【电子与信息科学／Electronics and Information Science】

基于BERT-BiLSTM模型的短文本自动评分系统

夏林中，叶剑锋，罗德安，管明祥，刘俊，曹雪梅

深圳信息职业技术学院人工智能技术应用工程实验室，广东深圳 518172

摘　要：针对短文本自动评分中存在的特征稀疏、一词多义及上下文关联信息少等问题，提出一种基于BERT-BiLSTM(bidirectional encoder representations from transformers - bidirectional long short-term memory)的短文本自动评分模型．使用BERT(bidirectional encoder representations from transformers)语言模型预训练大规模语料库习得通用语言的语义特征，通过预训练好的BERT语言模型预微调下游具体任务的短文本数据集习得短文本的语义特征和关键词特定含义，再通过BiLSTM(bidirectional long short-term memory)捕获深层次上下文关联信息，最后将获得的特征向量输入Softmax回归模型进行自动评分．实验结果表明，对比CNN(convolutional neural networks)、CharCNN(character-level CNN)、LSTM(long short-term memory)和BERT等基准模型，基于BERT-BiLSTM的短文本自动评分模型所获的二次加权kappa系数平均值最优．

关键词：信号与信息处理；自然语言处理；BERT语言模型；短文本自动评分；长短时记忆网络；二次加权kappa系数

中图分类号：TP18；H08　　**文献标志码：**A　　**doi：**10.3724/SP.J.1249.2022.03349

Short text automatic scoring system based on BERT-BiLSTM model

XIA Linzhong, YE Jianfeng, LUO De'an, GUAN Mingxiang, LIU Jun, and CAO Xuemei

Engineering Applications of Artificial Intelligence Technology Laboratory, Shenzhen Institute of Information Technology, Shenzhen 518172, Guangdong Province, P. R. China

Abstract: Aiming at the problems of sparse features, polysemy of one word and less context related information in short text automatic scoring, a short text automatic scoring model based on bidirectional encoder representations from transformers - bidirectional long short-term memory (BERT-BiLSTM) is proposed. Firstly, the large-scale corpus is pre-trained with bidirectional encoder representations from transformers (BERT) language model to acquire the semantic features of the general language. Then the semantic features of short text and the semantics of keywords in a specific context are acquired through the short text data for the pre-fine tuning downstream specific tasks set pre-fined by BERT. And then the deep-seated context dependency is captured through bidirectional long short-term memory (BiLSTM). Finally, the obtained feature vectors are input into Softmax regression model for automatic scoring. The experimental results show that compared with other benchmark models of convolutional neural networks (CNN), character-level CNN (CharCNN), long short-term memory (LSTM) and BERT, the short text automatic scoring

Received: 2021-09-19; **Accepted:** 2021-11-01; **Online (CNKI):** 2022-04-13
Foundation: Science and Technology Platform Construction for Universities of Department of Education of Guangdong Province (2020KTSCX301); Shenzhen Basic Research Foundation (JCYJ20190808093001772); Special Support Program of Leading Professional (Outstanding Teacher) of National High-Level Personnel of China (ZTZ [2018] No. 6)
Corresponding author: Associate professor XIA Linzhong. E-mail: 43966506@qq.com
Citation: XIA Linzhong, YE Jianfeng, LUO De'an, et al. Short text automatic scoring system based on BERT-BiLSTM model [J]. Journal of Shenzhen University Science and Engineering, 2022, 39(3): 349-354. (in Chinese)

model based on BERT-BiLSTM achieves the best average value of quadratic weighted kappa coefficient.

Key words: signal and information processing; natural language processing; BERT language model; short text automatic scoring; long short-term memory net; quadratic weighted kappa coefficient

