

基于不完全标注的自监督多标签文本分类

任俊飞, 朱桐, 陈文亮*
苏州大学计算机科学与技术学院
江苏, 苏州, 2150062

{jfrenjfren,tzhu7}@stu.suda.edu.cn, wlchen@suda.edu.cn

摘要

多标签文本分类(Multi-Label Text Classification, MLTC)旨在从预定义的候选标签集合中选择一个或多个文本对应的类别, 是自然语言处理(Natural Language Processing, NLP)的一项基本任务。前人工作大多基于规范且全面的标注数据集, 而这些规范数据集需要严格的质量控制, 一般很难获取。在真实的标注过程中, 难免会丢失掉一些相关标签, 进而导致不完全标注问题。为此本文提出了一种基于局部标注的自监督框架(Partial Self-Training, PST), 该框架利用教师模型自动地给大规模无标注数据打伪标签, 同时给不完全标注数据补充缺失标签, 最后再利用这些数据反向更新教师模型。在合成数据集和真实数据集上的实验表明, 本文提出的PST框架兼容现有的各类多标签文本分类模型, 并且可以缓解不完全标注数据对模型的影响。

关键词: 多标签文本分类; 不完全标注; 自监督学习

Self-Training With Incomplete Labeling For Multi-Label Text Classification

REN Junfei, ZHU Tong, CHEN Wenliang*
School of Computer Science and Technology, Soochow University
Suzhou, Jiangsu, 215006, China
{jfrenjfren,tzhu7}@stu.suda.edu.cn, wlchen@suda.edu.cn

Abstract

Multi Label Text Classification (MLTC) is a fundamental task of Natural Language Processing (NLP). It selects the most relevant labels from the predefined label set to annotate texts. Most of the previous studies are conducted on standardized and comprehensive datasets with manual annotations, which require strict quality control and are difficult to obtain. In the real annotation process, it is inevitable to lose some related labels, which leads to the problem of incomplete annotation. We propose a Partial Self-Training (PST) framework to address this problem. The teacher model not only generates pseudo labels on large-scale unlabeled data, but also provides supplement tags to incompletely labeled data. Finally, the teacher model is updated iteratively based on these data. Experiments on synthetic data sets and real data sets show that our proposed PST framework is compatible to different kinds of teacher models, and can alleviate the impact of incomplete labeled data.

Keywords: Multi-Label Text Classification, Incomplete Labeling, Self-Training

* 通讯作者 Corresponding Author.

1 引言

多标签文本分类作为自然语言处理中一项基本且实用的任务，可以自动地标注与文本相关的标签，在情感分析(Li et al., 2016)、话题识别(Dougrez-Lewis et al., 2021)、问答(Langton et al., 2020)和网页标记(Jain et al., 2016)等许多领域都有应用。然而，由于标注体系的复杂性，标注过程中可能存在标签缺失的情况，从而形成不完全标注的数据集。如表(1)所示，给定一段金融领域文本，标注员在标注过程中只标注了“重大赔付”和“财务造假”两个相关标签，而遗漏了“破产清算”标签，这条标注数据就是不完全标注数据。这种不完全标注问题在多标签分类数据集中尤为明显，会导致多标签分类模型无法准确地预测出相关标签。这种缺失标签对模型的影响主要分为两方面：1)退化影响：大量缺失标签的存在导致与文本相关的正例标签数量减少，模型在少量相关标签的训练下无法学到更加全面完整的信息；2)误导影响：大量缺失标签在模型训练过程中被当作与文本不相关的负例标签计算，从而误导模型学习到相反的信息。面向不完全标注的多标签文本分类旨在从不完全标注数据集中学习文本到相关标签的分类器，同时尽量缓解缺失标签对模型的影响，提升多标签分类的性能。

文本	标签
涉案的美国三大投行遭到重罚,花旗集团和摩根大通因涉嫌财务欺诈被判有罪, 向安然公司的破产受害者分别支付了20亿、22亿和6900万美元的赔偿罚款。	重大赔付 财务造假 破产清算

Table 1: 不完全标注样例，“破产清算”为缺失标签

现有针对多标签文本分类的方法主要集中在四个方面，分别是文本语义表示的研究、标签间关系的研究、标签分布的研究以及文本与标签语义链接的研究。文本语义表示的研究侧重于使用深度神经网络来提取出文本的深层语义表示(Liu et al., 2017)。标签间关系的研究通常利用标签式注意力机制(肖琳 et al., 2020)来建模标签相关性。标签分布的研究通过设计特殊的损失函数和数据采样策略来缓解样本标签分布不均衡等问题。还有一些研究(Du et al., 2019; Pappas and Henderson, 2019)通过对文本与标签联合建模来探索文本与标签的语义链接。然而，这些研究都是在人工标注的数据进行监督训练，无法解决不完全标注的标签缺失问题。

为此本文提出了一种基于局部标注的自监督框架(Partial Self-Training, PST)，该框架通过补充利用缺失标签来缓解缺失标签对模型造成的负面影响。具体地，PST框架首先利用基础的多标签文本分类模型在不完全标注数据集上训练以获取教师模型，然后利用教师模型自动地给大规模无标注数据和不完全标注数据打分，接着利用双阈值机制对标签按得分进行状态划分以获取正例标签、负例标签以及其它标签。最后通过联合训练充分利用三种不同状态的标签信息对教师模型进行更新。总的来说，PST框架通过对缺失标签的补充利用，可以从两个方面缓解缺失标签对模型的影响：一方面带有伪标签的无标注数据与补充了缺失标签的不完全标注数据大大增加了模型训练的正例标签，进而大大缓解缺失标签带来的退化影响；另一方面随着对缺失标签的补充，伪负例标签也会相应减少，进而缓解了缺失标签带来的误导影响。

为了更加全面地评估PST框架的性能，本文分别在合成数据集和真实数据集上进行实验。实验结果表明，随着不完全标注问题的加剧，多标签文本分类模型的性能急剧下降，而PST框架可以在一定程度上缓解下降的速度，缺失标签越多缓解越明显。同时从不同多标签分类教师模型上的实验结果发现，在不完全标注的数据集上，PST框架对不同的教师模型都有着不同程度的改善，充分证明PST框架的通用性。总体来说，本文贡献有以下三点：

- 本文提出了一种兼容现有多标签文本分类模型的自监督学习框架PST，该框架通过补充利用缺失标签来缓解不完全标注数据集对多标签文本分类模型造成的负面影响；
- 本文对CCKS2022 Task8面向金融领域的Few-Shot事件主体抽取的评测任务数据集进行修正，构建了一个新的真实数据集CCKS-IMLTC，该数据能够更好地模拟真实标注场景下的不完全标注问题。该数据集及本文所提出方法的实现代码将在GitHub上开源，供业界学习研究。相关代码与数据均已在GitHub开源⁰；

