Jia Guo was an intern at DAMO Academy, Alibaba Group.

[1]Our codes and datasets are available at https://github.com/guojiapub/PEMSCL.

contributions are three-fold:

- We propose an effective loss function that boosts the discriminability of both internal embeddings and probabilistic outputs.
- We achieve good integration of discriminability and robustness by incorporating a novel negative label sampling strategy.
- Experimental results consistently demonstrate that we achieve new state-of-the-art performance in a variety of settings.

2 Related Work

Document-level relation extraction (DocRE) Early works on DocRE focus on utilizing graph convolutional networks (GCNs) (Kipf and Welling, 2017) to conduct complex cross-sentence reasoning on a document graph (Sahu et al., 2019; Christopoulou et al., 2019; Wang et al., 2020; Zeng et al., 2021). Recently, methods fine-tuned on large pre-trained language models (Devlin et al., 2019; Liu et al., 2019) achieved significant performance gain. In particular, SSAN (Xu et al., 2021) encoded entity dependencies into the self-attention mechanism to strengthen context and entity reasoning. ATLOP (Zhou et al., 2021) employed the multi-head attention weights to generate entity-related context representations which enhanced the embeddings of entity pairs. To better address the multi-label classification problem, both ATLOP (Zhou et al., 2021) and NCRL (Zhou and Lee, 2022) proposed to treat the NA class as an adaptive threshold. DocuNet (Zhang et al., 2021) and KD-DocRE (Tan et al., 2022a) extended the ATLOP architecture by increasing interactions between entities and incorporating knowledge distillation, respectively. Besides, other DocRE models attempted to leverage auxiliary information for relation prediction, such as meta dependency paths (Nan et al., 2020), external knowledge bases (Li et al., 2021a), and evidences (Xie et al., 2022; Xiao et al., 2022). We additionally provide detailed comparison with existing works in Section 3.3.

Other related works Entropy Minimization technique was commonly seen in semi-supervised learning works (Grandvalet and Bengio, 2004; Vu et al., 2019). However, we are the first to employ entropy minimization in the challenging multi-label supervised learning framework. Besides, our entropy minimization takes effect in each customized probability distribution of the relation label and threshold class, which will encourage a larger distinction between them.

Supervised contrastive learning (SCL) (Khosla et al., 2020) extends self-supervised contrastive learning (He et al., 2020; Chen et al., 2020) to the fully supervised setting by constructing "positive" and "negative" examples based on their labels. ERICA (Qin et al., 2021) proposed a pre-training framework using contrastive learning to improve representations of entities and relations. However, this work samples positive pairs for relations proportionally to their total amount of examples, which will lead to biased optimization that favors primary relations over minor ones. Besides, they only maximize the similarity of one positive example pair each time, which may weaken the global effect of clustering. Instead, we give equal consideration to each relation and each positive example of anchors, and elaborately tailored the supervised contrastive learning to suit both the multi-label problem and long-tailed relation learning.

3 Methodology

In this section, we describe our model called PEM-SCL that is based on a **P**airwise moving-threshold loss, **E**ntropy **M**inimization, and **S**upervised **C**ontrastive **L**earning.

3.1 Problem Formulation

Let $D = \{w_l\}_{l=1}^{L}$ be a document containing L words and a set of entities $\mathcal{E}_D = \{e_i\}_{i=1}^{|\mathcal{E}_D|}$. Each entity e_i is associated with a set of mentions $\mathcal{M}_{e_i} = \{m_j^i\}_{j=1}^{|\mathcal{M}_{e_i}|}$ (i.e., a set of phrases referring to the same entity e_i). In document-level relation extraction, we predict the subset of relations in a predefined set $\mathcal{R} = \{r_k\}_{k=1}^{|\mathcal{R}|}$ that hold between each pair of entities $(e_h, e_t)_{h,t=1,\ldots,|\mathcal{E}_D|,h\neq t}$. We sometimes abbreviate an entity pair (e_h, e_t) as (h, t) to simplify notation. A relation is deemed to exist between the head entity e_h and tail entity e_t if it is expressed between any of their corresponding mentions. If no relation exists between any pair of their mentions, the entity pair is labeled `NA`. For each entity pair, we term a relation that holds between its constituent entities as *positive*, and the remaining relations in $\mathcal{R}$ as *negative*. An entity pair that is `NA` does not have any positive relation, and has the entire set $\mathcal{R}$ as negative relations (we could consider such a pair as having a special `NA` relation between them). Document-level relation extraction can be viewed as a multi-label problem, in which an entity

pair corresponds to a training/test *example*, and the relations in $\mathcal{R} \cup \{\texttt{NA}\}$ correspond to the possible *labels* or *classes* of the example.

3.2 Encoder Model

We leverage ATLOP (Zhou et al., 2021) as our encoder since recent work (Xie et al., 2022; Zhou and Lee, 2022) has borne out its usefulness as a backbone in neural architectures. For each entity pair (e_h, e_t), the encoder model generates the entity pair representation $\boldsymbol{x}_{h,t} \in \mathbb{R}^{d_x}$, and its unnormalized score vector $\boldsymbol{f}_{h,t} \in \mathbb{R}^{|\mathcal{R}|+1}$ for relation prediction, we briefly describe them as follows[2]:

$$\boldsymbol{x}_{h,t} = \text{Encoder}\big((e_h, e_t)|D, \mathcal{M}_{e_h}, \mathcal{M}_{e_t}\big) \quad (1)$$

$$\boldsymbol{f}_{h,t} = \text{Linear}(\boldsymbol{x}_{h,t}) \quad (2)$$

3.3 Pairwise Moving-Threshold Loss with Entropy Minimization

In document-level relation extraction, a fixed probability threshold (e.g. a hyperparameter tuned on the development dataset) is used to decide the boundary of positive and negative relations. However, such a threshold is only suitable for entity pairs *on average*, and may not be ideal for entity pairs with particular properties.

To address this problem, we design a loss function that utilizes the NA class as a dynamic threshold, learning how best to *move* the threshold in accordance with the regularities present in each entity pair. Specifically, we conduct a *pairwise* comparison between each relation and the NA class (separately for each relation), and encourage the prediction scores of each positive relation to be higher than that of the NA class, and incentivize the score of the NA class to be higher than those of negative relations. In this way, we induce a *partial order* over $\mathcal{R} \cup \{\texttt{NA}\}$ for each entity pair. Note that the positive relations are not compared against each other, and their relative rankings are not modeled (likewise for negative relations). This makes sense in the multi-label setting where we are interested in finding the set of relations that are true without being concerned about their relative degrees of veracity.