短文本自动评分是指使用计算机对人工问题的语言文本进行自动评分，由于回答问题的语言文本长度一般都较简短，所以称为短文本. 近年来，随着教育信息化水平的不断提升，学生借助各种智能终端、互联网及移动互联网进行同步学习，学习过程中会产生大量的过程性语言文本信息. 为进一步提升学习效果，需要对这些过程性语言文本信息进行实时分析，并及时向学生反馈分析评价结果，工作量非常大[1]. 文本自动评价的机器学习算法通过统计分析文本的字数、单词数、拼写错误数、单词平均字母数、句子数及词频-逆文本频率指数(term frequency-inverse document frequency, TF-IDF)等特征来评价文本，可用于实时分析学习中产生的大量过程性语言文本信息. 目前基于该思路的商用文本评分器主要包括PEG(project essay grader)[2]和E-rater(electronic essay rater)[3]，其应用效果好，但无法抽取文本的语义特征. 潜在语义分析(latent semantic analysis, LSA)算法[4]不仅能获取文本的语义特征，还能解决同义词问题，但计算消耗资源量大，无法获取文本词序信息. LDA(latent Dirichlet allocation)主题模型[5]很好解决了捕获文本词序信息的问题，其计算资源消耗也比LSA小. 然而，上述各类算法对于文本深层语义信息与上下文关联信息的挖掘能力非常有限.

近些年深度神经网络(deep neural networks, DNN)算法在文本分析中得到广泛应用[6]，其深层语义信息及上下文关联信息的挖掘能力促进了文本自动评分的发展[7]. DNN算法中最重要的是如何通过词向量表达文本[8]，使用较多的词向量获取方法包括Word2Vec[9]、C&W[10]及GloVe[11]等. 用获取的词向量表示文本并作为DNN的输入，从而实现对文本评分. DNN主要分为卷积神经网络(convolutional neural networks, CNN)和循环神经网络(recurrent neural networks, RNN)，RNN捕捉较短上下文依赖关系的能力非常强，但上下文之间的距离越长，RNN的捕捉能力就越弱. RNN改进型的长短时记忆(long short-term memory, LSTM)神经网络模型[12]更擅长捕捉长距离的上下文依赖关系信息[13]. 双向长短时记忆(bidirectional LSTM, BiLSTM)神经网络由前后两个方向的LSTM组合而成，能够捕捉当前词与上下文的依赖关系，从而更好地执行文本评分任务[14-15].

BERT(bidirectional encoder representations from transformers)[16]模型是Google公司提出的一种基于深度学习的语言表示模型，在11种不同的自然语言处理测试任务中效果最佳，其中也包括文本分类任务[17-18]. 基于BERT模型的文本分类主要包括预训练(pre-training)和预微调(fine-tuning)两个过程，前者利用大规模未经标注的文本语料进行自监督训练，有效学习文本语言特征及深层次文本向量表示，从而形成预训练BERT模型；预微调则直接通过预训练好的BERT模型作为起始模型，根据文本分类任务的特点，输入人工标注好的数据集，完成模型的进一步拟合与收敛. 本研究结合BiLSTM与BERT模型的优点，建立BERT-BiLSTM短文本自动评分模型，并针对已人工标注好的短文本数据集进行分析，克服了文本较短且因用语偏口语化带来的特征稀疏与一词多义问题.

1 基于BERT-BiLSTM短文本自动评分模型

为解决短文本特征稀疏的问题，采用BiLSTM模型捕获隐藏于上下文深度语义依赖关系中的更多特征；短文本的语义针对性强，一词多义现象普遍，采用BERT模型能够较好解决这一问题.

1.1 BiLSTM网络层

BiLSTM模型由1个正向LSTM与1个反向LSTM叠加而成，其具体结构如图1. 其中，$\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N$为输入词向量；$\overrightarrow{\boldsymbol{h}}_t\ (t = 1, 2, \cdots, N)$为$t$时刻正向LSTM隐藏层的输出向量，由当前时刻输入向量x_t和前一时刻的正向LSTM输出向量$\overrightarrow{\boldsymbol{h}}_{t-1}$共同确定，记为$\overrightarrow{\boldsymbol{h}}_t = f(\boldsymbol{x}_t, \overrightarrow{\boldsymbol{h}}_{t-1})$；$\overleftarrow{\boldsymbol{h}}_t(t = 1, 2, \cdots, N)$为$t$时刻反向LSTM隐藏层的输出向量，由当前$\boldsymbol{x}_t$和前一时刻反向LSTM输出$\overleftarrow{\boldsymbol{h}}_{t-1}$共同确定，记为$\overleftarrow{\boldsymbol{h}}_t = f(\boldsymbol{x}_t, \overleftarrow{\boldsymbol{h}}_{t-1})$；$\boldsymbol{h}_t$为$t$时刻的BiLSTM模型输出，由$\overrightarrow{\boldsymbol{h}}_t$和$\overleftarrow{\boldsymbol{h}}_t$共同确定的，记为$\boldsymbol{h}_t = \boldsymbol{w}_t\overrightarrow{\boldsymbol{h}}_t + \boldsymbol{v}_t\overleftarrow{\boldsymbol{h}}_t + \boldsymbol{b}_t$，其中，$\boldsymbol{w}_t$为正向LSTM输出的权重矩阵；$\boldsymbol{v}_t$为反向LSTM输出的权重矩阵；$\boldsymbol{b}_t$为权重矩阵的偏置.

