TCS:一种用于跨语言文本标签预测的“老师-课程-学生”学习框架

浦 通[1] 黄书剑[1),2)] 张洋铭[3)] 周祥生[3)] 屠要峰[3)] 戴新宇[1)] 陈家骏[1)]

[1)](计算机软件新技术国家重点实验室(南京大学) 南京 210023)

[2)](鹏城实验室 广东 深圳 518054)

[3)](中兴通讯股份有限公司 南京 210012)

摘 要 跨语言迁移旨在借助源语言的标注样本学习目标语言上的相应任务,是解决目标语言标记数据不足的重要途径.近期表现出色的方法多基于自训练,通过逐步自动标记无标注样本实现知识的迁移.然而自训练存在不准确监督的问题,即当前模型(称为老师模型)对目标语言无标注样本的错误预测会误导后续模型(称为学生模型)的学习.跨语言迁移中,源语言和目标语言样本之间存在的分布差异加重了这个问题.本文提出一种名为“老师-课程-学生”(TCS)的学习框架,综合使用三项技术解决自训练中的不准确监督的问题,包括软目标训练技术、渐进式样本选择技术、“从可信到可疑”的课程学习技术等.在跨语言文本分类和跨语言命名实体识别基准数据集上的实验表明,TCS 取得的平均结果在自训练的基础上分别提高了 2.51%和 3.25%,并分别比现有最佳结果高 1.51%和 4.45%.消融实验表明,TCS 使用的三项技术都能有效提升最终模型的性能,其中课程学习技术和“从可信到可疑”的课程顺序是取得出色结果的关键.相关代码和实验配置可以在 https://github.com/ericput/TCS 获取.

关键词 跨语言迁移;自训练;课程学习;文本分类;命名实体识别

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2022.01983

TCS: A Teacher-Curriculum-Student Learning Framework for Cross-Lingual Text Labeling

PU Tong[1)] HUANG Shu-Jian[1),2)] ZHANG Yang-Ming[3)] ZHOU Xiang-Sheng[3)]
TU Yao-Feng[3)] DAI Xin-Yu[1)] CHEN Jia-Jun[1)]

[1)](*National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing* 210023)

[2)](*Peng Cheng Laboratory, Shenzhen, Guangdong* 518054)

[3)](*ZTE Corporation, Nanjing* 210012)

Abstract In recent years, deep learning models have greatly promoted the development of English and Chinese natural language processing. However, most other languages in the world are unable to perform effective text processing and analysis because of the difficulty in obtaining labeled data. Cross-lingual transfer is the main way to solve this problem by using labeled samples in the source language to make the model learn the corresponding tasks in the target language, so it has been widely concerned. Recently, some works based on self-training achieve the best results in the cross-lingual text labeling tasks by using both labeled samples in the source language and unlabeled samples in the target language to fine-tune the multilingual BERT. However, self-training suffers

收稿日期:2021-06-02;在线发布日期:2022-02-22.本课题得到国家自然科学基金(U1836221,6217020152)、中兴通讯科研合作项目资助.**浦 通**,硕士,主要研究方向为自然语言处理、迁移学习.E-mail:putong7@outlook.com.**黄书剑**(通信作者),博士,副教授,主要研究方向为机器翻译、自然语言生成.E-mail:huangsj@nju.edu.cn.**张洋铭**,学士,算法预研工程师,主要研究方向为自然语言处理、知识图谱、语音处理等.**周祥生**,学士,资深研发经理,在 AI 平台及算法交付方面有多个电信级交付项目经验,在自然语言处理、机器学习方向有深入研究.**屠要峰**,博士,中国计算机学会(CCF)高级会员,主要研究方向为大数据、数据库和机器学习.**戴新宇**,博士,教授,主要研究领域为自然语言处理、推荐系统.**陈家骏**,博士,教授,主要研究领域为自然语言处理.

from the problem of inaccurate supervision, that is, the teacher model's inaccurate predictions on the target unlabeled samples (i. e. inaccurate samples) may mislead the subsequent student model. And in cross-lingual transfer scenarios, the natural distribution gap between the source labeled samples and the target unlabeled samples will make this problem even worse. In order to further improve the results of cross-lingual text labeling, in this paper, we utilize three techniques to address the inaccurate supervision problem in the self-training, and propose a learning framework called Teacher-Curriculum-Student (TCS). Firstly, we employ a soft-target training technique to reduce the impact of inaccurate samples at the level of loss function. Secondly, we employ a progressive sample selection technique to construct iterative training datasets containing more accurate samples. Finally, in order to handle the inaccurate samples in iterative training datasets, we propose a from-confident-to-suspicious curriculum learning technique: according to prediction confidences of the teacher model, a training dataset is organized into learning courses arranged from confident to suspicious, so as to enhance the role of accurate samples and reduce the role of inaccurate samples in the student model's training process. Experiments on the benchmark datasets of cross-lingual text classification and cross-lingual named entity recognition show that the average results obtained by TCS are improved by 2.51% and 3.25% respectively on the basis of self-training thanks to the three techniques TCS used, and are 1.51% and 4.45% higher than existing state-of-the-art results respectively. In addition, ablation experiments show that: all three techniques used in TCS can effectively improve the performance of the final model and the curriculum learning technique contributes to the largest increase; the from-confident-to-suspicious curriculum order is the key to the effectiveness of the curriculum learning technique in the self-training scenario. More interestingly, further analysis shows that in the whole iterative training process, the effect of TCS is always better than that of self-training, the sample selection technique plays a role in the initial iteration, and the effect of the curriculum learning technique is mainly reflected in the middle and later iteration. For reproducibility, we release the code and experimental configurations at https://github.com/ericput/TCS.

Keywords cross-lingual transfer; self-training; curriculum learning; text classification; named entity recognition

1 引 言

随着不同国家、地区和民族之间交流的日益频繁,自动地处理分析其他语言的文本变得越来越重要.深度学习模型大大推动了各种自然语言处理任务的发展,但训练这些模型往往需要大量标注样本.标注样本往往只存于英语、汉语等使用人口众多的语言,世界上绝大多数语言通常只有少量甚至没有标注样本,这导致现有的深度学习模型难以应用于这些低资源语言.跨语言迁移(Cross-Lingual Transfer,CLT)期望通过在语言之间迁移任务知识来解决这个问题,即借助一种语言(称为源语言)的带标注样本来让模型学会另一种语言(称为目标语言)的相应任务.当目标语言完全没有标注样本时,我们将面临零样本跨语言迁移(Zero-shot CLT)问题,这是 CLT 相关研究中最普遍和最困难的问题,也是本文关注的场景(为了描述上的简洁,本文讨论的 CLT 都是针对零样本的).本文针对文本分类(Text Classification)和命名实体识别(Named Entity Recognition)这两种具有代表性的文本标签预测任务,聚焦于提升它们在 CLT 场景下的结果.

随着一些工作[1-2]发现在 104 种语言的维基语料上预训练的多语言 BERT 模型(Multilingual BERT,M-BERT)[3]有着出色的 CLT 能力,使用源语言的带标注样本微调 M-BERT 逐渐成为跨语言文本标签预测的主流方法.考虑到在实际场景中,目标语言的无标注样本易于收集且包含目标分布的相关信息,因此本文认为综合利用源语言带标注样本和目标语言无标注样本去微调 M-BERT 是提升跨

语言文本标签预测性能的有效途径.

近期的一些工作属于此研究范畴，其主要分为两类：一类通过对抗训练(Adversarial Training)减少 M-BERT 中不同语言表示之间的差异[4-5]；另一类使用自训练(Self-Training)让模型学习自身在目标语言无标注样本上的伪标签[6-7]，即使用当前训练好的模型作为老师来监督后续学生模型的训练. 自训练多次迭代训练“老师-学生”，相比于对抗训练，自训练能更直接有效地利用目标语言无标注样本，也有着更好的性能. 然而，自训练存在不准确监督的问题[6,8-9]，即老师模型对无标注样本的错误预测所产生的错误样本会误导学生模型的学习. CLT 场景下该问题会更加严重，这是因为 CLT 的标注样本来自源语言，无标注样本来自目标语言，两者之间由于语言不同而存在较大分布差异，这意味着此时错误样本在训练集中的占比更大.

为了在 CLT 场景更有效地利用目标语言无标注样本，本文基于自训练提出“老师-课程-学生”(Teacher-Curriculum-Student，TCS)学习框架(如图 1 所示). 具体来说，TCS 综合使用如下三项技术来减轻不准确监督问题：

(1) 软目标训练. 为了从损失函数层面减轻错误样本的影响，TCS 使用老师预测的标签分布而不是取整之后的 0/1 标签作为学生模型的学习目标，因为当预测错误时，标签分布比 0/1 标签包含更少的错误信号.

