面向文档级关系抽取中可判别性与鲁棒性的融合

Jia Guo[*1, 2], **Stanley Kok**[1], **Lidong Bing**[2]
[1]新加坡国立大学计算机学院 [2]阿里巴巴集团达摩院
guojia@u.nus.edu, skok@comp.nus.edu.sg
l.bing@alibaba-inc.com

arXiv:2304.00824v1 [cs.CL] 3 Apr 2023

摘要

文档级关系抽取（DocRE）旨在预测文档中依赖长程上下文依赖推理的实体对之间的关系。作为一种典型的多标签分类问题，DocRE 面临着如何有效地区分少数正关系与大多数负关系的挑战。当数据集中存在大量标注错误时，这一挑战更难克服。在这项工作中，我们旨在更好地融合 DocRE 问题中的可判别性与鲁棒性。具体而言，我们首先设计了一种有效的损失函数，使概率输出和内部表示都具有较强的可判别性。我们创新性地将熵最小化和监督式对比学习应用于具有挑战性的多标签和长尾学习问题。为减轻标签错误的影响，我们在方法中加入了一种新颖的负标签采样策略，以增强模型鲁棒性。此外，我们引入两种新数据设置，以模拟更真实的标注错误场景并评估采样策略。实验结果验证了各组成部分的有效性，并表明我们的方法在 DocRED 数据集、其近期清理版本 Re-DocRED，以及所提出的数据设置上取得了当前最佳结果。 [1]

1 引言

文档级关系抽取（DocRE）问题受到研究社区 2017；Peng 等人，2017；Yao 等人，2019）的日益关注，因为它对现实世界应用至关重要。文档级关系抽取本质上是一个多标签问题，我们必须预测文档中每个实体对的预定义标签集中包含的一组关系。因此，文档级关系抽取模型必须采用有效的学习目标，以清晰地区分大量语义相近的关系。

近期，一些研究提出了新的损失函数，以学习自适应阈值，从而更好地区分正关系和负关系。然而，这些方法（Zhou等人，2021年；Tan 等人，2022年（a））要么强制学习所有关系之间的全序，导致产生多余的比较并使它们之间的差异逐渐缩小；要么在处理标签不平衡问题时，如果正实体对的平均间隔低于阈值，就不恰当地惩罚其所有预定义标签（Zhou和 Lee，2022）。与之相反，我们提出一种学习偏序的方法，分别将所有正关系排在阈值之上，而阈值又排在所有负关系之上。我们的方法不会浪费宝贵的数据和概率质量来建模正关系之间的顺序（负关系之间同理）。此外，我们通过有原则地使用熵最小化，进一步强化每个关系分布与阈值之间的区分。

此外，上述方法都未考虑内部表示的可判别性，也未考虑模型对标注错误的鲁棒性。为解决这些问题，我们对监督式对比学习进行了创新改进（Khosla 等人，2020年），以突出不同类别实体对的嵌入之间的差异，以及同一类别实体对嵌入之间的相似性。我们的方法能够更好地适应 DocRE 数据集中的多标签设定和长尾现象。为应对 Tan 等人 (2022b) 提出的标注错误问题，我们设计了两种新数据设置和一种新颖的负标签采样策略，即使在标注不完整的情况下也能持续取得良好效果。总之，我们

*这项工作部分是在 Jia Guo 于阿里巴巴集团达摩院实习期间完成的。[1]我们的代码和数据集可在以下地址获取：https://github.com/guojiapub/PEMSCL。

贡献有三点：

- 我们提出了一种有效的损失函数，可提升内部嵌入和概率输出的可判别性。
- 我们通过引入一种新颖的负标签采样策略，实现了判别性与鲁棒性的良好结合。
- 实验结果一再表明，我们在多种设置下均取得了新的最先进性能。

2 相关工作

文档级关系抽取（DocRE）早期的 DocRE 研究主要利用图卷积网络（GCNs）（Kipf 和 Welling，2017）在文档图上进行复杂的跨句推理（Sahu 等人，2019；Christopoulou 等人，2019；Wang 等人，2020；Zeng 等人，2021）。近年来，在大型预训练语言模型（Devlin 等人，2019；Liu 等人，2019）上进行微调的方法取得了显著的性能提升。具体而言，SSAN（Xu等人，2021）将实体依赖关系编码到自注意力机制中，以增强上下文和实体推理能力。ATLOP（Zhou et al., 2021）利用多头注意力权重生成实体相关上下文表示，从而增强实体对的嵌入。为更好地解决多标签分类问题，ATLOP（Zhou et al., 2021）和 NCRL（Zhou and Lee, 2022）均提出将 NA 类视为自适应阈值。DocuNet（Zhang等人，2021）和 KD-DocRE（Tan et al., 2022a）分别通过增加实体间的交互以及引入知识蒸馏，对 ATLOP 架构进行了扩展。此外，其他 DocRE 模型还尝试利用辅助信息进行关系预测，例如元依赖路径（Nan 等人，2020）、外部知识库（Li 等人，2021a）和证据（Xie 等人，2022；Xiao 等人，2022）。我们还在第 3.3 节中详细比较了现有工作。

其他相关工作 熵最小化技术常见于半监督学习工作中（Grandvalet 和 Bengio，2004；Vu 等人，2019）。然而，我们首次在具有挑战性的多标签监督学习框架中采用熵最小化。此外，我们的熵最小化作用于关系标签的每个定制概率分布，并阈值类，这将促使它们之间形成更大的差异。

监督式对比学习（SCL）（Khosla 等人，2020年）通过根据标签构造“正”和“负”样本，将自监督对比学习（He 等人，2020年；Chen 等人，2020年）扩展到完全监督设置。ERICA（Qin 等人，2021年）提出了一种基于对比学习的预训练框架，以提升实体和关系的表示。然而，该方法根据关系的样本总量，按比例采样正样本对，这会导致有偏优化，使主要关系优于次要关系。此外，该方法每次仅最大化一对正样本的相似度，这可能会削弱聚类的整体效果。与此不同，我们平等考虑每种关系以及每个锚点的正样本，并对监督式对比学习进行了精心调整，使其同时适用于多标签问题和长尾关系学习。