基金项目：2020-2024自然科学基金重点联合项目：自然语言对话交互的基础理论和方法(61936010)

⁰https://github.com/15962171082/Incomplete_MLTC

- 在合成数据集和真实数据集上的实验表明，本文提出的PST框架具有通用性，并且能在一定程度上缓解数据不完全标注问题带来的影响。

2 相关工作

多标签文本分类目的是通过分类器自动获取与文本相关的一个或多个预定义标签，关于该任务的研究主要分为四类，第一类是关于文本语义表示的研究，第二类是关于标签间关系的研究，第三类是关于标签分布的研究，第四类是关于文本与标签语义链接的研究。

文本语义表示：Kim (2014)将图像中效果显著的CNN模型迁移到文本，捕捉窗口局部信息来更好地表征文本。Nam et al. (2017);Yang et al. (2018)基于Seq2Seq的方法，采用RNN对输入文本进行编码，并采用基于注意力的RNN解码器，依次生成预测的标签。Devlin et al. (2018)利用预训练语言模型BERT直接编码获取文本的语义表示。

标签间关系：标签间通常存在依赖、相似、相反和层级等关系。肖琳et al. (2020)提出了标签特定注意力网络，可以在预测每个标签的同时关注到其它标签。Zhang et al. (2021)引入多任务学习方法来增强标签相关性反馈，并利用联合编码机制同时获得文本和标签的表示。Zhao et al. (2022)集成标签增强与模型训练，有效挖掘不同标签的隐含相对重要性信息。

标签分布：由于复杂的标签体系，多标签分类数据集通常伴随着样本标签分布不均衡等问题。Chawla et al. (2002)通过设计特殊的数据采样方式使得采样后的训练集样本标签尽量均衡。然而基于采样的方法无法充分利用标注数据，为此Lin et al. (2017);Wu et al. (2020)通过设计不同的损失函数来缓解标签分布不均衡。此外Xiao et al. (2021)利用迁移学习将头部标签与尾部标签联系起来，以缓解长尾分布中尾部标签数据量少的问题。

文本与标签语义链接：Xiao et al. (2019)在文档和标签之间共享单词表示，利用标签语义信息来确定标签和文档之间的语义链接，从而构建特定于标签的文档表示，并采用自注意力机制建立两者之间的联系。Ma et al. (2021)基于图神经网络来捕捉文本与相应标签的语义互动，它利用全局统计模式和局部动态关系来推导不同标签特定语义部分之间的依赖关系。

上述研究将训练集视为规范标注数据集，并通过全监督训练获取分类模型，而忽视了本文讨论的不完全标注问题。Self-Training(Scudder, 1965)作为一种半监督学习的方法有着非常悠久的历史，该方法的主要思想是利用教师模型对大规模无标注数据进行自动标注以增加训练集的数量，进而更新优化教师模型。随着神经网络模型的发展和对标注数据需求的增加，Self-Training一直是一个热门的研究方向。该方法已成功运用于各种任务如：机器翻译(Jiao et al., 2021)、问答(Sachan and Xing, 2018)、关系抽取(Yu et al., 2022)等。近几年关于Self-Training的研究主要集中在选择伪标签的策略和模型特征的设计两个方面(Triguero et al., 2015)，本文提出的PST框架主要针对如何有效准确地选择伪标签这一方面。

3 预备知识

3.1 任务定义

我们假设 $D = \{(x_i, y_i)\}_{i=1}^N$ 表示一个人工标注的多标签文本分类数据集，它由 N 个文本 x_i 及其对应人工标注的标签集合 y_i 组成。其中每个文本 x_i 由 n 个单词组成 $x_i = \{w_{i1}, \dots, w_{iq}, \dots, w_{in}\}$ ， w_{iq} 表示文本 x_i 的第 q 个单词。 x_i 对应的人工标注标签 $y_i \in \{0, 1\}^l$ ，其中 l 是预定义标签的总数，相关标签记为1，不相关记为0。由于不完全标注问题的存在，人工标注的标签集合 y_i 可能存在遗漏缺失，因此我们进一步定义 $\tilde{y}_i \in \{0, 1\}^l$ 表示真实标签集合，相关标签记为1，不相关记为0。人工标签集合 y_i 与真实标签集合 \tilde{y}_i 的异或即为缺失标签集合 $\hat{y}_i = \tilde{y}_i \oplus y_i$ ， $\hat{y}_i \in \{0, 1\}^l$ ，其中 l 是预定义标签的总数，缺失标签记为1其它记为0。

一般地，多标签文本分类任务需要学习一个从文本 x_i 到真实标签集合 \tilde{y}_i 的分类器，但在真实场景中人工标注的标签集合 y_i 可能存在不完全标注问题导致缺失标签集合 \hat{y}_i 的存在。而不完全标注的多标签文本分类的目的就是从已知的人工标签集合 y_i 出发，学习一个从文本 x_i 到相关标签的分类器，同时要尽可能地削弱未知的缺失标签 \hat{y}_i 给分类器带来的影响。

3.2 教师模型

本节将不同的多标签文本分类模型抽象为统一的整体，从编码到解码再到训练介绍了多标签文本分类的任务流程，同时公式化的损失函数方便后续更加直观地展示PST框架训练中对模型损失函数的改进。

3.2.1 编码

首先我们将文本 x_i 输入BERT得到句子的特征表示序列 $\mathbf{H} = \{\mathbf{h}_1, \dots, \mathbf{h}_n\}$, $\mathbf{H} \in \mathbb{R}^{L \times d}$, 计算如式(1)所示。

$$\{\mathbf{h}_1, \dots, \mathbf{h}_n\} = \text{BERT}(x_i) \quad (1)$$

其中每个字符的向量表示 $\mathbf{h}_i \in \mathbb{R}^d$, d 为词向量表示的维度, L 为预设的最大文本长度。

接着为进一步提取句子的语义特征, 多标签文本分类模型通过设计不同的网络架构对文本编码进一步操作以获取文本的向量表示 $\mathbf{v}_s \in \mathbb{R}^m$, 计算如式(2)所示。

$$\mathbf{v}_s = \text{Net}(\mathbf{H}) \quad (2)$$

其中 \mathbf{v}_s 为编码层得到的文本向量用于下一步编码, m 为向量维度, Net表示不同分类模型抽象为统一的网络架构。

3.2.2 解码

解码阶段我们将编码层获得的文本向量 \mathbf{v}_s 通过标签分类层得到最终的标签表示向量 $\mathbf{p}_i \in \mathbb{R}^l$, 计算如式(3)所示。

$$\mathbf{p}_i = \text{sigmoid}(\mathbf{W}_i \cdot \mathbf{v}_s + \mathbf{b}_i) \quad (3)$$

