基于不完全标注的自监督多标签文本分类

汇报人: 任俊飞

jfrenjfren@stu.suda.edu.cn

任俊飞 朱桐 陈文亮

苏州大学 计算机科学与技术学院

主要内容

- 任务定义
- 研究动机
- PST框架方法
- 实验结果

多标签文本分类 (MLTC)

- 定义: 多标签文本分类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，是自然语言处理的一项基本任务。输入文本**X**和标签集合**Y**，要求识别出与文本**X**相关的集合**y**∈**Y**

- 应用:
  - **新闻分类：**一篇新闻文章可能与多个主题相关。
  - **电影分类：**一部电影可能既是"喜剧"，又是"浪漫"，并可能含有"冒险"元素。
  - **事件检测：**一段文本可能同时包含多种事件

不完全标注

- 定义: 在标注好的数据集中，可能存在部分数据相关的标签标注不全，即数据存在不完全标注的情况。而不完全标注的多标签文本分类旨在从不完全标注数据集出发，学习一个文本到相关标签的分类模型，同时要尽可能地缓解缺失标注标签对分类模型的影响。

- 举例: 金融领域中无触发词事件检测任务标注数据集中，只标注两个相关事件，而漏标了“*破产清算*”一个相关标签。

文本	标签
涉案的美国三大投行遭到重罚,花旗集团、摩根大通和美洲银行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*

研究动机

point1: 缺失标注标签影响模型性能

point2: 大规模标注数据中不完全标注情况严重

point3: 在多标签文本分类任务上研究尚不完善

point1: 缺失标注标签影响模型性能

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；

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。

point2: 大规模标注数据中不完全标注情况严重

大规模数据集标注：

- 远程监督 ✚ 人工修正
  - 源知识库局限性与时效性使得远程监督标注的标签更偏向流行大众，而一些其他并不“显眼”的标签容易忽略
  - 部分研究表明[1]，人工修正阶段标注人员往往更加倾向于对远程监督获取的标签进行复查删除错误标注标签，而很少会补充远程监督缺失标注的标签
  - 因此大规模标注规范的数据集仍存在比较严重的标签缺失标注的问题。

1.Huang Q, Hao S, Ye Y, et al. Does Recommend-Revise Produce Reliable Annotations? An Analysis on Missing Instances in DocRED[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022: 6241-6252.

point3: 在多标签文本分类任务上研究尚不完善

NER：Li等人基于负采样的方法降低负例实体被采样训练的可能，进而缓解缺失实体对模型的影响

RE：Tan等人[3]提出类自适应重采样自训练框架，通过决策和回忆得分对每个类别的伪标签进行重新采样，进而缓解缺失标签对模型训练的影响。

MLTC：尚未有成熟的研究

2. Li Y, Shi S. Empirical Analysis of Unlabeled Entity Problem in Named Entity Recognition[C]//International Conference on Learning Representations. 2020.
3. Tan Q, Xu L, Bing L, et al. Class-Adaptive Self-Training for Relation Extraction with Incompletely Annotated Training Data[J]. arXiv preprint arXiv:2306.09697, 2023.