Formally, we split the predefined relation set $\mathcal{R} = \mathcal{P}_{h,t} \cup \mathcal{N}_{h,t}$ into two mutually exclusive sets for each entity pair (h,t) in a training set, where $\mathcal{P}_{h,t}$ and $\mathcal{N}_{h,t}$ respectively denote the positive and negative relations of (h,t). As mentioned in Section 3.2, we make use of $\boldsymbol{f}_{h,t} \in \mathbb{R}^{|\mathcal{R}|+1}$ that is

[2]Please refer to Appendix D for the computation details.

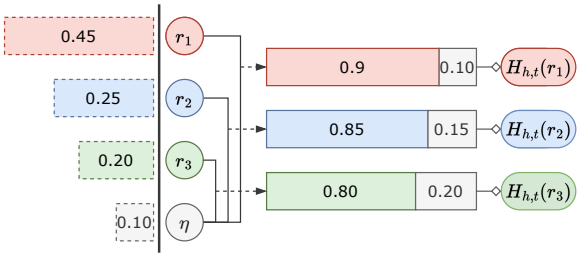

Figure 1: Each positive relation (colored rectangles with solid lines) exhibits a large probability difference from the threshold class (white rectangle with solid lines) when they are separately compared (like what is achieved with our $\mathcal{L}^{h,t}_{pmt}$ loss. We further expand this difference by minimizing $H(r)$ as stated in Eq. 5). However, the probabilities are diminished when each positive relation (colored rectangles with dashed lines) is made to compete with the other, reducing the disparity between the probability of each positive relation and that of the threshold class (white rectangle with dashed lines).

computed by Equation 1. We denote the elements of $\boldsymbol{f}_{h,t}$ that correspond to relation $r \in \mathcal{R}$ and to the NA class as f_r and f_η respectively. (Both f_r and f_η represent unnormalized prediction scores (logits).) Using f_r and f_η, we compute the probability that the label C of entity pair (h,t) is r (or η) *conditioned* on C being either r or η, i.e., $P_{h,t}(C = r|C = \{r, \texttt{NA}\})$ and $P_{h,t}(C = \eta|C = \{r, \texttt{NA}\})$ respectively, as follows.

$$P^r_{h,t}(r) = \frac{\exp(f_r)}{\exp(f_r) + \exp(f_\eta)},$$
$$P^\eta_{h,t}(r) = 1 - P^r_{h,t}(r) = \frac{\exp(f_\eta)}{\exp(f_r) + \exp(f_\eta)}, \quad (3)$$

where we have abbreviated $P_{h,t}(C = r|C = \{r, \texttt{NA}\})$ and $P_{h,t}(C = \eta|C = \{r, \texttt{NA}\})$ as $P^r_{h,t}(r)$ and $P^\eta_{h,t}(r)$ respectively.

Our *pairwise moving-threshold* loss $\mathcal{L}^{h,t}_{pmt}$ that maximizing the joint probability of all relations for an entity pair (h,t) is defined as:

$$\begin{aligned}\mathcal{L}^{h,t}_{pmt} &= -\log\Big(\prod_{r\in\mathcal{P}_{h,t}} P^r_{h,t}(r) \prod_{r\in\mathcal{N}_{h,t}} \big(1 - P^r_{h,t}(r)\big)\Big) \\ &= -\sum_{r\in\mathcal{P}_{h,t}} \log P^r_{h,t}(r) - \sum_{r\in\mathcal{N}_{h,t}} \log P^\eta_{h,t}(r) \\ &= \sum_{r\in\mathcal{P}_{h,t}} \log(1 + \exp(f_\eta - f_r)) \\ &+ \sum_{r\in\mathcal{N}_{h,t}} \log(1 + \exp(f_r - f_\eta)). \quad (4)\end{aligned}$$

正关系没有相互比较，它们的相对排名也没有建模（负关系也一样）

因为我们只寻找真实的关系集，而不关心它们的相对真实性程度

In Equation 4, note that the same threshold f_η is used for all $r \in \mathcal{R}$ for an entity pair (h, t). From the equation, we see that minimizing $\mathcal{L}^{h,t}_{pmt}$ equates to learning scores such that $f_r > f_\eta$ when r is a positive relation, and such that $f_\eta > f_r$ when r is negative relation. The (relative) scores for relations f_r and for the threshold f_η are fully learned from training data, and are tailored to individual entity pairs. Hence, they can better model the peculiarities specific to each entity pair.

虽然之前的工作ATLOP,KD-DocRE采用类似的阈值机制，但他们学习了所有关系（或一组关系）和阈值类的总顺序

Although previous work (Zhou et al., 2021; Tan et al., 2022a) employed a similar thresholding mechanism, they learn a *total order* for all relations (or a set of relations) and the threshold class. This wastes finite probability mass (total value of 1.0) in modeling the superfluous ordering among the relations that is not beneficial to multi-label problem, and inevitably diminishes the difference between the probability of each relation and that of the threshold. See Figure 1 for illustration.

这不可避免地减少了每个关系概率与阈值的概率之间的差异（因为总概率和为1）

Intuitively, a desirable trait of a loss function is that it reduces the uncertainty about whether a relation is positive or negative, thereby allowing its value to be discerned easily. To achieve this in a principled manner, we employ the principle of entropy minimization (Grandvalet and Bengio, 2005). Entropy minimization is typically used on unlabeled data in unsupervised or semi-supervised learning (Berthelot et al., 2020). In our case, we apply it on *labeled* data in a supervised setting. The information entropy for each pairwise probability distribution between relation r and the threshold class NA for entity pair (h, t) is defined as:

为了减少一个关系是正是负的不确定性，从而使其值很容易被识别，采用熵最小化的原理

让熵进入损失函数，可以让每个关系的概率与阈值的概率之间的差异变大，便于更好区分

$$H_{h,t}(r) = -P^r_{h,t}(r) \log P^r_{h,t}(r) - P^\eta_{h,t}(r) \log P^\eta_{h,t}(r). \quad (5)$$

（即：关系概率和阈值概率差距大，损失越小）

In Equation 5, information entropy decreases as the absolute difference between $P^r_{h,t}(r)$ and $P^\eta_{h,t}(r)$ increases, attaining a maximum when $P^r_{h,t}(r) = P^\eta_{h,t}(r) = 0.5$ and a minimum when either probability is 1.0 (and the other is 0.0). Thus, incorporating entropy into our loss function would help to accentuate the disparity between the pair $P^r_{h,t}(r)$ and $P^\eta_{h,t}(r)$ for all relations, making it easier to distinguish a positive (or negative) relation from the threshold NA.

