• 计算机工程与应用 •

基于梯度相似性的自动作文评分多主题联合预训练方法

李晨亮[1,2*]，吴鸿涛[1,2]

(1. 武汉大学空天信息安全与可信计算教育部重点实验室 武汉 430072；2. 武汉大学国家网络安全学院 武汉 430072)

【摘要】提出了一种基于梯度相似性的自动加权方法，用于作文评分的多主题联合预训练。在预训练阶段同时使用多个主题的数据，通过计算外部主题的训练样本的梯度向量与目标主题的梯度向量之间的相似度作为该样本的损失权重。将深度学习与特征工程相结合，手工设计了 3 类特征。在公开数据集上进行对比实验表明，与现有的基线模型相比，提出的多主题联合预训练方法和手工特征均能有效提升作文评分模型的评分准确性。

关 键 词 自动作文评分； 深度学习； 特征工程； 预训练

中图分类号 TP391.1 **文献标志码** A doi:10.12178/1001-0548.2022061

A Gradient-Similarity Based Multi-Topic Jointly Pre-Training Method for Automated Essay Scoring

LI Chenliang[1,2*] and WU Hongtao[1,2]

(1. Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, Wuhan University Wuhan 430072;
2. School of Cyber Science and Engineering, Wuhan University Wuhan 430072)

Abstract This paper proposes a gradient-similarity based multi-topic jointly pre-training method for automated essay scoring (AES). Specifically, in the pre-training stage, the training data of multiple topics are used at the same time, and the similarity between the gradient vector of a sample from other topics and the gradient vector of target topic is calculated as the loss weight for this sample. Besides, this paper also designs three types of handcrafted features, combining deep learning with feature engineering. Comparative experiments are conducted on publicly available datasets, and the results show that compared with the existing baselines, both proposed multi-topic jointly pre-training method and handcrafted features can effectively improve the scoring accuracy of the AES model.

Key words automated essay scoring; deep learning; feature engineering; pre-training

自动作文评分 (automated essay scoring, AES) 是自然语言处理领域的一项重要任务，目标是采用自动化的方法来评估作文的写作质量并给作文评分。自动作文评分系统在教育领域有着广阔的应用前景，如托福考试的写作题判卷就使用了机器判卷系统 E-rater[1]。最早的 AES 系统可以追溯到 1966 年开发的 Project Essay Grader 系统[2]，并在此后一直备受关注。早期的 AES 系统主要依赖手工提取的特征，包括长度特征、词汇特征、句法特征及语义特征等类别。然而，手工提取特征费时费力，还可能需要专家的参与。

随着深度学习技术的兴起，目前主流的 AES 系统大多是基于深度神经网络，或者神经网络与特征工程相结合的方式来实现。文献 [3] 最早尝试使用神经网络来构建 AES 系统，通过卷积神经网络 (convolutional neural network, CNN) 和循环神经网络 (recurrent neural network, RNN) 来学习作文的向量表征，再使用全连接层来预测作文分数。文献 [4] 对作文的层次结构进行建模，分别使用一个单词级的 CNN+池化层来学习句子表示，以及一个句子级的 CNN+池化层用于根据句子表示来学习全文的表示。文献 [5] 在文献 [4] 的基础上引入了注意力机

收稿日期：2022－02－25；修回日期：2022－03－30
基金项目：国家自然科学基金 (61872278)
作者简介：李晨亮 (1983－)，男，博士，教授，主要从事信息检索、数据挖掘、机器学习、社交网络分析及自然语言处理等方面的研究.
*通信作者：李晨亮，E-mail：cllee@whu.edu.cn

