计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(3):143−149 [doi: 10.15888/j.cnki.csa.008384]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于孪生网络和 BERT 模型的主观题自动评分系统①

钱升华[1,2]

[1](北京师范大学 人工智能学院, 北京 100875)
[2](天津财经大学珠江学院 数据工程学院, 天津 301811)
通信作者: 钱升华, E-mail: qsh0709@163.com

摘　要: 由于现在缺乏多语言教学中的主观题自动评分, 针对这一问题提出了一种基于孪生网络和 BERT 模型的主观题自动评分系统. 主观题的问题文本和答案文本通过自然语言预处理 BERT 模型得到文本的句向量, BERT 模型已经在大规模多种语言的语料上经过训练, 得到的文本向量包含了丰富的上下文语义信息, 并且能处理多种语言信息. 然后把问题文本和答案文本的句向量再通过深度网络的孪生网络进行语义相似度的计算, 最后连接逻辑回归分类器, 完成主观题的自动评分. 实验所使用数据集由 Hewlett 基金提供的英文数据集以及翻译后的中文数据集, 并以二次加权 Kappa 系数作为模型的评估指标. 实验结果表明, 对比其他基准模型, 基于孪生网络和 BERT 模型的自动评分系统在各个数据子集得到的结果最好.

关键词: 自然语言处理; 主观题自动评分; 孪生网络; 基于 transformer 的双向编码器表示; 二次加权 Kappa 系数

引用格式: 钱升华.基于孪生网络和 BERT 模型的主观题自动评分系统.计算机系统应用,2022,31(3):143–149. http://www.c-s-a.org.cn/1003-3254/8384.html

Automatic Short Answer Grading Based on Siamese Network and BERT Model

QIAN Sheng-Hua[1,2]

[1](School of Artificial Intelligence, Beijing Normal University, Beijing 100875, China)
[2](School of Data Engineering, Tianjin University of Finance and Economics Pearl River College, Tianjin 301811, China)

Abstract: To make up the gap of short answer grading systems in multilingual teaching, this study proposes an automatic short answer grading system based on siamese network and bidirectional encoder representations from transformers (BERT) model. First, the question and answer texts of short answers yield sentence vectors of texts with the natural language-preprocessed BERT model. The BERT model has been trained on a large-scale multilingual corpus, and the obtained text vectors contain rich contextual semantic information and can deal with multilingual information. Then, the sentence vectors of question and answer texts are subjected to the calculation of the semantic similarity in the siamese network of a deep network. Finally, a logistic regression classifier is employed to complete automatic short answer grading. The datasets used for automatic short answer grading tasks are provided by the Hewlett Foundation, and the quadratic weighted Kappa coefficient is used as the evaluation index of the model. The experimental results show that the proposed method outperforms other baseline models for automatic short answer grading in each data subset.

Key words: natural language processing; automatic short answer grading; siamese network; bidirectional encoder representation from transformers (BERT); quadratic weighted Kappa coefficient

① 基金项目: 天津财经大学珠江学院教学改革重点项目 (ZJJG20-04Z)
收稿时间: 2021-05-28; 修改时间: 2021-07-01; 采用时间: 2021-07-09; csa 在线出版时间: 2022-01-24

主观题自动评分处于发展阶段, 是自然语言理解、模式识别、人工智能和教育技术等相融合的研究问题. 主观题短答案自动评分 (automatic short answer grading, ASAG)[1] 是智能教育系统的重要组成部分, 教学评价中的关键环节. 主观题开放程度更高, 更好考查学生对知识掌握情况与学生学习能力. 但是一般老师教的学生比较多, 频繁考试以随时掌握学生学习情况, 假如采用人工阅卷方式, 老师工作量大大增加. 自动评分不仅可以帮助老师提高工作效率, 还可以避免阅卷过程中主观因素的影响.