其中, l 为标签总数, sigmoid为激活函数, $\mathbf{W}_i \in \mathbb{R}^{l \times m}$ 为可学习权重矩阵, \mathbf{b}_i 为偏置。标签表示向量 \mathbf{p}_i 的第 n 个数值表示第 n 个标签与文本相关的概率, 如果该概率大于我们设定的阈值 θ , 则判定该标签与文本相关。

3.2.3 训练

在多标签分类任务中, 通常使用二元交叉熵损失函数计算损失, 计算如式(4)所示。

$$L_{BCE} = \begin{cases} -\log(p_i^k) & \text{if } y_i^k = 1 \\ -\log(1 - p_i^k) & \text{otherwise} \end{cases} \quad (4)$$

然而该交叉熵损失对每个标签的计算有着相同的权重, 当标签分布不平衡时优化效果降低。而本文实验对比的部分多标签分类模型通过设计不同的损失函数缓解标签分布不平衡问题, 例如Cui et al. (2019)设计Class-balanced focal loss(CBloss)作为损失函数, 计算如式(5-6)所示。

$$r_{CB} = \frac{1 - \epsilon}{1 - \epsilon^{freq}} \quad (5)$$

$$L_{CB} = \begin{cases} -r_{CB} (1 - p_i^k)^\gamma \log(p_i^k) & \text{if } y_i^k = 1 \\ -r_{CB} (p_i^k)^\gamma \log(1 - p_i^k) & \text{otherwise} \end{cases} \quad (6)$$

其中 $\epsilon \in [0, 1)$ 为人为设置的超参数, $freq$ 为训练集中每个标签对应的频率, p_i^k 为上一步得到的标签向量 \mathbf{p}_i 的第 k 个值, $\gamma \geq 0$ 为可调的聚焦参数。本文对比的其它模型的不同损失函数公式详见附录。

4 本文方法

本节详细介绍了本文所提出的基于局部标注的自监督框架(PST), 如图1所示。首先介绍PST框架的整体流程, 然后描述PST框架中最重要的伪标签生成与选择算法, 最后提出联合训练来优化教师模型。

4.1 PST框架

标准的Self-Training框架利用在人工标注数据集上训练得到的教师模型自动地对大量无标注样本进行标注得到伪标签, 最后将这些伪标签样本与人工标注数据集混合起来训练得到最终模型。具体流程如图1中红色框线内所示: (1)使用人工标注数据集来训练教师模型; (2)使用教师模型对无标注数据集进行伪标签预测; (3)通过预先定义的阈值将预测的标签分为相关标签与不相关标签; (4)将含有相关标签的样本与人工标注数据集混合起来训练更新教师模型; (5)重复2至4步直到教师模型的性能不再提高或满足停止条件。

然而在本文提到的不完全标注应用场景中，人工标注数据可能存在缺失标签，教师模型可能会受缺失标签的负面影响给无标注数据打上错误标签。因此本文提出的PST框架对标准Self-Training框架的(2)(3)两步进行了补充与修正。如图1所示，在第二步利用教师模型对无标注数据标签预测时，同时对人工标注数据进行补充预测，以补充人工标注所遗漏的缺失标签。为了能够同时预测这两部分数据，我们为每一个样本的每一个标签定义了一个状态标记(详见节4.2)，防止教师模型在预测标签时破坏原有的人工标注标签。在第三步通过阈值确定伪标签时，我们利用双阈值策略将标签分为三类：正例标签、负例标签和其它标签(详见节4.2)。同时根据预测标签类别反向更新该样本对应的标签状态集合。双阈值的设定一方面能够缓解教师模型因缺失标签影响而错误标签预测的数量，另一方面通过双阈值获取的其它标签在训练时不对模型产生影响，进而缓解缺失标签导致的错误信息影响模型性能。

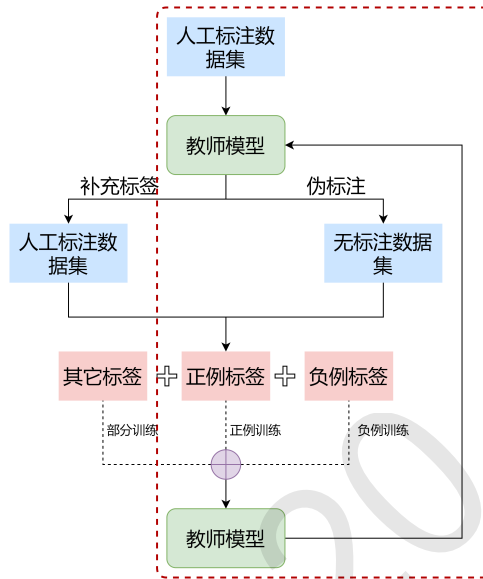


Figure 1: PST框架结构图

4.2 伪标签生成与选择

算法1是本文所提的伪标签选择算法，该算法利用双阈值策略和全局标签状态集合，通过教师模型的标签预测对缺失标签进行补充利用。

具体的，我们定义了一个全局标签状态集合 $State = \{-2, -1, 0, 1, 2\}^{N \times l}$ ，其中 N 表示人工标注数据和无标注数据的总和， l 表示标签的数量。全局标签状态集合 $State$ 用来表示每个样本当前对应的每个标签的状态，该状态用于辅助后续标签的选择，其中 -2 到 2 表示标签的状态标记， 2 代表确定正例，即确定该标签与文本相关，且教师模型不再对该标签预测得分，并直接选择该状态的标签当作正例相关标签； -2 代表确定负例，即确定该标签与文本不相关，且教师模型不再对该标签预测得分，并直接选择该状态的标签当作负例不相关标签； $-1, 0, 1$ 代表中间状态，处于该状态的标签暂未确定是否与文本相关，PST框架会利用教师模型对这些状态的标签预测打分，并根据得分和预定义的阈值来动态地改变标签的状态。首先我们初始化一个全局标签状态集合：把人工标注样本的相关标签状态初始化为 2 ，其它所有标签的状态初始化为 0 (算法1第1行)。接着，我们利用教师模型对所有数据进行标签预测 (算法1第3行)。之后，利用人为定义的正例阈值 T_{Pos} 和负例阈值 T_{Neg} 与模型预测每个标签的得分进行比较来将标签分为正例、负例或其它三类，同时更新标签状态集合 (算法1第13-19行)，其中正例阈值 T_{Pos} 和负例阈值 T_{Neg} 的取值选择是根据多组取值组合实验中最佳实验结果确定 (详见节5.4.3)。当某样本的某个标签状态为 2 时，即模型连续两次给予该标签高分，则该标签在后续Self-Training中无需模型预测打分；若某样本的某个标签状态为 -2 ，即教师模型连续两次给予该标签低分，则该标签无需模型预测打分 (算法1第7-12行)。最后通过筛选返回用于下一轮训练更新教师模型的数据 (算法1第21-25行)。附录中的案例分析详细展示分析了样例标签在PST过程中的变化。