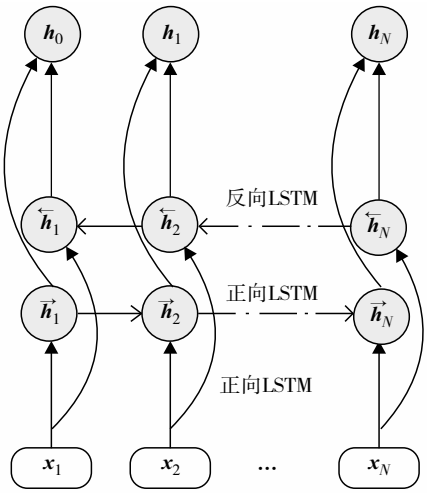

图1　BiLSTM模型结构

Fig. 1　The architecture of the BiLSTM model

1.2　BERT语言模型

BERT模型是一种旨在取代或改进RNN或CNN的全新架构，其基于注意力机制对文本数据进行建模[19]．如图2所示，BERT模型采用12或24层双向Transformer编码结构，其中，$\boldsymbol{E}_1, \boldsymbol{E}_2, \cdots, \boldsymbol{E}_N$为输入向量；$\boldsymbol{T}_1, \boldsymbol{T}_2, \cdots, \boldsymbol{T}_N$为经过多层Transformer编码器后的输出向量．BERT通过大规模语料对模型进行预训练，获取适应通用自然语言处理任务的模型网络参数，再使用当前任务的文本数据对预训练网络参数进行预微调，使模型适应当前任务．

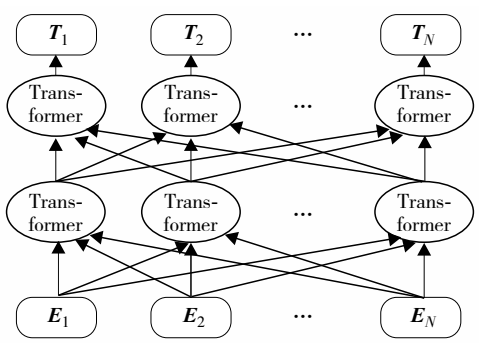

图2　BERT模型结构

Fig. 2　The architecture of the BERT model

1.2.1　Transformer编码器结构

如图3所示，Transformer编码器结构包括自注意力机制(self-attention)和前馈(feed forward)神经网络单元，单元之间设计残差连接层（add&normal）．Transformer是一个基于自注意力机制的Seq2seq模型，BERT模型主要使用Seq2seq的Encoder部分．自注意力机制单元是Transformer编码器的核心，其计算每个词与其所在句子中所有词的相互关系，据此调整每个词的权重，从而获取每个词新的向量表达式．Encoder的输入是文本的词向量表示($\boldsymbol{X}_1$、$\boldsymbol{X}_2$)及每个词的位置信息，将自注意力机制单元的输出进行相加和归一化处理，使输出具有固定均值(大小为0)和标准差(大小为1)．归一化后的向量传入前馈神经网络进行残差处理和归一化输出．