主观题的短答案自动评分方法总结有 3 类: 一是规则匹配的方法, 根据参考答案建立评分规则[2], 按照规则进行自动评分. 二是传统机器学习的方法, 一般通过人工构建特征和监督机器学习算法完成自动评分. Saha 等人[3] 提出了基于字符特征和句子级特征相结合的自动评分模型. Sultan 等人[4] 提出根据文本相似度、词项权重等特征构建了高精度的评分分类器. 三是深度神经网络的学习方法, 从数据中自动学习特征, 以端到端的方式实现自动评分. Riordan 等人[5] 使用卷积神经网络 (convolutional neural network, CNN) 与长短期记忆网络 (long short-term memory, LSTM) 的神经网络完成自动评分. Huang 等人[6] 提出了使用连续词袋模型 (continuous bag-of-words model, CBOW) 与 LSTM 神经网络方法, 在中文自动评分任务得到了很好的结果.

上述文献为下步研究奠定了良好的基础, 提供了必要且充分的理论与实践支撑. 不过, 目前的主观题的短答案自动评分系统都是基于一种语言实现的 (比如英文、中文等), 但是在实际的评分系统中, 可能让学生答题的语言不止有一种, 可能是多种语言答题, 所以对自动评分系统提出了更高的要求. 随着深度学习技术的发展, 自然语言处理 (natural language processing, NLP) 也取得了巨大的突破, 特别是预训练技术的出现让多种语言使用同一个模型进行训练成为可能. 2018 年基于引入了注意力的 Transformer 架构的 (bidirectional encoder representations from Transformers, BERT)[7] 预训练出现, 刷新众多 NLP 任务, 使得预训练技术的发展迎来了一个高潮. 孪生网络 (siamese network)[8] 是进行相似度量的神经网络, 最初应用于图像处理中. 由于结构具有鲜明的对称性, 就像两个孪生兄弟, 这种神经网络结构被研究人员称作孪生网络. 在 NLP 中孪生网络基本是用来计算句子间的语义相似度的. Kenter 等人[9] 提出了 Siamese CBOW, 基于孪生网络和 CBOW 的方式来无监督式的训练句子的向量表示. Mueller 等人[10] 提出了一种 MaLSTM 的网络结构, 通过两个 LSTM 孪生网络来处理句子对, 取 LSTM 最后时刻的输出作为句子的向量, 用曼哈顿距离来计算两个句子向量的相似度. Neculoiu 等人[11] 提出用 LSTM 来处理句子对, 将句子对的关系看作是一个二分类的问题. 后来出现预训练模型 BERT 后, Reimers 等人[12] 提出 Sentence-BERT 模型, 把 BERT 预训练模型应用到孪生网络结构中.

由于现在许多主观题的发散性很强, 根据问题提供所有的标准答案比较困难, 不能单纯通过学生答案与标准答案进行自动评分[13], 而是提出了问题与学生答案相匹配的自动评分. 基于孪生网络和 BERT 模型的主观题自动评分系统, 可以把多种语言 (主要是中文和英文) 的问题文本和学生的答案文本转化自然语言处理的文本匹配问题, 使用预训练语言模型 BERT, 它已经在大型语料库学习了通用语义表示, 学习到的先验知识, 直接应用到具体任务中, 提高目标任务的效果. 利用 BERT 模型的优势, 使用孪生网络结构, 完成问题文本和学生的答案文本的相似度计算, 设计主观题自动评分系统.

1 自动评分模型

1.1 孪生网络

自动评分模型结构如图 1 所示, 整个模型沿用了孪生网络的结构, 主要由 2 层神经网络构成, 分别为编码层和输出层. 编码层通过同一个预训练模型 BERT 来处理, 直接用 BERT 的原始权重初始化, 在具体数据集上微调, 这种训练方式能更好地捕捉句子之间的关系, 生成更优质的句向量. 在预训练 BERT 模型中, 生成动态的两个文本对的向量, 然后接一层池化层, 得到包含整个文本含义的向量. 在输出层, 对得到的句向量, 使用计算文本对相似度进行相似度计算, 最后接逻辑回归分类器, 得到学生的分数.