制，来区分不同词和不同句子的重要性，并且通过实验证明 CNN 更适合对句子建模，RNN 更适合对全文建模。文献 [6] 从词向量入手，首先在 C&W 词向量模型[7] 的基础上增加作文评分任务来学习任务增强的词向量，然后将训练好的词向量输入两层双向 LSTM 网络[8] 学习预测作文分数。文献 [9] 使用神经张量网络对 LSTM 不同时间步的隐状态向量之间的关系进行建模，以提升对作文连贯性的表达能力。以上工作都是将作文评分任务视为一个回归任务，损失函数为模型预测分数和人工标注分数之间的均方误差 (mean-square error, MSE)。文献 [10] 将自动作文评分看作一项分类任务，使用强化学习来解决，将模型与人工专家评分之间的 kappa 一致性系数作为 agent 的奖励值。最新的工作开始在作文评分任务中应用大规模预训练语言模型 (如 BERT[11])。文献 [12] 将特征工程与 BERT 相结合用于作文评分，使用长度特征、句法特征、词语特征和可读性特征 4 类特征，将特征向量与神经网络模型输出的作文向量拼接后输入全连接层预测分数。文献 [13] 提出了一个基于 BERT 的 AES 方法，将作文评分同时视为回归任务和排序任务，使用一个随训练过程动态调整的权重来平衡回归损失和排序损失。大规模预训练语言模型的应用已经将评分准确性提升到更高的水平。

基于神经网络的模型摆脱了对手工设计特征的依赖。然而，深层神经网络需要大量数据来训练才能发挥其强大的学习能力。但是，自动作文评分任务面临较严重的数据不足的问题，因为现有的 AES 数据集由多个主题 (即题目) 组成，每个主题下只有较少量的样本。目前的 AES 方法大多是单独考虑一个主题的，即分别用每个主题的训练数据训练一个模型，该模型只用于对该主题的作文进行评分。这些方法只使用目标主题的少量训练样本，而不利用剩余其他主题的样本。本文通过实验证明，引入外部主题的样本来增加训练数据量能进一步提升模型的表现。考虑到不同主题之间存在差异，直接加入其他主题的训练样本，训练模型会引入训练噪声，为此本文提出了一种基于梯度相似性的自动加权方法。该方法的训练包含预训练和微调两个阶段。在预训练阶段，同时使用多个主题的作文数据来训练模型。对于其他主题的训练样本，计算该样本的梯度向量与目标主题的梯度向量之间的相似度作为该样本的损失权重。在微调阶段，使用目标主题的训练数据来微调模型。最后，将训练好的模型用于对目标主题的测试样本进行评分。相比于传统的单主题 AES 方法，本文方法引入了其他主题的训练样本，有效缓解了数据不足的问题，并且基于梯度相似性进行加权，降低了直接使用其他领域的样本带来的训练噪声。此外，本文还将神经网络与特征工程相结合，手工设计了 3 类特征，进一步提升了模型的评分表现。在公开数据集上进行的对比实验表明，与传统的基于单主题样本的方法相比，本文提出的多主题联合预训练方法能有效提升 AES 模型的评分准确性，结合特征工程后能实现进一步的提升。

1　研究方法

1.1　作文评分模型

本文模型基于预训练 BERT 实现，模型结构如图 1 所示。BERT 是由谷歌提出的大规模预训练语言模型，基于 Transformer 网络实现[14]，由多个 Transformer 编码器层堆叠组成。每个 Transformer 层包含一个多头自注意力模块和前馈神经网络，并且使用残差连接[15] 将输入和输出结合。Transformer 的这种结构能有效抽取文本的特征表示，因此在自然语言处理领域得到广泛应用。原始的 BERT 分为 BERT-base 和 BERT-large 两个版本，base 版本包含 12 个 Transformer 编码器层，每层的注意力头数为 12 个，总的参数量约 1.1 亿。本文采用 BERT-base 来构建作文评分模型。