主要内容

- 任务定义
- 研究动机
- **PST框架方法**
- 实验结果

(Partial Self-Training, PST)框架

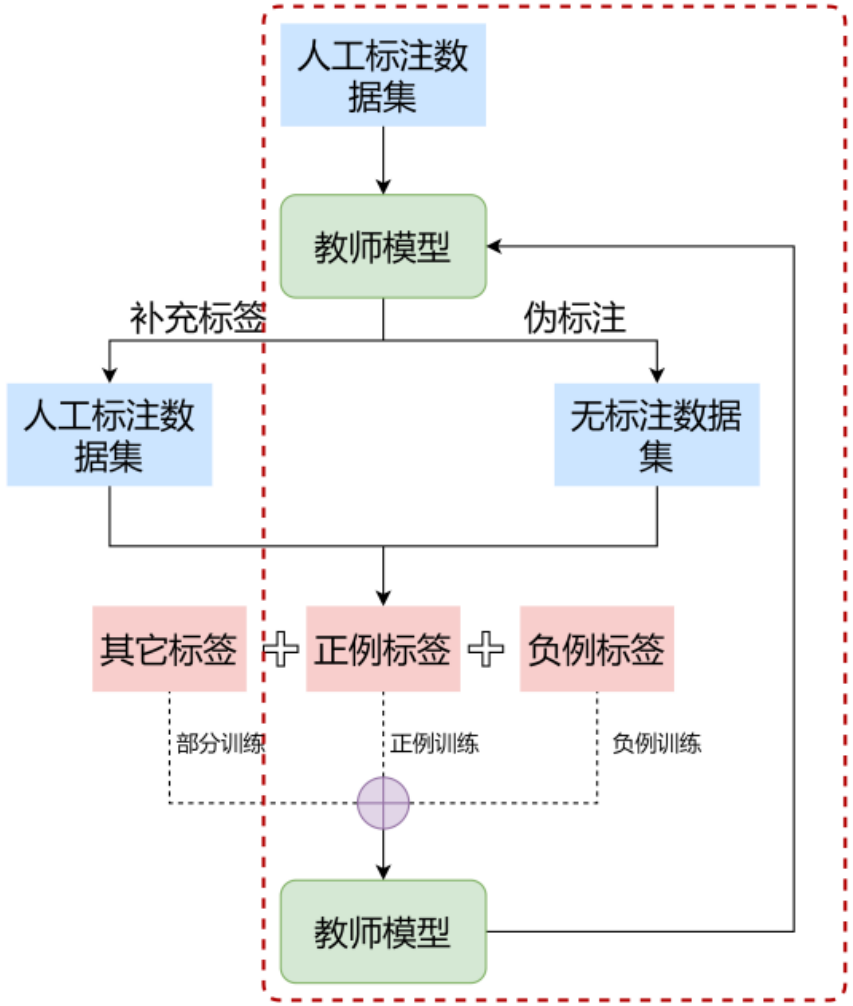


Figure 1: PST框架结构图

1. 原始标注数据训练教师模型
2. 利用教师模型预测标注数据每个标签上的得分
3. 根据得分选择合适的伪标签补充至标注数据
4. 新的标注数据更新教师模型
5. 重复2-4步直到满足退出条件

伪标签生成

动态全局状态标签：

- State = $\{-2, -1, 0, 1, 2\}^{N\times l}$，其中N为样本数，l为标签数
- 初始化State，样本标注标签为2，其他标签为0

双阈值策略：

- 正例阈值：0.6
- 负例阈值：0.4

教师模型打分：

- 为每个样本每个标签预测得分，不计算状态为-2或2的标签得分
- 得分大于正例阈值则标签状态加一
- 得分小于负例阈值则标签状态减一
- 得到新的State = $\{-2, -1, 0, 1, 2\}^{N\times l}$

伪标签选择

引入Partial Label：计算损失时忽略这部分标签的Loss

根据全局状态标签State = $\{-2, -1, 0, 1, 2\}^{N\times l}$确定数据的标签：

- 样本X，对应的状态标签为 $\{-2, -1, 0, 1, 2\}^{l}$
- 若标签y_i状态为2，则记y_i为样本X的正例标签
- 若标签y_i状态为-2，则记y_i为样本X的负例标签
- 若标签y_i状态为-1, 0, 1，则记y_i为样本X的Partial Label

新一轮训练

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

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

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

主要内容

- 任务定义
- 研究动机
- PST框架方法
- **实验结果**

实验: 数据集

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

真实数据CCKS-IMLTC

1. 人工补全测试集缺失的标签
2. 部分缺失严重标签记为Few标签

合成数据 AAPD

1. 按比例随机删除标签，模拟构建不完全标注数据。
2. 方案一：保证每条样本至少存在一个标签，删除标签比例上限为0.585。
3. 方案二：严格按删除比例丢失标签，同时无相关标签样本当纯负例训练。

实验: 评价指标

$$P = \frac{\text{预测正确的标签数}}{\text{预测的标签总数}}$$

$$R = \frac{\text{预测正确的标签数}}{\text{标注的标签总数}}$$

$$F1 = \frac{2*P*R}{P+R}$$

$$\text{退化率}: \alpha_p = \frac{f_0^a - f_p^a}{f_0^a}$$

$$\text{误导率}: \beta_p = \frac{f_p^a - f_p}{f_p^a}$$

实验: 不同分类模型PST框架前后性能对比

模型	Teacher			Self-Training			PST(Ours)				
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	Δ_T(F1)	Δ_{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

