基于深度学习的中文语法错误诊断方法研究

王 辉[1],潘俊辉[1],王浩畅[1],张 强[1],张 岩[1],Marius. Petrescu[2]
(1.东北石油大学 计算机与信息技术学院,黑龙江 大庆 163318;
2.普罗莱斯蒂石油天然气大学,罗马尼亚 什蒂市 100680)

摘 要:随着中国国际影响力的日益提高和汉语国际地位的提升,学习和使用汉语的国际学者越来越多。中文文本校对技术有助于各个领域处理所涉及到的文本错误,其中中文语法错误诊断是中文计算机辅助学习的研究热点之一。鉴于此,根据中文语法错误诊断的特点,通过分析现有中文语法错误诊断方法存在的问题,提出一种基于注意机制的双向长短期记忆网络(BI-LASM-ATT)与条件随机场(CRF)相结合的模型应用于中文语法错误诊断研究。该模型采用 jieba 分词技术对数据进行分词和词性标注等预处理工作,利用 Skip-gram 模型得到词向量表示,作为 BI-LSTM-ATT 模型的词嵌入层,获取到两个方向上的长距离信息提供给 CRF 模型进行序列标注。在 NLPCC2018 的 TASK2 提供的数据集上的实验结果表明,该模型对比传统语法错误诊断模型,在中文语法错误诊断的 Accuracy、精确率、召回率和 F_meature 方面均有明显提高。

关键词:深度学习;条件随机场;长短期记忆网络;注意机制;语法错误诊断
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2020)11-0069-05
doi:10.3969/j.issn.1673-629X.2020.11.013

Research on Chinese Grammar Error Diagnosis Method Based on Deep Learning

WANG Hui[1],PAN Jun-hui[1],WANG Hao-chang[1],ZHANG Qiang[1],
ZHANG Yan[1],Marius. Petrescu[2]
(1.School of Computer and Information Technology,Northeast Petroleum University,Daqing 163318,China;
2.Petroleum-Gas University of Ploiesti,Ploiesti 100680,Romania)

Abstract:With the increasing international influence of China and the promotion of the international status of Chinese,more and more international scholars are learning and using Chinese. The Chinese text automatic proofreading technology is helpful to deal with the text errors involved in various fields,among which Chinese grammar error diagnosis is one of the research hotspots in Chinese computer-aided learning. Based on this,according to the characteristics of Chinese grammar error diagnosis,after the problems existing in the existing Chinese grammar error diagnosis methods are analyzed,a model of bidirectional long-term memory network (BI-LSTM-ATT) and conditional random field (CRF) is proposed based on attention mechanism for Chinese grammar error diagnosis. In this model,the jieba is used as data preprocessing of word segmentation and POS tagging,and the Skip-gram model is used to get word vector representation as word embedding layer of BI-LSTM-ATT to capture long-distance context information in two directions for sequence labeling of the CRF. The experiments are carried out in the data set from NLPCC2018 TASK2,which show that the proposed model has significantly higher accuracy,precision,recall,and F_meature than traditional model of grammar error diagnosis.

Key words:deep learning;CRF;LSTM;attention mechanism;grammar error diagnosis

0 引 言

随着“一带一路”倡议的提出,风靡全球的“汉语热”承载了当今世界各国人民对中华文明的深切渴望。汉语水平考试(HSK)的萌出,以及中文在国际舞台的迅猛传播,逐渐彰显了中国国际地位的提升及中国文化的重要性。教育部、国家语委印发的《国家中长期语言文字事业改革和发展规划纲要(2012-2020年)》[1]中提出必须纠正语言文字的不规范使用,构建

收稿日期:2020-01-03 **修回日期**:2020-05-07
基金项目:国家自然科学基金(61402099,61702093);黑龙江省自然科学基金项目(2018003);东北石油大学青年科学基金(2018QNL-8,2018QNL-49)
作者简介:王 辉(1979-),女,副教授,硕士,CCF 专业会员(B3489M),通讯作者,研究方向为自然语言处理。

和谐语言生活,服务社会主义文化强国建设。然而,中文作为最难学的语言之一,与英文有很大的不同,其在发音、语法、语义、形态上有着极强的复杂性和灵活性,无形中加大了中文语法错误诊断的难度。包含语法错误的中文自然语言描述,往往会导致计算机做出错误的回应,直接影响着人工智能的水平。因此,中文语法错误诊断的研究意义重大,已然成为了计算机处理自然语言领域的一个重要研究方向。