(2) 渐进式样本选择. 为了构建包含更多正确样本的迭代训练数据集，TCS 将老师模型对无标注样本的预测置信度作为无标注样本正确概率的度量，并在每轮迭代前选择一部分高置信度的无标注样本加入到训练集.

(3) “从可信到可疑”的课程学习. 然而样本选择也无法避免迭代训练数据集中出现错误样本，为了进一步处理这些错误样本，TCS 根据样本置信度将训练集组织成“从可信到可疑”的课程①，以此在学生模型的学习过程中提升正确样本的作用并降低错误样本的作用.

本文在跨语言文本分类基准数据集 MLDoc[10] 和跨语言命名实体识别基准数据集 CoNLL-2002[11]/2003[12] 上评估 TCS 学习框架的效果，实验结果显示：TCS 在两个数据集的共 10 个目标语言任务上都取得了新的最佳结果；TCS 在 MLDoc 上的平均准确率达到 88.10%，比自训练高 2.51%、比现有性能最佳的方法[6] 高 1.51%②；TCS 在 CoNLL 上的平均 $F1$ 值达到 80.45%，比自训练高 3.25%，比现有性能最佳的方法[7] 高 4.45%.

2　TCS 学习框架

在本节中，我们将详细描述 TCS 学习框架，如图 1 所示，TCS 通过不断迭代训练来获得 CLT 性能更好的模型，其通过在自训练框架中加入软目标训练、渐进式样本选择以及“从可信到可疑”的课程学习共三项技术来达到这一目标.

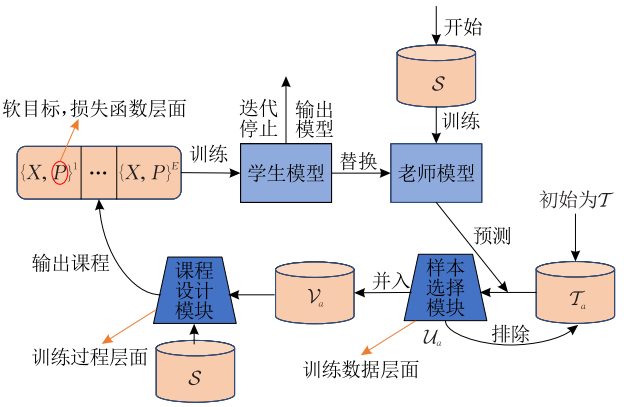

图 1　TCS 学习框架示意图(老师模型和学生模型都是相同结构的任务模型(见 2.1 节)，整体上通过自训练(见 2.2 节)迭代式地训练模型. $\mathcal{S}$ 表示全体源语言带标注样本，$\mathcal{T}$ 表示全体目标语言无标注样本. 每 a 轮迭代，样本选择模块(见 2.4 节)先从 $\mathcal{T}_{a-1}$ 中选取部分高置信度样本 $\mathcal{U}_a$ 加入到 $\mathcal{V}_a$，然后课程设计模块(见 2.5 节)处理 $\mathcal{S}$ 和 $\mathcal{V}_a$ 并输出训练课程，最后根据课程使用软目标(见 2.3 节)对学生模型进行训练)

2.1　任务模型

BERT(Bidirectional Encoder Representations from Transformers)[3] 是一种通过在大规模无标注语料上进行自监督预训练而得到的 Transformer[13] 表示模型，其大大推动了下游各项语言理解任务上的发展. M-BERT③ 是 BERT 的多语言版本，它的训练数据采样自 104 种语言的维基语料. M-BERT 可以在同一表示空间处理多种语言的文本，近期的一些工作[1-2] 发现 M-BERT 有着出色的 CLT 性能，在众多 CLT 任务上超越了之前的方法. 因此，本文通过在 M-BERT(记为 f_θ)上添加线性分类层(记为 $\langle W,b\rangle$)来构建任务模型 Θ，即 $\Theta=\{f_\theta,W,b\}$.

给定源语言带标注样本集合 $\mathcal{S}=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{N}$

① 以置信度最高的样本子集为始，在训练过程中逐渐加入置信度更低的样本子集.
② 基于本文复现的结果.
③ https://github.com/google-research/bert/blob/master/multilingual.md

(在文本分类中,y_i表示样本标签,在命名实体识别中,y_i表示 token 标签序列),我们可以使用$\mathcal{S}$训练模型得到初始老师模型Θ_0.对于文本分类:

$$\Theta_0=\arg\min_{\Theta}\sum_{x_i\in\mathcal{S}}\mathrm{CE}(p(\boldsymbol{x}_i,\Theta),y_i) \tag{1}$$

对于命名实体识别:

$$\Theta_0=\arg\min_{\Theta}\sum_{x_i\in\mathcal{S}}\frac{1}{L}\sum_{j=1}^{L}\mathrm{CE}(p(x_{ij},\Theta),y_{ij}) \tag{2}$$

其中函数 CE 表示交叉熵(Cross Entropy,CE),模型输出样本的标签分布$p(\boldsymbol{x},\Theta)\in\mathbb{R}^{|\mathcal{Y}|}$,$\mathcal{Y}$为标签集合.$\Theta_0$直接应用于目标语言任务就可以取得很好的结果[1,2,14].

2.2 自训练

TCS 是一种自训练类型的算法,其输入包括基于 M-BERT 的任务模型Θ、源语言带标注样本集合$\mathcal{S}=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{N}$和目标语言无标注样本集合$\mathcal{T}=\{(\boldsymbol{x}_i)\}_{i=1}^{M}$,其输出训练好的目标模型$\Theta_A$.

通常来说,自训练的第a轮迭代是指使用该轮训练集$\mathcal{D}_a$将任务模型Θ训练至收敛并输出该轮模型Θ_a,其可以写成如下形式:

$$\Theta_a=\arg\min_{\Theta}\sum_{x_i\in\mathcal{D}_a}\mathrm{LF}(p(\boldsymbol{x}_i,\Theta),p(\boldsymbol{x}_i,\Theta_{a-1})) \tag{3}$$

其中函数 LF 表示某种示损失函数(Loss Function,LF),Θ_{a-1}为老师模型,Θ为学生模型,学生模型学习老师模型在训练集$\mathcal{D}_a$的预测.当Θ_a相比于Θ_{a-1}的性能不再提高或a达到预设值,自训练便会停止.

对于无标注样本$\boldsymbol{x}_i$,我们假设其真实标签为y_i,当老师模型预测出错时,即$y_i\neq\arg\max p(\boldsymbol{x}_i,\Theta_{a-1})$,优化算法会用一个带错误标签的样本去更新学生模型$\Theta_a$,这就是自训练存在的不准确监督问题.接下来我们介绍三项技术来减轻这一问题.

2.3 软目标训练

经典的自训练[6,8]通常会把老师模型预测的标签分布取整转化为伪标签(硬目标),然后使用交叉熵作为损失函数.然而当老师模型预测出错时,硬目标会将标签分布中原来小于 1 的错误分量放大到 1,即放大不准确的监督信号.

受知识蒸馏相关工作[7,15]的启发,TCS 使用软目标(即标签分布)作为学生模型的学习目标并且使用均方误差(Mean Square Error,MSE)作为损失函数.对于文本分类,式(3)可以写成:

$$\Theta_a=\arg\min_{\Theta}\sum_{x_i\in\mathcal{D}_a}\mathrm{MSE}(p(\boldsymbol{x}_i,\Theta),p(\boldsymbol{x}_i,\Theta_{a-1})) \tag{4}$$

对于命名实体识别,式(3)可以写成:

$$\Theta_a=\arg\min_{\Theta}\sum_{x_i\in\mathcal{D}_a}\frac{1}{L}\sum_{j=1}^{1}\mathrm{MSE}(p(x_{ij},\Theta),p(x_{ij},\Theta_{a-1})) \tag{5}$$

2.4 样本选择方案

针对$\mathcal{T}$中存在错误样本的问题,一个自然的想法是从$\mathcal{T}$中筛选出正确的样本,接下来介绍一种渐进式的样本选择方案.