We formulate our final pairwise moving-threshold loss with entropy minimization as follows:

$$\mathcal{L}^{h,t}_{em} = \frac{1}{\gamma_1} \sum_{r \in \mathcal{P}_{h,t}} H_{h,t}(r) + \frac{1}{\gamma_2} \sum_{r \in \mathcal{N}_{h,t}} H_{h,t}(r), \quad (6)$$

$$\mathcal{L}_1 = \sum_{(h,t) \in \mathcal{B}} \mathcal{L}^{h,t}_{pmt} + \mathcal{L}^{h,t}_{em}, \quad (7)$$

where $\mathcal{B}$ refers to a training batch, and $\gamma_1 = \{1, |\mathcal{P}_{h,t}|\}$ and $\gamma_2 = \{1, |\mathcal{N}_{h,t}|\}$ are hyperparameters weighting the effect of entropy minimization.

Lpmt不足：当负关系的数量远远超过正关系时(Nh,t的总和可能会压倒Ph,t的总和)

这可能导致在优化过程中，为了最小化Lpmt，需要将fη设置为远高于每个负关系r的fr值。这种情况下，模型可能会过于关注负关系，而不够关注正关系

It is noted that using L_{pmt} on its own would lead to poor optimization for positive relations, in the situation where there is a preponderance of negative relations, the sum over $\mathcal{N}_{h,t}$ in Equation 4 might overwhelm the sum over $\mathcal{P}_{h,t}$ to such an extent that pushing f_η to a large value far above that of f_r for every negative relation r in order to minimize $\mathcal{L}^{h,t}_{pmt}$ (the same issue that also affects previous work (Zhou and Lee, 2022) without being properly addressed). Instead, our entropy minimization via $\mathcal{L}^{h,t}_{em}$ in Equation 6 provides a *principled* means to "balance" the sharp disparity between the probability of r and that of η across *all* relations. Empirically, $\mathcal{L}^{h,t}_{em}$ also demonstrates its efficacy in an ablation study (see Section 4.4.)

Lem弥补了不足：公式6中Lem提供了一种有原则的方法来“平衡”所有关系r的概率fr和η的概率fη之间的巨大差异

因为关系概率和阈值概率差距大，损失越小

3.4 Supervised Contrastive Learning for Multi-Labels and Long-Tailed Relations

Rather than focusing only on sharpening the disparity of probability outputs as stated in Equation 7, we also seek to accentuate the disparities for the embeddings of entity pairs that are labeled with different relations. To do so, we take inspiration from supervised contrastive learning (Khosla et al., 2020) which aims to "pull" the embeddings of similar examples together, and "push" those of dissimilar examples apart.

However, the original supervised contrastive learning technique only deals with single-label data, and does not handle long-tail distributions. We have to introduce some novel modifications for it to work on our multi-label problem. We make use of the embedding $\boldsymbol{x}_{h,t}$ that is computed by Equation 1, and normalized it by L2 normalization before using it in the loss function below. After transplanting the loss function of supervised contrastive learning for our multi-label problem, we obtain the following loss function for an entity pair (h, t):

$$\mathcal{L}^{h,t}_{scl} = -\log \Big\{ \frac{1}{|\mathcal{S}_{h,t}|} \sum_{p \in \mathcal{S}_{h,t}} \frac{\exp(\boldsymbol{x}_{h,t} \cdot \boldsymbol{x}_p / \tau)}{\sum_{d \in \mathcal{B}, d \neq (h,t)} \exp(\boldsymbol{x}_{h,t} \cdot \boldsymbol{x}_d / \tau)} \Big\}, \quad (8)$$

In Equation 8, $\mathcal{B}$ is a batch of examples (entity pairs) including (h, t). $\mathcal{S}_{h,t} \subseteq \mathcal{B}$ is such that each

entity pair $p=(h',t')$ in $\mathcal{S}_{h,t}$ has at least one positive relation in common with (h,t), and p is termed a *positive* example of (h,t) (also $(h,t) \notin \mathcal{S}_{h,t}$). The *negative* examples of (h,t) are the remaining examples in the batch, i.e., $\mathcal{B} \setminus (\mathcal{S}_{h,t} \cup \{(h,t)\})$. The operator $\cdot$ refers to the dot product, and $\tau \in \mathbb{R}^+$ is a temperature parameter. To minimize $\mathcal{L}_{scl}^{h,t}$, we maximize the numerator in Equation 8 by learning embeddings for (h,t) and its positive examples that are close to each other (according to cosine similarity), and minimize the denominator by learning embeddings for (h,t) and its negative examples that are far apart.

公式8不适用于DocRE数据集中通常存在的长尾现

Equation 8 would work for document-level relation extraction (DocRE) if not for the long-tail phenomenon that is typically present in DocRE datasets. For example, in the datasets used for our experiments, the top 10 relations account for about 60% of entity pairs in the dataset. Thus, we often find that an entity pair (h,t) with only long-tailed positive relations does not have any other entity pair in the same batch that has that relation in common, i.e., $|\mathcal{S}_{h,t}| = 0$. This means that Equation 8 could not be applied to such entity pairs. To take such an entity pair (h,t) into account, we design the following loss term:

$$\mathcal{L}_{lt}^{h,t} = \log \sum_{d \in \mathcal{B}, d \neq (h,t)} \exp(\boldsymbol{x}_{h,t} \cdot \boldsymbol{x}_d / \tau), \quad (9)$$

in which we solely maximize the dissimilarities between the embedding of (h,t) and those of other entity pairs in the same batch $\mathcal{B}$. The final loss function for supervised contrastive learning is:

L2含义：若实体对(h, t)在bs中有正例，让其接近正例嵌入，远离负例嵌入

如果没有正例，则远离该bs中其他所有实体对的嵌入

$$\mathcal{L}_2 = \sum_{(h,t) \in \mathcal{B}_\mathcal{P}} \mathbb{I}_{\{|\mathcal{S}_{h,t}| \neq 0\}} \mathcal{L}_{scl}^{h,t} + \mathbb{I}_{\{|\mathcal{S}_{h,t}| = 0\}} \mathcal{L}_{lt}^{h,t}, \quad (10)$$

where $\mathbb{I}_{\{\}}$ is an indicator function that takes the value of 1 if the condition in $\{\}$ is satisfied, and the value of 0 otherwise. In Equation 10, $\mathcal{B}_\mathcal{P} \subseteq \mathcal{B}$ is a subset of entity pairs in a batch that is labeled with at least one relation in $\mathcal{R}$. In other words, $\mathcal{B}_\mathcal{P}$ does not contain any entity pair that is labeled with the NA class (i.e., all relations in $\mathcal{R}$ are considered negative for the entity pair), since it does not make sense to minimize the embedding distance between two entity pairs that are labeled NA, and thus have no relation in common.