Algorithm 1 伪标签生成与选择算法

Input: 教师模型自动对人工标注数据集和无标准数据集进行打分, 得到评分数据 $D_{auto} = \{x_i, p_i, Y\}_{i=1}^N$, 其中 p_i 为模型预测文本 x_i 的标签向量表示, Y 为预定义的标签集合, N 为训练集和无标注数据集的总和数。正例阈值 T_{Pos} 和负例阈值 T_{Neg} 。

Output: 挑选出下一轮训练的标注数据。

```

1: 初始化全局变量  $State = \{-2, -1, 0, 1, 2\}^{N \times l}$  ▷ 详见节4.2.
2:  $D_{train} = []$ 
3: for  $(x_i, p_i) \in D_{auto}$  do
4:    $S_i = State[i]$ ;
5:    $y_i = \{0, 1, 2\}^l$ ;
6:   for  $(p_i^k, y_i^k, S_i^k) \in p_i, y_i, S_i$  do
7:     if  $S_i^k == 2$  then
8:        $y_i^k = 1$ ; Continue; ▷ 状态标记为2, 无需改动, 直接添加正例
9:     end if
10:    if  $S_i^k == -2$  then
11:       $y_i^k = 0$ ; Continue; ▷ 状态标记为-2, 无需改动, 直接添加负例
12:    end if
13:    if  $p_i^k > T_{Pos}$  then
14:       $S_i^k = S_i^k + 1$ ;  $y_i^k = 1$ ; ▷ 本轮标签预测为正例, 标签状态加一
15:    else if  $p_i^k < T_{Neg}$  then
16:       $S_i^k = S_i^k - 1$ ;  $y_i^k = 0$ ; ▷ 本轮标签预测为负例, 标签状态减一
17:    else if  $T_{Pos} > p_i^k > T_{Neg}$  then
18:       $y_i^k = 2$ ; ▷ 本轮标签预测为其它, 标签状态不变
19:    end if
20:  end for
21:  if  $y_i$  not all 0 then
22:     $(x_i, y_i) \rightarrow D_{train}$ ;
23:  end if
24: end for
25: Return:  $D_{train}$ 

```

4.3 联合训练

通过伪标签选择算法获取的新数据将直接用于训练更新教师模型。不同于第一阶段直接利用人工标注数据集训练教师模型, 由于双阈值的设定, 新数据集中添加了其它标签的额外信息。为此PST框架通过修改教师模型的损失函数以引入其它标签的额外信息, 削弱错误信息对模型的误导。同样的这里以CBLoss(Cui et al., 2019)为例对式(6)进行修改, 计算如式(7)所示, 其中符号变量与式(6)相同。本文用到的其它损失函数修正后引入其他标签信息的计算公式详见附录。

$$L_{CB-Part} = \begin{cases} -r_{CB} (1 - p_i^k)^\gamma \log(p_i^k) & \text{if } y_i^k = 1 \\ 0 & \text{if } y_i^k = 2 \\ -r_{CB} (p_i^k)^\gamma \log(1 - p_i^k) & \text{otherwise} \end{cases} \quad (7)$$

5 实验与分析

在本节中, 我们分别在合成数据集和真实数据集上进行实验, 以证明所提出方法的有效性与通用性。本节首先介绍两个数据集和实验的评估策略。然后, 简要描述实验的相关设置和用于对比的基线模型。最后, 我们列出不同数据下的实验结果并进行相应的分析。

5.1 实验数据

5.1.1 合成数据集

本文采用多标签文本分类任务中常见的英文数据集AAPD(Yang et al., 2018)作为合成数据, 该数据集是由网络上收集的55,840 篇论文的摘要和相应学科类别组成。一篇学术论文属于一个或者多个学科, 总共由54个学科组成, 目的是根据给定的摘要来预测学术论文相对应的学科。为了模拟不完全标注的数据集, 我们在标注规范的AAPD数据集上, 对训练集按照不同的缺失比例 p 来随机的删除一些标签。同时为了更好的评估模型的性能, 我们并未对验证集和测试集进行随机删除操作。

为了更全面地分析不同场景下的不完全标注问题, 我们采用两种不同的方案来人为删除标签构造不完全标注的合成数据集。方案一: 我们确保删除标签后的数据集每个样本至少仍保留一个相关标签。我们首先统计平均每个样本包含的标签数为2.41, 并进一步确定该方案下标签缺失的上限概率为 $(2.41 - 1.0) \div 2.41 = 0.585$, 因此我们按照0.1、0.2、0.3、0.4、0.5、0.585的概率来按标签分布等比例删减相关标签并且始终保证每个样本至少一个标签; 方案二: 对于标注数据集中一条未标注标签的样本, 我们假设该样本并不是无标注数据, 而是标注为与所有标签都不相关的“其它”标签, 因此该样本被当作完全负例加入模型训练。按照这种假设, 我们采取0.1至0.9的概率随机丢失样本标签, 当概率较高时样本的所有标签可能都会丢失, 对于此类数据我们并未抛弃仍当作负例训练教师模型。我们按照上述两种方案人为构造了共15组训练数据(方案一6组, 方案二9组), 加上原始完整的数据共构建了16组合成数据进行相关实验。

5.1.2 真实数据集

我们对CCKS2022 Task8¹面向金融领域的Few-Shot事件主体抽取学术评测提供的数据集进行人为修正, 将其构建为多标签文本分类任务的中文数据集CCKS-IMLTC作为真实场景下不完全标注的数据集。

CCKS-IMLTC数据集具体的构建流程如下: (1)原CCKS数据为事件主体抽取数据集, 每条样本由一段文本和该文本包含的一个或多个事件类型与事件主体对儿构成, 但其标注质量不高, 存在比较严重的事件类型缺失问题。因此我们删除原数据集中的事件主体只保留事件类型, 同时删去了部分数据量极少的类别, 将其修正为多标签文本分类任务的中文数据集, 并按8:1:1的比例将数据切分为训练集、验证集和测试集。(2)为了更加精确地评估模型性能, 我们对划分后的验证集和测试集进行人工补充, 并对每个补充的标签打上标记便于后续统计每个标签的缺失情况。(3)对比补充前后的测试集与验证集, 估算出整体标签的平均缺失比例为9.2%。虽然整体数据集上的缺失比例不高, 但在我们对测试集和验证集进行缺失标签补充的过程中发现, 有一部分关联性与共现性比较强的标签的缺失率可达60%左右, 据我们人工统计, 共有12个标签缺失率较高, 我们将这些标签当作一个集合称为Few, 并在实验中单独对这些标签进行评价。