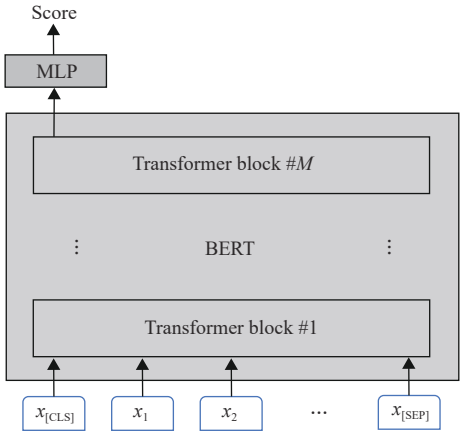

图 1　作文评分模型结构

给定一篇作文s，首先通过 BERT 的分词器将其转化为子词 (sub-token) 序列，并在序列开头和结尾分别插入“[CLS]”和“[SEP]”两个特殊词：

$$s = [s_{[CLS]}, s_1, \cdots, s_N, s_{[SEP]}] \tag{1}$$

然后通过词向量层转换为词向量序列：

$$x = [x_{[CLS]}, x_1, \cdots, x_N, x_{[SEP]}] \tag{2}$$

再将 x 输入到 Transformer 编码器层。每个编码器层由一个多头自注意力模块和一个前馈神经网络组成，如图 2 所示。

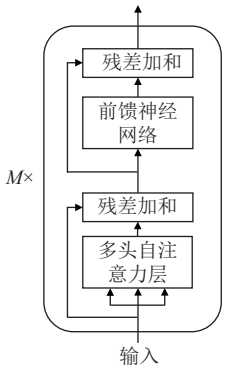

图 2 Transformer 编码器结构

多头自注意力模块将输入的向量序列组成的矩阵$\boldsymbol{X} \in \mathbb{R}^{L\times d}$($L$是序列长度，$d$是向量维度) 映射到多个不同的语义空间进行自注意力操作，抽取不同方面的语义特征。对于每个自注意力头，首先将$\boldsymbol{X}$分别经过 3 个不同的线性变换得到$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$这 3 个矩阵，然后计算自注意力头的输出：

$$\mathrm{Att}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d/h}})\boldsymbol{V} \tag{3}$$

式中，$h = 12$是注意力头的数量；$\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}} \in \mathbb{R}^{L\times L}$计算每个词对序列中所有词的注意力分数；使用 softmax(·)函数将注意力分数归一化为权重，再将注意力权重乘上$\boldsymbol{V}$得到自注意力模块的输出。

然后，将 12 个自注意力头的输出拼接后再经过一层线性变换得到多头自注意力层的最终输出：

$$\boldsymbol{Z} = \mathrm{Concat}(\mathrm{Att}_1, \mathrm{Att}_2, \cdots, \mathrm{Att}_h)\boldsymbol{W}_l + \boldsymbol{b}_l \tag{4}$$

式中，$\boldsymbol{W}_l$和$\boldsymbol{b}_l$是参数矩阵和偏置向量。

然后通过残差连接将$\boldsymbol{Z}$与$\boldsymbol{X}$相加，再进行层归一化操作。将结果输入到后面的全连接前馈神经网络FFN中。FFN 具有相同的残差连接和层归一化机制，FFN 的最终输出将传递给下一个编码器模块。

经过$M = 12$个编码器模块后，BERT 输出一个隐状态向量序列：

$$\boldsymbol{H} = [h_{[CLS]}, h_1, \cdots, h_N, h_{[SEP]}] \tag{5}$$

最后，$\boldsymbol{r} = h_{[CLS]}$被用作作文的表示向量。将$\boldsymbol{r}$与通过特征工程抽取的人工特征向量$\boldsymbol{f}$(将在下一节介绍) 拼接后输入预测层 (两层的全连接网络)，然后输出最终的预测分数$\hat{y}$：

$$a = \tanh(\mathrm{Concat}(\boldsymbol{r}, \boldsymbol{f})\boldsymbol{W}_0 + \boldsymbol{b}_0) \tag{6}$$

$$\hat{y} = \sigma(a\boldsymbol{W}_1 + \boldsymbol{b}_1) \tag{7}$$

式中，σ是 sigmoid 激活函数；$\hat{y} \in (0,1)$。在模型的推理阶段，将$\hat{y}$缩放到作文的分数区间即得到预测的作文分数。

本文将作文评分视为回归任务，因此损失函数为预测分数$\hat{y}$与归一化后的人工评分y之间的均方误差：

$$L = \frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2 \tag{8}$$

式中，m是一批样本的样本数。

1.2 特征工程

为了进一步增强作文评分的准确性和可解释性，本文将神经网络与手工设计的特征相结合，设计了 3 类特征，如表 1 所示。其中，Flesch-Kincaid (FK) 可读性分数是衡量文章是否易于理解的指标，包括 FK 可读性分数 (Flesch-Kincaid reading ease formula) 和 FK 可读性年级水平 (Flesch-Kincaid grade level formula) 两个指标。FK 可读性分数 (s) 和 FK 年级水平 (g) 的计算公式分别为：

$$s = 206.835 - 1.015a - 84.6b \tag{9}$$

$$g = 0.39a + 11.8b - 15.59 \tag{10}$$

式中，a和b 分别指该文章中句子的平均词数以及单词的平均音节数；FK 可读性分数在 0～100 之间，分数越高表示文章越容易理解。一般来说，分数在 60 分以上即可视为可读性较好。FK 年级水平反映的是要能理解这篇文章至少需要达到的美国学校年级。FK 年级水平越高，说明文章更难理解。