该模型的整体运作流程是: 首先输入问题文本和答案文本数据对, 文本编码部分用同一个预训练模型 BERT, 通过 BERT 得到动态的字/词向量, 对向量进一步特征提取、压缩; 然后进行池化, 获得文本句子的整个语义的 embedding 向量 V_1、V_2; 计算这两个文本句向

量的相似度, 最后连接到一个分类器得到学生的分数.

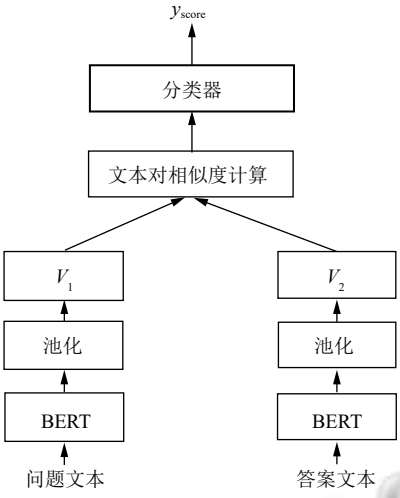

图 1 自动评分模型结构图

1.2 BERT 模型

自然语言预训练模型, 是先在足够大的数据集上采用无监督方法对复杂神经网络进行训练而产生的模型, 训练好的模型可以直接运用到目标任务中, 这样避免每次针对不同任务从头开始训练, 并且减少了对标注数据的依赖. 预训练的语言模型解决了针对小数据集, 提供好的模型初始化, 训练速度大大提升, 也避免对小数据的过度拟合.

BERT 模型也是经过大规模无标注语料训练后的模型, 得到文本的表示向量包含了丰富的语义表示, 直接把这些文本表示和模型参数运用到下游目标任务中. 本文利用训练好的模型作为文本向量嵌入到其他任务模型中. BERT 模型的结构图如图 2, E_1, E_2, …, E_N 是文本输入向量, 中间层是多个 Transformer 模块 (Trm), T_1, T_2, …, T_N 是输出向量, 代表文本的向量化表示.

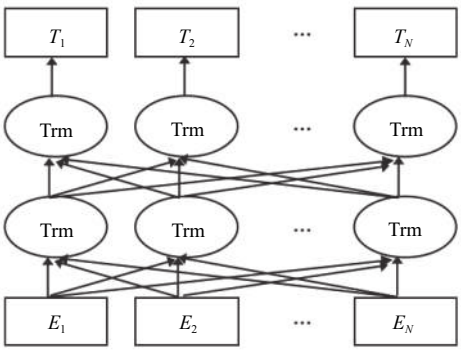

图 2 BERT 结构图

BERT 模型的输出如图 3 所示, 有两种形式输出: 一是每个 token 对应的向量表示, token 可以是字母、字、词等; 二是每个句子对应的向量表示, 模型最左边的输出符号 [CLS], 表示输出向量, 这个向量表示整个句子.

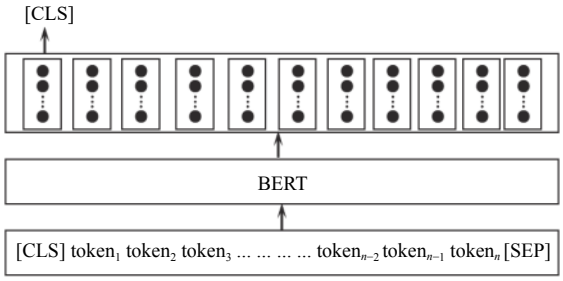

图 3 BERT 模型输出

参考图 1 整个模型图, BERT 输出后接一层池化层, 在这里池化层使用了 3 种策略: 一种 [CLS] token, 用这个向量表示整个文本; 另外两种是池化策略, 一种对文本中的所有的字/词向量进行平均, 叫做平均池化 (mean-pooling), 来表示整个文本; 另外一种对所有的字/词向量取最大值平均, 叫做最大值池化 (max-pooling), 来表示整个文本. 在 BERT-sentence[12] 对相应文本匹配的实验和本研究对数据集进行自动评分实验, 使用 mean-pooling 效果最好.