3 方法

本节介绍我们的模型 PEM-SCL，该模型基于**成对**移动阈值损失、**熵最小化**和**监督式对比学习**。

3.1 问题表述

设 $D = \{w_l\}_{l=1}^{L}$ 是一个包含 L 个词和一组实体 $\mathcal{E}_D = \{e_i\}_{i=1}^{|\mathcal{E}_D|}$ 的文档。每个实体 e_i 都关联着一组提及 $\mathcal{M}_{e_i} = \{m_j^i\}_{j=1}^{|\mathcal{M}_{e_i}|}$（即指代同一实体 e_i 的一组短语）。在文档级关系抽取中，我们预测预定义集合 $\mathcal{R} = \{r_k\}_{k=1}^{|\mathcal{R}|}$ 中在每对实体（e_h，e_t）$h,\ t=1,\ \ldots\ldots,\ |\mathcal{E}_D|,\ h \neq t$ 之间成立的关系子集。为简化记法，我们有时将实体对（e_h，e_t）简写为（h，t）。如果头实体 e_h 与尾实体 e_t 的任意对应提及之间表达了某种关系，就认为这两个实体之间存在该关系。如果它们的提及之间任意一对都不存在关系，则该实体对被标记为无关系。对于每个实体对，我们将其组成实体之间成立的关系称为正关系，而将 $\mathcal{R}$ 中其余关系称为负关系。无关系的实体对不具有任何正关系，并将整个集合 $\mathcal{R}$ 视为负关系（也可以认为这类实体对之间存在一种特殊的无关系关系）。文档级关系抽取可以视为一个多标签问题，其中一个实体

对对应一个训练集/测试集样例，而 $\mathcal{R} \cup \{\text{无关系}\}$ 中的关系则对应于该样例可能的标签 或类别。

3.2 编码器模型

我们采用 ATLOP（Zhou等人，2021年）作为编码器，因为近期研究（Xie等人，2022；Zhou和Lee，2022）已证明它在神经网络架构中作为骨干网络的有效性。对于每个实体对（e_h，e_t），编码器模型会生成实体对表示 $\boldsymbol{x}_{h,t} \in \mathbb{R}^{d_x}$，以及用于关系预测的未归一化分数向量 $\boldsymbol{f}_{h,t} \in \mathbb{R}^{|\mathcal{R}|+1}$，具体如下[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 带熵最小化的成对移动阈值损失

在文档级关系抽取中，通常使用固定的概率阈值（例如，在开发集上调优的超参数）来确定正关系和负关系的分界。然而，这样的阈值只适用于平均而言的实体对，对于具有特定属性的实体对而言可能并不理想。

为解决这一问题，我们设计了一种损失函数，将 NA 类作为动态阈值，学习如何根据每个实体对所呈现的规律移动 阈值。具体而言，我们对每个关系分别将其与 NA 类进行成对比较，并促使每个正关系的预测分数高于 NA 类的预测分数，同时促使 NA 类的分数高于负关系的分数。通过这种方式，我们为每个实体对建立了相对于 $\mathcal{R} \cup \{\text{NA}\}$ 的偏序 。请注意，正关系之间不会相互比较，也不会对它们的相对排序进行建模（负关系同理）。这符合多标签设定，因为我们关心的是找出真实成立的关系集合，而不关心它们各自的真实性程度。

具体而言，对于训练集中的每个实体对（h，t），我们将预定义的关系集合$\mathcal{R} = \mathcal{P}_{h,t} \cup \mathcal{N}_{h,t}$ 划分为两个互斥集合，其中$\mathcal{P}_{h,t}$ 和 $\mathcal{N}_{h,t}$ 分别表示（h，t）的正关系和负关系。如第 3.2 节所述，我们使用$\boldsymbol{f}_{h,t} \in \mathbb{R}^{|\mathcal{R}|+1}$，它是

[2]计算细节请参阅附录D 。

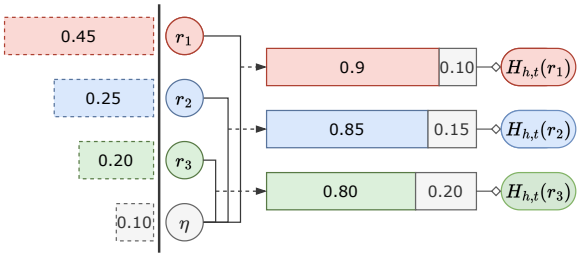


图 1：分别比较时，每个正关系（带实线边框的彩色矩形）与阈值类（带实线边框的白色矩形）之间的概率差异较大（这与我们的 $\mathcal{L}^{h,t}_{pmt}$ 损失函数所达到的效果相同）。如式 $H(r)$ 所述，我们进一步通过最小化5来扩大这一差异。然而，当每个正关系（带虚线边框的彩色矩形）相互竞争时，其概率会降低，从而缩小每个正关系与阈值类（带虚线边框的白色矩形）之间的概率差异。

根据公式 1计算得到的。我们将 $\boldsymbol{f}_{h,t}$ 中对应关系 $r \in \mathcal{R}$和 NA 类的元素分别记为 f_r 和 f_η。（f_r和 f_η 均表示未归一化预测分数（逻辑值）。）利用 f_r 和 f_η，我们计算实体对（h，t）的标签 C 为 r（或 η）的概率，条件是 C 为 r 或 η，即分别为$P_{h,t}(C = r|C = \{r, \text{NA}\})$ 和 $P_{h,t}(C = \eta|C = \{r, \text{NA}\})$，如下所示。

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

其中，我们分别将 $P_{h,t}(C = r|C = \{r, \text{无关系}\})$ 和 $P_{h,t}(C = \eta|C = \{r, \text{无关系}\})$ 简写为 $P^r_{h,t}(r)$ 和$P^\eta_{h,t}(r)$。

我们的成对移动阈值损失函数 $\mathcal{L}^{h,t}_{pmt}$ 定义为：最大化实体对（h，t）所有关系的联合概率。

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

在公式4中，注意，实体对（h，t）的所有 $r \in \mathcal{R}$ 都使用相同的阈值 f_η。从公式中可以看出，最小化 $\mathcal{L}^{h,t}_{pmt}$ 等价于学习一组分数，使得当 r 为正关系时，$f_r > f_\eta$；当 r 为负关系时，$f_\eta > f_r$。关系f_r 的（相对）分数以及阈值 f_η 完全从训练数据中学习得到，并针对各个实体对进行调整。因此，它们能够更好地建模每个实体对特有的差异。