表 1 手工设计的特征

特征类型	特征介绍
统计特征	单词数，去重之后的单词数，句子数，句子的平均词数，不同长度的单词数量，各种连接词的数量，拼写错误的数量
语法特征	不同类型从句的数量，包括主语从句、宾语从句、状语从句、定语从句、同位语从句等
可读性特征	Flesch-Kincaid可读性分数[16]

将抽取的特征值按照最小–最大归一化方法缩放为介于 [0,1] 的值，然后将所有特征拼接为一个特征向量$\boldsymbol{f}$，与神经网络抽取的表示向量$\boldsymbol{r}$拼接后输入到预测层预测分数。

1.3　多主题联合预训练

公开的作文评分数据集由多个主题构成，每个主题下平均只有大约一千篇作文，这样的数据量用来训练深层神经网络模型容易造成过拟合，限制了模型的泛化能力。为此，本文提出一种基于梯度相似性来自动加权的多主题联合预训练的方法。模型的训练分为两个阶段：预训练和微调阶段。在预训练阶段，使用所有主题的全部训练样本来训练模型；在微调阶段，只使用目标主题的样本来对模型进行精调训练。

假设作文评分数据集D一共有 n 个主题，现在需要训练一个模型用于对其中一个主题的作文进行评分，称该主题为目标主题。将目标主题的训练集记为D_t，其他主题组成的混合训练集记为D_o。

考虑到不同主题之间存在差异，直接使用其他主题的训练样本未必能获得较大收益。本文提出根据梯度相似性来对其他主题的样本进行加权。首先根据式 (8) 计算目标主题的样本的平均损失L_t，然后计算L_t对模型参数θ的梯度向量$\boldsymbol{g}_t=\nabla_\theta L_t$，将其称为标准梯度。然后对一个来自其他主题的训练样本b_i计算其损失L_i，并求梯度向量$\boldsymbol{g}_i=\nabla_\theta L_i$。由于梯度向量反映了模型优化的方向，本文假设，如果一个样本产生的梯度方向与标准梯度方向越接近，则该样本与目标主题越“接近”，或者说产生的优化效果与目标主题的样本产生的效果越相似。于是，计算$\boldsymbol{g}_i$和$\boldsymbol{g}_t$之间的余弦相似度：

$$\mathrm{sim}(\boldsymbol{g}_t,\boldsymbol{g}_i)=\frac{\boldsymbol{g}_t^{\mathrm{T}}\boldsymbol{g}_i}{\|\boldsymbol{g}_t\|_2\|\boldsymbol{g}_i\|_2} \tag{11}$$

如果$\mathrm{sim}(\boldsymbol{g}_t,\boldsymbol{g}_i)>0$，则认为该样本对模型优化有一定的正面效果，将$\mathrm{sim}(\boldsymbol{g}_t,\boldsymbol{g}_i)$作为该样本的损失权重。于是，第 i 个其他主题的样本的权重为：

$$\boldsymbol{w}_i=\mathrm{relu}(\mathrm{sim}(\boldsymbol{g}_t,\boldsymbol{g}_i)) \tag{12}$$

完整的训练过程如下：

1) 从D_t、D_o中分别采样一批样本记为B_t、B_o；

2) 根据式 (8) 计算B_t的平均损失$L(B_t)$，并求得其梯度向量$\boldsymbol{g}_t$；

3) 对B_o中的每个样本b_i，计算对应损失$L(b_i)$，并求梯度向量$\boldsymbol{g}_i$，根据式 (12) 求得权重$\boldsymbol{w}_i$；