Combining Equations 7 and 10, we obtain the final loss function that is used for training our model:

$$\mathcal{L} = \mathcal{L}_1 + \lambda \mathcal{L}_2, \quad (11)$$

where $\lambda \in \mathbb{R}^+$ is a hyperparameter.

3.5 Negative Label Sampling

As reported by recent works (Huang et al., 2022; Tan et al., 2022b), the DocRE benchmark suffers from the severe false-negative problem, which means that quite a few entity pairs previously labeled as NA class should have at least one relation label. Blithely ignoring this issue will greatly harm the performance of the method and cause ill-defined evaluation. To enhance the robustness of our method, we propose a novel negative label sampling strategy, which only samples a small fraction of negative relations for each entity pair with NA label when computing the loss function. We assume that the true relation labels for those false-negative examples are hard to be sampled from the massive negative relations, thus we could avoid erroneously treating the correct labels as negative relations in the loss function.

DocRE基准测试存在严重的假阴性问题

这意味着之前被标记为NA类的不少实体对应该至少有一个关系标签

在计算损失函数时，只对每个具有NA标签的实体对采样一小部分 负关系

且假设这些假阴性例子的真实关系标签不在采样的样本中，因此可以避免在损失函数中将正确的标签作为负关系

Let $\mathcal{B}_\mathcal{N} \subseteq \mathcal{B}$ denote the subset of all entity pairs that are labeled NA in a current batch $\mathcal{B}$. For each entity pair (h,t) in $\mathcal{B}_\mathcal{N}$, we uniformly sample a subset of negative relations $\mathcal{N}'_{h,t} \subseteq \mathcal{N}_{h,t}$, and define the following loss function:

$$\mathcal{L}' = \sum_{(h,t) \in \mathcal{B}_\mathcal{N}} \sum_{r \in \mathcal{N}'_{h,t}} -\log P_{h,t}^{\eta}(r) + \frac{1}{\gamma_2} \sum_{r \in \mathcal{N}'_{h,t}} H_{h,t}(r), \quad (12)$$

where $P_{h,t}^{\eta}(r)$ and $H_{h,t}(r)$ are defined in Equation 3 and Equation 5 respectively.

Let $\mathcal{B}_\mathcal{P} = \mathcal{B} \setminus \mathcal{B}_\mathcal{N}$ denote the subset of entity pairs in the current batch $\mathcal{B}$ that is labeled with at least one positive relation. Combining terms in Equations 7, 10, and 12, we obtain the final loss function $\mathcal{L}^{\text{NA}}$ that incorporates our sampling approach:

$$\mathcal{L}_1^{\text{NA}} = \mathcal{L}' + \sum_{(h,t) \in \mathcal{B}_\mathcal{P}} \mathcal{L}_{pmt}^{h,t} + \mathcal{L}_{em}^{h,t},$$
$$\mathcal{L}^{\text{NA}} = \mathcal{L}_1^{\text{NA}} + \lambda \mathcal{L}_2. \quad (13)$$

Observe that $\mathcal{L}_1^{\text{NA}}$ has modified $\mathcal{L}_1$ (Equation 7) by changing the latter's sum over negative relations for entity pairs that are labeled NA. Also note that the loss $\mathcal{L}_2$ due to supervised contrastive learning remains unchanged in Equation 13 because it operates at the level of entities (specifically their embeddings) rather than at the level of relation labels.

Although previous papers (Li et al., 2021b, 2022) seem to adopt a similar negative sampling strategy, our approach has significant differences from them. The previous works sampled negative instances

(i.e., entire entity pairs with NA labels in our case) and removed those unselected negative instances from the training dataset. In our approach, we sample negative *labels* of NA entity pairs, and do not discard any entity pairs, making our approach potentially more data efficient.

4 Experiments

4.1 Benchmark Description

DocRED (Yao et al., 2019) is a large-scale dataset constructed from Wikipedia and Wikidata, and is widely used as a benchmark for document-level relation extraction (DocRE). However, recent studies (Huang et al., 2022; Tan et al., 2022b) have found that many entity pairs (or examples) that are labeled NA are erroneous, and should be instead labeled with at least one positive relation in $\mathcal{R}$. To ameliorate this problem, Re-DocRED (Tan et al., 2022b) relabels the original training and development sets of DocRED and splits DocRED's development set into two equal halves as new development and test sets, respectively. Instead of comparing models on the faulty DocRED dataset, the results on Re-DocRED should be regarded as a fair comparison.

4.2 Two New Data Regimes

To evaluate the models in a more realistic experimental setting in which their resilience to noisy data is carefully tested, we propose two new data regimes, OOG-DocRE and OGG-DocRE, that are based on the above DocRED and Re-DocRED benchmarks. Every "O" represents the **O**riginal labels obtained from the original unclean, noisy DocRED dataset; similarly, every "G" represents the **G**old labels in the new, cleaned Re-DocRED dataset. Each letter in "OOG" and "OGG" represent different sources of labels for training, validation, and testing, respectively. Both regimes reflect the real-world scenario where training data is noisy, and manual effort can only be expended on cleaning a relatively small validation/test set. All models are trained and tuned only on the training and validation sets respectively, and evaluated on the test set. Note that in both regimes the cleaned training set from Re-DocRed is not used. Table 1 contains details about the datasets.

[3]500 documents share the same titles as the development set of Re-DocRED, but labeled by DocRED.