实验: Few标签上模型性能对比

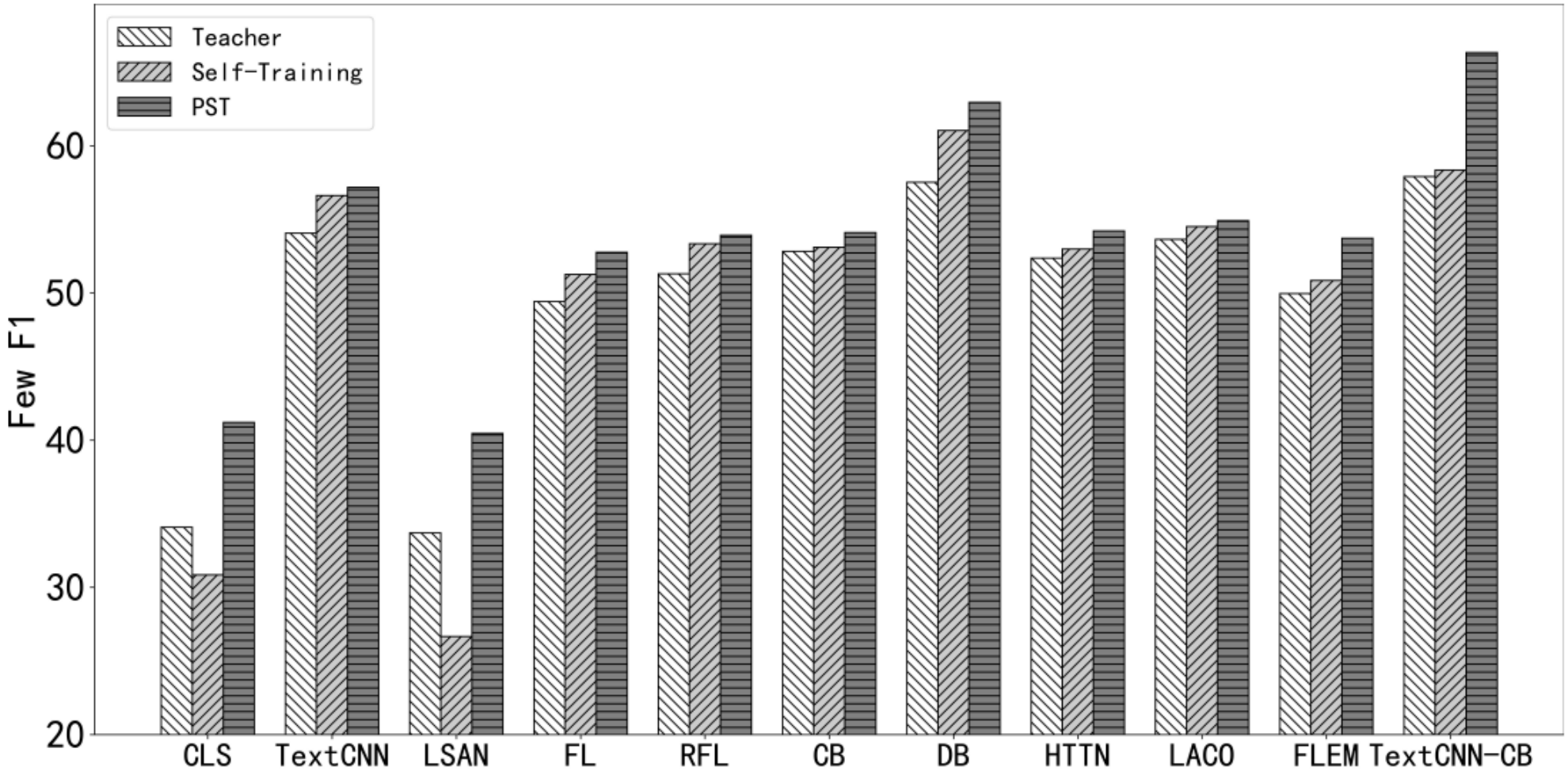