尽管以往的工作（Zhou等人，2021年；Tan 等人，2022年（a））采用了类似的阈值机制，但它们为所有关系（或一组关系）与阈值类学习一个全序。这会在建模关系之间多余的排序时浪费有限的概率质量（总值为 1.0），而这种排序对多标签问题并无益处，并且不可避免地缩小每种关系的概率与阈值概率之间的差异。图 1对此作了说明。

直观而言，理想的损失函数应能降低对某种关系为正还是为负的不确定性，从而使其取值易于判别。为以有原则的方式实现这一点，我们采用熵最小化原理（Grandvalet 和 Bengio，2005）。熵最小化通常用于无监督或半监督学习中的无标签数据（Berthelot 等人，2020年）。在我们的场景中，我们将其应用于监督设置下的有标签数据。实体对 r 的每个关系与阈值类 无关系之间的成对概率分布，其信息熵定义为：$(h,\ t)$

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

在公式 5 中，信息熵随着 $P^r_{h,t}(r)$ 与 $P^\eta_{h,t}(r)$ 之间绝对差值的增大而减小，在 $P^r_{h,t}(r) = P^\eta_{h,t}(r) = 0.5$ 时达到最大值，并在其中任一概率为 1.0（另一个为 0.0）时达到最小值。因此，将熵纳入我们的损失函数有助于凸显所有关系中 $P^r_{h,t}(r)$ 与 $P^\eta_{h,t}(r)$ 之间的差异，从而更容易将正关系（或负关系）与阈值“无关系”区分开来。

我们将最终的成对移动阈值损失与熵最小化结合，形式如下：

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

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

其中，$\mathcal{B}$ 指训练批次，$\gamma_1 = \{1, |\mathcal{P}_{h,t}|\}$ 和 $\gamma_2 = \{1, |\mathcal{N}_{h,t}|\}$ 是用于调节熵最小化效果的超参数。

需要注意的是，单独使用 L_{pmt} 会导致正关系的优化效果不佳：当负关系占绝大多数时，公式 4 中对 $\mathcal{N}_{h,t}$ 的求和可能远远压过对$\mathcal{P}_{h,t}$ 的求和，以至于为了最小化 $\mathcal{L}^{h,t}_{pmt}$，每个负关系 r 的 f_η 都会被推至远高于 f_r 的较大值（先前的工作（Zhou and Lee, 2022）也存在同样的问题，且未得到妥善解决）。相比之下，我们通过公式 6 中的 $\mathcal{L}^{h,t}_{em}$ 进行熵最小化，为“平衡”所有关系中 r 与 η 的概率之间的显著差异提供了一种有理论依据的方法。在实证方面，$\mathcal{L}^{h,t}_{em}$ 也在消融实验中展现了其有效性（见第 4.4 节）。

3.4 面向多标签和长尾关系的监督式对比学习

我们不只关注于强化概率输出的差异（如公式7所示），还希望突出标注为不同关系的实体对嵌入之间的差异。为此，我们借鉴了监督式对比学习（Khosla 等人，2020年），其目标是将相似样本的嵌入“拉近”，并将不相似样本的嵌入“推远”。

然而，原始的监督式对比学习方法仅适用于单标签数据，无法处理长尾分布。要使其适用于我们的多标签问题，我们需要进行一些创新性修改。我们使用公式$\boldsymbol{x}_{h,\ t}$ 计算得到的嵌入，并在将其用于下文的损失函数之前进行L2归一化。将监督式对比学习的损失函数移植到我们的多标签问题后，我们得到实体对$(h,\ t)$的如下损失函数：

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

在公式 8 中，$\mathcal{B}$ 是一个示例批次（实体对），包含（h,t）。$\mathcal{S}_{h,t} \subseteq \mathcal{B}$ 满足以下条件：每个

实体对 $p=(h',t')$ 在 $\mathcal{S}_{h,t}$ 中，至少与 (h,t) 共享一个正向关系，p 被称为 (h,t) 的正样本（也就是 $(h,t) \notin \mathcal{S}_{h,t}$）。$(h,t)$ 的负样本是批次中其余的示例，即 $\mathcal{B} \setminus (\mathcal{S}_{h,t} \cup \{(h,t)\})$。运算符 $\cdot$ 表示点积，$\tau \in \mathbb{R}^+$ 是温度参数。为了最小化 $\mathcal{L}^{h,t}_{scl}$，我们通过学习彼此接近（根据余弦相似度）的 (h,t) 及其正样本的嵌入来最大化公式 8 中的分子，并通过学习彼此远离的 (h,t) 及其负样本的嵌入来最小化分母。

公式 8 适用于文档级关系抽取（DocRE），但 DocRE 数据集中通常存在长尾现象，因此例外。在我们的实验所用的数据集中，排名前 10 的关系约占数据集中实体对的 60%。因此，我们经常会发现，某个实体对 (h,t) 仅具有长尾正关系，而同一批次中没有其他实体对与该关系共有关系，即 $|\mathcal{S}_{h,t}| = 0$。这意味着公式 8无法应用于此类实体对。为考虑此类实体对 (h,t)，我们设计了以下损失项：

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

其中，我们仅最大化 (h,t) 的嵌入与同一批次中其他实体对 $\mathcal{B}$ 的嵌入之间的差异。监督式对比学习的最终损失函数为：

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

其中，$\mathbb{I}_{\{\}}$ 是一个指示函数：当 $\{\}$ 中的条件满足时取值为 1，否则取值为 0。公式 10 中，$\mathcal{B}_\mathcal{P} \subseteq \mathcal{B}$是批次中至少标注了一个关系 $\mathcal{R}$的实体对的子集。换言之，$\mathcal{B}_\mathcal{P}$不包含任何被标注为无关系 类别的实体对（即，对于该实体对，$\mathcal{R}$ 中的所有关系均被视为负关系），因为最小化两个被标注为无关系、因而没有共同关系的实体对之间的嵌入距离并无意义。