Dataset	Train #Doc / #Example	Dev #Doc / #Example	Test #Doc / #Example
DocRED	3,053 / 1,198,650	1,000 / 396,790	1,000 / 392,158
Re-DocRED	3,053 / 1,193,092	500 / 193,232	500 / 198,670
Our new data regimes			
OOG-DocRE	3,053 / 1,198,650	500[3] / 195,682	500 / 198,670
OGG-DocRE	3,053 / 1,198,650	500 / 193,232	500 / 198,670

Table 1: Dataset statistics. We construct two new data regimes based on the **O**riginal labels from DocRED and **G**old labels from Re-DocRED. The total number of predefined relation labels for all datasets is 96 (i.e., $|\mathcal{R}| = 96$).

4.3 Results on DocRE Benchmarks

From Table.2, we see that our PEMSCL model performs the best on both development and test sets of the original DocRED dataset and the cleaned Re-DocRED dataset (the models are trained on their corresponding training sets). It is worth noting that the results among recent models (e.g., DocuNet, KD-DocRE, NCRL) are almost indistinguishable on DocRED (see Appendix C), especially after considering their standard deviations. However, the performance gaps between models become significant when we validate and test on the Re-DocRED dataset. This strongly suggests that the original DocRED's (overly erroneous) development and test sets cannot truly ascertain the performance differences between models. In contrast, the cleaned version Re-DocRED provides a more faithful comparison of the models. Henceforth, we analyze model performances based solely on Re-DocRED's development and test sets.

Compared with ATLOP (upon which our model is developed), our PEMSCL model achieves around a 3-point improvement in terms of both Ign F_1 and F_1 scores on Re-DocRED's development and test sets. When compared against the recent strong baseline NCRL, our PEMSCL model continues to do better, achieving about a 1-point improvement in terms of F_1 score on Re-DocRED's development set. After taking the standard deviations into account, the results still show that PEMSCL outperforms NCRL. In sum, the above results demonstrate the effectiveness of our proposed model, and ascertain that it has achieved new state-of-the-art performances.

4.4 Ablation Study

In addition to the main metrics F_1 and Ign F_1, we also report the F_1 scores for different types of relations. We first rank in descending order all pre-

Model	DocRED Dev		DocRED Test	
	Ign F_1	F_1	Ign F_1	F_1
Implemented on DeBERTa$_{Large}$				
ATLOP (Zhou et al., 2021)	62.16±0.15	64.01±0.12	62.12	64.08
ATLOP + BCE (Zhou and Lee, 2022)	61.92±0.13	63.96±0.15	61.83	63.92
NCRL (Zhou and Lee, 2022)	62.98±0.18	64.79±0.13	63.03	64.96
PEMSCL (Ours)	**63.25±0.09**	**65.15±0.10**	**63.40**	**65.41**
	Re-DocRED Dev		Re-DocRED Test	
	Ign F_1	F_1	Ign F_1	F_1
Implemented on RoBERTa$_{Large}$				
JEREX (Eberts and Ulges, 2021)	69.12	70.33	68.97	70.25
ATLOP + BCE* (Zhou and Lee, 2022)	75.86±0.13	75.25±0.11	75.91	75.36
ATLOP (Zhou et al., 2021)	76.88	77.63	76.94	77.73
DocuNet (Zhang et al., 2021)	77.53	78.16	77.27	77.92
KD-DocRE (Tan et al., 2022a)	77.92	78.65	77.63	78.35
NCRL* (Zhou and Lee, 2022)	78.41±0.21	79.15±0.20	78.45	79.19
PEMSCL (Ours)	**79.02±0.20**	**79.89±0.17**	**79.01**	**79.86**

Table 2: Results on DocRED and Re-DocRED. Ign F_1 stands for the F1 score excluding relational facts in the training set. Results for baseline models on the test and dev set of DocRED are taken from their original papers. The results on RE-DocRED for NCRL and ATLOP + BCE (Zhou and Lee, 2022) (i.e., marked with *) are reproduced by us with their default code[4] and our implementation, respectively; other results of baselines on Re-DocRED are taken from (Tan et al., 2022b). We report the mean and standard deviation on the development set of 5 runs with different random initialization for our PEMSCL model and the reproduced baselines, and report the test scores using the best-performing model on the development set. For implementation details, please refer to Appendix A.

Model	Dev Ign F_1	Dev F_1	Head F_1	Mid F_1	Tail F_1
Ours	**79.02**	**79.89**	**82.99**	**75.70**	**63.51**
– $\mathcal{L}_{em}^{h,t}$	78.38	79.17	82.35	74.75	62.35
– $\mathcal{L}_2$	78.36	79.10	82.40	74.50	62.22
– $\mathcal{L}_{em}^{h,t}$ and $\mathcal{L}_2$	77.92	78.63	81.92	74.06	61.16

Table 3: Ablation study of our PEMSCL model on Re-DocRED. "–" represents the removal of our model's components. We also report the F_1 scores for the top 10 relations (Head F_1), the middle 70 relations (Mid F_1), and the last 20 relations (Tail F_1) ranked by the number of entity pairs that are related by them. The mean result of 3 runs with different random initialization on the development set of Re-DocRED are reported.

defined relations by the number of entity pairs that are labeled with them. Next, we classify them into three categories: *head* relations (the top 10 relations, accounting for 64% of Re-DocRED's training data), *tail* relations (the bottom 20 relations, accounting for 2% of training data), and *middle* relations (the remaining relations).

From Table 3, we see that each component plays a pivotal role in the effectiveness of our PEMSCL model – removing a component or a combination of them compromises performance. Removing the $\mathcal{L}_{em}^{h,t}$ and $\mathcal{L}_2$ components individually results in a performance decline of 0.90% and 0.99% in terms of F_1 score respectively. When either of these two components is removed, we see a sharper decline in terms of Tail F_1 (1.82% and 2.03%) than in terms of Head F_1 (0.77% and 0.71%). This shows that both components are useful for long-tailed relations, and highlights the effectiveness of $\mathcal{L}_2$, part of which is designed to cater to long-tailed relations. After removing both $\mathcal{L}_{em}^{h,t}$ and $\mathcal{L}_2$ together, the performances on Head F_1, Mid F_1, and Tail F_1 all significantly drop by 1.29%, 2.17%, and 3.70% respectively. Even with only one loss term remaining (i,e, $\mathcal{L}_{pmt}^{h,t}$), our EMSCL model still surpasses the baseline ATLOP on Re-DocRED in Table 2 by 1.3% on Dev F_1, reflecting the usefulness of our pairwise moving-threshold loss.