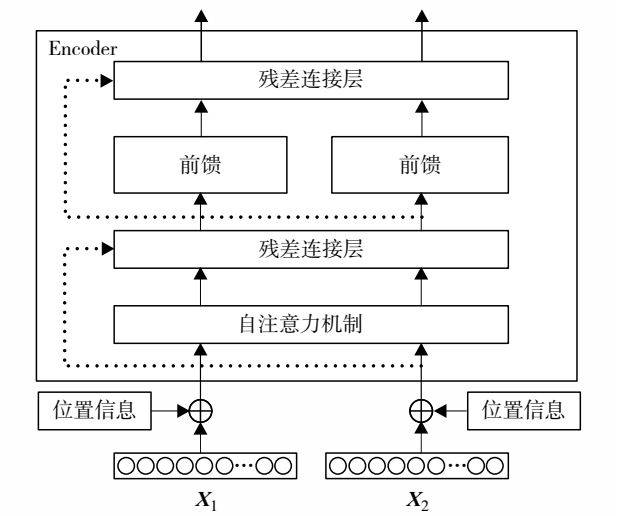

图3　Transformer编码器结构

Fig. 3　The architecture of the Transformer encoder

1.2.2　BERT模型的输入表示

BERT模型的输入由标记嵌入、分段嵌入和位置嵌入部分叠加表示，如图4．其中，标记嵌入为第1个标志是$\boldsymbol{E}_{[\mathrm{CLS}]}$的词向量，其初始值可随机产生，$\boldsymbol{E}_{[\mathrm{SEP}]}$为句子间的分隔标志；分段嵌入为区分不同句子向量；位置嵌入表示文本中每个词的位置信息．可见，BERT模型的输入向量不仅含有短文本语义信息，还包括了不同句子之间的区分信息与每个词的位置信息．

1.2.3　BERT模型的预训练与预微调

BERT模型的预训练过程使用大规模未经标注过的文本语料，经过充分自监督训练后有效学习文本的通用语言特征，得到深层次文本词向量表示，并获得预训练模型．具有逻辑关系与的优点．预训练过程的掩藏语言模型(masked language model，MLM)依据上下文语义信息对随机掩盖的词进行预测，可以更好学习上下文内容特征；下一句预测(next sentence predication，NSP)[20]为每个句子的句首和句尾分别插入[CLS]和[SEP]标签，通过学习句子间的关系特征预测两个句子的位置是否相邻．

预微调过程直接将预训练获取的网络参数作为模型起始，根据下游任务输入人工标注好的数据

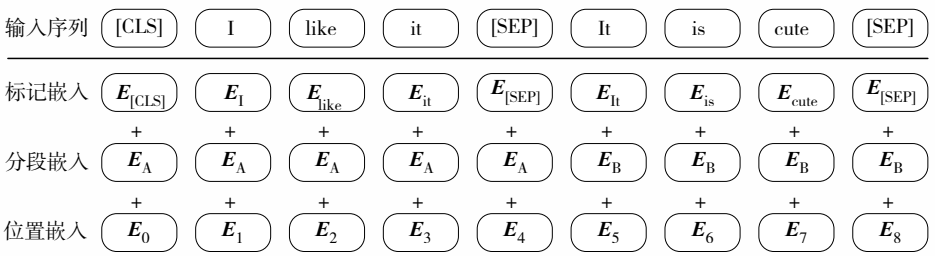

图4 BERT模型的输入

Fig. 4 Input of the BERT model

集，使BERT模型得到进一步拟合与收敛，得到可用于下游任务的深度学习模型.

1.3 BERT-BiLSTM模型

本研究设计基于BERT-BiLSTM的短文本自动评分模型，由BERT层和BiLSTM层构成，如图5. 其中，$\overleftarrow{\boldsymbol{h}}_1, \overleftarrow{\boldsymbol{h}}_2, \cdots, \overleftarrow{\boldsymbol{h}}_N$为BiLSTM层反向LSTM隐藏层状态；$\overrightarrow{\boldsymbol{h}}_1, \overrightarrow{\boldsymbol{h}}_2, \cdots, \overrightarrow{\boldsymbol{h}}_N$为BiLSTM层正向LSTM隐藏层状态；$\boldsymbol{T}_1, \boldsymbol{T}_2, \cdots, \boldsymbol{T}_N$为BERT层输出向量；$\boldsymbol{C}$为BERT短文本级输出向量；$\boldsymbol{E}_1, \boldsymbol{E}_2, \cdots, \boldsymbol{E}_N$为BERT层词向量输入；$\boldsymbol{E}_{[\mathrm{CLS}]}$为BERT层自动添加的短文本开头表示符号；$\mathrm{Tok}_1, \mathrm{Tok}_2, \cdots, \mathrm{Tok}_N$为短文本输入；[CLS]为随机赋予初值的短文本开头；Softmax回归模型通过分析输入特征向量来对短文本进行分类. BERT模型针对大规模语料库的预训练可以习得通