结合公式7 和10，我们得到用于训练模型的最终损失函数：

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

其中，$\lambda \in \mathbb{R}^+$ 是一个超参数。

3.5 负标签采样

近期研究报告（Huang 等人，2022；Tan 等人，2022b）指出，文档级关系抽取基准存在严重的假负例问题，这意味着此前被标记为无关系类的许多实体对实际上至少应具有一个关系标签。忽视这一问题会严重损害方法的性能，并导致评估定义不明确。为增强方法的鲁棒性，我们提出一种新颖的负标签采样策略：在计算损失函数时，仅为每个标记为无关系的实体对采样少量负关系。我们假设，这些假负例所对应的真实关系标签很难从海量负关系中被采样到，因此可以避免在损失函数中错误地将正确标签视为负关系。

令 $\mathcal{B}_\mathcal{N} \subseteq \mathcal{B}$ 表示当前批次 $\mathcal{B}$ 中所有被标记为无关系的实体对的子集。对于 $\mathcal{B}_\mathcal{N}$ 中的每个实体对（h，t），我们均匀采样一个负关系子集 $\mathcal{N}'_{h,t} \subseteq \mathcal{N}_{h,t}$，并定义如下损失函数：

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

其中$P^\eta_{h,t}$（r）和$H_{h,t}$（r）分别定义于公式 3 和公式5 中。

令 $\mathcal{B}_\mathcal{P} = \mathcal{B} \setminus \mathcal{B}_\mathcal{N}$ 表示当前批次 $\mathcal{B}$ 中至少被标记为一个正关系的实体对的子集。合并公式7、10和12中的各项，我们得到最终损失函数 $\mathcal{L}^{\texttt{NA}}$，其中融入了我们的采样方法：

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

注意，$\mathcal{L}^{\texttt{NA}}_1$通过将针对被标记为无关系的实体对的负关系求和项替换，修改了 $\mathcal{L}_1$（公式7）。此外请注意，由监督式对比学习产生的损失 $\mathcal{L}_2$ 在公式13中保持不变，因为它作用于实体层面（具体而言，是实体的嵌入），而非关系标签层面。

尽管以往的论文（Liet al.，2021b，2022）似乎采用了类似的负采样策略，但我们的方法与它们存在显著差异。以往的工作会对负例进行采样

（即在我们的研究中，标签为无关系的完整实体对），并将未选中的负例从训练数据集中移除。我们的方法对无关系实体对的标签进行负采样，而不会丢弃任何实体对，因此可能更加节省数据。

4 实验

4.1 基准描述

DocRED（Yao 等人，2019）是一个基于维基百科和维基数据构建的大规模数据集，广泛用作文档级关系抽取（DocRE）的基准。然而，近期研究（Huang 等人，2022；Tan 等人，2022b）发现，许多被标记为无关系 的实体对（或示例）存在错误，本应至少标记为一种正关系$\mathcal{R}$。为缓解这一问题，Re-DocRED（Tan 等人，2022b）重新标注了 DocRED 原始训练集和开发集，并将 DocRED 的开发集平均拆分为两部分，分别作为新的开发集和测试集。相比在有缺陷的 DocRED 数据集上比较模型，Re-DocRED 上的结果应被视为更公平的比较依据。

4.2 两种新的数据制度

为了在更贴近现实的实验设置中评估模型，并仔细检验其应对噪声数据的能力，我们基于上述 DocRED 和 Re-DocRED 基准数据集提出了两种新数据设置：OOG-DocRE 和 OGG-DocRE。每个“O”都表示**原**始标签，这些标签来自未经清理、含有噪声的原始 DocRED 数据集；同样，每个“G”都表示新近清理的 Re-DocRED 数据集中的**金**标准标签。“OOG”和“OGG”中的每个字母分别代表训练、验证和测试所用标签的不同来源。这两种设置都反映了真实场景：训练数据含有噪声，而人工精力只能用于清理规模相对较小的验证集/测试集。所有模型仅在训练集上训练、仅在验证集上调优，并在测试集上评估。请注意，这两种设置均未使用 Re-DocRED 中经过清理的训练集。表1 列出了数据集的详细信息。

[3]500 这些文档的标题与 Re-DocRED 开发集相同，但标签由 DocRED 标注。

数据集	训练集 #文档 / #示例	Dev #文档 / #示例	Test #文档 / #示例
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
我们的新数据范式			
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

表 1：数据集统计。我们基于 DocRED 的**原始**标签和 Re-DocRED 的**黄金**标签构建了两种新数据范式。所有数据集预定义关系标签的总数为 96（即 $|\mathcal{R}| = 96$）。

4.3 DocRE基准上的结果

从表2可以看出，我们的 PEMSCL 模型在原始 DocRED 数据集和清理后的 Re-DocRED 数据集的开发集和测试集上均表现最佳（模型均使用各自对应的训练集进行训练）。值得注意的是，近期模型（如 DocuNet、KD-DocRE、NCRL）在 DocRED 上的结果几乎难以区分（见附录C），尤其是在考虑其标准差之后。然而，在 Re-DocRED 数据集上进行验证和测试时，模型间的性能差异变得显著。这有力地表明，原始 DocRED（错误过多）的开发集和测试集无法真正反映模型之间的性能差异。相比之下，清理后的 Re-DocRED 能够更忠实地比较各模型。下文将仅基于 Re-DocRED 的开发集和测试集分析模型性能。

与 ATLOP（我们的模型以其为基础开发）相比，我们的 PEMSCL 模型在 Re-DocRED 的开发集和测试集上，Ign F_1 和F_1 分数均提高了约 3 个百分点。与近期表现强劲的基线 NCRL 相比，我们的 PEMSCL 模型仍然表现更好，在 Re-DocRED 开发集上的 F_1 分数提高了约 1 个百分点。考虑到标准差后，结果仍表明 PEMSCL 优于 NCRL。总之，上述结果证明了我们所提出模型的有效性，并确认其已取得新的最先进性能。