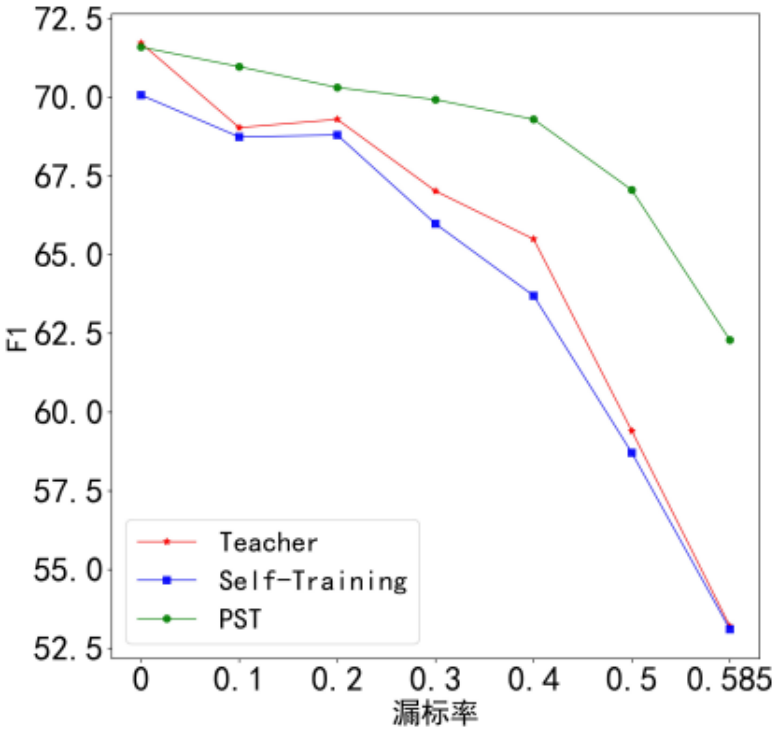
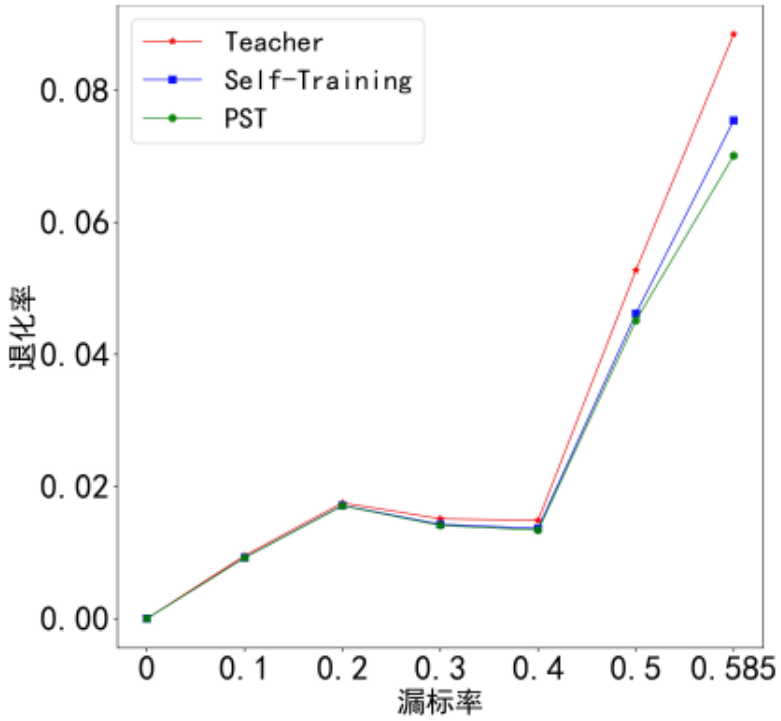