4) 按下式计算总的损失，并更新模型参数

$$L=L(B_t)+\frac{1}{K}\sum_{i=1}^{K}\boldsymbol{w}_iL(b_i) \tag{13}$$

5) 重复前 4 步直至训练收敛。

预训练完成后，再用目标主题的样本对模型微调训练，然后在目标主题测试集上评估模型表现。

2　实验及结果分析

2.1　数据集

本文在 ASAP 作文评分数据集 (https://www.kaggle.com/c/asap-aes/data) 上进行实验。ASAP 数据集是目前公开的最大、最常用的作文评分数据集，被广泛用于评测 AES 模型的评分准确性。ASAP 数据集由 8 个主题的子集构成，包含 3 种作文体裁：记叙文、议论文和回复类作文。数据集的统计结果如表 2 所示。

表 2　ASAP 数据集的统计结果

主题编号	作文数	体裁	作文平均长度	分数范围
1	1 783	议论文	350	2～12
2	1 800	议论文	350	1～6
3	1 726	回复文	150	0～3
4	1 772	回复文	150	0～3
5	1 805	回复文	150	0～4
6	1 800	回复文	150	0～4
7	1 569	记叙文	250	0～30
8	723	记叙文	650	0～60

2.2　基线模型和评估指标

本文比较了以下主流的基线方法：1) EASE；2) LSTM、BiLSTM；3) CNN-LSTM[3]；4) NN-LSTM-Att[5]；5) RL1[10]；6) SKIPFLOW[9]；7) HISK+BOSWE[17]；8) R^{2}BERT[13]；9) BERT。

其中，1) 和 7) 是传统统计学习方法，其余是神经网络模型。EASE 是一个开源的作文评分系统 (https://github.com/edx/ease)，使用手工设计的特征 (如长度特征、词袋特征等) 来训练机器学习算法，如 SVM，进行评分预测。HISK+BOSWE 分别使用直方图相交字符串核 (histogram intersection string kernel, HISK) 和 BOSWE 词向量来构造核矩阵，并将两个核矩阵结合后训练一个 ν-SVR 模型；LSTM、BiLSTM 表示分别训练一个两层的 LSTM 和双向的 LSTM，并且将所有隐向量进行平均池化作为作文表示；CNN-LSTM 使用 CNN 来抽取句内的词组特征，用 LSTM 建模句子间的语义关联特征；CNN-LSTM-Att 在 CNN-LSTM 的基础上引入了注意力机制；SKIPFLOW 在学习作文表示的同时显式地加入了对连贯性的建模；RL1 是基于强化学习的 AES 模型，使用 Dilated LSTM[18] 网络来抽取作文的表示并执行决策，使用策略梯度直接优化 QWK 指标；BERT 是本文方法的基础版本，与

本文模型的区别在于：没有使用特征工程，没有使用其他主题的作文数据；R²BERT 是文献 [13] 提出的基于 BERT 的 AES 模型，同时优化回归损失 (MSE 损失) 和排序损失 (list-wise 排序损失)，是目前最强的 AES 神经网络模型。

与其他相关工作一样，本文使用 Quadratic Weighted Kappa (QWK) 系数作为模型评分准确性的衡量指标。QWK 系数用于衡量两组评分 (机器评分和人工评分) 之间的一致性，取值通常介于 0～1 之间，分数越高表示一致性越高。如果小于 0，说明一致性甚至不如随机评分。QWK 的计算可以参考文献 [10,13]。

2.3 实现细节

本文的模型使用 Python 语言和 Pytorch 框架实现，预训练 BERT 采用 HuggingFace 发布的“bert-base-uncased”模型。在预训练阶段，初始学习率设为 4×10^{-5}，使用 Adam 优化器，训练 50 轮，取验证集上结果最好的模型用于微调。在微调阶段，初始学习率设为 3×10^{-6}，仍使用 Adam 优化器，训练 30 轮。对于超出 BERT 长度限制的作文，将超出长度限制的部分裁减掉。由于每个主题的样本数较少，因此本文采用 5-折交叉验证方法，每折按照 3:1:1 的比例划分训练集、验证集、测试集。将 5 个测试集的结果求平均作为最终结果。