4.4 消融实验

除主要指标 F_1 和 Ign F_1外，我们还报告了不同类型关系的 F_1 得分。我们首先按降序排列所有预-

模型	DocRED开发集		DocRED测试集	
	忽略 F_1	F_1	忽略 F_1	F_1
基于 *DeBERTa* 实现*Large*				
ATLOP(Zhou等人，2021年)	62.16±0.15	64.01±0.12	62.12	64.08
ATLOP +二元交叉熵 (Zhou和Lee，2022)	61.92±0.13	63.96±0.15	61.83	63.92
NCRL (Zhou和Lee,2022)	62.98±0.18	64.79±0.13	63.03	64.96
PEMSCL（我们的方法）	**63.25±0.09**	**65.15±0.10**	**63.40**	**65.41**
	Re-DocRED 开发集		Re-DocRED 测试集	
	忽略 F_1	F_1	忽略 F_1	F_1
基于 *RoBERTa* 实现*Large*				
JEREX（Eberts 和 Ulges，2021年）	69.12	70.33	68.97	70.25
ATLOP +二元交叉熵* (Zhou和Lee, 2022)	75.86±0.13	75.25±0.11	75.91	75.36
ATLOP(Zhou等人,2021年)	76.88	77.63	76.94	77.73
DocuNet (Zhang等人, 2021年)	77.53	78.16	77.27	77.92
KD-DocRE (Tan 等人,2022年（a）)	77.92	78.65	77.63	78.35
NCRL* （Zhou和Lee，2022）	78.41±0.21	79.15±0.20	78.45	79.19
PEMSCL（我们的方法）	**79.02±0.20**	**79.89±0.17**	**79.01**	**79.86**

表 2：DocRED 和 Re-DocRED 上的结果。Ign F_1 表示不包括训练集中的关系事实时的 F1分数。DocRED 测试集和开发集上基线模型的结果取自其原始论文。NCRL 和 ATLOP + BCE（Zhou和Lee，2022）（即以 *标记的结果）由我们分别使用其默认代码[4] 和我们的实现复现；Re-DocRED 上其他基线的结果取自（Tan 等人，2022b）。对于我们的 PEMSCL 模型和复现的基线，我们报告使用不同随机初始化运行 5 次后在开发集上的均值和标准差；测试分数则采用在开发集上表现最佳的模型。实现细节请参阅附录A。

模型	开发集 Ign F_1	开发集 F_1	头部 F_1	中部 F_1	尾部 F_1
Ours	**79.02**	**79.89**	**82.99**	**75.70**	**63.51**
– $\mathcal{L}^{h}_{em}$	78.38	79.17	82.35	74.75	62.35
– $\mathcal{L}_2$	78.36	79.10	82.40	74.50	62.22
– $\mathcal{L}^{h}_{em}$ 和 $\mathcal{L}_2$	77.92	78.63	81.92	74.06	61.16

表 3：我们在 Re-DocRED 上对 PEMSCL 模型进行的消融实验。“–”表示移除模型中的相应组件。我们还报告了按与其相关的实体对数量排序的前 10 个关系（头部 F_1）、中间 70 个关系（中部 F_1）和最后 20 个关系（尾部 F_1）的 F_1 得分。报告的是在 Re-DocRED 的开发集上使用不同随机初始化运行 3 次所得的平均结果。

按当时被标记为这些关系的实体对数量定义关系。接下来，我们将其分为三类：头部关系（前 10 个关系，占 Re-DocRED 训练集的 64%）、尾部关系（后 20 个关系，占训练数据的 2%）以及中部关系（其余关系）。

从表3可以看出，每个组件对我们的 PEMSCL 模型的有效性都起着关键作用——移除某个组件或其组合都会损害模型性能。移除 $\mathcal{L}^{h}_{em}$ 和 $\mathcal{L}_2$ 组件时，F_1 得分分别下降了 0.90% 和 0.99%。移除这两个组件中的任意一个时，Tail F_1 的降幅（1.82% 和 2.03%）比 Head F_1（0.77% 和 0.71%）更明显。这表明这两个组件对长尾关系均有帮助，也凸显了 $\mathcal{L}_2$ 的有效性，其中一部分专为长尾关系而设计。同时移除 $\mathcal{L}^{h}_{em}$ 和 $\mathcal{L}_2$ 后，Head F_1、Mid F_1 和 Tail F_1的性能均大幅下降，分别降低了 1.29%、2.17% 和 3.70%。即使只保留一个损失项（即 $\mathcal{L}^{h}_{pmt}$），我们的 EMSCL 模型在表2的 Re-DocRED 开发集上仍比基线 ATLOP 高出 1.3% F_1，这体现了我们提出的成对移动阈值损失的实用性。

4.5 新数据范式上的结果

表4 展示了我们的 PEMSCL 模型及其基线模型在我们提出的数据机制 OOG-DocRE 和 OGG-DocRE 上的性能。考虑到 NCRL 在 Re-DocRED 上具有竞争力的表现，我们从近期的基线模型中选取 NCRL 作为重点基线

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

	Orig-Dev		Gold-Dev		Gold-Test	
	忽略 F_1	F_1	忽略 F_1	F_1	忽略 F_1	F_1
关于 *OOG-DocRE* 设置						
ATLOP（Zhou等人，2021年）	60.94	62.95	46.99	47.14	47.52	47.65
NCRL（Zhou和Lee，2022）	61.42	63.52	49.06	49.21	48.41	48.53
PEMSCL（本文方法）	**62.05**	**64.19**	50.82	50.99	50.92	51.10
PEMSCL† （本文方法）	46.07	49.51	**62.05**	**63.39**	**62.76**	**64.03**
在 *OGG-DocRE* 设置下						
ATLOP（Zhou等人，2021年）	-	-	48.23	48.54	48.50	48.77
NCRL（Zhou和Lee，2022）	-	-	49.92	50.08	50.10	50.25
PEMSCL（我们的方法）	-	-	50.43	50.62	51.09	51.25
PEMSCL† （我们的方法）	-	-	**62.40**	**63.72**	**62.47**	**63.73**

表 4：两种新数据设置下的结果。最佳结果以**粗体**标出，次佳结果加下划线。PEMSCL† 表示采用我们提出的负标签采样策略（第3.5节）后，在开发集上表现最佳的模型。采样比例 0.1 是在开发集上设定的。