2.4.1 选择标准

考虑到$\mathcal{T}$中样本的真实标签未知,我们只能借助其他指标去估计样本为正确的概率.在分类任务中,样本预测置信度表示分类器对预测结果的信心,和预测是否准确有较强的相关性.对于文本分类,样本预测置信度可以直接定义为

$$c_i=\max p(\boldsymbol{x}_i,\Theta_{a-1}) \tag{6}$$

对于命名实体识别,其样本中所有 token 预测置信度的平均值可以反映样本整体的准确程度.考虑到算术平均值易受极端值的影响,本文使用的是几何平均数:

$$c_i=\sqrt[L]{\prod_{j=1}^{L}\max p(x_{ij},\Theta_{a-1})} \tag{7}$$

2.4.2 选择策略

考虑到简单的阈值筛选[8]存在阈值难以设置的问题,本文采用一种渐进式的选择策略[6]:每轮迭代从$\mathcal{T}$中选择置信度最高的一部分样本加入到训练集.

记第a轮迭代新选中的无标注样本集合为$\mathcal{U}_a$,已选中的无标注样本集合为$\mathcal{V}_a$,剩余的无标注样本集合为$\mathcal{T}_a$,该选择策略的形式化描述如下:

$$\begin{aligned}\mathcal{T}_a&=\mathcal{T}_{a-1}-\mathcal{U}_a,\\ \mathcal{V}_a&=\mathcal{V}_{a-1}\cup\mathcal{U}_a,\\ \mathcal{D}_a&=\mathcal{S}\cup\mathcal{V}_a\end{aligned} \tag{8}$$

对于文本分类,为了平衡$\mathcal{U}_a$中不同类别样本的数量,本文选择每种类别样本中置信度最高的K个:

$$\mathcal{U}_a=\bigcup_{\tilde{y}=1}^{|\mathcal{Y}|}\mathrm{sorted}(\mathcal{T}_{a-1}^{(\tilde{y})})[:K] \tag{9}$$

其中函数 sorted 的作用是按置信度从大到小对无标注样本进行排序,$\mathcal{T}_{a-1}^{(\tilde{y})}$表示$\mathcal{T}_{a-1}$中老师预测标签为$\tilde{y}$的样本集合.对于命名实体识别,一则样本往往包含多种实体标签,因此本文未采用平衡选择策略,直接选择$\mathcal{T}_{a-1}$中置信度最高的K个样本:

$$\mathcal{U}_a=\mathrm{sorted}(\mathcal{T}_{a-1})[:K] \tag{10}$$

2.5 课程学习方案

2.5.1 动机

由于$\mathcal{T}$中样本的真实标签未知,我们无法通过样本选择技术来确保$\mathcal{D}_a$中只包含正确样本.而且为了尽可能保留$\mathcal{T}$中的正确样本以确保模型获取更完

整的目标分布信息，我们通常需要保留$\mathcal{T}$的多数样本，这也意味着相当数量的错误样本也一同被保留. 2.4 节的样本选择方案表现出的具体问题为：$\mathcal{D}_a$中错误样本数量会随着迭代的进行而逐步上升. 因此在样本选择之外，我们还需要其他技术来减轻$\mathcal{D}_a$中错误样本对模型Θ_a训练的影响.

课程学习相关工作[16-17]表明按“从简单到困难”的策略组织训练样本可以使模型收敛得更快更好. 在我们的问题中，正确样本可以看作是模型易于建模的样本，它们的置信度通常较高；错误样本可以认为是模型难以建模的样本，它们的置信度往往较低. 因此本文作出如下假设：

假设 1.　按“从可信到可疑”的策略组织$\mathcal{D}_a$可以训练出性能更好的Θ_a.

2.5.2　数据分片

为了实现“从可信到可疑”的课程，本文先将全集$\mathcal{V}_a$划分成B个正确程度从高到低排列的子集序列$\{\mathcal{V}_{a,b}\}_{b=1}^{B}$. 考虑到常用的均匀划分①不能动态适应$\mathcal{V}_a$中样本置信度分布，容易把置信度相差不大的样本分到不同子集，进而导致子集之前正确程度相差不大. 受 Zhang 等人[18]的启发，本文使用 Jenks 自然划分算法(Jenks Natural Breaks，JNB)[19]，该算法通过最小化子集内部的距离并最大化子集之间的距离来寻找全集的最优分割点. 对于文本分类，我们同样需要平衡$\mathcal{V}_{a,b}$中不同类别样本的数量，因此数据分片方法设计如下：

$$\{\mathcal{V}_{a,b}^{(\bar{y})}\}_{b=1}^{B}=\mathrm{JNB}(\mathcal{V}_a^{(\bar{y})},B)\ \text{s.t.}\ \min(\mathcal{V}_{a,b}^{(\bar{y})})>\max(\mathcal{V}_{a,b+1}^{(\bar{y})})$$

$$\mathcal{V}_{a,b}=\bigcup_{\bar{y}=1}^{|\mathcal{Y}|}\mathcal{V}_{a,b}^{\bar{y}} \tag{11}$$

其中函数 min 和 max 分别表示取集合的最小和最大置信度. 对于命名实体识别，数据分片方法设计如下：

$$\{\mathcal{V}_{a,b}\}_{b=1}^{B}=\mathrm{JNB}(\mathcal{V}_a,B)\ \text{s.t.}\ \min(\mathcal{V}_{a,b})>\max(\mathcal{V}_{a,b+1}) \tag{12}$$

2.5.3　课程安排

在常规深度模型训练方案中，Θ_a需要在$\mathcal{D}_a$上重复训练E个时期(Epoch)，本文在此基础上进行拓展，通过为每个Epoch设置不同的训练数据来实现课程学习. 记第e个 Epoch 的训练数据为$\mathcal{D}_{a,e}$，本文按照“从可信到可疑”的策略设计如下课程安排方案：

$$\mathcal{D}_{a,e}=\mathcal{S}\cup\bigcup_{b=1}^{\min(e,B)}\mathcal{V}_{a,b},\ e\in[1,E] \tag{13}$$

与常规深度模型训练方案一致，训练集$\mathcal{D}_{a,e}$中的样本是无序的，即优化器从$\mathcal{D}_{a,e}$中随机采样数据并按式(4)或式(5)对Θ_a进行优化.

2.6　框架小结

事实上，如果从课程学习角度来看，2.4 节的样本选择方案可以看做是迭代间的课程学习方案，2.5 节则是迭代内的课程学习方案，因此“课程”是 TCS 的核心，也是 TCS 和其他自训练类型的方法的主要区别.

综上，用于跨语言文本分类的 TCS 学习框架如算法 1 所示，用于跨语言命名实体识别的 TCS 学习框架如算法 2 所示. 算法输出了每一轮训练好的学生模型，本文只使用最后一轮学生模型作为目标模型Θ_A来进行目标语言任务上的推理和预测. 这意味着 TCS 虽然在训练过程中涉及多个模型，但在推理时只涉及一个模型.

算法 1.　用于跨语言文本分类的 TCS 算法.

输入：M-BERT f_θ，源语言带标注样本集合$\mathcal{S}$，目标语言无标注样本集合$\mathcal{T}$

输出：目标模型Θ_A

1. $a=0$，$\mathcal{T}_0=\mathcal{T}$，$\mathcal{V}_0=\varnothing$
2. REPEAT
3.　初始化本轮模型Θ_a：加载f_θ并随机初始化$\langle W,b\rangle$
4.　IF $a==0$ THEN
5.　按式(1)在$\mathcal{S}$上微调Θ_a E 个 Epoch
6.　ELSE
7.　使用式(8)和式(9)计算$\mathcal{T}_a$和$\mathcal{V}_a$
8.　使用式(11)计算$\{\mathcal{V}_{a,b}\}_{b=1}^{B}$
9.　FOR $e=1$ to E DO
10.　使用式(13)计算$\mathcal{D}_{a,e}$
11.　按式(4)在$\mathcal{D}_{a,e}$上微调Θ_a 1 个 Epoch
12.　END FOR
13.　END IF
14.　输出Θ_a
15.　$a=a+1$
16. UNTIL Θ_a性能不再提高或达到最大迭代轮次

算法 2.　用于跨语言命名实体识别的TCS算法.

输入：M-BERT f_θ，源语言带标注样本集合$\mathcal{S}$，目标语言无标注样本集合$\mathcal{T}$

输出：目标模型Θ_A

1. $a=0$，$\mathcal{T}_0=\mathcal{T}$，$\mathcal{V}_0=\varnothing$
2. REPEAT
3.　初始化本轮模型Θ_a：加载f_θ并随机初始化$\langle W,b\rangle$
4.　IF $a==0$ THEN
5.　按式(2)在$\mathcal{S}$上微调Θ_a E 个 Epoch
6.　ELSE
7.　使用式(8)和式(10)计算$\mathcal{T}_a$和$\mathcal{V}_a$

① 先进行置信度排序，再通过保子集样本数量相等来选取分割点.