2.4 实验结果

表 3 列出了不同模型在 ASAP 数据集的 8 个主题上的 QWK 分数，粗体字表示最优结果。

表 3 不同方法在 ASAP 数据集的 8 个主题上的结果

方法	主题1	主题2	主题3	主题4	主题5	主题6	主题7	主题8	平均
LSTM	0.582	0.517	0.516	0.702	0.604	0.670	0.661	0.566	0.602
BiLSTM	0.591	0.491	0.498	0.702	0.643	0.692	0.683	0.563	0.608
EASE(SVR)	0.781	0.630	0.621	0.749	0.782	0.771	0.727	0.534	0.699
CNN+LSTM[3]	0.821	0.688	0.694	0.805	0.807	0.819	0.808	0.644	0.761
LSTM-CNN-Att[5]	0.822	0.682	0.672	0.814	0.803	0.811	0.801	0.705	0.764
RL1[10]	0.766	0.659	0.688	0.778	0.805	0.791	0.760	0.545	0.724
SKIPFlOW[9]	0.832	0.684	0.695	0.788	0.815	0.810	0.800	0.697	0.764
HISK+BOSWE[17]	**0.845**	**0.729**	0.684	0.829	0.833	0.830	0.804	0.729	0.785
BERT	0.806	0.680	0.698	0.808	0.810	0.814	0.835	0.705	0.770
R²BERT[13]	0.803	0.692	0.694	0.811	0.811	0.811	0.830	0.738	0.774
All-prompt	0.819	0.710	0.686	0.830	0.838	0.834	0.829	0.740	0.786
本文方法	0.835	0.721	**0.702**	**0.841**	**0.842**	**0.846**	**0.837**	**0.749**	**0.797**

从结果来看，BiLSTM 优于 LSTM，这是符合直觉的，因为 BiLSTM 能够捕捉更全面的上下文信息。EASE 的表现优于 LSTM 和 BiLSTM 这两个神经网络模型，说明简单的神经网络结构未必比使用精心设计的手工特征来训练传统机器学习算法的效果更好。CNN-LSTM 的表现相比单纯使用 LSTM 的方法有较大提升，这主要是由于 LSTM、BiLSTM 这两个基线模型直接将整篇作文作为一个长序列，而忽视了句子级的结构信息，且 LSTM 不适合捕捉长距离依赖信息，在长文本的表现较差。而 CNN-LSTM 是先用 CNN 来学习句子的向量表示，再用 LSTM 来学习全文的表示，这种建模方式考虑了文章的层次结构，保留了句子级的结构信息，且 CNN 天然适合学习相邻词之间的关联特征，更适合对句子语义建模[5]。CNN-LSTM-Att 在 CNN-LSTM 的基础上引入了注意力机制，使评分准确性获得进一步提升。注意力机制能区分不同词和不同句子的重要性，能够关注到更重要的局部特征，是目前的神经网络模型的基础结构。RL1 使用强化学习的方法来直接优化 QWK 指标，但结果并不是特别优异，这跟它也是使用简单的 LSTM 网络来学习文章表示有关。SKIPFlOW 通过增加前后句子间连贯性进行建模，获得了进一步的表现提升，这说明对于 AES 任务来说，评估文章的连贯性也很重要。基于核方法的 HISK+BOSWE 超越了所有神经网络基线模型，尤其是在议论文体裁的主题 1 和主题 2 上仍保持目前最优秀的性能，这也说明传统的机器学习方法的潜力不应被低估。而 BERT、R²BERT 这两个基于预训练 BERT 的模型并未取得最好的基线模型结果，这可能是由于数据量太小，容易导致

过拟合，限制了大模型的学习能力。本文方法在 6 个主题上都取得了最优结果，说明了本文提出的特征工程和多主题联合预训练的方法的有效性。All-prompt 是本文方法的基线版本，与本文方法的区别在于，没有使用基于梯度相似性的加权方法，而是直接将其他主题的数据加到训练集中。由结果可知，使用本文的损失加权方案可以在引入外部主题数据的同时，降低跟目标主题差异较大的样本的权重，进一步提升模型表现。

2.5　消融实验

为了验证本文设计的 3 类特征的效果，本文通过实验比较了不使用人工特征 (none)、只使用统计特征 (st)、使用统计特征+语法特征 (st+gm) 和使用所有特征的结果 (all)。结果如表 4 所示，粗体字表示最优结果。