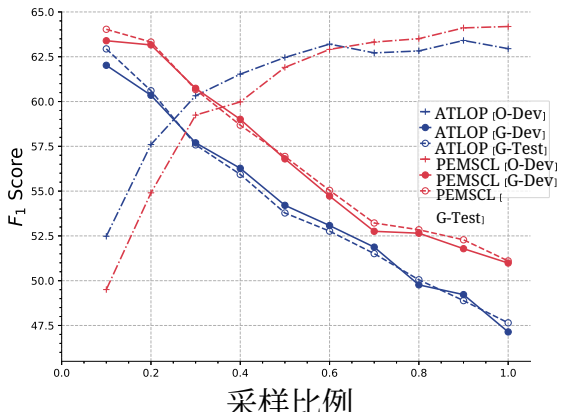


图 2：OOG-DocRE 设置中负标签采样比例的影响。

这是因为它在 Re-DocRED 上表现出色。我们看到，使用原始训练数据集进行训练时，所有模型在 Orig-Dev（DocRED 原始开发集）上的表现都超过 62 分 F_1。然而，当我们在 Gold-Dev 和 Gold-Test（Re-DocRED 的干净开发集和测试集）上评估所有模型时，包括我们的方法在内的模型性能在 Gold-Dev 和 Gold-Test 上都大幅下降了约 15 分 F_1。经检查，我们发现模型将大量正样本（实体对）错误分类为无关系，这正是 DocRED 中错误的假阴性标签造成误导后预期会出现的结果。

然而，采用我们提出的负标签采样损失（公式 13）后，我们的 PEMSCL 模型在 Gold-Dev 和 Gold-Test 上的 F_1分数分别提高了 24% 和 25%。这表明，我们的负标签采样策略能够有效应对实体对被标记为无关系时所存在的噪声。从 OGG-DocRE 设置的结果中也可以得出相同的结论。此外，与 OOG 设置相比，ATLOP 和 NCRL 在 OGG 设置上的表现均至少提高了 1 个百分点 F_1。这表明，即使金标准标签数量很少，它们仍然有用。然而，我们的模型在两种设置下的表现相当，说明模型在不同设置下具有稳定性。

我们还研究了采样比例对所提出策略的影响。采样比例指的是在训练过程中，对于每个被标注为无关系的实体对，我们保留的负标签所占的比例。我们将负标签采样方法应用于ATLOP和我们的 PEMSCL 模型。如图 2 所示，PEMSCL 的表现始终明显优于 ATLOP。我们还发现，随着采样比例提高，模型在 Gold-Dev 和 Gold-Test 上的表现逐渐下降。这是因为在损失函数中保留更多（据称为）负标签，会相应增加错误惩罚潜在真实标签的风险。

我们观察到，采样比例对 Orig-Dev 的影响恰好相反。随着采样比例增加，Orig-Dev 上的 F_1 也随之增加，这可能会让人误以为应采用较大的采样率。这有力地证明了 DocRED 的数据质量较差，并表明

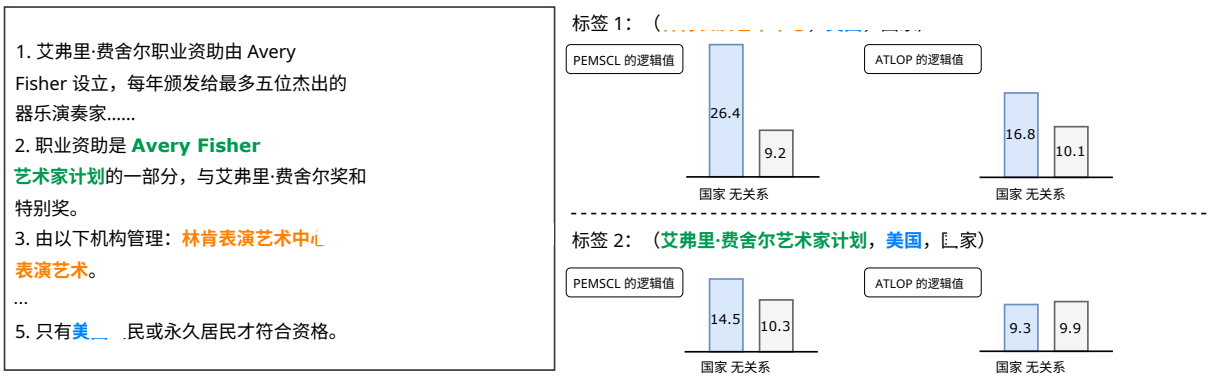


图 3：案例分析。

它会如何误导训练并导致较差的结果。OGG-DocRE 方案下的结果也类似（详见附录B ）。

4.6 案例分析

图 3展示了我们提出的 PEMSCL 模型与基线 ATLOP 模型的一个案例分析。可以看到，对于实体对（林肯表演艺术中心，美国），两个模型都成功检测出了正确的关系标签，即国家。然而，在我们的模型中，国家关系与阈值标签无关系之间的逻辑值差异远大于 ATLOP 模型（26.4-9.2 > 16.8-10.1）。这表明，我们的模型能够学习到最终预测分数区分度更高的分布。对于实体对（艾弗里·费舍尔艺术家计划，美国），ATLOP 模型未能正确预测其标签，而是将其归为无关系类，因为国家的逻辑值低于阈值类（9.3 < 9.9）。然而，我们的模型不仅正确预测了标签，还最大限度地提高了预测分数的可判别性（14.5 对比 10.3）。

5 结论

本文提出了一种用于文档级关系抽取问题的新方法 PEMSCL。该方法包含结合熵最小化的成对移动阈值损失、改进的监督式对比学习，以及一种新颖的负采样策略，旨在兼顾可判别性与鲁棒性。实验结果表明，我们的方法取得了新的当前最佳结果。

致谢

本研究得到新加坡教育部提供给 Stanley Kok 的 AcRF Tier 1 资助（R-253-000-146-133）的支持。本材料中表达的任何观点、研究发现、结论或建议均为作者所持，不代表资助机构的观点。

局限性

首先，由于我们处理的是大型基于文档的数据集（其输入文本明显长于传统句子级任务中的文本），因此开展实验需要大量 GPU资源。其次，我们在两个大型预训练语言模型 RoBERTa-large（Liu 等人，2019）和 DeBERTa-large（He 等人，2021年）上实现了我们的模型，而这两个模型同样需要大量 GPU资源。第三，我们改进后的监督式对比学习组件的性能取决于 GPU批量大小（更大的批次大小可以使用更多对比样本，从而学习到更好的嵌入）。

参考文献

David Berthelot、Nicholas Carlini、Ekin D. Cubuk、Alex Kurakin、Kihyuk Sohn、Han Zhang 和 Colin Raffel。2020年。Remixmatch：结合分布对齐与增强锚定的半监督学习。载于*ICLR* 论文集。Ting Chen、Simon Kornblith、Mohammad Norouzi 和 Geoffrey E. Hinton。2020年。一种用于视觉表征对比学习的简单框架。载于国际机器学习会议论文集。Fenia Christopoulou、Makoto Miwa 和 Sophia Ananiadou。2019。连点成线：利用边导向图进行文档级神经关系抽取。载于*EMNLP/IJCNLP* 论文集。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019。BERT：用于语言理解的深层双向 Transformer 预训练。载于*NAACL-HLT* 论文集。