Figure 2: 不同模型在CCKS-IMLTC上Few标签的实验结果

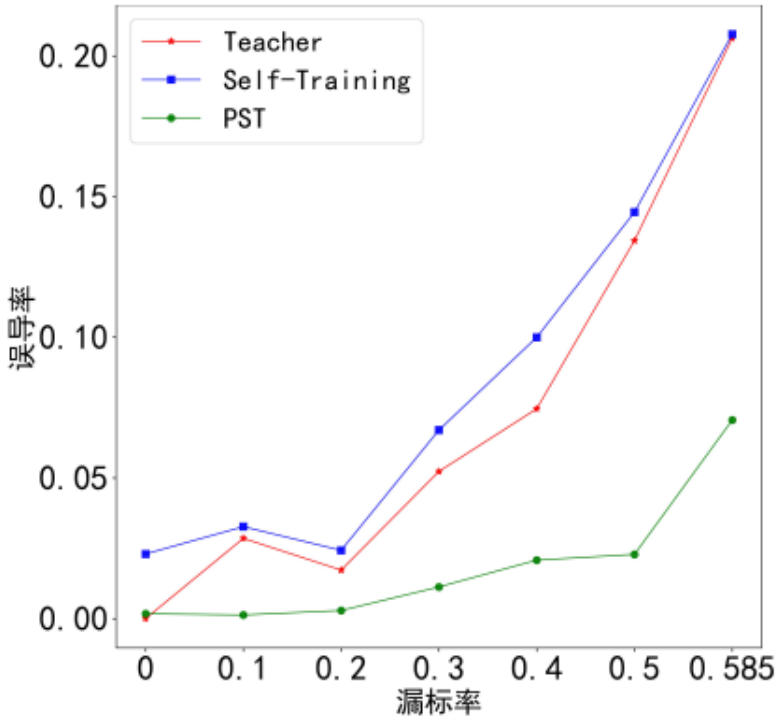
实验: 退化、缺失影响对比



(a) 方案一F1曲线图

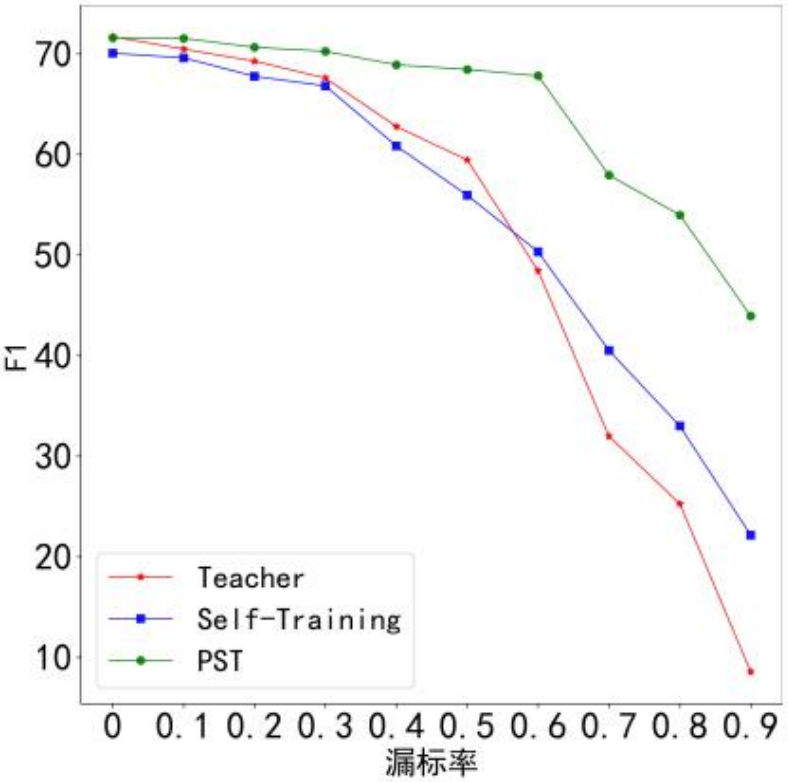


(b) 方案一退化率曲线图

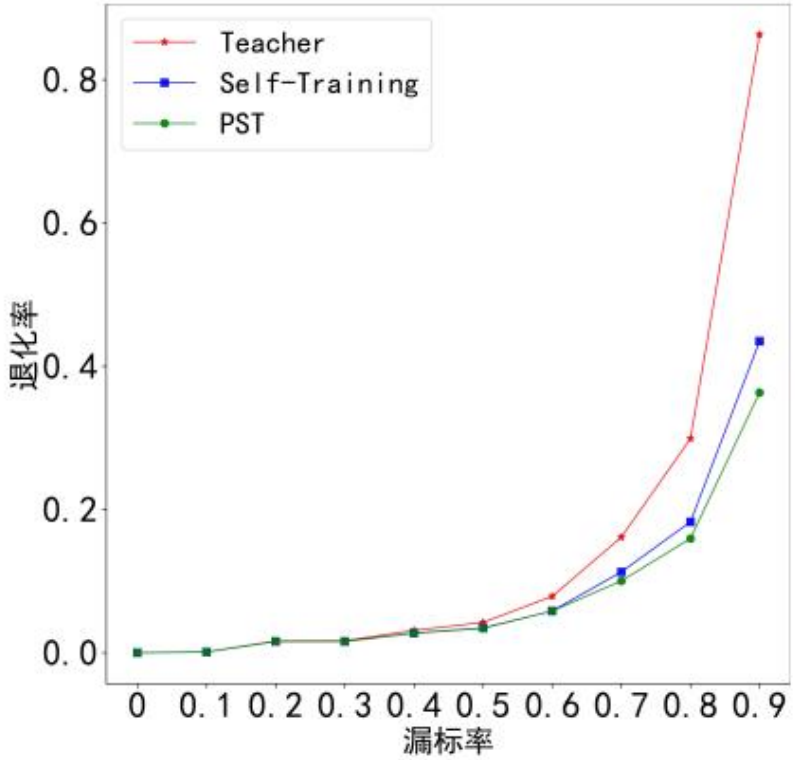


(c) 方案一误导率曲线图