数据集	标签数	训练集	验证集	测试集	无标注	平均标签数
AAPD	54	26,920	1,000	1,000	26,920	2.41
CCKS-IMLTC	96	40,000	5,000	5,000	43,147	1.21

Table 2: AAPD与CCKS-IMLTC数据集规模

表(2)记录了两个数据集的规模大小和标签数量。其中表格中无标注一列表表示Self-Training过程中无标注文本数量。CCKS-IMLTC数据对应的无标注文本主要来自CCKS2021和CCKS2020中相关任务的纯文本数据。AAPD数据共55,840条标注样本, 首先按标签分布等比例各拿出1,000条当作测试集与验证集, 接着把剩下的标注样本一半当作训练集, 另一半去除掉标注标签当作无标注数据。

5.2 评价指标

本文用三种指标全面评价模型性能。第一种指标为精确率(Precision, P)、召回率(Recall, R)和F1值(F1-measure, F1), 该指标常用于分类任务的模型性能的直观评估。后两种评价指标

¹https://www.biendata.net/competition/ccks2022_eventext/data/

分别为退化率 α_p 和误导率 β_p ，计算如式(8-9)所示。该指标由负采样NER(Li et al., 2020)提出，用于评估分析缺失实体对模型的退化和误导两方面的影响。本文将该指标从NER任务中迁移到多标签文本分类中，以评估缺失标签对模型的影响，同时验证PST框架对两方面影响的削弱效果。

$$\alpha_p = \frac{f_0^a - f_p^a}{f_0^a} \quad (8)$$

$$\beta_p = \frac{f_p^a - f_p}{f_p^a} \quad (9)$$

其中 f_p 和 f_p^a 分别表示在缺失概率 p 数据上是否经过调整缺失标签损失训练后模型分类的F1值。具体地，我们通过将人工构建的合成数据中的缺失标签当作其它标签不计算其loss来调整缺失标签损失。而在真实数据集中，我们并不知道具体的缺失标签，因此实验部分我们只将该指标作用到合成数据AAPD上。 f_0^a 表示在原始完整数据上训练模型分类的F1值。从公式(8)可以发现，退化率 α_p 指标消除了缺失标签对loss的影响，主要是为了评价正例标签数量减少导致分类器训练不充分的退化影响。从公式(9)可以发现，误导率 β_p 指标计算相同缺失概率下是否消除缺失标签作为负例对loss的影响，从而进一步评价不完全标注数据对分类器的误导影响。

5.3 实验设置与对比模型

我们将PST框架运用到多种常见的多标签文本分类模型上进行实验，相关的模型如下：1)BERT-CLS(CLS)(Devlin et al., 2018):首先利用预训练语言模型将文档编码到向量空间，然后独立地输出每个标签的概率。2)BERT-TextCNN(TextCNN)(Kim, 2014):将BERT作为文本编码结合CNN卷积提取单词信息。3)LSAN(Xiao et al., 2019):借助注意力机制获得特定标签的文本表示。4)Focal loss(FL)(Lin et al., 2017):一种简单但广泛使用的加权损失分类策略。5)Rebalanced focal loss(R-FL)(Wu et al., 2020):重平衡加权与focal loss的组合。6)Class-balanced loss(CB)(Cui et al., 2019):以每个类的有效数量为指导的按类重新加权的损失函数。7)Distribution-balance loss(DB)(Wu et al., 2020):利用负容忍正则化缓解模型对负类的过度抑制而造成的分类边界偏移。8)HTTN(Xiao et al., 2021):利用元学习将头部标签与尾部标签联系起来，以缓解长尾分布中尾部标签数据量少的问题，是一种基于迁移学习的多标签分类器。9)LACO(Zhang et al., 2021):通过引入多任务学习方法来增强标签相关性反馈，并利用联合编码机制同时获得文本和标签的向量表示。10)FLEM(Zhao et al., 2022):集成标签增强与模型训练，从而有效地挖掘不同标签的隐含相对重要性信息。11)BERT-TextCNN+CBLoss(TextCNN-CB)，作为PST框架的基础教师模型，用于合成数据集上补充分析实验。

我们使用开源的预训练模型bert-base-cased²和bert-base-chinese³分别作为英文和中文的编码层，通过教师模型在验证集的性能，设定超参。具体地，我们设置最大句长为256，学习率为2e-5，batch-size为16，教师模型的训练轮次为20，self-training的轮次为10，随机种子为1227，dropout为0.5，正例阈值为0.6，负例阈值为0.4，线性层维度大小为300，FLEM模型中 α 和 β 均为0.01，HTTN模型中头部标签数量为84，TextCNN中滤波器为200、窗口大小为[1, 3, 5, 7]，LSTM隐层大小为256，其它模型超参均遵循其模型原论文中的设置。

5.4 实验结果与分析

5.4.1 真实数据集上的实验

表3展示了在CCKS-IMLTC数据集上不同系统的实验结果，其中Teacher表示采用标准有监督训练，Self-Training表示采用标准Self-Training框架，PST(Ours)表示采用本文所提PST框架。横向对比分类模型在不同框架下的结果发现，标准Self-Training框架对某些模型的性能有所提升，然而这种提升并不稳定，而本文提出的PST框架对所有教师模型都有较高的性能提升，充分证明了PST框架的有效性与通用性。对比不同框架下的性能发现，标准Self-Training框架通过优化模型的准确率(P值)提升性能，相反PST框架主要优化模型的召回率(R值)以提升整体性能。并且相较于Self-Training，PST框架对教师模型性能的提升更加稳定且有效。进一步分

²<https://huggingface.co/bert-base-cased>

³<https://huggingface.co/bert-base-chinese>

析实验结果发现，教师模型与传统的Self-Training框架训练中大都将缺失标签当作负例标签训练，因此其预测结果大都是训练中见过的正例标签，进而导致预测出的正例标签准确率较高即整体的P值较高。而PST框架通过双阈值策略和全局标签状态补充利用缺失的正例标签，进而使得可以预测出更多的正例标签，即使得模型整体的R值升高。

图2展示了不同模型在CCKS-IMLTC中标签缺失严重的Few标签集合上的实验结果，我们发现相较于表3整体标签上的实验结果，PST框架在缺失比例高的标签上对教师模型的提升更为明显。同时我们发现当教师模型性能过低时，标准Self-Training框架反而会给模型带来负优化，而PST框架对所有教师模型都有不同程度的提升，更加充分地证明了PST框架的通用性，可以兼容现有多种不同的多标签文本分类模型。进一步对比不同模型在整体标签和Few标签集合上的实验结果，我们发现相较于注重标签间关系的模型FLEM和LACO，针对标签分布设计的模型如CB，DB等随着标签缺失率升高性能下降的较为缓慢。