8. 使用式(12)计算$\{\mathcal{V}_{a,b}\}_{b=1}^{B}$
9. FOR $e=1$ to E DO
10. 使用式(13)计算$\mathcal{D}_{a,e}$
11. 按式(5)在$\mathcal{D}_{a,e}$上微调Θ_a 1 个 Epoch
12. END FOR
13. END IF
14. 输出Θ_a
15. $a=a+1$
16. UNTIL Θ_a性能不再提高或达到最大迭代轮次

2.7 算法分析

接下来,我们对算法 1 和 2 的复杂度进行简单分析.考虑到它们是深度模型训练算法,因此本文选取模型的更新步(Step)为时间复杂度分析中的基本操作单元.至于空间复杂度,本文考虑算法运行过程中内存和显存的占用.需要说明的是,每轮迭代前,老师模型的预测结果是被预先计算并保存的,因此在学生模型的训练过程中,老师模型无需保存在内存和显存中进行实时计算.算法 1 和 2 的时间复杂度为$O(A\times E\times(|\mathcal{S}|+|\mathcal{T}|)/C)$,其中$A$为最大迭代次数,$E$为每次迭代训练的 Epoch 数,$C$为批大小(Batch Size),$|\mathcal{S}|$和$|\mathcal{T}|$分别是源语言带标注样本数量以及目标语言无标注样本数量.相比常规的深度模型训练方法,算法 1 和 2 因为需要迭代A次从而时间复杂度更高.算法 1 和 2 的空间复杂度与常规的深度模型训练方法是一致的,即只与待优化的模型以及所使用的梯度下降算法有关.

3 实 验

本文在跨语言文本分类和跨语言命名实体识别的基准数据集上进行了详细实验来评估和分析 TCS 的性能.

3.1 实验设置

在本节中,我们从数据集、评价指标以及模型和训练细节 3 个方面介绍实验设置.

3.1.1 数据集

对于跨语言文本分类,本文使用的是包含英语(en)、德语(de)、西班牙语(es)、法语(fr)、意大利语(it)、日语(ja)、俄语(ru)和汉语(zh)的多语言文档分类数据集 MLDoc[10].MLDoc 是基于路透社 RCV1 和 RCV2 新闻语料①构建的类别平衡数据集,它的每种语言数据都拥有相同的类别标签:CCAT(公司/产业)、ECAT(经济)、GCAT(政府/社会)和 MCAT(市场),且被分为训练集、验证集和测试集.每种语言的验证集样本数量都为 4000,测试集样本数量都为 1000.每种语言的训练集都有 4 种配置,分别是 *language.train.*1000、*language.train.*2000、*language.train.*5000 和 *language.train.*10000,总体上每种配置的训练集的样本数量等于其尾标,但下列配置会缺少一些样本:*spanish.train.*10000 样本数量为 9458;*russian.train.*10000 样本数量为 5216.遵循之前工作[4-5]的设置,本文使用英语作为源语言,本文使用英语作为源语言,其余 7 种语言作为目标语言;使用 *english.train.*1000 作为$\mathcal{S}$,使用去掉标签的 *target_language.train.*10000 作为$\mathcal{T}$.

对于跨语言命名实体识别,本文使用的是包含西班牙语(es)和荷兰语(nl)的 CoNLL-2002 数据集[11]以及包含英语(de)和德语(de)的 CoNLL-2003 数据集[12].CoNLL 的每种语言数据拥有相同的实体标签:PER(人物)、LOC(地点)、ORG(组织)和 MISC(杂项),且被分为训练集、验证集和测试集.CoNLL 数据集的具体统计信息如表 1 所示.遵循之前工作[1,4-5,7]的设置,本文使用 BIO 实体标签格式;使用英语作为源语言,其余 3 种语言作为目标语言;使用英语训练集作为$\mathcal{S}$,使用去掉标签的目标语言训练集作为$\mathcal{T}$.

表 1 CoNLL 数据集统计信息

语言	种类	训练集	验证集	测试集
英语-en	句子	14987	3466	3684
(CoNLL-2003)	实体	23499	5942	5648
德语-de	句子	12705	3068	3160
(CoNLL-2003)	实体	11851	4833	3673
西班牙语-es	句子	8323	1915	1517
(CoNLL-2002)	实体	18798	4351	3558
荷兰语-nl	句子	15806	2895	5195
(CoNLL-2002)	实体	13344	2616	3941

3.1.2 评价标准

CLT 任务的核心评价标准是最终模型在目标语言测试集上的性能.具体来说,本文使用目标语言测试集的准确率(Accuracy,Acc)作为 MLDoc 实验的性能指标,使用目标语言测试集实体短语级别 $F1$(Phrase-based $F1$)作为 CoNLL 实验的性能指标.

3.1.3 训练细节

在训练 MLDoc 和 CoNLL 上的模型时,本文大部分遵循 Wu 和 Dredze[1]的超参设置:M-BERT 的词嵌入层和底部共 3 层在训练时固定;输入样本的最大长度为 128 tokens②;Batch Size 为 32;Epoch

① http://trec.nist.gov/data/reuters/reuters.html
② MLDoc 样本超出最大长度部分将被舍弃;使用大小为最大长度的滑动窗口处理 CoNLL 样本.

数量为 5；优化器使用 AdamW[20] 且 Weight Decay 为 0.01；学习率为 3e-5 且在前 10% 训练步进行学习率线性热身(Linear Warmup)；Dropout 概率为 0.1.

对于 TCS 涉及的超参，本文基于目标语言验证集进行经验性地设置：算法停止条件统一设为迭代 10 次，即 $A=10$；数据分片超参 B 统一设为 4；对于 MLDoc 中除 ru 的每种目标语言，样本选择超参 K 统一设为 500，因为实际上 *russian.train.*10000 的样本数只有 5216，所以设置 $K_{ru}=250$；对于 CoNLL 的每种目标语言，样本选择超参分别设置为 $K_{de}=6000$、$K_{es}=2000$ 和 $K_{nl}=4000$.

本文基于 PyTorch-1.3.1 框架实现所有模型和算法并进行实验，其中 Jenks 划分算法使用 jenkspy 包①实现.

3.2　性能对比与分析

在本节中，我们分别在 MLDoc 和 CoNLL 上对比已有 CLT 方法和 TCS 的性能，这些方法包括：

(1) 未使用 M-BERT 的方法：在 MLDoc 上，Schwenk 和 Li[10] 通过跨语言词向量和卷积神经网络进行模型迁移；Artetxe 和 Schwenk[21] 预训练一个多语言神经翻译模型并通过翻译模型的编码器进行模型迁移. 在 CoNLL 上，Ni 等人[22] 综合使用了基于平行语料的标注映射和跨语言词向量；Xie 等人[23] 使用跨语言词向量进行词翻译并以此进行标注映射.

(2) 只使用源语言带标注样本微调 M-BERT，记为 Base.

(3) 通过对抗训练使用目标语言无标注样本提升 Base 的方法，包括 Keung 等人[4] 和 Zhang 等人[5].

(4) 通过自训练使用目标语言无标注样本提升 Base 的方法：在 MLDoc 上，Dong 等人[6] 使用带样本选择机制的自训练来综合利用源语言带标注样本和目标语言无标注样本，由于原论文的结果是在 *target_language.train.*1000 上得到的，我们在本文的实验设置下复现了该方法并汇报复现的结果；在 CoNLL 上，Wu 等人[7] 先在源语言带标注样本上微调 M-BERT 得到老师模型，然后在目标语言无标注样本上蒸馏老师模型，该方法可以看成是使用软目标训练的单轮自训练.

此外，为了探究 TCS 每个部分的效果，我们还进行了如下消融实验：

(1) ST(Self-Training). Base 加上最基础的自训练.

(2) SD(Self-Distillation). ST 加上软目标训练.

(3) SDS(SD with Selection). SD 加上渐进式样本选择.

(4) TCS(Teacher-Curriculum-Student). SDS 加上式(13)生成的“从可信到可疑”的课程.

进一步，为了验证假设 1 中“从可信到可疑”课程顺序的必要性，我们还设置了使用了逆课程(Reversed Curriculum)的对照实验 SDS+RC. 逆课程，即“从可疑到可信”的课程，其安排方案如下式所示：

$$\mathcal{D}_{a,e}=\mathcal{S}\cup\bigcup_{b=1}^{\max(1,B-e+1)}\mathcal{V}_{a,b},\ e\in[1,E] \tag{14}$$

对比式(13)和(14)，易知 TCS 和 SDS+RC 有着相反的课程顺序但一致的课程内容.