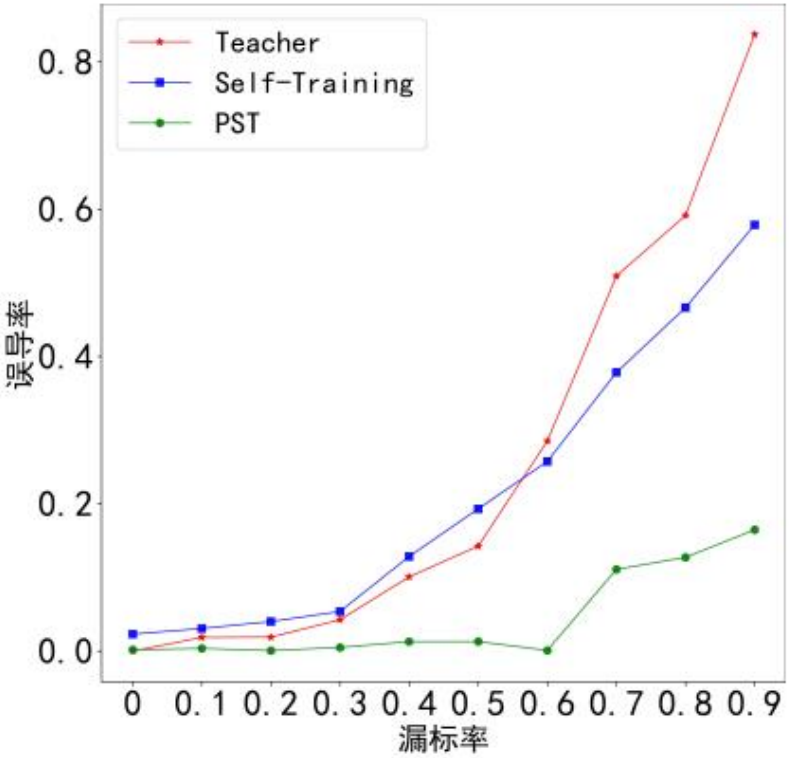
实验: 退化、缺失影响对比



(d) 方案二F1曲线图



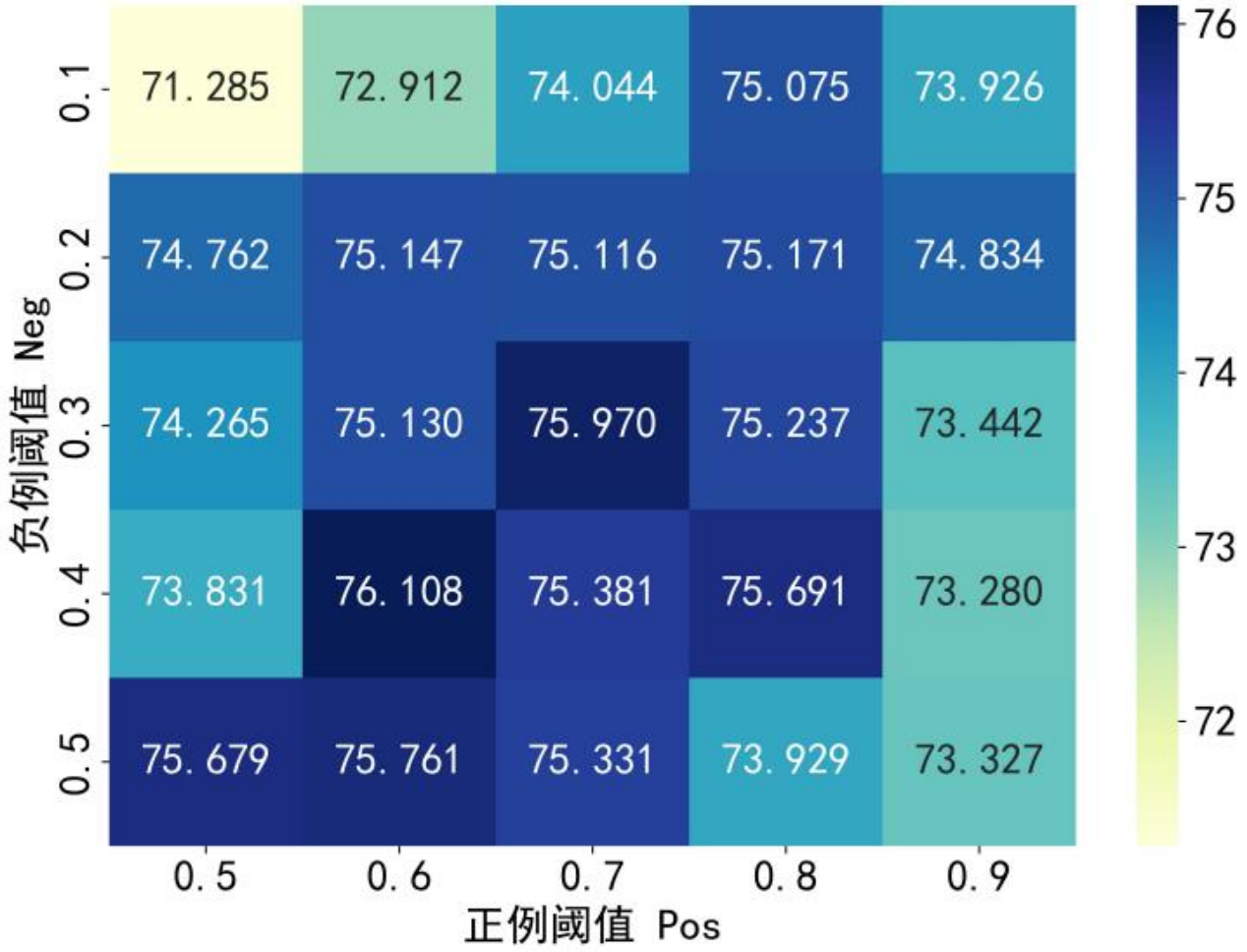
(e) 方案二退化率曲线图



(f) 方案二误导率曲线图

实验: 正负阈值选择实验

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



总结

- 提出一种面向不完全标注的多标签文本分类任务的通用框架
- 全面分析了不完全标注数据对不同分类模型的影响
- 在真实数据与合成数据上全面实验验证PST框架有效性与通用性
- 开源代码和数据集 https://github.com/15962171082/Incomplete_MLTC

Thanks

Q&A

任俊飞
jfrenjfren@stu.suda.edu.cn

https://github.com/15962171082/Incomplete_MLTC