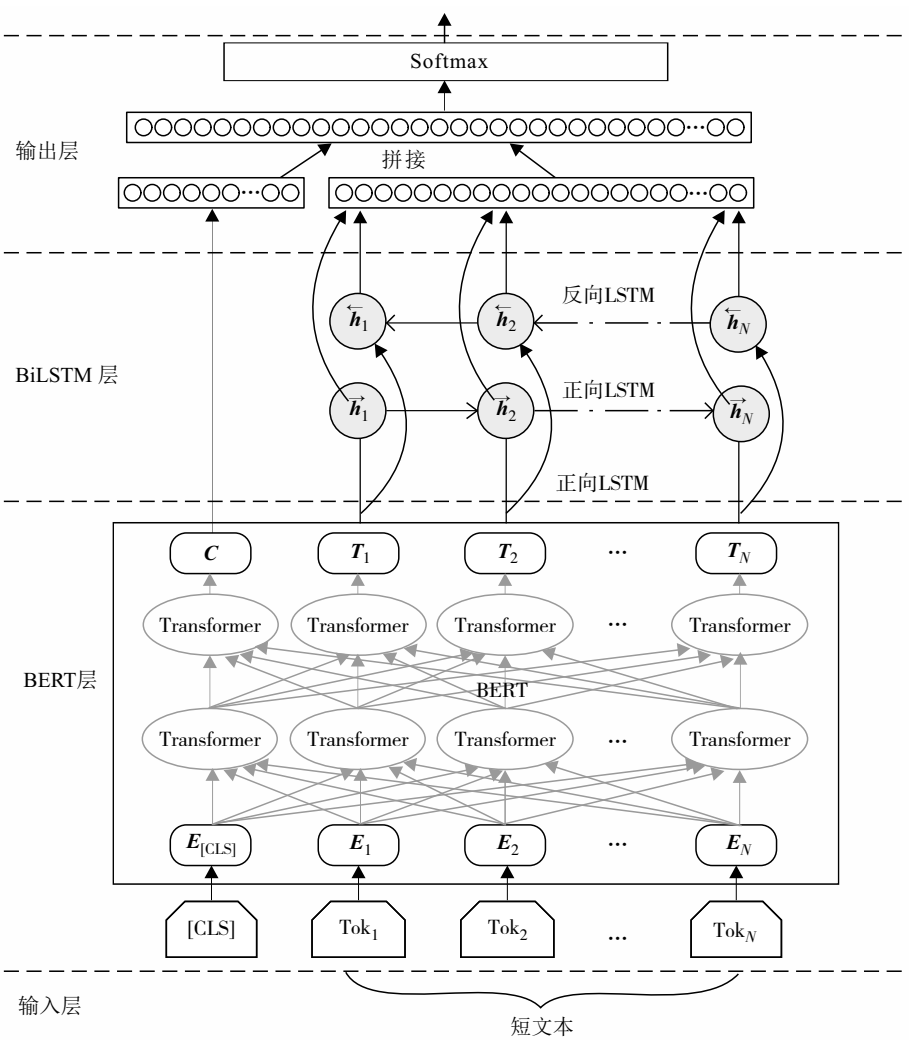

图5 基于BERT-BiLSTM的短文本自动评分模型结构

Fig. 5 The architecture of the short text automatic scoring model based on BERT-BiLSTM

用语言的语义特征，针对已标注的短文本数据集微调BERT模型参数，以适应短文本数据集的语义特点．本研究使用的短文本数据集中，有些词的语义与其在通用语言中的语义存在差异，而BERT模型的输出可以根据上下文信息特点进行调整，即同一词语在不同上下文中对应的向量编码输出不同，这样就可以解决一词多义的问题．BiLSTM模型能够从前向和后向获取字词与上下文的语义关联信息，进而捕获深层次的上下文依赖关系．

2　实验与分析

2.1　实验设置

2.1.1　数据集

实验使用的短文本数据集包括训练集和测试集，由Hewlett基金会提供．训练集和测试集分别包含17 207篇和5 224篇已人工评分的短文本，实验从训练集中随机抽取20%的短文本作为校验集．训练集和测试集分别由10个子集组成，见表1．