Markus Eberts 和 Adrian Ulges。2021年。一种利用多实例学习进行实体级关系抽取的端到端模型。载于*EACL* 论文集。

Yves Grandvalet 和 Yoshua Bengio。2004。通过熵最小化进行半监督学习。载于神经信息处理系统论文集。

Yves Grandvalet 和 Yoshua Bengio。2005。通过熵最小化进行半监督学习。载于神经信息处理系统论文集。

何恺明、Haoqi Fan、Yuxin Wu、Saining Xie 和 Ross B. Girshick。2020年。用于无监督视觉表征学习的动量对比。载于*CVPR* 论文集。

何鹏程、刘晓东、高建峰和陈卫祖。2021年。DeBERTa：采用解耦注意力的解码增强型 BERT。载于*ICLR* 论文集。

Quzhe Huang、Shibo Hao、Yuan Ye、Shengqi Zhu、冯岩松和赵东岩。2022年。推荐-修订能否产生可靠标注？对 DocRED 中缺失实例的分析。收录于*ACL* 论文集。Prannay Khosla、Piotr Teterwak、Chen Wang、Aaron Sarna、Yonglong Tian、Phillip Isola、Aaron Maschinot、Ce Liu 和 Dilip Krishnan。2020年。监督式对比学习。收录于*NeurIPS* 论文集。Thomas N. Kipf 和 Max Welling。2017。基于图卷积网络的半监督分类。收录于*ICLR* 论文集 Bo Li、Wei Ye、Canming Huang 和 Shikun Zhang。2021a。利用不确定知识进行关系抽取的多视图推理。收录于*AAAI* 论文集。Yangming Li、Lemao Liu 和 Shuming Shi。2021b。命名实体识别中未标注实体问题的实证分析。收录于*ICLR* 论文集。Yangming Li、Lemao Liu 和 Shuming Shi。2022。重新思考负采样，以处理缺失的实体标注。收录于*ACL* 论文集。Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Mandar Joshi、陈丹琦、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。2019。RoBERTa：一种经过稳健优化的 BERT 预训练方法。*arXiv* 预印本 *arXiv:1907.11692*。Guoshun Nan、Zhijiang Guo、Ivan Sekulic 和 Wei Lu。2020年。通过潜在结构精炼进行文档级关系抽取推理。收录于*ACL* 论文集。

Nanyun Peng、Hoifung Poon、Chris Quirk、Kristina Toutanova 和 Wen-tau Yih。2017。利用图长短期记忆网络进行跨句子 n 元关系抽取。计算语言学协会汇刊，5：101–115。

Yujia Qin、Yankai Lin、Ryuichi Takanobu、Zhiyuan Liu、Peng Li、Heng Ji、Minlie Huang、Maosong Sun 和 Jie Zhou。2021年。ERICA：通过对比学习提升预训练语言模型的实体与关系理解能力。载于*ACL/IJCNLP* 论文集。

Chris Quirk 和 Hoifung Poon。2017。超越句子边界的关系抽取远程监督。载于*EACL* 论文集。

Sunil Kumar Sahu、Fenia Christopoulou、Makoto Miwa 和 Sophia Ananiadou。2019。基于文档级图卷积神经网络的跨句关系抽取。载于*ACL* 论文集。

Qingyu Tan、Ruidan He、Lidong Bing 和 Hwee Tou Ng。2022年（a）。基于自适应焦点损失和知识蒸馏的文档级关系抽取。载于*ACL* 研究成果。

Qingyu Tan、Lu Xu、Lidong Bing、Hwee Tou Ng 和 Sharifah Mahani Aljunied。2022b。重新审视 DocRED：解决关系抽取中的假阴性问题。收录于*EMNLP* 论文集。Tuan-Hung Vu、Himalaya Jain、Maxime Bucher、Matthieu Cord 和 Patrick Pérez。2019。ADVENT：用于语义分割领域自适应的对抗式熵最小化。收录于*CVPR* 论文集。Difeng Wang、Wei Hu、Ermei Cao 和 Weijian Sun。2020年。用于文档级关系抽取的全局到局部神经网络。收录于*EMNLP* 论文集。Yuxin Xiao、Zecheng Zhang、Yuning Mao、Carl Yang 和 Jiawei Han。2022。SAIS：监督并增强文档级关系抽取的中间步骤。收录于*NAACL* 论文集。Yiqing Xie、Jiaming Shen、Sha Li、Yuning Mao 和 Jiawei Han。2022。Eider：通过高效证据抽取和推理阶段融合赋能文档级关系抽取。收录于 *ACL* 研究成果。Benfeng Xu、Quan Wang、Yajuan Lyu、Yong Zhu 和 Zhendong Mao。2021年。实体结构的内部与整体：对文档级关系抽取中的提及依赖关系进行建模。收录于*AAAI* 论文集。Yuan Yao、Deming Ye、Peng Li、Xu Han、Yankai Lin、Zhenghao Liu、Zhiyuan Liu、Lixin Huang、Jie Zhou 和 Maosong Sun。2019。DocRED：大规模文档级关系抽取数据集。收录于 *ACL* 论文集。

Deming Ye、Yankai Lin、杜佳举、Zhenghao Liu、Peng Li、Maosong Sun 和 Zhiyuan Liu。2020年。用于语言表示的共指推理学习。收录于 *EMNLP* 论文集。

曾爽、吴雨婷和常宝宝。2021年。SIRE：将文档级关系抽取的句内推理与句间推理分离。载于 *ACL/IJCNLP* 研究成果。

Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si 和 HuajunChen。2021年。将文档级关系抽取视为语义分割。载于国际人工智能联合会议论文集。

Wenxuan Zhou, Kevin Huang, Tengyu Ma 和 Jing Huang。2021年。采用自适应阈值和局部上下文池化的文档级关系抽取。载于*AAAI* 论文集。