1.3 文本相似度计算

文本相似度计算是对输入的问题文本和答案文本的文本对的句向量进行计算, 输出文本对的相似程度, 通常用 [0, 1] 区间内的小数来表示. 主要用于文本相似度的计算方法有余弦相似度、欧几里得距离 (Euclidean distance) 相似度.

余弦相似度 (cosine similarity), 通过计算两个向量的夹角余弦值来评估他们的相似度. 此余弦值就可以用来表征这两个向量的方向性, 夹角越小, 余弦值越接近于 1, 它们的方向越吻合, 则越相似. 在文本相似度计算中, 可以用文本对的句向量的夹角余弦值来表示它们的差异. 这个余弦值通常被称为余弦距离. 假设空间中有两个向量, 它们的余弦相似度计算公式如式 (1).

$$similarity_{\cos} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (1)$$

欧几里得距离相似度, 通过两个向量的欧几里德距离表示相似度, 一般采用式 (2) 进行转换规约到 (0, 1] 之间, 距离越小, 相似度越大. 假设空间中有两个向量, 它们的欧几里得距离相似度计算公式如式 (2).

$$similarity_{\text{euclidean}} = \frac{1}{1+\sum_{i=1}^{n}(A_i - B_i)^2} \tag{2}$$

1.4 损失函数

损失函数采用均方误差, 均方误差 (*MSE*) 是实际值和预测值之间距离平方之和, 如式 (3), 用来作为损失函数.

$$MSE = \frac{\sum_{i=1}^{n}(y_i - y_i^p)^2}{n} \tag{3}$$

其中, n为样本数, y_i为目标值, y_i^p为预测值.

2 实验结果与分析

2.1 评价指标和实验数据集

2.1.1 评价指标

实验结果的评价指标使用平方加权 Kappa (quadratic weighted Kappa, *QWK*) 进行评估, Kappa 用于评价不同测量者对同一事物的判断是否准确. 平方加权是将线性加权的权值平方, 放大级别距离大的判定不一致程度, 权值计算如式 (4).

$$\omega_{i,j} = \frac{(i-j)^2}{(N-1)^2} \tag{4}$$

其中, i 和 j 分别是参考评分和预测评分, N 为可能评分的数量.

QWK 计算公式如式 (5).

$$QWK = 1 - \frac{\sum_{i,j}\omega_{i,j}O_{i,j}}{\sum_{i,j}\omega_{i,j}E_{i,j}} \tag{5}$$

其中, 矩阵$O_{i,j}$表示参考评分为 i 分和预测评分为 j 分的答案文本数量; 期望矩阵$E_{i,j}$表示随机生成的参考评分和预测评分两个向量的直方图外积. *QWK* 是评估不同评分之间一致性的重要标准, *QWK* 的计算是基于混淆矩阵, 取值为−1 到 1 之间, 通常大于 0. 当 *QWK*=1 时, 代表不同评分之间的完全一致性; 当 *QWK*=0 时, 表示不同评分之间的一致性完全随机.

2.1.2 数据集

本文使用的英文数据集是 Hewlett 基金会支持的学生文本答案自动评分大赛 (automated student assessment prize, ASAP) 提供, 该数据集涵盖不同主题, 包括 4 个学科, 由 10 个子集组成, 每个子集约 2 000 个学生答案, 每个学生成绩由 2 个评分员给出, 所以有判断评分一致性的 *QWK*, 具体情况如表 1.