2014 年,Yu 发布的中文语法错误诊断共享任务中,将语法错误诊断重点集中在语料中出现的四类语法错误,即词语冗余(redundant words,R)、词语缺失(missing words,M)、词语误用(word selection errors,S)和词语乱序(word ordering errors,W)[2]。近年来,传统的中文语法错误诊断方法多采用基于规则、统计、语料、特征等策略,结合机器学习判定中文语法错误,这些方法的最大缺陷在于难以关联上下文语义,导致中文语法检错效果并不是很好。随着深度学习技术在计算机视觉、语音识别、自然语言处理等诸多领域的迅猛发展和优异表现,相比于传统的检错方法,基于深度学习的中文语法检错方法逐渐占据了中文语法错误诊断技术的主导地位。

2014 年,Shuk-Man Cheng 等提出应用 CRF 和基于排序 SVM 算法的模型检测词语乱序错误[3]。2015 年,Jui-Feng Yeh 等先采用 CKIP 自动标注系统分词,再应用 CRF 模型进行中文语法错误诊断,取得了较好的召回率和精准率[4]。同年,韩文颖构建了基于序列标注的 CRF 语法错误检测模型,提高了识别层的精度[5]。2016 年,在参加中文语法纠错任务的队伍提交方法中多次出现了深度学习相关算法,在各级别的评估下都取得了不错的成绩。例如,北京大学采用了 Bi-LSTM 模型诊断方法,云南大学提出了基于字向量的 CNN 模型和 LSTM 模型的诊断方法[6]。同年,Zheng 等也采用了 LSTM 模型进行句子语法错误标注[7]。2017 年,Yang 等提出的 LSTM-CRF 模型很大程度上提高了句子标注的准确性,获得了同年国际自然语言处理联合会议(IJCNLP 2017)中文语法纠错第一名[8]。2019 年,杨劲男探讨和对比了现有机器学习模型以及其他神经网络模型,提出一种基于门控递归单元与条件随机场的组合模型(GRU_CRF),提高了文本特征拟合度,同时证明 CRF 在判错定位时确实具有较好的效果[9]。然而,现有的方法往往需要大量人工标注特征,同时忽略了特征词的上下文信息的影响。

鉴于此,该文提出一种基于 BI-LASM-ATT 与 CRF 相结合的模型应用于中文语法错误诊断研究。首先,对句子进行断句、按固定长度补全句子、采用 jieba 分词技术进行数据预处理;其次,为反映词对全文信息的重要程度,以及词位置的影响,利用 Skip-gram 模型标注词向量;最后,将生成的向量表示作为基于注意机制的双向 LASM 模型的词嵌入层数据,利用 CRF 模型进行序列标注。

1 相关工作

1.1 分词技术

词语是自然语言处理最基本的一个元素,在进行语法错误检错之前,首先需要对测试文本进行词语划分,即将句子分割成独立的词语。目前,英文分词技术主要采用规则模型主导和统计模型主导的分词技术,中文分词技术主要采用基于字符串匹配、基于理解和基于统计的分词方法,其中最为常用的是 Python 的 jieba 分词组件。

jieba 分词支持以下四种分词模式:

(1)精确模式。将句子精确切分以作文本分析。

(2)全模式。快速扫描出句子中所有可成词的词语。

(3)搜索引擎模式。在精确模式的基础上,再次切分长词,以提高召回率。

(4)paddle 模式。利用深度学习框架,训练序列标注网络模型实现分词与词性标注。

jieba 分词技术能够实现高效的词图扫描,生成句中汉字所有可能的成词情况的有向无环图;并采用了动态规划算法以查找最大概率路径的方式,找出以词频为基础的最大切分组合;对于未登录到词库的词,使用了基于汉字成词能力的隐马尔可夫(hidden Markov model,HMM)模型和维特比(Viterbi)算法得到分词结果。同时,jieba 分词技术能够实现词性标注工作,为语法检错提供了更详尽的数据。

1.2 词嵌入技术

分词之后得到的每个词语相互独立,一定程度上忽略了上下文影响因素,使得在语法错误诊断过程中遗漏了很多语法错误,因此需要找出其相关的所有信息。词嵌入模型可以使用词向量的方式来描述词语的相关信息,2013 年 Mikolov 等学者提出了 Word2vec 方法来解决这个问题[10]。Word2vec 是从大量文本语料中以无监督方式学习语义知识,一种用于训练词向量的模型工具,作用是将所有词语投影到 K 维的向量空间,每个词语都可以用一个 K 维向量表示。Word2vec 可使用连续词袋 CBOW(continuous bag-of-words)和 Skip-gram 模型来学习词向量表达,CBOW 主要通过上下文预测词的方式学习,Skip-gram 主要以词来预测周围上下文方式学习。对于没有标注的训练数据集,Skip-gram 模型作为一种无监督学习技术,可根据样本间的规律统计对样本进行分析,查找给定词的最