表 4　比较使用不同特征的结果

方法	主题1	主题2	主题3	主题4	主题5	主题6	主题7	主题8	平均
Ours(none)	0.811	0.701	0.701	0.833	0.835	0.834	0.822	0.741	0.785
Ours(st)	0.832	0.716	0.700	0.835	0.836	0.842	0.829	0.745	0.792
Ours(st+gm)	**0.836**	0.718	0.699	**0.845**	**0.842**	0.845	**0.838**	**0.749**	**0.797**
Ours(all)	0.835	**0.721**	**0.702**	0.841	**0.842**	**0.846**	0.837	**0.749**	**0.797**

从结果可知，统计类特征和语法类特征均对提升评分准确性有帮助，尤其是在议论文体裁的主题 1 和主题 2 上提升更加显著。这说明在当今深度神经网络成为主流机器学习方法的时期，考虑融合传统人工设计的特征对提升模型表现仍可能是有帮助的。加入可读性特征尽管未获得明显提升，但也不能说明不需要考虑作文的可读性，这也可能是使用的可读性特征不能很好反映文章是否易于理解，未来可以尝试设计更有效的可读性特征。

3　结 束 语

针对自动作文评分领域当前面临的训练数据不足的问题，本文提出了一种多主题联合预训练方法，通过引入外部主题的作文数据，增加了训练样本的数量。为了降低直接引入差异较大的外部主题的样本带来的训练噪声，本文提出了一种基于梯度相似性的损失重加权方案。此外，为了进一步提升深度学习模型的性能，本文还将深度学习与传统特征工程相结合，并设计了 3 类作文特征。在公开数据集上的实验结果表明，本文方法显著优于传统的单主题内训练方法。

进一步的工作可以考虑以下 3 个方向：1) 多维度质量评分。现有的 AES 方法绝大多数都是对作文质量进行整体评分，而很少关注作文质量的不同维度。文献 [19] 列举了作文质量的不同方面，包括语法、遣词、标点、组织结构、连贯性、说服力等 11 个方面。研究从不同维度对作文质量进行评分能够给用户提供更详细的信息，使用户知晓作文的不足之处。2) 评分反馈。现有的 AES 方法都只能对作文评分，而无法提供类似教师评语的反馈，如果能生成对应的反馈意见将有助于帮助学生提高写作水平。3) 少样本学习。由于 AES 领域缺少大规模数据集，因此可以研究少样本学习的方法，减轻对监督数据的依赖。

参　考　文　献

[1] ATTALI Y, BURSTEIN J. Automated essay scoring with e-rater[r]v.2[J]. Journal of Technology, Learning, and Assessment, 2006, 4(3): 1-21.

[2] PAGE E B. The imminence of... grading essays by computer[J]. The Phi Delta Kappan, 1966, 47(5): 238-243.

[3] TAGHIPOU K, NG H T. A neural approach to automated essay scoring[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas: Association for Computational Linguistics, 2016: 1882-1891.

[4] DONG F, ZHANG Y. Automatic features for essay scoring: An empirical study[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas: Association for Computational Linguistics, 2016, 435: 1072-1077.

[5] DONG F, ZHANG Y, YANG J. Attention-based recurrent convolutional neural network for automatic essay scoring[C]//CoNLL. Vancouver: Association for Computational Linguistics, 2017: 153-162.

[6] ALIKANIOTIS D, YANNAKOUDAKIS H, REI M. Automatic text scoring using neural networks[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: Association for Computational Linguistics, 2016: 715-725.

[7] COLLOBERT R, WESTON J. A unified architecture for natural language processing: Deep neural networks with multitask learning[C]//Proceedings of the 25th International Conference on Machine Learning. New York: Association for Computing Machinery, 2008: 160-167.

[8] SAK H, SENIOR A, BEAUFAYS F. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition[EB/OL]. [2014-02-05]. https://arxiv.org/abs/1402.1128.

[9] TAY Y, PHAN M, TUAN L A, et al. SkipFlow: Incorporating neural coherence features for end-to-end automatic text scoring[C]//Proceedings of the AAAI Conference on Artificial Intelligence. California: AAAI Press, 2018: 5948-5955.