表 1 ASAP 数据集概况

子集	科目	数量	*QWK*	平均长度	得分分布 (Score 0 / Score 1 / Score 2 / Score 3)
1	科学	2 229	0.90	50	22% / 27% / 31% / 20%
2	科学	1 704	0.83	50	14% / 25% / 36% / 25%
3	英语语言艺术	2 297	0.71	50	24% / 53% / 23%
4	英语语言艺术	2 033	0.67	50	39% / 53% / 8%
5	生物学	2 393	0.90	60	77% / 18% / 3% / 2%
6	生物学	2 396	0.92	50	84% / 9% / 4% / 3%
7	英语	2 398	0.88	50	51% / 25% / 24%
8	英语	2 398	0.77	50	30% / 26% / 44%
9	英语	2 397	0.80	40	24% / 41% / 35%
10	科学	2 186	0.85	60	17% / 47% / 36%

注: 数据来源于https://www.kaggle.com/c/asap-sas

中文数据集由于没有找到免费公开的评分的中文数据集, 使用了 ASAP 翻译的中文数据集. ASAP 数据集中子集 1、子集 2 和子集 10 是科学学科的任务, 它们没有很深的文化背景, 因此被认为适合转移到其他语言.

ASAP-ZH 中文数据集是由数据集提供商 BasicFinder 帮助一个研究项目[14] 收集的中文数据集. 根据 ASAP 数据集中的子集 1、子集 2 和子集 10 的问题, 针对每个子集收集了 314 个答案, 总共收集了 942 个答案. 它们是从 9–12 年级的高中学生那里收集的, 与 ASAP 数据集中的英语答案集相当. 答案以手写形式收集后, 把手动转录为数字文本形式. 再根据原始 ASAP 的一系列答案, 达成可接受的协议后, 两位中文的评分老师以 0 到 3 分 (子集 1 和子集 2) 或 0 到 2 分 (子集 10) 的分数对 ASAP-ZH 数据评分, 平均 *QWK* 为 0.7. 表 2 是该数据集的关键统计信息, 学科是科学学科.

ASAP-ZH^{MP} 中文数据集, 和 ASAP-ZH 中文数据集来自于同一个项目[14], 为了进行比较, 由 Google Translate API 将原始 ASAP 数据集中的子集 1、子集 2 和子集 10 的英文答案翻译为中文. 由两位中文的评分老师修正了部分错误, 重新给出了原始 ASAP 数据集相同评分指标的评分, 表 3 是该数据集的关键统计信息, 学科也是科学学科. 对照表 2 和表 3 所示, 表 3 翻译后答案的平均长度大于表 2 重新收集的在同一子集的原始中文答案的长度.

表 2 ASAP-ZH 数据集概况

子集	数量	*QWK*	平均长度	得分分布 (Score 0 / Score 1 / Score 2 / Score 3)
1	314	0.72	35.3	20% / 28% / 34% / 18%
2	314	0.70	38.2	36% / 43% / 18% / 3%
10	314	0.69	37.6	54% / 32% / 14%

表 3 ASAP-ZH^{MT} 数据集概况

子集	数量	*QWK*	平均长度	得分分布 (Score 0 / Score 1 / Score 2 / Score 3)
1	2229	0.96	68	22% / 27% / 31% / 20%
2	1704	0.94	94	22% / 27% / 31% / 20%
10	2186	0.91	61	22% / 27% / 31% / 20%

2.2 基线模型

HR1-HR2 代表由 2 个人工评分员进行评分, 本研究还给出了 5 种自动评分模型作为基准模型, 对比研究基于孪生网络和 BERT 模型的自动评分系统 (Siamese-BERT) 性能. 其中, Siamese-CNN[8] 和 Siamese-LSTM[11] 两个模型使用 GloVe 方法[15] 生成文本的词向量 (英文直接使用单词作为输入, 中文先使用 jieba 分词工具进行分词, 然后使用词作为输入), 然后使用孪生网络结构, 对两个文本进行匹配, 这两个模型分别在孪生网络的编码层使用 CNN 结构和 LSTM 结构. BERT、BERT-CNN、BERT-LSTM 三个模型直接使用文本对作为输入, 然后通过预训练 BERT 模型后连接线性分类器、CNN 网络、LSTM 网络按评分进行分类.