模型	Teacher			Self-Training			PST(Ours)				
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	$\Delta_T(F1)$	$\Delta_{ST}(F1)$
CLS	79.91	63.81	70.95	77.37	63.08	69.50	76.60	67.25	71.62	+0.67	+2.12
TextCNN	76.69	71.91	74.22	75.91	73.87	74.88	78.06	72.22	75.03	+0.81	+0.15
LSAN	75.51	59.89	66.80	77.24	60.89	68.10	75.34	63.92	69.16	+2.36	+1.06
FL	80.99	68.83	74.41	82.14	67.83	74.30	80.44	70.38	75.07	+0.66	+0.77
RFL	80.65	69.67	74.76	81.97	69.37	75.15	80.36	70.58	75.15	+0.39	+0.00
CB	81.16	70.28	75.33	81.83	69.39	75.10	80.51	71.03	75.48	+0.15	+0.38
DB	74.76	74.50	74.63	73.99	76.66	75.30	77.32	74.68	75.97	+1.34	+0.67
HTTN	81.46	67.81	74.01	81.56	68.38	74.39	80.79	69.88	74.94	+0.93	+0.55
LACO	78.19	71.45	74.65	78.85	70.46	74.42	79.53	70.68	74.84	+0.19	+0.42
FLEM	80.54	69.39	74.55	83.91	67.81	75.01	82.05	69.57	75.30	+0.75	+0.29
TextCNN-CB	76.56	74.35	75.44	77.00	73.53	75.22	77.59	74.68	76.11	+0.67	+0.89

Table 3: 不同模型在CCKS-IMLTC数据集上的实验结果

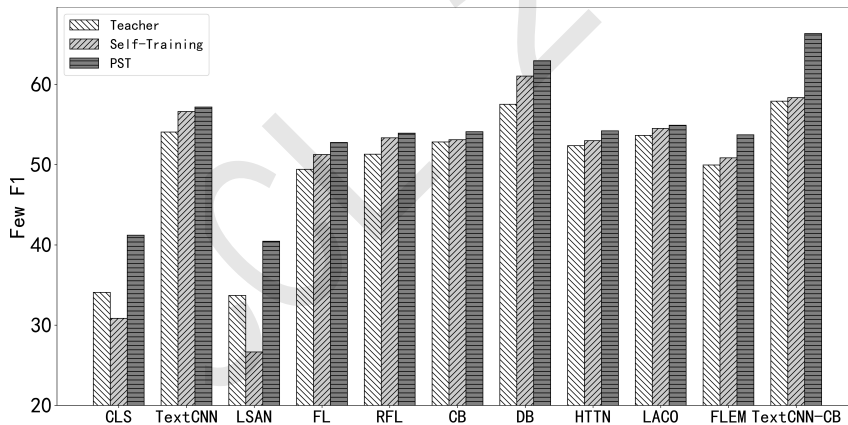


Figure 2: 不同模型在CCKS-IMLTC上Few标签的实验结果

5.4.2 合成数据集上的实验

图3展示了两种不同标签缺失方案(详见节5.1.1)下TextCNN-CB在人工合成数据集AAPD上的实验结果。图3(a)与图3(d)为不同框架下F1值随标签缺失比例变化的折线图，可以发现随着缺失率的增加所有框架下模型的性能都会有所下降，但相较于教师模型和标准Self-Training框架，本文提出的PST框架下降得更加缓慢，即PST框架可以有效缓解缺失标签对模型的负面影响进而提升模型性能，并且标签缺失问题越严重提升效果越明显。同时我们发现传统的Self-Training框架在两种不同标签缺失方案下性能基本都不如教师模型，说明教师模型受到缺失标签的影响在对训练集打分时可能会导致标注错误标签进而影响模型下一轮训练，而我们提出的PST半监督学习框架可以通过双阈值策略和全局标签状态来缓解这种错误标签的出现，同时提升教师模型补充缺失标签的能力。