4.5 Results on New Data Regimes

Table 4 shows the performance of our PEMSCL model and baselines on our proposed data regimes: OOG-DocRE and OGG-DocRE. We select NCRL as a focal baseline from among the recent baselines

[4] https://github.com/yangzhou12/NCRL

	Orig-Dev		Gold-Dev		Gold-Test	
	Ign F_1	F_1	Ign F_1	F_1	Ign F_1	F_1
On OOG-DocRE Regime						
ATLOP (Zhou et al., 2021)	60.94	62.95	46.99	47.14	47.52	47.65
NCRL (Zhou and Lee, 2022)	61.42	63.52	49.06	49.21	48.41	48.53
PEMSCL (Ours)	**62.05**	**64.19**	50.82	50.99	50.92	51.10
PEMSCL† (Ours)	46.07	49.51	**62.05**	**63.39**	**62.76**	**64.03**
On OGG-DocRE Regime						
ATLOP (Zhou et al., 2021)	-	-	48.23	48.54	48.50	48.77
NCRL (Zhou and Lee, 2022)	-	-	49.92	50.08	50.10	50.25
PEMSCL (Ours)	-	-	50.43	50.62	51.09	51.25
PEMSCL† (Ours)	-	-	**62.40**	**63.72**	**62.47**	**63.73**

Table 4: Results on two new data regimes. The best results are **bolded**, and the second best results are underlined. PEMSCL† refers to our best-performing model on the development set after using our proposed negative label sampling strategy (Section 3.5). The sampling ratio of 0.1 is set with a development set.

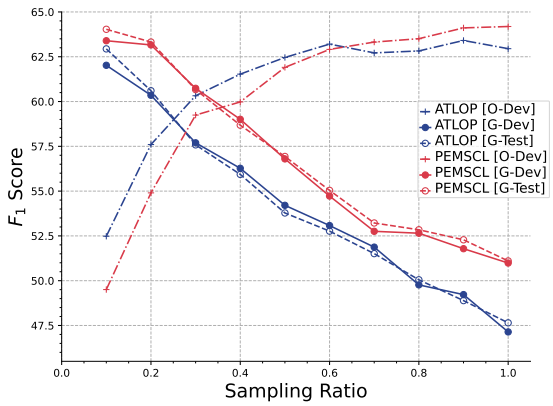

Figure 2: The effect of negative label sampling ratio in the OOG-DocRE regime.

due to its competitive performance on Re-DocRED. We see that all models perform above 62-point F_1 on Orig-Dev (the original development set from DocRED) when trained on the original training dataset. However, when we evaluate all models on Gold-Dev and Gold-Test (the clean development and test sets from Re-DocRED), the performances of the models (including ours) dramatically decrease by around 15-point F_1 on both Gold-Dev and Gold-Test. Upon inspection, we find that the models misclassify a lot of positive examples (entity pairs) as NA, which is an expected outcome of being misguided by the erroneous false-negative labels in DocRED.

However, after using our proposed negative label sampling loss (Equation 13), our PEMSCL model exhibits tremendous improvement on both Gold-Dev and Gold-Test by 24% and 25% on the F_1 scores respectively. This demonstrates the effectiveness of our negative label sampling strategy in countering the noise present in entity pairs that are labeled NA. The same conclusion can be drawn from the results for the OGG-DocRE regime. Moreover, we notice that both ATLOP and NCRL improve by at least 1-point F_1 on the OGG regime compared with the OOG regime. This demonstrates the usefulness of gold labels even in a small amount. However, our model performs comparably on both regimes, indicating the stability of our model in different regimes.

We also investigate the effect of the sampling ratio on our proposed strategy. The sampling ratio refers to the ratio of negative labels that we keep during the training for each entity pair that is labeled NA. We apply our negative label sampling approach on both ATLOP and our PEMSCL model. As seen from Figure 2, PEMSCL consistently performs better than ATLOP by a clear margin. We also find that the performances of the models on Gold-Dev and Gold-Test gradually decrease as the sampling ratio is increased. This is because as we keep more (purportedly) negative labels in our loss function, the risk of wrongly penalizing potentially true labels increases concomitantly.

We observe that the sampling ratio has the opposite effect on Orig-Dev. As the sampling ratio increases, the F_1 on Orig-Dev increases, leading one to mistakenly conclude that a large sampling rate should be used. This provides strong evidence of the poor data quality in DocRED, and shows

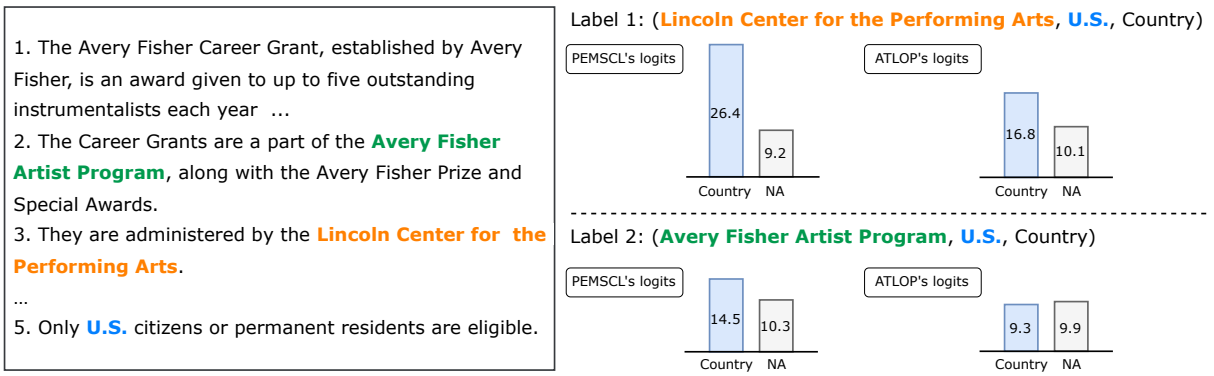

Figure 3: Case Study.

how it can misguide training and lead to poor results. The results on the OGG-DocRE regime are similar (see Appendix B for details).