2.3 参数设置

Siamese-BERT 模型使用 Adam 随机优化器及均方根误差损失函数. 其中 BERT 模型使用 Google 提供的 BERT-base 预训练模型, 英文、中文都是使用的是 BERT-base-multilingual. 这个模型包括 12 层网络, 网络内部维度有 768 维, multi-head self-attention (heads=12), 共有 110 M 个参数. Siamese-BERT 模型的超参数设置见表 4.

表 4 超参数设置

层	参数名	参数值
BERT	模型名	BERT-base
池化层	池化策略	Mean-pooling
输出层	相似度计算公式	cos similarity
	分类器	线性回归分类器
	Dropout随机失活率	0.1
其他	模型迭代次数	4–6
	学习率	5e–5
	每批训练集的数据量	16

2.4 实验结果与分析

表 5 是英文数据集 ASAP 的实验结果, 在所有模型上都使用对实验结果行进评估, 表中加粗字体表示最好的实验结果. 本文研究的 Siamese-BERT 模型, 对比其他深度神经网络模型, 在各个数据子集上结果都是最好. Siamese-CNN、Siamese-LSTM 这 2 个模型适合进行文本相似度计算, 但是由于数据集规模很小, 在对小数据集结果都不是很理想. BERT、BERT-CNN、BERT-LSTM 这 3 个模型使用了预训练模型的优势, 把问题和答案直接根据分值作为文本分类问题, 整体效果也不是很理想; BERT-CNN、BERT-LSTM 在经过 BERT 与训练后的向量接复杂网络模型, 有时结果还不如 BERT 模型, 可能在小数据集上有过拟合的情况. Siamese-BERT 和人工评分比较, 在数据集的子集 3 结果很差, 以及在子集 5、子集 10 结果略差外, 其余结果和平均值都超过了人工评分 HR1-HR2.

表 6 是中文数据集 ASAP-ZH 和 ASAP-ZH^{MT} 的实验结果, 在所有模型上都使用 *QWK* 对实验结果行进评估, 表中加粗字体表示最好的实验结果. 实验结果与

英文数据集的结果类似, Siamese-BERT 实验结果最好. 但是在 ASAP-ZH^{MT} 是数据集上 Siamese-BERT 的结果比人工评分 HR1-HR2 结果有差距, 可能这个数据集后续有经过人工修正的原因.

表 5 基于 ASAP 数据集的不同模型的实验结果

模型	*QWK*										平均值
	1	2	3	4	5	6	7	8	9	10	
HR1-HR2	0.897	0.832	0.710	0.675	0.898	0.923	0.880	0.775	0.804	0.848	0.824
Siamese-CNN	0.712	0.703	0.454	0.543	0.687	0.623	0.584	0.532	0.643	0.653	0.614
Siamese-LSTM	0.734	0.714	0.502	0.512	0.594	0.631	0.653	0.623	0.675	0.687	0.633
BERT	0.745	0.725	0.543	0.767	0.714	0.687	0.648	0.645	0.776	0.710	0.696
BERT-CNN	0.872	0.549	0.533	0.522	0.701	0.571	0.512	0.688	0.767	0.727	0.644
BERT-LSTM	0.711	0.648	0.574	0.387	0.619	0.532	0.493	0.614	0.696	0.716	0.599
Siamese-BERT	**0.947**	**0.922**	**0.597**	**0.933**	**0.888**	**0.944**	**0.919**	**0.851**	**0.838**	**0.837**	**0.867**