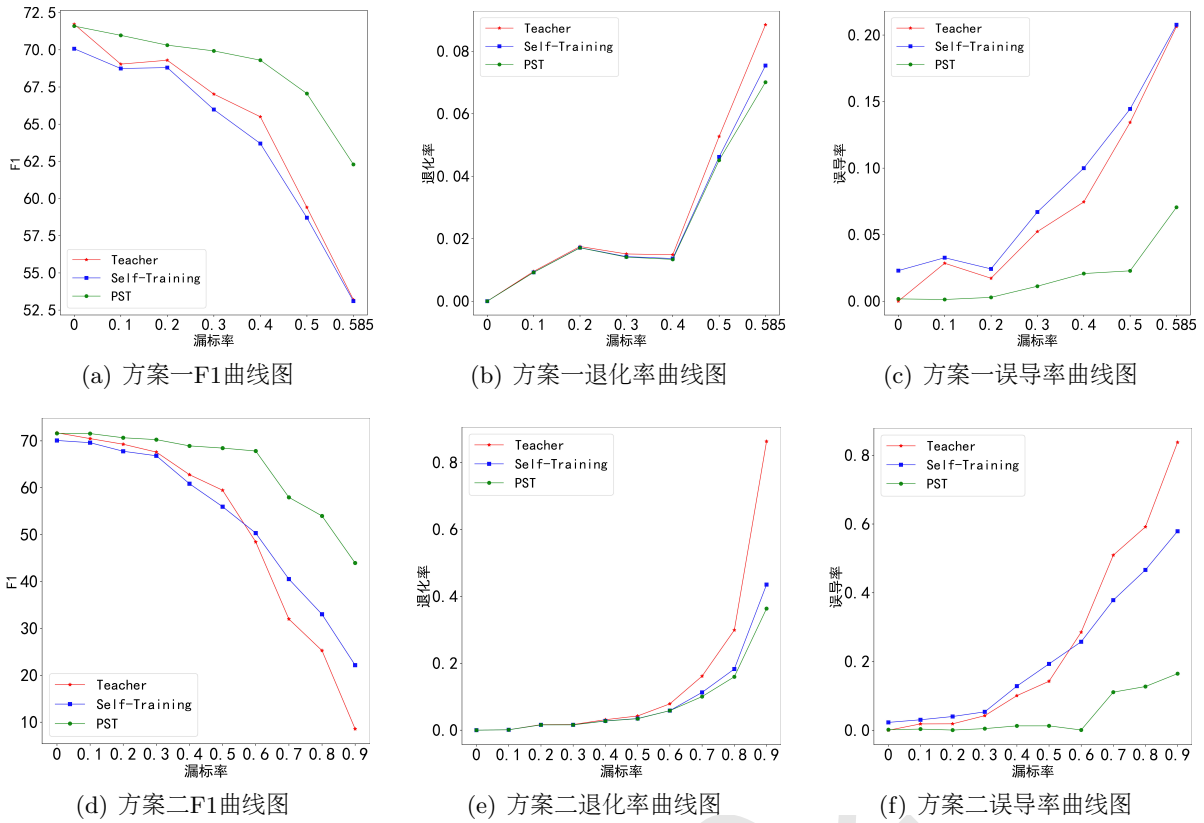


Figure 3: 不同标签缺失方案下人工合成数据AAPD上的实验结果

图3(b)与图3(e)为不同框架下退化率随标签缺失比例变化的折线图，可以发现标签缺失比例不足50%时不同框架下缺失标签带来的退化影响都很小最高仅为5%。当缺失率达到70%时，因缺失标签带来的退化影响陡增，而PST框架可以明显地缓解，尤其是缺失率为90%时PST框架可以将退化率从90%降至40%。图3(c)与图3(f)为不同框架下误导率随标签缺失比例变化的折线图，可以发现随着标签缺失率的增加越来越多的缺失标签被当作负例学习误导模型，而PST框架将可能产生误导的缺失标签转为其它标签忽略其损失，进而有效的减弱模型被误导的概率。总的来说，实验结果表明PST框架可以缓解缺失标签对模型误导和退化两方面的影响。

5.4.3 不同正负例阈值选择对PST框架的影响

图4展示了PST框架中正负例阈值不同组合时TextCNN-CB在CCKS-IMLTC数据上的实验结果热力图，颜色越暗模型的F1值越高，相反颜色越淡模型的F1值越低，我们发现PST框架的正例阈值设为0.6，负例阈值设为0.4时模型性能最优。

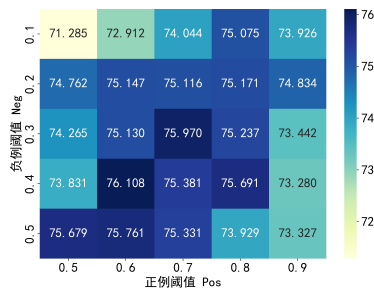


Figure 4: 不同正负例阈值组合在CCKS-IMLTC数据上的实验结果

5.4.4 消融实验

为了进一步分析PST框架中各个组件对整个框架的影响，我们在CCKS-IMLTC上进行了消

融实验，实验结果如表4所示，其中Few和All分别指的是标签缺失严重的Few标签集合上的结果与整体全部标签上的结果，右下角正负数表示较于教师模型性能的变化。首先我们发现去除外部无标注数据只在人工标注数据集上使用PST框架，教师模型的性能仍有明显提升，侧面说明PST框架的确可以对不完全标注数据补充部分缺失标签；接着我们去掉双阈值策略，只保留正例阈值整体框架退化为标准Self-Training，此时整体性能有所降低，但Few标签上的性能仍有微弱提升；最后保留双阈值去掉其它标签，此时Few标签性能提升1.36%但较完整PST框架仍有很大差距。总的来说，PST框架可以无需外部无标注数据辅助直接在不完全标注的数据上提升教师模型性能，同时PST框架中的双阈值策略，其他标签设置等各组件都是整体框架中不可或缺的部分。

	Few(F1%)	All(F1%)
TextCNN-CB	57.90	75.44
TextCNN-CB w/PST	66.33 _{+8.43}	76.11 _{+0.67}
-Unlabel data	63.99 _{+6.09}	75.82 _{+0.38}
-负例阈值 T_{Neg}	58.35 _{+0.45}	75.22 _{-0.22}
-Partial label	59.26 _{+1.36}	75.37 _{-0.07}

Table 4: TextCNN-CB在CKS-IMLTC上的消融实验

6 总结与展望

本文提出了一种全新的基于局部标注的自监督框架(PST)，以缓解不完全标注问题在多标签文本分类中的影响。该框架是一种模型无关的插件式框架，可以兼容多种不同的教师模型。充分利用外部无标注数据来优化教师模型的同时，对不完全标注数据的缺失标签补充利用，进而削弱了缺失标签给模型带来的影响。实验结果表明，我们提出的框架具有通用性，并且能一定程度缓解数据不完全标注问题带来的影响。

我们发现教师模型的选择也影响着我们框架的上限，效果更好的教师模型通过PST框架可以更好的补充缺失标签，从而更大程度地缓解不完全标注的问题。因此如何设计更优雅、高效的教师模型，提升该任务的整体性能，是我们今后的研究方向。

参考文献

- Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.
- Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277.
- 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*.
- John Dougrez-Lewis, Maria Liakata, Elena Kochkina, and Yulan He. 2021. Learning disentangled latent topics for twitter rumour veracity classification. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 3902–3908.
- Cunxiao Du, Zhaozheng Chen, Fuli Feng, Lei Zhu, Tian Gan, and Liqiang Nie. 2019. Explicit interaction model towards text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6359–6366.
- Himanshu Jain, Yashoteja Prabhu, and Manik Varma. 2016. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 935–944.
- Wenxiang Jiao, Xing Wang, Zhaopeng Tu, Shuming Shi, Michael Lyu, and Irwin King. 2021. Self-training sampling with monolingual data uncertainty for neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2840–2850.

- Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October. Association for Computational Linguistics.
- John Langton, Krishna Srihasam, and Junlin Jiang. 2020. Comparison of machine learning methods for multi-label classification of nursing education and licensure exam questions. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 85–93.
- Xin Li, Haoran Xie, Yanghui Rao, Yanjia Chen, Xuebo Liu, Huan Huang, and Fu Lee Wang. 2016. Weighted multi-label classification model for sentiment analysis of online news. In *2016 International conference on big data and smart computing (bigcomp)*, pages 215–222. IEEE.
- Yangming Li, Lemaoy Liu, and Shuming Shi. 2020. Empirical analysis of unlabeled entity problem in named entity recognition.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.
- Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 115–124.
- Qianwen Ma, Chunyuan Yuan, Wei Zhou, and Songlin Hu. 2021. Label-specific dual graph neural network for multi-label text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3855–3864.
- Jinseok Nam, Eneldo Loza Mencía, Hyunwoo J Kim, and Johannes Fürnkranz. 2017. Maximizing subset accuracy with recurrent neural networks in multi-label classification. *Advances in neural information processing systems*, 30.
- Nikolaos Pappas and James Henderson. 2019. Gile: A generalized input-label embedding for text classification. *Transactions of the Association for Computational Linguistics*, 7:139–155.
- Mrinmaya Sachan and Eric Xing. 2018. Self-training for jointly learning to ask and answer questions. 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 629–640.
- Henry Scudder. 1965. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371.
- Isaac Triguero, Salvador García, and Francisco Herrera. 2015. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information systems*, 42:245–284.
- Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. 2020. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 162–178. Springer.
- Lin Xiao, Xin Huang, Boli Chen, and Liping Jing. 2019. Label-specific document representation for multi-label text classification. 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 466–475.
- Lin Xiao, Xiangliang Zhang, Liping Jing, Chi Huang, and Mingyang Song. 2021. Does head label help for long-tailed multi-label text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14103–14111.
- Pengcheng Yang, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. 2018. Sgm: Sequence generation model for multi-label classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3915–3926.
- Junjie Yu, Xing Wang, Jiangjiang Zhao, Chunjie Yang, and Wenliang Chen. 2022. Stad: Self-training with ambiguous data for low-resource relation extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2044–2054.