相关词语,更适合完成中文语法检错的词嵌入层向量的生成。

1.3　深度学习模型

深度学习是机器学习研究中一个崭新的领域,其模型属于一种多隐藏层、多感知层的神经网络结构,并具备优秀的数据表示。近年来,深度学习模型及其各种改进形式的模型层出不穷,纷纷被应用在自然语言处理研究中,得到了不错的效果。为了更好地捕捉词语上下文相关信息,保留语序特征信息,该文采用了基于注意机制的双向长短期记忆网络模型。

1.3.1　双向长短期记忆网络模型

长短期记忆网络(long short - term memory networks,LSTM)模型是递归神经网络(recurrent neural network,RNN)模型的一种,可以更好地解决中文语法检错中的长距离依赖问题。近年来,LSTM 作为 RNN 的一种优化,利用其所具备的学习长距离文本依赖的特点,被广泛应用在情感分类、机器阅读理解等研究中,已然成为了深度学习的一个重要研究热点[11]。双向长短期记忆网络(Bi-LSTM)由前向和后向 LSTM 组合而成,由于其具备神经网络拟合非线性的能力,可以更好地捕捉上下文的双向语义信息,为语法检错提供上下文依赖性更强的文本信息。

1.3.2　注意机制

注意(attention)机制强调把注意力集中放在重要的点上,忽略其他不重要的因素。神经网络注意机制是具备专注于其输入(或特征)的神经网络,它可选择特定的输入。目前,基于注意机制深度学习网络模型的改进研究,已在机器翻译等领域取得了很好的应用效果。

1.4　序列标注技术

中文语法检错任务可以视为序列化标签标注任务,在做标注时给定特定的标签集合,即可完成序列标注。常见的解决方案往往借助于隐马尔可夫(HMM)[12]、最大熵马尔可夫(MEMM)[13]和条件随机场(conditional random field,CRF)模型。其中,2001 年由 Lafferty 等提出的 CRF,是一种无向图判别式概率模型,作为解决序列标注问题的主流方法,很好地解决了 HMM 和 MEMM 的标注偏差以及标签之间的依赖关系信息问题,在分词、词性标注和命名实体识别等序列标注任务中取得了很好的应用效果。

2　基于 BI-LSTM-ATT 与 CRF 的中文语法错误诊断模型

该文构建了基于 BI-LSTM-ATT 与 CRF 的中文语法错误诊断模型,由词嵌入层、BI-LSTM-ATT 层、CRF 层构成,结构如图 1 所示。

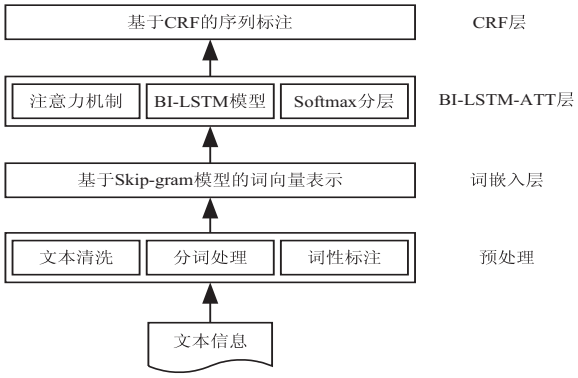

图 1　基于 BI-LSTM-ATT 与 CRF 的中文语法错误诊断研究框架

2.1　词嵌入层

词嵌入层的作用在于通过大量样本训练 Word2vec 输入词向量,提供给下一层使用。应用 Skip-gram 模型分别将输入词和 POS 标签作为中心词,预测出其上下文词和上下文 POS 标签为输出词。训练前,将语料库中的所有 n 个词语进行独热编码为 $w_{(i)} \in R^n$,同时对输出也进行独热编码。Skip-gram 模型结构如图 2 所示。