表1　短文本数据集

Table 1　Short text dataset

子集	文本主题	短文本数量/篇		平均长度/单词数	得分/分
		训练集	测试集		
1	科学1	1 672	557	50	0～3
2	科学2	1 278	426	50	0～3
3	英语语言文学1	1 891	406	50	0～2
4	英语语言文学2	1 738	295	50	0～2
5	生物学1	1 795	598	60	0～3
6	生物学2	1 797	599	50	0～3
7	英语1	1 799	599	50	0～2
8	英语2	1 799	599	50	0～2
9	英语3	1 798	599	40	0～2
10	科学3	1 640	546	60	0～2

2.1.2　评价指标与参数设置

本研究采用二次加权kappa（quadratic weighted kappa，QWK）系数[21] κ评估预测分数与专家打分的一致性，且$0 \leqslant \kappa \leqslant 1$．$\kappa = 0$表示作文不同评分之间的一致性完全随机；$\kappa = 1$表示作文不同评分之间的一致性完全相同．在预训练好的BERT模型基础上，采用Adam优化器对短文本数据集进行预微调，学习率设置为2×10^{-5}，权重衰减系数设置为1×10^{-5}．

2.2　实验结果与讨论

将本研究BERT-BiLSTM模型的κ值与基准模型CharCNN（character-level CNN）、CNN、LSTM和BERT进行对比，结果见表2．所有模型均使用相同的数据集，训练集与测试集的短文本篇数相同．

表2　BERT-BiLSTM模型与基准模型的κ值对比[1)]

Table 2　The quadratic weighted kappa coefficients comparison between BERT-BiLSTM model and benchmark models

子集	CharCNN	CNN	LSTM	BERT	BERT-BiLSTM
1	0.68	0.68	0.70	0.79	0.85
2	0.58	0.67	0.66	0.70	0.79
3	0.29	0.27	0.28	0.37	0.31
4	0.56	0.55	0.59	0.69	0.69
5	0.72	0.75	0.78	0.75	0.83
6	0.80	0.74	0.74	0.84	0.83
7	0.55	0.58	0.60	0.66	0.65
8	0.43	0.50	0.54	0.60	0.64
9	0.67	0.64	0.70	0.80	0.82
10	0.61	0.67	0.71	0.74	0.75
均值	0.60	0.62	0.65	0.71	0.72
合并集	0.70	0.73	0.75	0.79	0.80

[1)]灰底数字代表最优实验结果

表2中的合并集表示将子集1至子集10合并为1个大集合．可见，对比CharCNN、CNN和LSTM模型，BERT与BERT-BiLSTM的模型的κ值最优；相比BERT模型，BERT-BiLSTM模型在子集1、2、5、8、9及10上的κ值分别提升了6%、9%、8%、4%、2%及1%；对比其他所有模型，BERT-BiLSTM模型的κ平均值最高．因此，BERT-BiLSTM模型短文本自动评分的整体性能最优．

由表2还可见，子集3的κ值最低，这是因为子集3为开放式英语语言文学问题，回答短文本多为学生根据自己的理解对相关语句进行的解释，因此，特征不明显．该子集的上下文关联信息较少，人工评分员在该子集上的评分一致性也很低．

结　语

本研究提出基于BERT-BiLSTM的短文本自动评分模型，通过BERT语言模型表示短文本向量、BiLSTM捕获短文本的上下文信息深层依赖关系，提升了短文本自动评分性能．实验结果表明，本模型不仅在短文本数据集的子集上取得最好的自动评分效果，其整体自动评分性能也优于其他基准模型．后续研究将在本模型的句子表征上融入标点符

号及情感词等位置信息，以丰富短文本的句子向量特征表示，并设计出更高效、简洁的神经网络结构，进一步提高短文本自动评分效果.

基金项目：广东省教育厅高校科研平台资助项目(2020KTSCX301)；深圳市基础研究计划资助项目（JCYJ20190808093001772)；国家高层次人才特殊支持计划领军人才（教学名师）资助项目（组厅字[2018]6号）
作者简介：夏林中（1980—），深圳信息职业技术学院副教授、博士. 研究方向：自然语言处理.
E-mail: 43966506@qq.com
引　文：夏林中，叶剑锋，罗德安，等. 基于BERT-BiLSTM模型的短文本自动评分系统[J]. 深圳大学学报理工版，2022，39(3)：349-354.