3.2.1　MLDoc 上的结果

结果如表 2 所示，TCS 在 MLDoc 的 7 种目标语言上都显著超越了 ST 和现有方法，比 ST 平均高 2.16 *Acc*，比 Dong 等人[6] 的方法平均高 1.31 *Acc*，取得了新的最佳结果.

表 2　各方法在 MLDoc 测试集上 *Acc*(%)的对比(粗体表示该目标语言上的最佳结果)

	de	es	fr	it	ja	ru	zh	平均
Schwenk 和 Li[10]	81.20	72.50	72.40	69.40	67.60	60.80	74.70	71.23
Artetxe 和 Schwenk[21]	84.80	77.30	77.90	69.40	60.30	67.80	71.90	72.77
Base	86.50	78.15	81.60	67.90	76.83	70.83	80.63	77.49
Keung 等人[4]	88.10	80.80	85.70	72.30	76.80	77.40	84.70	80.83
Zhang 等人[5]	91.90	87.20	87.90	77.90	77.10	70.10	87.50	82.80
Dong 等人[6]	94.63	89.53	93.08	80.55	79.65	82.03	88.08	86.79
ST	92.33	89.78	91.13	79.50	79.98	80.60	88.28	85.94
SD	93.50	89.73	92.15	80.33	80.73	80.65	88.10	86.46
SDS	**95.20**	89.28	92.88	**81.58**	80.83	82.20	88.08	87.15
SDS+RC	95.13	87.50	92.85	80.05	80.93	82.18	87.78	86.63
TCS	95.18	**92.35**	**94.33**	81.35	**82.03**	**82.65**	**88.83**	**88.10**

① https://pypi.org/project/jenkspy

3.2.2 CoNLL 上的结果

结果如表 3 所示，TCS 在 CoNLL 的 3 种目标语言上都显著超越了 ST 和现有方法，比 ST 平均高 2.53 $F1$，比 Wu 等人[7]的方法平均高 3.43 $F1$，取得了新的最佳结果.

表 3 各方法在 CoNLL 测试集上 $F1$(%)的对比（粗体表示该目标语言上的最佳结果）

	de	es	nl	平均
Ni 等人[22]	65.10	65.40	58.50	63.00
Xie 等人[23]	72.37	71.25	57.76	67.13
Base	70.00	75.10	80.39	75.16
Keung 等人[4]	71.90	74.30	77.60	74.60
Zhang 等人[5]	72.10	75.00	79.40	75.50
Wu 等人[7]	73.22	76.94	80.89	77.02
ST	73.39	78.5	81.88	77.92
SD	75.05	79.11	82.56	78.91
SDS	75.18	79.73	83.74	79.55
SDS+RC	74.95	79.87	84.12	79.65
TCS	**76.22**	**80.56**	**84.57**	**80.45**

3.2.3 结果分析

Base 的性能显著优于 Schwenk 和 Li[10]、Artetxe 和 Schwenk[21]、Ni 等人[22]和 Xie 等人[23]的方法，这验证了基于 M-BERT 的模型迁移性能优于基于跨语言词向量或标注映射的方法，同时也表明本文使用的基线是可靠的. 相比 Base，ST 在 MLDoc 上平均提升 8.45 Acc，在 CoNLL 上平均提升 2.76 $F1$，这说明在微调 M-BERT 的过程中通过自训练引入目标语言无标注样本信息可以大大提升 M-BERT 的 CLT 性能. 相比 Keung 等人[4]和 Zhang 等人[5]的方法，ST 在 MLDoc 上平均高 5.11 Acc 和 3.14 Acc，在 CoNLL 上平均高 3.32 $F1$ 和 2.42 $F1$，这说明自训练方案优于现有的对抗训练方案.

在 MLDoc 上，SD 比 ST 平均高 0.52 Acc、SDS 比 SD 平均高 0.69 Acc、TCS 比 SDS 平均高 0.95 Acc；在 CoNLL 上，SD 比 ST 平均高 0.99 $F1$、SDS 比 SD 平均高 0.64 $F1$、TCS 比 SDS 平均高 0.90 $F1$. 这些对比结果说明 TCS 使用的软目标训练、渐进式样本选择和“从可信到可疑”的课程学习都是有效的. 软目标训练在 CoNLL 上比在 MLDoc 上有更显著的效果，这可能是因为硬目标的标签分布错误放大作用会随着标签数变多而更严重(CoNLL 数据集经 BIO 编码，共 9 个标签；MLDoc 数据集共 4 个标签). “从可信到可疑”的课程学习在 MLDoc 和 CoNLL 上的效果都很显著，而且综合来看是三项技术方案中效果最好的.

相比无课程(或随机课程顺序)的 SDS，使用“从可信到可疑”课程顺序的 TCS 在 MLDoc 和 CoNLL 上都有显著提升；而使用“从可疑到可信”课程顺序的 SDS+RC 在 CoNLL 上无明显提升，在 MLDoc 上甚至出现明显下降. 这个结果说明“从可信到可疑”的课程顺序在提升模型性能上起到关键作用，同时也说明模型训练初期的样本质量更为重要.

3.3 实验分析

在 3.3.1～3.3.3 节中，我们基于 MLDoc 中英语迁移到法语(记为 MLDoc-fr)和 CoNLL 中英语迁移到德语(记为 CoNLL-de)这两个子任务上的实验现象来进一步分析 TCS 的有效性以及验证技术方案的动机. 在 3.3.4 节，我们基于 MLDoc 简单讨论 TCS 在不同语言任务上效果的差异.

3.3.1 模型性能随迭代的变化

在本节中，我们通过分析 ST、SD、SDS 和 TCS 在运行过程中模型性能随迭代的变化来说明 TCS 相比于其它方法的优势. 图 2 体现了各方法中学生模型在 MLDoc-fr 测试集上的性能(即泛化性能)随迭代的变化情况. 由于在每轮迭代中，学生模型的性能受老师模型在目标无标注样本(在本文的实验中就是去掉标签的目标训练集)上的性能影响，本节使用图 3 来体现各方法中老师模型在 MLDoc-fr 训练集上的性能随迭代的变化情况. 类似的，图 4 和图 5 则体现 CoNLL-de 的情况. 图 2～图 5 中，图例的名称对应相应的方法的名称，需要特别说明图 3 和图 5 的图例含义：图例 ST、SD、SDS 和 TCS 表示对应方法在$\mathcal{T}$上的性能，图例 SDS-V 和 TCS-V 表示对应方法在选中的无标注样本集合$\mathcal{V}_a$上的性能.