Yang Zhou 和 Wee Sun Lee。2022。用于文档级关系抽取的无关系类别排序损失。载于国际人工智能联合会议论文集。

A 实现细节

我们主要使用预训练的 RoBERTa-large（Liu 等人，2019）作为编码器模型。由于计算资源有限，我们仅在 DocRED 基准上使用更大的 DeBERTa-large（He 等人，2021年）。我们对温度参数 τ 和损失系数 λ （{0.1, 0.2, 0.5, 1.0, 2.0}）、学习率（{1e-5, 2e-5, 3e-5}）以及优化器的预热比例（{0.02, 0.06, 0.10}）进行网格搜索。我们使用 Huggingface Transformers 的 PyTorch 版本实现模型[5]，，并在 NVIDIA Quadro RTX 8000 GPU 上运行所有实验。实验中使用的最佳超参数见表5。

B 采样比例对 OGG-DocRE 设置的影响

我们分析负标签采样比例对 OGG-DocRE 设置的影响，如图 4 所示。其呈现的模式与第 4.5 节所述的 OOG-DocRE 设置相似。

C DocRED 数据集上的结果

我们在表 6 中提供了基于 RoBERTa-large 的模型在 DocRED 上的结果，以便进行完整比较。然而，这些结果无法反映出

[5]https://huggingface.co/

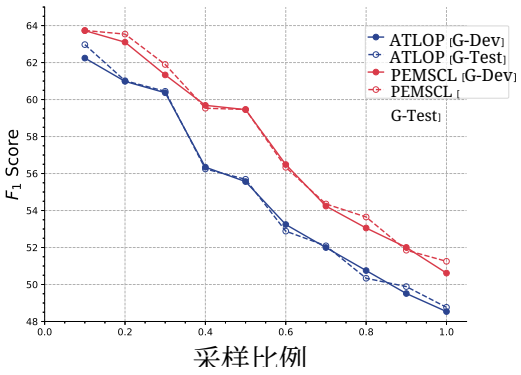


图 4：负标签采样比例对 OGG-DocRE 设置的影响。

由于 DocRED 数据集中错误标签占比过高，无法据此进行忠实的性能比较。相比之下，Re-DocRED 数据集上的结果应被视为可靠、公平的比较依据。

D 背景：ATLOP 编码器

对于每篇文档，编码器模型首先在每个实体提及的起始位置和结束位置分别标记特殊标记“*”，然后将得到的文档 $D = \{w_l\}_{l=1}^{L}$ 输入预训练语言模型（PLM），以获得文档中每个 L 词元的上下文嵌入：$\mathbf{H} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_L] = \text{PLM}([w_1, \ldots, w_L])$，其中$\boldsymbol{h}_l \in \mathbb{R}^d$。ATLOP将提及 m_j^i 起始位置处“*”的嵌入作为其表示，即 $\boldsymbol{h}_{I(m_j^i)}$，其中 $I(.)$ 是一个将提及 m_j^i 映射到 $\mathbf{H}$ 中其代表性 “*” 索引的函数。接下来，通过 logsumexp 池化获得每个实体 e_i 的嵌入 $\boldsymbol{h}_{e_i} \in \mathbb{R}^d$：

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

对于每个实体对（e_h，e_t），ATLOP利用其多头自注意力机制中存在的词元级依赖关系，计算局部化上下文嵌入 $\boldsymbol{c}_{h,\,t} \in \mathbb{R}^d$，以捕获与两个实体 e_h 和 e_t相关的上下文信息。由于篇幅所限，关于如何计算$\boldsymbol{c}_{h,\,t}$ 的详细信息，请参阅Zhou等人 （2021年）。

对于每个实体对（e_h，e_t），编码器将为该实体对生成最终表示 $\boldsymbol{x}_{h,\,t}$，以及其对应的所有关系的未归一化预测分数向量 $\boldsymbol{f}_{h,\,t} \in \mathbb{R}^{|\mathcal{R}|+1}$，包括

数据集	批次大小	训练轮次	预热比例	学习率	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	$\lvert\mathcal{P}_{h,t}\rvert$	$\lvert\mathcal{N}_{h,t}\rvert$

表5：基准数据集的最佳超参数。

	DocRED开发集		DocRED测试集	
模型	忽略 F_1	F_1	忽略 F_1	F_1
基于 *RoBERTa* 实现*Large*				
共指消解 (Ye等人，2020年)	57.35	59.43	57.90	60.25
SSAN(Xu等人，2021年)	60.25	62.08	59.47	61.42
ATLOP（Zhou等人，2021年）	61.32±0.14	63.18±0.19	61.39	63.40
DocuNet（Zhang等人，2021年）	62.23±0.12	64.12±0.14	62.39	**64.55**
KD-DocRE(Tan 等人，2022年（a）)	62.16±0.10	64.19±0.16	**62.57**	64.28
NCRL(Zhou和Lee，2022)	62.21±0.22	64.18±0.20	61.94	64.14
PEMSCL（我们的方法）	**62.31±0.19**	**64.21±0.17**	62.17	64.28

表 6：DocRED 数据集上的实验结果。

$\mathcal{R} \cup \{\text{无关系}\}$ 如下：

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

其中，$\boldsymbol{z}_h$、$\boldsymbol{z}_t \in \mathbb{R}^{d_1}$ 被划分为 P 个大小相等的组 $[\boldsymbol{z}_h^1;\ \ldots\ldots;\ \boldsymbol{z}_h^P]$ 和 $[\boldsymbol{z}_t^1;\ \ldots\ldots;\ \boldsymbol{z}_t^P]$ 分别如此；$\boldsymbol{W}_{\{h,t,c_1,c_2\}} \in \mathbb{R}^{d_1 \times d}$、$\boldsymbol{W}_o \in \mathbb{R}^{(|\mathcal{R}|+1) \times d_x}$、$\boldsymbol{x}_h$、$_t \in \mathbb{R}^{d_x}$（$d_x = \frac{d_1 \times d_1}{P}$）和 $\boldsymbol{b}_o \in \mathbb{R}^{|\mathcal{R}|+1}$ 是可学习参数（我们的模型中也是如此）；$\otimes$ 是外积运算符；运算符；和 $||$ 分别表示向量和矩阵的拼接。$\boldsymbol{f}_h$、$_t$ 中的元素是逻辑值，我们的模型将它们成对输入（不一定相同的）softmax 函数，以获得关系之间的相对概率（第 3.3 节）。