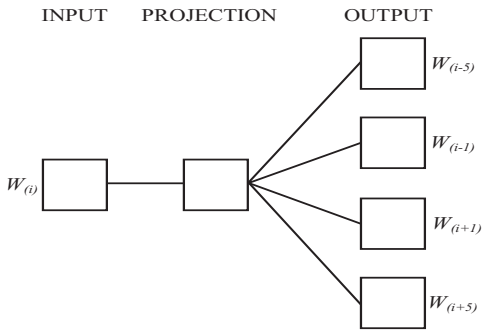

图 2　Skip-gram 模型

Skip-gram 模型分为三层:第一层为输入层(INPUT),$w_{(i)}$ 为输入语句的每个词语,以此作为输入词;第二层为投影层(PROJECTION);第三层为输出层(OUTPUT),窗口为 c,假设 c 取 5,得到目标单词 $w_{(i)}$ 上下文中的 $5c$ 个词向量 $(w_{(i-5)},\cdots,w_{(i-1)},w_{(i+1)},\cdots,w_{(i+5)})$,即输入词的邻近词的概率分布。同理,再将每个词 POS 标签 $p_{(i)}$ 作为输入,得到其上下文 POS 标签向量 $(p_{(i-5)},\cdots,p_{(i-1)},p_{(i+1)},\cdots,p_{(i+5)})$,即输入词 POS 标签的邻近 POS 标签的概率分布。

给定一个大小为 $|W|$ 的词集,将每个词 $w \in W$ 映射到 d_w 维嵌入空间。同理,给定大小为 $|P|$ 的 POS 标签集,将每个 POS 标签 $p \in P$ 映射到 d_p 维嵌入空间。最后,将得到的词向量和 POS 标签向量嵌入连接到单个向量 $x_i \in R^{H_c}$ 中,其中 $H_c = c \times (d_w + d_p)$,作为 BI-LSTM-ATT 层的输入。

2.2　基于注意机制的 BI-LSTM 模型

1997 年,Hochreiter 与 Schmidhuber 提出对递归神经网络(recurrent neural network,RNN)进行优化,得到

了长短期记忆网络(long short term memory networks, LSTM)[14],解决了长序列学习的梯度消失问题[15]。近年来,很多学者在自然语言处理的研究应用中,针对具体问题,对 LSTM 模型进行各种形式的改进都取得了不错的应用效果[16]。该文采用 BI-LSTM-ATT 模型,通过充分利用序列上下文中所有可能对标记有用的信息,即提取词过去和未来的特征来提升标签的准确度,如图 3 所示。

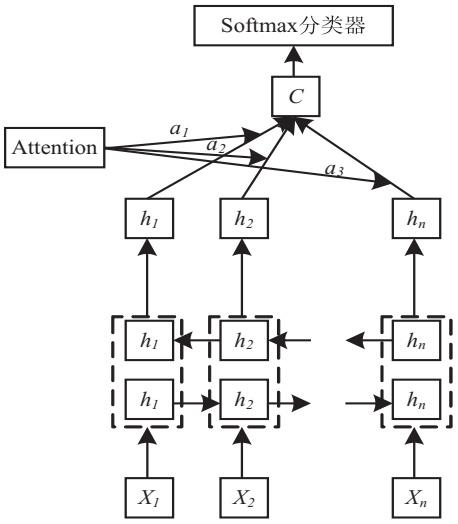

图 3 基于注意机制的 BI-LSTM 模型

图 3 所用的模型在传统 LSTM 模型的基础上增加了注意机制。向量 x_i 表示一个句子中的每个词语,使用双向 LSTM 模型得到 $h_i = [\overrightarrow{h_i}, \overleftarrow{h_i}]$ 的集合表示整个句子样本的句子向量。模型中的注意机制计算出每个元素的概率权重 α_{it},抽取对整句重要的词向量构成最终特征向量,相关计算公式如式(1)所示。

$$u_i = \tanh(W_w h_i + b_w)$$
$$\alpha_i = \frac{\exp(u_i^{\mathrm{T}} u_w)}{\sum \exp(u_i^{\mathrm{T}} u_w)} \tag{1}$$
$$C = \sum \alpha_i h_i$$

其中,$\boldsymbol{W}$ 是权重矩阵,将输入的 h_i 进行线性转换,u_w 是词水平的上下文矢量,C 是第 i 个词语的向量。

最终得到每个词的所有标签的各自得分,即每个词映射到标签的概率值。

2.3 CRF 层

BI-LSTM-ATT 模型充分考虑了输入序列的上下文信息,但忽略了标签之间存在的依赖关系,而相邻文字之间的信息对于语法识别很重要。在 BI-LSTM-ATT 模型之后再加入一个 CRF 层,可针对不同任务设计特征,所有特征可进行全局归一化,求得全局最优解,通过选取有效特征,生成相应的标签序列。CRF 模型可以把前后标记依赖约束考虑进去,使用标记状态转换概率作为评分。