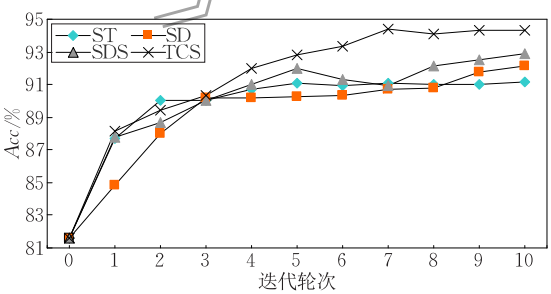

图 2 学生模型在 MLDoc-fr 测试集上的性能随迭代的变化

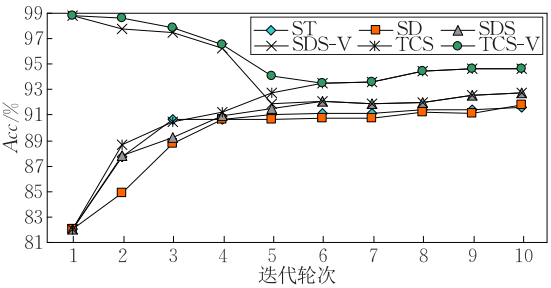

图 3 老师模型在 MLDoc-fr 训练集上的性能随迭代的变化

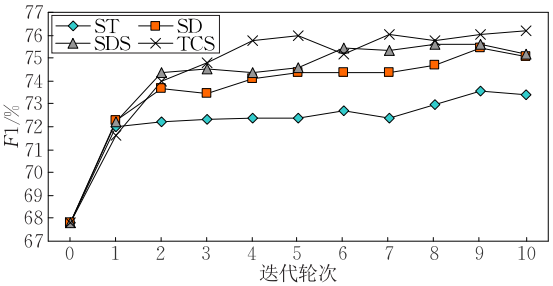

图 4　学生模型在 CoNLL-de 测试集上的性能随迭代的变化

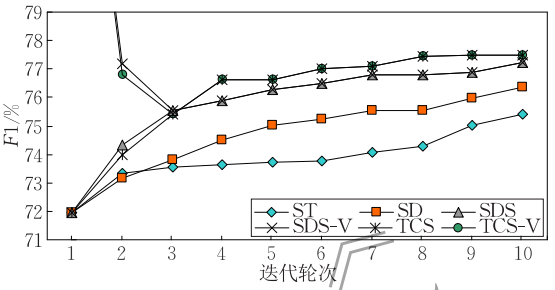

图 5　老师模型在 CoNLL-de 训练集上的性能随迭代的变化

如图 2 和图 4 所示，4 种方法中的学生模型的泛化性能会随着迭代而逐步提高并最终收敛；同时图 3 和图 5 表明，随着迭代的进行，4 种方法中的老师模型在$\mathcal{T}$上的性能也在逐步提高并最终收敛. 这说明在一定范围内，自训练类型算法会随着迭代不断提升目标无标注样本的标注质量从而不断提升学生模型的泛化性能. 多次迭代可以显著提升自训练类型算法的性能，这也是只进行单轮迭代的 Wu 等人[7]的性能结果不如方法 ST 的原因. 此外，图中的折线变化情况也表明 TCS 使用的三项技术方案都有效地扩大了这个增长范围，最终获得了更好的结果.

如图 3 和图 5 所示，在前期的迭代中，折线 SDS-V 和 TCS-V 分别大幅高于折线 SDS 和 TCS，这说明样本选择方案成功地从全集$\mathcal{T}$中选出平均标注质量更高的子集$\mathcal{V}_a$. 但如图 2 和图 4 所示，最初两轮迭代，方法 SDS 的泛化性能并不如 ST 或 SD；之后的迭代中，方法 SDS 的泛化性能才超越了 ST 和 SD. 本文认为造成该现象的原因是：最初的两轮迭代，虽然$\mathcal{V}_a$的平均标注质量很高，但规模太小，提供的目标分布信息严重不足；而随着迭代的进行，$\mathcal{V}_a$规模变大，虽然其平均标注质量出现了下降，但能提供的目标分布信息在逐渐增加. 该现象说明样本选择存在选择数量和选择质量的权衡，相比于简单地设置置信度阈值，本文使用的动态选择策略随着迭代由侧重质量到侧重数量，更好地兼顾了该权衡.

如图 2～图 5 所示，折线 TCS 在前期的迭代中和 SDS 十分接近，直到中后期折线 TCS 才显著高于 SDS，且折线 TCS-V 和 SDS-V 的情况也相同. 这一现象说明课程学习方案在$\mathcal{V}_a$中错误样本数量足够多时起到了显著效果，支持了本文引入课程学习的动机.

3.3.2　数据分片方法对比

在本节中，我们通过对比 Jenks 划分和均匀划分①这两种数据分片方法的效果来说明 Jenks 划分的优势.

首先对比使用上述两种数据分片方法的课程学习方案的性能. TCS 表示使用 Jenks 划分的课程学习方案，而 TCS-BS 表示使用均匀划分的课程学习方案，*-test 表示该方法中学生模型在目标测试集上的性能，*-train 表示该方法中老师模型在$\mathcal{V}_a$上(见式(8))的性能. 结果如图 6 和图 7 所示，在整个迭代过程中，TCS 的泛化性能要一直强于 TCS-BS，TCS 在$\mathcal{V}_a$上的性能总体上也要优于 TCS-BS.

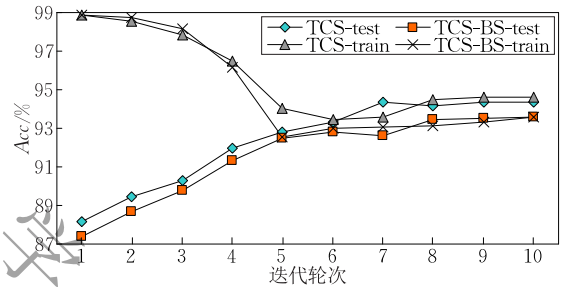

图 6　TCS 和 TCS-BS 在 MLDoc-fr 上的性能

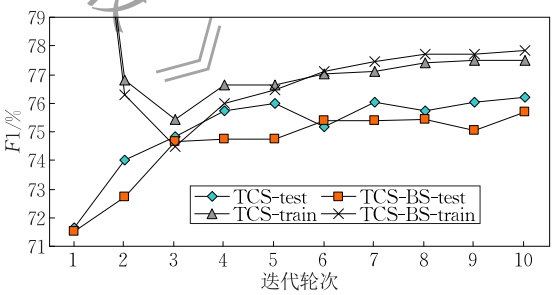

图 7　TCS 和 TCS-BS 在 CoNLL-de 上的性能

接下来我们以 TCS 实验和 TCS-BS 实验的第 4 轮迭代为例，统计并分析 Jenks 划分和均匀划分在 MLDoc-fr 和 CoNLL-de 上具体的划分结果. 表 4 和表 5 中，$\mathcal{V}_{4,1}$是在全集$\mathcal{V}_4$上划分出的第 1 子集，“占比”指该集合占全集的比例，“置信度”指老师模型在该集合上的平均预测置信度，“质量”指老师模型在该集合上的平均性能. 结果显示，两种划分方法

① 与 Jenks 划分保持一致，均匀划分在文本分类任务上也使用了类别均衡策略.

都成功地根据置信度将$\mathcal{V}_4$划分成质量不同的子集，并且排在前面的子集的质量要高于全集的质量. Jenks 划分生成的子集序列满足质量从高到低的要求；但均匀划分在 MLDoc-fr 和 CoNLL-de 上都出现了第 1 子集质量低于第 2 子集的现象，这使得后续的课程规划无法满足“从可信到可疑”的课程顺序. Jenks 划分生成的子集之间的差异比均匀划分的大得多：在 MLDoc-fr 上，两者的子集置信度标准差分别为 0.013 和 0.006，子集质量标准差分别为 4.34 和 2.26；在 CoNLL-de 上，两者的子集置信度标准差为分别为 0.158 和 0.016，子集质量标准差分别为 16.94 和 6.83. 这个结果说明 Jenks 划分的样本划分效果更好.

表 4 实验的第 4 轮迭代，MLDoc-fr 上的数据分片结果

	TCS			TCS-BS		
	占比/%	置信度	质量	占比/%	置信度	质量
$\mathcal{V}_4$	100.00	0.9875	96.49	100.00	0.9823	96.16
$\mathcal{V}_{4,1}$	67.80	0.9915	98.36	25.00	0.9987	97.85
$\mathcal{V}_{4,2}$	21.79	0.9834	94.66	25.00	0.9865	98.20
$\mathcal{V}_{4,3}$	8.38	0.9740	88.21	25.00	0.9834	96.70
$\mathcal{V}_{4,4}$	2.04	0.9537	87.73	25.00	0.9707	91.90

表 5 实验的第 4 轮迭代，CoNLL-de 上的数据分片结果

	TCS			TCS-BS		
	占比/%	置信度	质量	占比/%	置信度	质量
$\mathcal{V}_4$	100.00	0.9894	76.62	100.00	0.9868	75.99
$\mathcal{V}_{4,1}$	92.63	0.9957	79.63	25.00	0.9984	78.70
$\mathcal{V}_{4,2}$	6.22	0.9381	60.64	25.00	0.9978	87.71
$\mathcal{V}_{4,3}$	0.94	0.7921	40.67	25.00	0.9957	82.93
$\mathcal{V}_{4,4}$	0.20	0.5800	40.00	25.00	0.9552	67.42

3.3.3 无标注样本规模的影响

在本节中，我们基于 MLDoc-fr 探究目标语言无标注样本$\mathcal{T}$的规模对 TCS 效果的影响. 我们分别将去掉标签的 *french.train.*1000、*french.train.*2000、*french.train.*5000 和 *french.train.*10000 设置为$\mathcal{T}$并进行实验，实验的模型和超参除样本选择参数 K 之外保持与 3.1.3 节一致，样本选择参数按照$\mathcal{T}$的规模进行放缩，分别为 50、100、250 和 500.

图 8 展示的是在各种规模的$\mathcal{T}$下，学生模型在目标测试集上的性能随迭代的变化情况. 在实验的模型和超参设置基本一致的情况下，各个实验中 TCS 的收敛情况也基本一致，这显示了 TCS 的稳定性. 当$\mathcal{T}$的规模仅为 1000 条样本时，TCS 就可以将模型的性能由 Base 的 81.6 *Acc* 提升到 90.88 *Acc*，且随着$\mathcal{T}$规模变大，模型的性能进一步增长，这显示出 TCS 有着很高的样本利用效率.