参考文献 / References：

[1] DIKLI S. An overview of automated scoring of essays [J]. Journal of Technology, Learning, and Assessment, 2006, 5 (1): 1-35.

[2] PAGE E B. The imminence of grading essays by computer [J]. Phi Delta Kappan, 1966, 48: 238-243.

[3] CLAUDIA L, MARTIN C. C-rater: automated scoring of short-answer questions [J]. Computers and the Humanities, 2003, 37(4): 389-405.

[4] DEERWESTER S, DUMAIS S T, FURNAS G W, et al. Indexing by latent semantic analysis [J]. Journal of the American Society for Information Science, 1990, 41(6): 391-407.

[5] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3 (4/5): 993-1022.

[6] TANG D. Sentiment-specific representation learning for document-level sentiment analysis [C]// Proceedings of the 8th International Conference on Web Search and Data Mining. Shanghai, China: ACM, 2015: 447-452.

[7] PANG B, LEE L. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales [C]// Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. Ann Arbor, USA: ACL, 2005: 115-124.

[8] LEE K, HAN S, MYAENG S H. A discourse-aware neural network-based text model for document-level text classification [J]. Journal of Information Science, 2018, 44(6): 715-735.

[9] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [EB/OL]. (2013-09-07) [2021-08-05]. https: //arxiv. org/abs/1301. 3781.

[10] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch [J]. Journal of Machine and Learning Research, 2011, 12: 2493-2537.

[11] PENNINGTON J, SOCHER R, MANNING C D. Glove: global vectors for word representation [C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Doha, Qatar: ACL, 2014: 1532-1543.

[12] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735-1780.

[13] RAN Xiangdong, SHAN Zhiguang, FANG Yufei, et al. An LSTM-based method with attention mechanism for travel time prediction [J]. Sensors, 2019, 19(4): 861.

[14] GRAVES A, SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures [J]. Neural Networks, 2005, 18(5/6): 602-610.

[15] BIN Yi, YANG Yang, SHEN Fumin, et al. Describing video with attention-based bidirectional LSTM [J]. IEEE Transactions on Cybernetics, 2019, 49(7): 2631-2641.

[16] 刘欢，张智雄，王宇飞. BERT模型的主要优化改进方法研究综述[J]. 数据分析与知识发现，2021，5(1)：3-15.
LIU Huan, ZHANG Zhixiog, WANG Yufei. A review on main optimization methods of BERT [J]. Data Analysis and Knowledge Discovery, 2021, 5(1): 3-15. (in Chinese)

[17] 方晓东，刘昌辉，王丽亚，等. 基于BERT的复合网络模型的中文文本分类[J]. 武汉工程大学学报，2020，42(6)：688-692.
FANG Xiaodong, LIU Changhui, WANG Liya, et al. Chinese text classification based on BERT′s composite network model [J]. Journal of Wuhan Institute of Technology, 2020, 42(6): 688-692. (in Chinese)

[18] 段丹丹，唐加山，温勇，等. 基于BERT模型的中文短文本分类算法[J]. 计算机工程，2021，47(1)：79-86.
DUAN Dandan, TANG Jiashan, WEN Yong, et al. Chinese short text classification algorithm based on BERT model [J]. Computer Engineering, 2021, 47(1): 79-86. (in Chinese)

[19] DEVLIN J, CHANG Mingwei, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C]// Proceedings of NAACL-HLT. Minneapolis, USA: ACL, 2019: 4171-4186.

[20] SU Jing, DAI Qingyun, GUERIN F, et al. BERT-hLSTMs: BERT and hierarchical LSTMs for visual storytelling [J]. Computer Speech & Language, 2021, 67: 1-14.

[21] 夏林中，罗德安，刘俊，等. 基于注意力机制的双层LSTM自动作文评分系统[J]. 深圳大学学报理工版，2020，37(6)：559-566.
XIA Linzhong, LUO De'an, LIU Jun, et al. Attention-based two-layer long short-term memory model for automatic essay scoring [J]. Journal of Shenzhen University Science and Engineering, 2020, 37(6): 559-566. (in Chinese)

【中文责编：方圆；英文责编：溯心】