Ximing Zhang, Qian-Wen Zhang, Zhao Yan, Ruifang Liu, and Yunbo Cao. 2021. Enhancing label correlation feedback in multi-label text classification via multi-task learning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1190–1200.

Xingyu Zhao, Yuexuan An, Ning Xu, and Xin Geng. 2022. Fusion label enhancement for multi-label learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*.

肖琳, 陈博理, 黄鑫, 刘华锋, 景丽萍, and 于剑. 2020. 基于标签语义注意力的多标签文本分类. 软件学报, 31(4):1079–1089.

附录.不同损失函数经过其它标签修正前后的计算公式

- **BCE loss**

$$L_{BCE} = \begin{cases} -\log(p_i^k) & \text{if } y_i^k = 1 \\ -\log(1 - p_i^k) & \text{otherwise} \end{cases}$$

$$L_{BCE}^{Partial} = \begin{cases} -\log(p_i^k) & \text{if } y_i^k = 1 \\ 0 & \text{if } y_i^k = 2 \\ -\log(1 - p_i^k) & \text{otherwise} \end{cases}$$

- **Focal loss(FL):** γ 为超参

$$L_{FL} = \begin{cases} -(1 - p_i^k)^\gamma \log(p_i^k) & \text{if } y_i^k = 1 \\ -(p_i^k)^\gamma \log(1 - p_i^k) & \text{otherwise} \end{cases}$$

$$L_{FL}^{Partial} = \begin{cases} -(1 - p_i^k)^\gamma \log(p_i^k) & \text{if } y_i^k = 1 \\ 0 & \text{if } y_i^k = 2 \\ -(p_i^k)^\gamma \log(1 - p_i^k) & \text{otherwise} \end{cases}$$

- **Rebalanced-Focal loss(RFL):** \hat{r}_{DB} 详见(Wu et al., 2020)

$$L_{RFL} = \begin{cases} -\hat{r}_{DB}(1 - p_i^k)^\gamma \log(p_i^k) & \text{if } y_i^k = 1 \\ -\hat{r}_{DB}(p_i^k)^\gamma \log(1 - p_i^k) & \text{otherwise} \end{cases}$$

$$L_{RFL}^{Partial} = \begin{cases} -\hat{r}_{DB}(1 - p_i^k)^\gamma \log(p_i^k) & \text{if } y_i^k = 1 \\ 0 & \text{if } y_i^k = 2 \\ -\hat{r}_{DB}(p_i^k)^\gamma \log(1 - p_i^k) & \text{otherwise} \end{cases}$$

- **Class-balanced focal loss(CB):** ϵ, γ 为超参

$$r_{CB} = \frac{1 - \epsilon}{1 - \epsilon^{freq}}$$

$$L_{CB} = \begin{cases} -r_{CB}(1 - p_i^k)^\gamma \log(p_i^k) & \text{if } y_i^k = 1 \\ -r_{CB}(p_i^k)^\gamma \log(1 - p_i^k) & \text{otherwise} \end{cases}$$

$$L_{CB}^{Partial} = \begin{cases} -r_{CB}(1 - p_i^k)^\gamma \log(p_i^k) & \text{if } y_i^k = 1 \\ 0 & \text{if } y_i^k = 2 \\ -r_{CB}(p_i^k)^\gamma \log(1 - p_i^k) & \text{otherwise} \end{cases}$$

- **Distribution-balanced loss(DB)**: λ 为超参 \hat{r}_{DB}, q_i^k 详见(Wu et al., 2020)

$$L_{DB} = \begin{cases} -\hat{r}_{DB}(1 - q_i^k)^\gamma \log(q_i^k) & \text{if } y_i^k = 1 \\ -\hat{r}_{DB}\frac{1}{\lambda}(q_i^k)^\gamma \log(1 - q_i^k) & \text{otherwise} \end{cases}$$

$$L_{DB}^{Partial} = \begin{cases} -\hat{r}_{DB}(1 - q_i^k)^\gamma \log(q_i^k) & \text{if } y_i^k = 1 \\ 0 & \text{if } y_i^k = 2 \\ -\hat{r}_{DB}\frac{1}{\lambda}(q_i^k)^\gamma \log(1 - q_i^k) & \text{otherwise} \end{cases}$$

附录.案例分析

文本: 涉案的美国三大投行遭到重罚,花旗集团和摩根大通因涉嫌财务欺诈被判有罪,向安然公司的破产受害者分别支付了20亿、22亿和6900万美元的赔偿罚款。					
文本的相关标签: 重大赔付, 财务造假, 破产清算					
标注员标注标签: 重大赔付, 财务造假					
部分不相关标签: 债务违约, 股东减持					
正例阈值 T_{Pos} : 0.6, 负例阈值 T_{Neg} : 0.4					
PST过程中样本标签状态变化:					
初始状态值:	重大赔付_2	财务造假_2	破产清算_0	债务违约_0	股东减持_0
epoch1打分:	/	/	0.53	0.47	0.23
epoch1状态:	重大赔付_2	财务造假_2	破产清算_0	债务违约_0	股东减持_-1
epoch2打分:	/	/	0.65	0.44	0.16
epoch2状态:	重大赔付_2	财务造假_2	破产清算_1	债务违约_0	股东减持_-2
epoch3打分:	/	/	0.71	0.46	/
epoch3状态:	重大赔付_2	财务造假_2	破产清算_2	债务违约_0	股东减持_-2
⋮					
epoch10打分:	/	/	/	/	/
epoch10状态:	重大赔付_2	财务造假_2	破产清算_2	债务违约_-2	股东减持_-2

Table 5: 一条不完全标注样例的标签状态在PST过程中的变化。标签数量过多这里只选取了两个不相关标签演示, 其中绿色标签作为下一轮次训练的正例标签, 红色标签作为下一轮次训练的负例标签, 黑色标签作为其它标签不用来下一轮模型训练。