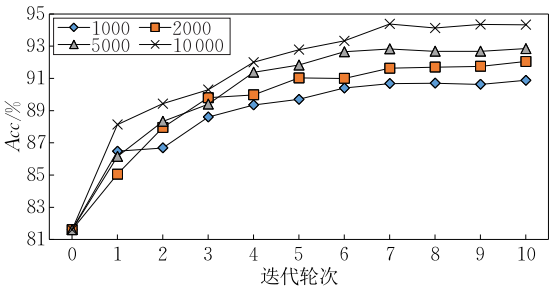

图 8 TCS 使用不同规模的$\mathcal{T}$在 MLDoc-fr 上的性能

3.3.4 不同语言任务上效果的差异

3.2.3 节在平均意义上对不同语言任务的实验结果进行分析，发现 TCS 使用的软目标训练、渐进式样本选择和“从可信到可疑”的课程学习都是有效的，且课程学习方案对最后结果的贡献最大. 在本节中，我们基于 MLDoc 数据集从个体角度来分析 TCS 在不同语言任务上效果的差异.

我们分别考察三项技术方案在不同语言上的效果. 相比 ST，SD 在 de、fr、it 和 ja 上取得了显著提升，相对提升幅度最大的是 de(1.27%)；在 es 和 zh 上出现了微弱下降，相对下降幅度最大的是 zh(0.20%)；相比 SD，SDS 在 de、fr、it 和 ru 上取得了显著提升，相对提升幅度最大的是 ru(1.92%)；在 es 和 zh 上出现了微弱下降，相对下降幅度最大的是 es(0.50%)；相比 SDS，TCS 在 es、fr、ja、ru 和 zh 上取得了显著提升，相对提升幅度最大的是 es(3.44%)；在 de 和 it 上出现了微弱下降，相对下降幅度最大的是 it(0.28%). 上述结果表明，三项技术方案在多数语言上的提升效果都是比较显著的；在个别语言上出现了微弱下降，但并没有表现出在特定语言上持续下降的规律. 此外，相比 ST，TCS 在全部 7 种目标语言上都取得了显著提升，相对提升幅度最大的是 fr(3.51%). 这说明即使某个数据集中错误样本的分布不契合其中一项技术方案，同时使用三项技术方案就可以避免在该任务上的下降风险.

4 相关工作

本文关注的跨语言文本分类和跨语言命名实体识别主要有两种技术路线：基于标注映射的样本迁移；基于跨语言表示的模型迁移. 此外，本文提出的 TCS 主要涉及自训练和课程学习两种技术.

4.1 基于标注映射的样本迁移

这类方法借助某种双语间的平行资源完成从源

语言标注到目标语言标注的映射以此来构造目标语言带标注样本，然后在构造的目标语言样本上训练目标模型.

一些工作利用平行语料来进行标注映射[22,24-25]，它们先在源语言带标注样本上训练源模型，再使用源模型预测平行语料源语言侧的标签，最后将源语言侧的标签映射到对齐的目标语言侧. 对于跨语言文本分类来说，平行语料本身提供的句子或篇章级别对齐信息已经可以完成映射；但因为命名实体识别需要词级别标签，所以这类方法往往还需要先在平行语料上学习词对齐，再根据词对齐信息映射词级别标签[22].

然而很多场景中任务相关的平行语料难以获得，一些工作转而利用机器翻译系统来进行标注映射[26-28]，它们使用机器翻译系统对源语言带标注样本进行翻译并映射标注. 也有工作利用资源需求更低的词典代替完整的机器翻译系统来进行词翻译[23,29]. 然而，低资源语言上较低的翻译性能以及命名实体识别需要词级别标注的特点都会降低基于翻译的标注映射方法的性能.

4.2　基于跨语言表示的模型迁移

这类方法使用源语言带标注样本训练带有跨语言表示模块的模型，模型借助跨语言表示就可以直接泛化到目标语言任务上.

跨语言词向量[30]被广泛应用于跨语言文本分类[10,25]和跨语言命名实体识别[22]. Artetxe 等人[21]通过训练多语言神经机器翻译模型来获得跨语言句子表示. 此外，在跨语言命名实体识别中，词聚类[31]、地名词典[32]以及维基百科[33]等特征也被用来提供跨语言信息.

随着一些工作发现[1-2] M-BERT[3]有着出色的 CLT 能力，M-BERT 逐渐成为近期跨语言文本分类和跨语言命名实体识别工作的基础. Wu 等人[34]应用一种元学习方法来完成 M-BERT 跨语言迁移，在跨语言命名实体识别上取得了更好的迁移效果，但该方法需要在推理阶段根据每条测试样本重新微调模型，导致模型推理阶段开销非常大. 下面这些工作与本文更为相关，它们也利用目标语言无标注样本来提升 M-BERT 的 CLT 性能：Keung 等人[4]使用语言对抗训练来促使 M-BERT 产生更加语言独立的特征；Zhang 等人[5]将 CLT 任务看成领域适应问题并使用一种基于对抗训练的无监督领域适应方法来降低领域之间的差异；Wu 等人[7]先使用源语言带标注样本微调模型，然后使用目标语言无标注样本对模型进行知识蒸馏来提升单源和多源跨语言命名实体识别的性能，该方法可以看成是使用软目标的单轮自训练；Dong 等人[6]使用带有样本选择机制的自训练来提升跨语言文本分类的性能. 前两种基于对抗训练的方法在利用目标语言无标注样本上不同后两种基于自训练的方法简单直接，在性能上也弱于后者. 本文整合 Wu 等人[7]和 Dong 等人[6]的方法并提出一种“从可信到可疑”的课程学习方案来进一步减轻老师模型错误预测对学生模型的影响.

此外，XLM[14]和 XLM-R[35]等工作在 M-BERT 的基础上探索更佳的多语言预训练表示. 由于 XLM 和 XLM-R 的模型结构和预训练方法和 M-BERT 十分相似，且为了和现有的跨语言文本标签预测工作进行比较，本文还是基于 M-BERT 开展研究，但需要说明的是，本文提出的 TCS 也可以用于 XLM 或 XLM-R 的跨语言迁移.

4.3　自训练

自训练作为经典的包裹式半监督学习方法[8]，可以和任意基础模型结合使用，因此适用于各种任务. 一些早期工作已将其成功应用于词义消歧[36]和句法分析[37]等自然语言处理任务. 近期，Artetxe 等人[38]表明自训练可以迭代式地提升跨语言词向量的质量；He 等人[9]则表明自训练可以提升机器翻译和文本摘要等文本生成任务的性能；与本文的任务一致，Dong 等人[6]也使用自训练类型方法来提升 M-BERT 在跨语言文本分类上的性能. 另外，从使用哪种技术减轻自训练中不准确监督问题的角度来看，He 等人[9]和本文都使用了软目标训练；Dong 等人[6]和本文都使用了渐进式的样本选择；不同的是，本文还根据自训练和跨语言文本标签预测的特点为学生模型设计一种“从可信到可疑”的课程学习方案.

4.4　课程学习

课程学习是一种通过设计模型训练过程中样本顺序来引导模型更快更好训练的技术，Bengio 等人[16]提出了这一术语并在语言建模任务上证明“从简单到困难”的课程顺序是有效的. 随后，Kocmi 等人[17]将课程学习应用于机器翻译任务同时也验证了“从简单到困难”的课程顺序可以使模型产生更好的翻译结果；Zhang 等人[18]在机器翻译任务上试验了多种样本复杂度度量方法以及课程安排方法. 这

些工作使用课程学习在有正确标注的训练集上提升模型训练效果,它们往往使用语言学特征来衡量样本的困难程度,比如词频[16-18]和句子长度[17-18].与这些工作不同的是,本文使用课程学习来减轻自训练的每轮训练集中错误样本对学生模型训练的影响,因此本文使用更加相关的预测置信度来衡量样本的困难程度.此外,在课程学习的实现方式上,本文与 Zhang 等人[18]较为相似,都是先使用聚类算法根据样本困难度对数据集进行分片,再通过控制训练不同阶段不同片区数据的可用性来实现课程安排.