4.6 Case Study

Figure 3 shows a case study of our proposed PEMSCL model and the baseline ATLOP model. We can see that for the entity pair (*Lincoln Center for the Performing Arts*, *U.S.*), both models successfully detect the correct relation label, i.e., `Country`. However, the logit difference between the `Country` relation and the threshold label `NA` in our model is much larger than that of the ATLOP model (26.4-9.2 > 16.8-10.1). This demonstrates that our model is capable of learning a more differentiated distribution of the final probability scores. For the entity pair of (*Avery Fisher Artist Program*, *U.S.*), the ATLOP model fails to correctly predict its label and classifies it as `NA` class since the logit of `Country` is lower than that of the threshold class (9.3 < 9.9). However, our model not only correctly predicts its correct label, but also maximizes the discriminability of the prediction scores (14.5 vs 10.3).

5 Conclusions

In this paper, we propose a novel method for DocRE problem called PEMSCL, which contains a pairwise moving-threshold loss with entropy minimization, adapted supervised contrastive learning, and a novel negative sampling strategy, to achieve good integration of both discriminability and robustness. Experimental results show that our method achieves new state-of-the-art results.

Acknowledgements

This research is supported by Singapore Ministry of Education's AcRF Tier 1 Grant (R-253-000-146-133) to Stanley Kok. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors, and do not reflect the views of the funding agencies.

Limitations

First, we require a large amount of GPU resources to conduct our experiments because we deal with large document-based datasets (whose input text is significantly longer than those of traditional sentence-level tasks). Second, we implement our model on two large pre-trained language models, Roberta-large (Liu et al., 2019) and Deberta-large (He et al., 2021), both of which also have a large GPU footprint. Third, the performance of our adapted supervised contrastive learning component is dependent on GPU batch size (a larger batch size allows more contrastive examples to be used to learn better embeddings).

References

David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. 2020. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. In *Proceedings of ICLR*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of ICML*.

Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs. In *Proceedings of EMNLP/IJCNLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Markus Eberts and Adrian Ulges. 2021. An end-to-end model for entity-level relation extraction using multi-instance learning. In *Proceedings of EACL*.

Yves Grandvalet and Yoshua Bengio. 2004. Semi-supervised learning by entropy minimization. In *Proceedings of Neural Information Processing Systems*.

Yves Grandvalet and Yoshua Bengio. 2005. Semi-supervised learning by entropy minimization. In *Proceedings of Neural Information Processing Systems*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of CVPR*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *Proceedings of ICLR*.

Quzhe Huang, Shibo Hao, Yuan Ye, Shengqi Zhu, Yansong Feng, and Dongyan Zhao. 2022. Does recommend-revise produce reliable annotations? an analysis on missing instances in docred. In *Proceedings of ACL*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Proceedings of NeurIPS*.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of ICLR*.

Bo Li, Wei Ye, Canming Huang, and Shikun Zhang. 2021a. Multi-view inference for relation extraction with uncertain knowledge. In *Proceedings of AAAI*.

Yangming Li, Lemao Liu, and Shuming Shi. 2021b. Empirical analysis of unlabeled entity problem in named entity recognition. In *Proceedings of ICLR*.

Yangming Li, Lemao Liu, and Shuming Shi. 2022. Rethinking negative sampling for handling missing entity annotations. In *Proceedings of ACL*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Guoshun Nan, Zhijiang Guo, Ivan Sekulic, and Wei Lu. 2020. Reasoning with latent structure refinement for document-level relation extraction. In *Proceedings of ACL*.

Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. 2017. Cross-sentence n-ary relation extraction with graph lstms. *Trans. Assoc. Comput. Linguistics*, 5:101–115.

Yujia Qin, Yankai Lin, Ryuichi Takanobu, Zhiyuan Liu, Peng Li, Heng Ji, Minlie Huang, Maosong Sun, and Jie Zhou. 2021. ERICA: improving entity and relation understanding for pre-trained language models via contrastive learning. In *Proceedings of ACL/IJCNLP*.

Chris Quirk and Hoifung Poon. 2017. Distant supervision for relation extraction beyond the sentence boundary. In *Proceedings of EACL*.

Sunil Kumar Sahu, Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Inter-sentence relation extraction with document-level graph convolutional neural network. In *Proceedings of ACL*.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022a. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of ACL*.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022b. Revisiting docred - addressing the false negative problem in relation extraction. In *Proceedings of EMNLP*.

Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. 2019. ADVENT: adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of CVPR*.

Difeng Wang, Wei Hu, Ermei Cao, and Weijian Sun. 2020. Global-to-local neural networks for document-level relation extraction. In *Proceedings of EMNLP*.

Yuxin Xiao, Zecheng Zhang, Yuning Mao, Carl Yang, and Jiawei Han. 2022. SAIS: supervising and augmenting intermediate steps for document-level relation extraction. In *Proceedings of NAACL*.

Yiqing Xie, Jiaming Shen, Sha Li, Yuning Mao, and Jiawei Han. 2022. Eider: Empowering document-level relation extraction with efficient evidence extraction and inference-stage fusion. In *Findings of ACL*.

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 AAAI*.

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 ACL*.

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 EMNLP*.

Shuang Zeng, Yuting Wu, and Baobao Chang. 2021. SIRE: separate intra- and inter-sentential reasoning for document-level relation extraction. In *Findings of ACL/IJCNLP*.

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 IJCAI*.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of AAAI*.

Yang Zhou and Wee Sun Lee. 2022. None class ranking loss for document-level relation extraction. In *Proceedings of IJCAI*.

A Implementation Details

We mainly implement our method using the pre-trained RoBERTa-large (Liu et al., 2019) as the encoder model. Due to limited computational resources, we only use a larger model DeBERTa-large (He et al., 2021) on the DocRED benchmark. We conduct grid search for the temperature parameter τ and the loss coefficient λ ({0.1, 0.2, 0.5, 1.0, 2.0}), learning rate ({1e-5, 2e-5, 3e-5}), and warmup ratio of optimizer ({0.02, 0.06, 0.10}). We implement our model in the PyTorch version of Huggingface Transformers[5], and run all experiments on a NVIDIA Quadro RTX 8000 GPU. The best hyperparameters used in our experiments are shown in Table5.

B The Effect of Sampling Ratio on the OGG- DocRE Setting

We analyze the effect of the negative label sampling ratio in the OGG-DocRE regime, which is shown in Figure 4. It presents a similar pattern with that of the OOG-DocRE regime as described in Section 4.5.