[10] WANG Y, WEI Z, ZHOU Y, et al. Automatic essay

scoring incorporating rating schema via reinforcement learning[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels: [s.n.], 2018: 791-797.
[11] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//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, Minnesota: [s.n.], 2018: 4171-4186.
[12] UTO M, XIE Y, UENO M. Neural automated essay scoring incorporating handcrafted features[C]//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona: International Committee on Computational Linguistics, 2020: 6077-6088.
[13] YANG R, CAO J, WEN Z, et al. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. [S.l.]: Association for Computational Linguistics, 2020: 1560-1569.
[14] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: Curran Associates Inc, 2017: 6000-6010.
[15] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV: IEEE, 2016: 770-778.
[16] FLESCH R. A new readability yardstick[J]. Journal of Applied Psychology, 1948, 32(3): 221-233.
[17] COZMA M, BUTNARU A M, IONESCU R T. Automated essay scoring with string kernels and word embeddings[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018: 503-509.
[18] CHANG S, ZHANG Y, HAN W, et al. Dilated recurrent neural networks[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. CA: Curran Associates Inc, 2017: 77-87.
[19] KE Z, NG V. Automated essay scoring: A survey of the state of the art[C]//Proceedings of the 28th International Joint Conference on Artificial Intelligence Survey Track. Macao, China: IJCAI, 2019: 6300-6308.

编 辑 蒋 晓

(上接第 541 页)

[6] NETTO E J, PAULICENA H E, SILVA A R, et al. Analysis of energy consumption using HTTP and FTP protocols over IEEE 802.11[J]. IEEE Latin America Transactions, 2014, 12(4): 668-674.
[7] RAJAH K, RANKA S, XIA Y. Advance reservations and scheduling for bulk transfers in research networks[J]. IEEE Transactions on Parallel and Distributed Systems, 2008, 20(11): 1682-1697.
[8] DU D H C, TONG S R. Multilevel extendible hashing: A file structure for very large databases[J]. IEEE Transactions on Knowledge and Data Engineering, 1991, 3(3): 357-370.
[9] ZHAO X, LAM K Y, ZHU C J, et al. MVLevelDB: Using log-structured tree to support temporal queries in IoT[J]. IEEE Internet of Things Journal, 2021, 21(9): 1109-1120.
[10] FU S, HE L, HUANG C, et al. Performance optimization for managing massive numbers of small files in distributed file systems[J]. IEEE Transactions on Parallel and Distributed Systems, 2014, 26(12): 3433-3448.
[11] 王荣德, 荆一楠, 王欢, 等. 基于时间戳索引的 log 文件并行检索技术研究[J]. 计算机应用与软件, 2011, 28(2): 145-147.
WANG R D, JING Y N, WANG H, et al. Research on log file parallel research technology based on timestamp indexing[J]. Computer Applications and Software, 2011, 28(2): 145-147.
[12] MOHAMMADI H, SOLTANOLKOTABI M, JOVANOVIC M R. On the linear convergence of random search for discrete-time LQR[J]. IEEE Control Systems Letters, 2020, 5(3): 989-994.
[13] ALESSANDRI A, GAGGERO M. Fast moving horizon state estimation for discrete-time systems using single and multi iteration descent methods[J]. IEEE Transactions on Automatic Control, 2017, 62(2): 4499-4511.
[14] BOZORGI A M, FARASAT M, MAHMOUD A. A time and energy efficient routing algorithm for electric vehicles based on historical driving data[J] IEEE Transactions on Intelligent Vehicles, 2017, 2(4): 308-320.
[15] MARCHETTI M, STABILI D. READ: Reverse engineering of automotive data frames[J]. IEEE Transactions on Information Forensics and Security, 2019, 14(4): 1083-1097.
[16] MACHAEL K G. FTP access as a user-defined file system[J]. Operating Systems Review, 1994, 2(28): 73-80.
[17] WANG S, DA X, LI M, et al. Adaptive backtracking search optimization algorithm with pattern search for numerical optimization[J]. Journal of Systems Engineering and Electronics, 2016, 27(2): 395-406.
[18] 孙韩林, 金跃辉, 高雪松, 等. FTP 协议的测试及分析[J]. 计算机工程, 2008, 34(23): 133-138.
SUN H L, JIN Y H, GAO X S, et al. Measurement and analysis of FTP protocol[J]. Computer Engineering, 2008, 34(23): 133-138.
[19] LIU K. TCP performance over mobile data networks[D]. Hong Kong, China: The Chinese University of Hong Kong, 2004.

编 辑 叶 芳