5 总结与展望

为了在微调 M-PLM 过程中引入目标语言无标注样本信息,本文在自训练的基础上提出 TCS 学习框架,该框架利用软目标训练、渐进式样本选择和"从可信到可疑"的课程学习三项技术来减轻老师模型的错误预测对学生模型训练的影响.在跨语言文本分类和跨语言命名实体识别上的实验表明,TCS 可以大幅提升 M-BERT 的 CLT 性能且效果显著超越现有方法.当同时有多个迁移目标时,为每个目标语言单独优化一个模型会造成模型冗余,所以后续工作准备拓展 TCS,使之产生的统一模型可以处理多种目标且性能不弱于单独的模型.

参 考 文 献

[1] Wu S, Dredze M. Beto, Bentz, Becas: The surprising cross-lingual effectiveness of BERT//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China, 2019: 833-844

[2] Pires T, Schlinger E, Garrette D. How multilingual is multilingual BERT?//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2019: 4996-5001

[3] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, USA, 2019: 4171-4186

[4] Keung P, Bhardwaj V, et al. Adversarial learning with contextual embeddings for zero-resource cross-lingual classification and NER//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China, 2019: 1355-1360

[5] Zhang D, Nallapati R, Zhu H, et al. Unsupervised domain adaptation for cross-lingual text labeling//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. Online, 2020: 3527-3536

[6] Dong X L, de Melo G. A robust self-learning framework for cross-lingual text classification//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China, 2019: 6307-6311

[7] Wu Q, Lin Z, Karlsson B, et al. Single-/multi-source cross-lingual NER via teacher-student learning on unlabeled data in target language//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online, 2020: 6505-6514

[8] Zhu X. Semi-supervised learning literature survey. World, 2005, 10: 10

[9] He J, Gu J, Shen J, et al. Revisiting self-training for neural sequence generation//Proceedings of the International Conference on Learning Representations. Online, 2020

[10] Schwenk H, Li X. A corpus for multilingual document classification in eight languages//Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan, 2018

[11] Tjong Kim Sang E F. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition//Proceedings of the COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002). Taipei, China, 2002

[12] Tjong Kim Sang E F, De Meulder F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition//Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003. Edmonton, Canada, 2003: 142-147

[13] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Advances in Neural Information Processing Systems. Long Beach, USA, 2017: 5998-6008

[14] Conneau A, Lample G. Cross-lingual language model pretraining//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada, 2019: 7059-7069

[15] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531, 2015

[16] Bengio Y, Louradour J, Collobert R, et al. Curriculum learning//Proceedings of the 26th Annual International

Conference on Machine Learning. Montréal, Canada, 2009: 41-48

[17] Kocmi T, Bojar O. Curriculum learning and minibatch bucketing in neural machine translation//Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2017). Varna, Bulgaria, 2017: 379-386

[18] Zhang X, Kumar G, Khayrallah H, et al. An empirical exploration of curriculum learning for neural machine translation. arXiv preprint arXiv:1811.00739, 2018

[19] Jenks G F. Optimal data classification for choropleth maps. Department of Geographiy, University of Kansas Occasional Paper, 1977

[20] Loshchilov I, Hutter F. Decoupled weight decay regularization //Proceedings of the International Conference on Learning Representations. New Orleans, USA, 2019

[21] Artetxe M, Schwenk H. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Transactions of the Association for Computational Linguistics, 2019, 7: 597-610

[22] Ni J, Dinu G, Florian R. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vancouver, Canada, 2017: 1470-1480

[23] Xie J, Yang Z, Neubig G, et al. Neural cross-lingual named entity recognition with minimal resources//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 369-379

[24] Mihalcea R, Banea C, Wiebe J. Learning multilingual subjective language via cross-lingual projections//Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Prague, Czech Republic, 2007: 976-983

[25] Rasooli M S, Farra N, Radeva A, et al. Cross-lingual sentiment transfer with limited resources. Machine Translation, 2018, 32(1): 143-165

[26] Wan X. Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis//Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. Honolulu, USA, 2008: 553-561

[27] Salameh M, Mohammad S, Kiritchenko S. Sentiment after translation: A case-study on Arabic social media posts//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, USA, 2015: 767-777

[28] Jain A, Paranjape B, Lipton Z C. Entity projection via machine translation for cross-lingual NER//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China, 2019: 1083-1092

[29] Mayhew S, Tsai C T, Roth D. Cheap translation for cross-lingual named entity recognition//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2017: 2536-2545

[30] Ruder S, Vulić I, Søgaard A. A survey of cross-lingual word embedding models. Journal of Artificial Intelligence Research, 2019, 65: 569-631

[31] Täckström O, McDonald R, Uszkoreit J. Cross-lingual word clusters for direct transfer of linguistic structure//Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Montréal, Canada, 2012: 477-487

[32] Zirikly A, Hagiwara M. Cross-lingual transfer of named entity recognizers without parallel corpora//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Beijing, China, 2015: 390-396

[33] Tsai C T, Mayhew S, Roth D. Cross-lingual named entity recognition via Wikification//Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning. Berlin, Germany, 2016: 219-228

[34] Wu Q, Lin Z, Wang G, et al. Enhanced metalearning for cross-lingual named entity recognition with minimal resources //Proceedings of the AAAI Conference on Artificial Intelligence: Volume 34. Online, 2020: 9274-9281

[35] Conneau A, Khandelwal K, Goyal N, et al. Unsupervised cross-lingual representation learning at scale//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online, 2020: 8440-8451

[36] Yarowsky D. Unsupervised word sense disambiguation rivaling supervised methods//Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics. Cambridge, USA, 1995: 189-196

[37] McClosky D, Charniak E, Johnson M. Effective self-training for parsing//Proceedings of the Human Language Technology Conference of the NAACL, Main Conference. New York, USA, 2006: 152-159

[38] Artetxe M, Labaka G, Agirre E. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia, 2018: 789-798

[39] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958

PU Tong, M. S. His research interests include natural language processing and transfer learning.

HUANG Shu-Jian, Ph. D., associate professor. His research interests include machine translation and natural language generation.

ZHANG Yang-Ming, B. S., algorithm engineer. His research interests include natural language processing, knowledge graph and speech processing.

ZHOU Xiang-Sheng, B. S., senior R&D manager. His research interests include natural language processing and machine learning.

TU Yao-Feng, Ph. D. His research interests include big data, database and machine learning.

DAI Xin-Yu, Ph. D., professor. His research interests include natural language processing and recommender systems.

CHEN Jia-Jun, Ph. D., professor. His main research field is natural language processing.

Background

In recent years, deep learning models have greatly promoted the development of various natural language processing tasks, but training these models often requires a large number of labeled samples. In order to perform natural language processing in languages where labeled samples are unavailable, cross-lingual transfer (CLT) utilizes the labeled samples in the source language to enable the model to perform the corresponding task of the target language. Text classification and named entity recognition are two representative text analysis tasks and this paper aims at improving their performance in the Zero-shot CLT (for simplicity, Zero-shot is omitted) setting.

There are two main methodologies to CLT problems: sample transfer based on annotation projections and model transfer based on cross-lingual representations. The former requires task related parallel corpus or high-performance machine translation system, therefore, its application scenarios are limited and the latter has attracted more attention. In particular, with the discovery that multilingual BERT (M-BERT) can produce powerful cross-lingual representations, fine-tuning M-BERT with source labeled samples has become the mainstream method of CLT tasks. In real scenarios, unlabeled samples in the target language are easy to collect and contain target distribution information, therefore, how to utilize target unlabeled samples in the M-BERT's fine-tuning process becomes a problem worthy of investigation. Recent research on this problem can be divided into two categories: one narrows the gap between different language representations in M-BERT through adversarial training (AT); the other uses Self-Training (ST) to let the model learn the pseudo labels of target unlabeled samples. Compared with AT based methods, ST based methods utilize target unlabeled samples more simply and directly, and achieve better results. However, ST suffers from the inaccurate supervision problem, that is, the teacher model's inaccurate predictions may mislead the student model. And in CLT, the natural distribution gap between the source labeled samples and the target unlabeled samples will worsen this problem.

In this paper, we utilize three techniques to alleviate the inaccurate supervision problem and propose a learning framework called Teacher-Curriculum-Student (TCS) based on ST. Specifically, we first combine the soft-target training and progressive sample selection techniques used in existing works. Further, we introduce a from-confident-to-suspicious curriculum between the teacher and the student to enhance the role of accurate samples and reduce the role of inaccurate samples in the student's training process. Experiments on the benchmark datasets of cross-lingual text classification and cross-lingual named entity recognition show that the average results of TCS is 2.51% and 3.25% higher than that of ST, and 1.51% and 4.45% higher than that of existing state-of-the-art methods, respectively.

This work is supported by the National Natural Science Foundation of China (U1836221, 6217020152) and the ZTE scientific research cooperation project.