C Results on the DocRED Dataset

We provide the results of RoBERTa-large based models on DocRED in Table 6 for a complete comparison. However, these results can not reflect a faithful performance comparison due to the preponderance of erroneous labels in the DocRED dataset. Instead, the results on the Re-DocRED dataset should be taken as a reliable fair comparison.

[5] https://huggingface.co/

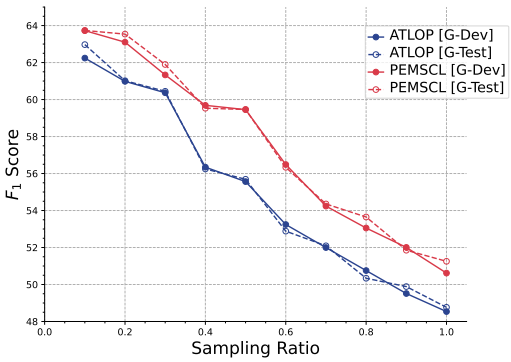

Figure 4: The effect of negative label sampling ratio in the OGG-DocRE regime.

D Background: ATLOP Encoder

For every document, the encoder model first marks each entity mention with a special token "*" at its start and end positions, and then feeds the resulting document $D = \{w_l\}_{l=1}^{L}$ into a pre-trained language model (PLM) to obtain contextual embeddings for each of the document's L tokens: $\mathbf{H} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_L] = \text{PLM}([w_1, \ldots, w_L])$ where $\boldsymbol{h}_l \in \mathbb{R}^d$. ATLOP regards the embedding of "*" at the start position of mention m_j^i as its representation, i.e., $\boldsymbol{h}_{I(m_j^i)}$, where $I(.)$ is a function mapping a mention m_j^i to the index of its representative "*" in $\mathbf{H}$. Next, the embedding $\boldsymbol{h}_{e_i} \in \mathbb{R}^d$ of each entity e_i is obtained with logsumexp pooling:

$$\boldsymbol{h}_{e_i} = \log \sum_{j=1}^{\mathcal{M}_{e_i}} \exp\left(\boldsymbol{h}_{I(m_j^i)}\right).$$

For each entity pair (e_h, e_t), ATLOP uses the token-level dependencies present within its multi-head self-attention mechanism to compute a *localized contextual* embedding $\boldsymbol{c}_{h,t} \in \mathbb{R}^d$, capturing the contextual information that is relevant to *both* entities e_h and e_t. Due to space constraints, we refer readers to Zhou et al. (2021) for details on how $\boldsymbol{c}_{h,t}$ is computed.

For each entity pair (e_h, e_t), the encoder will generate the final representation $\boldsymbol{x}_{h,t}$ for the pair, and its corresponding vector of unnormalized prediction scores $\boldsymbol{f}_{h,t} \in \mathbb{R}^{|\mathcal{R}|+1}$ for all relations in

Dataset	Batch size	Epoch	Warmup ratio	Learning rate	τ	λ	γ_1	γ_2
DocRED	4	5	0.10	2e-5	2.0	2.0	1	1
Re-DocRED	4	8	0.06	2e-5	0.2	0.1	$\lvert\mathcal{P}_{h,t}\rvert$	$\lvert\mathcal{N}_{h,t}\rvert$

Table 5: Best hyperparameters for benchmarks.

	DocRED Dev		DocRED Test	
Model	Ign F_1	F_1	Ign F_1	F_1
Implemented on RoBERTa$_{Large}$				
Coref (Ye et al., 2020)	57.35	59.43	57.90	60.25
SSAN (Xu et al., 2021)	60.25	62.08	59.47	61.42
ATLOP (Zhou et al., 2021)	61.32±0.14	63.18±0.19	61.39	63.40
DocuNet (Zhang et al., 2021)	62.23±0.12	64.12±0.14	62.39	**64.55**
KD-DocRE (Tan et al., 2022a)	62.16±0.10	64.19±0.16	**62.57**	64.28
NCRL (Zhou and Lee, 2022)	62.21±0.22	64.18±0.20	61.94	64.14
PEMSCL (Ours)	**62.31±0.19**	**64.21±0.17**	62.17	64.28

Table 6: Experimental results on the DocRED dataset.

$\mathcal{R} \cup \{\texttt{NA}\}$ as follows:

$$[\boldsymbol{z}_h^1; \ldots; \boldsymbol{z}_h^P] = \boldsymbol{z}_h = \tanh\left(\boldsymbol{W}_h \boldsymbol{h}_{e_h} + \boldsymbol{W}_{c_1} \boldsymbol{c}_{h,t}\right),$$
$$[\boldsymbol{z}_t^1; \ldots; \boldsymbol{z}_t^P] = \boldsymbol{z}_t = \tanh\left(\boldsymbol{W}_t \boldsymbol{h}_{e_t} + \boldsymbol{W}_{c_2} \boldsymbol{c}_{h,t}\right),$$
$$\boldsymbol{x}_{h,t} = ||_{p=1}^{P}(\boldsymbol{z}_h^p \otimes \boldsymbol{z}_t^p), \quad (14)$$
$$\boldsymbol{f}_{h,t} = \boldsymbol{W}_o \boldsymbol{x}_{h,t} + \boldsymbol{b}_o, \quad (15)$$

where $\boldsymbol{z}_h, \boldsymbol{z}_t \in \mathbb{R}^{d_1}$ are split into P equal-sized groups $[\boldsymbol{z}_h^1; \ldots; \boldsymbol{z}_h^P]$ and $[\boldsymbol{z}_t^1; \ldots; \boldsymbol{z}_t^P]$ respectively; $\boldsymbol{W}_{\{h,t,c_1,c_2\}} \in \mathbb{R}^{d_1 \times d}$, $\boldsymbol{W}_o \in \mathbb{R}^{(|\mathcal{R}|+1) \times d_x}$, $\boldsymbol{x}_{h,t} \in \mathbb{R}^{d_x}$ ($d_x = \frac{d_1 \times d_1}{P}$), and $\boldsymbol{b}_o \in \mathbb{R}^{|\mathcal{R}|+1}$ are *learnable* parameters (in our model too); $\otimes$ is the outer product operator; and the operators ; and $||$ respectively represent the concatenation of vectors and matrices. The elements in $\boldsymbol{f}_{h,t}$ are logits that our model feeds pairwise into (not necessarily the same) softmax functions to obtain relative probabilities between relations (Section 3.3).