表 6 基于中文数据集 ASAP-ZH、ASAP-ZH^{MT} 的不同模型的实验结果

模型	ASAP-ZH				ASAP-ZH^{MT}			
	1	2	10	平均值	1	2	10	平均值
HR1-HR2	0.720	0.700	0.690	0.703	0.957	0.937	0.912	0.935
Siamese-CNN	0.432	0.473	0.414	0.440	0.614	0.642	0.593	0.616
Siamese-RNN	0.457	0.456	0.431	0.448	0.628	0.637	0.621	0.629
BERT	0.519	0.542	0.527	0.529	0.590	0.671	0.703	0.655
BERT-CNN	0.543	0.523	0.508	0.525	0.409	0.453	0.603	0.488
BERT-LSTM	0.521	0.492	0.513	0.509	0.393	0.659	0.465	0.506
Siamese-BERT	**0.875**	**0.744**	**0.846**	**0.822**	**0.950**	**0.886**	**0.907**	**0.914**

3 结语

基于孪生网络和 BERT 模型的主观题自动评分 Siamese-BERT 模型通过使用孪生网络和自然语言处理的预训练模型 BERT, 比单纯地使用孪生网络和 BERT 模型性能提升很多. 实验结果也表明 对比其他基准模型, Siamese-BERT 模型在中、英文数据集上取得了最好的评分效果. 接下来的研究将着重对于学生答案的自动纠错, 完成整个智能教育系统的开发.

参考文献

1 Madnani N, Loukina A, Cahill A. A large scale quantitative exploration of modeling strategies for content scoring. Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications. Copenhagen: Association for Computational Linguistics, 2017. 457–467.

2 Ramachandran L, Cheng J, Foltz P. Identifying patterns for short answer scoring using graph-based lexico-semantic text matching. Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications. Denver: Association for Computational Linguistics, 2015. 97–106.

3 Saha S, Dhamecha TI, Marvaniya S, *et al*. Sentence level or token level features for automatic short answer grading? Use both. Proceedings of the 19th International Conference on Artificial Intelligence in Education. London: Springer, 2018. 503–517.

4 Sultan MA, Salazar C, Sumner T. Fast and easy short answer grading with high accuracy. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: Association for Computational Linguistics, 2016. 1070–1075.

5 Riordan B, Horbach A, Cahill A, *et al*. Investigating neural architectures for short answer scoring. Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications. Copenhagen: Association for Computational Linguistics, 2017. 159–168.

6 Huang YW, Yang X, Zhuang FZ, *et al*. Automatic Chinese reading comprehension grading by LSTM with knowledge adaptation. Proceedings of the 22nd Pacific-Asia Conference on Knowledge Discovery and Data Mining. Melbourne: Springer, 2018. 118–129.

7 Devlin J, Chang MW, 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. Minneapolis: Association for Computational Linguistics, 2019. 4171–4186.

8 Bertinetto L, Valmadre J, Henriques JF, *et al*. Fully-convolutional siamese networks for object tracking. European Conference on Computer Vision. Cham: Springer, 2016: 850–865.

9 Kenter T, Borisov A, de Rijke M. Siamese CBOW: Optimizing word embeddings for sentence representations. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: Association for Computational Linguistics, 2016. 941–951.

10 Mueller J, Thyagarajan A. Siamese recurrent architectures for learning sentence similarity. Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix: AAAI Press, 2016. 2786–2792.

11 Neculoiu P, Versteegh M, Rotaru M. Learning text similarity with Siamese recurrent networks. Proceedings of the 1st Workshop on Representation Learning for NLP. Berlin: Association for Computational Linguistics, 2016. 148–157.

12 Reimers N, Gurevych I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. arXiv: 1908.10084, 2019.

13 Burstein J, Tetreault J, Madnani N. The e-rater® automated essay scoring system. In: Shermis MD, Burstein J, eds. Handbook of Automated Essay Evaluation: Current Applications and New Directions. New York: Routledge, 2013. 55–67.

14 Ding YN, Horbach A, Wang HS, *et al*. Chinese content scoring: Open-access data sets and features on different segmentation levels. Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing. Suzhou: Association for Computational Linguistics, 2020. 347–357.

15 Pennington J, Socher R, Manning C. GloVe: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha: Association for Computational Linguistics, 2014. 1532–1543.