在自然语言处理问题中,普遍采用线性链条件随机场解决序列标注问题。设 X 为线性链表示的输入观测序列,Y 为对应的状态序列,$X = (X_1, X_2, \cdots, X_n)$,$Y = (Y_1, Y_2, \cdots, Y_n)$,则 Y 的条件概率分布 $P(Y \mid X)$ 构成条件随机场,模型定义为:

$$P(Y \mid X) = \frac{1}{Z(X)} \exp\left(\sum\nolimits_k \lambda_k f_k\right) \tag{2}$$

其中,f_k 和 λ_k 分别表示特征集合和对应权重,$Z(X)$ 表示归一化因子,表示所有可能状态的条件概率之和,公式如下:

$$Z(X) = \sum_Y \exp\left(\sum\nolimits_k \lambda_k f_k\right) \tag{3}$$

该文按中文语法错误的四种类型加无错误类型,将错误类型标签 Y 定义为{R,M,S,W,N},分别对应词语冗余、词语缺失、词语误用、词语乱序和无错误。

3 实 验

3.1 数据来源

实验选取 2017 年“汉语水平考试(HSK)”写作部分数据的 10 000 句作为训练数据,其中正确句子总数为 3 658 句;从 NLPCC2018(CCF 国际自然语言处理与中文计算会议,Natural Language Processing and Chinese Computing)新增的 TASK2 中文语法错误修正任务提供的数据集中,随机选取 3 000 句作为测试数据。

3.2 数据预处理

(1)文本清洗。

实验过程中,为了保证不会因为句子分割或字向量过于稀疏等因素影响检测结果,定义句子长度为 100。对长于 100 的句子,采用人工分割的方式,尽量保证文本特征集中;对短于 100 的句子,采用在句后以 0 字补全的方式。

(2)分词和词性标注。

利用 python 自带的 jieba 分词器对训练集所有语句进行分词处理和词性标注。使用 Skip-gram 模型训练得到 300 维的字向量。

3.3 评价指标

根据常用的几个性能评价指标对模型进行评价,主要有精确率 Precision(P)、召回率 Recall(R)和 F_meature(F),公式如下:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{F\_meature} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

其中,TP 表示正确的句数,FP 表示错误的句数,FN 表示没有被检测出的句数。F_meature 作为标准测度,考

虑了 Precision 和 Recall 的综合影响。

3.4　实验结果及分析

在数据进行了预处理的前提下,将提出的方法与常用的人工智能方法进行实验对比,包括 LSTM、Bi-LSTM、CRF、Bi-LSTM-CRF。实验结果如表 1 所示,包括不同算法模型对应的 Accuracy、Precision、Recall 和 F_meature 值。从表 1 可以看到,LSTM 与 Bi-LSTM 虽然能够解决中文语法检错中的长距离依赖问题,甚至 Bi-LSTM 能够更好地捕捉上下文相关信息,实验效果明显优于 LSTM,但在更加复杂的真实数据实验情况下,CRF 模型的 Accuracy、Recall 和 F_meature 指标明显高于 LSTM 与 Bi-LSTM。Bi-LSTM-CRF 模型由于结合了获取上下文信息和局部特征条件概率的能力,在位置级别上比以上模型效果都要好,但由于梯度下降等原因,仍不能完美发挥作用。因 Bi-LSTM-ATT 能够通过注意机制捕捉句中关键部分,优化了语法检错任务,再通过 CRF 层对标签序列概率分布建模,得到了更高的 Precision、Recall 和 F_meature,在位置级别上与 Bi-LSTM-CRF 实验效果相差无几。实验表明,提出的基于 Bi-LSTM-ATT 与 CRF 相结合的模型,可有效提高中文语法错误诊断效果。

表 1　实验结果对比

算法模型	Accuracy /%	Precision /%	Recall /%	F_meature /%
LSTM	8.86	8.52	5.36	5.81
Bi-LSTM	9.52	11.69	7.83	7.25
CRF	11.10	9.41	9.67	6.42
Bi-LSTM-CRF	16.88	13.96	17.52	16.75
Bi-LSTM-ATT+CRF	16.83	14.23	18.17	17.33

4　结束语

提出了一种基于 BI-LSTM-ATT 与 CRF 相结合的中文语法错误诊断模型,应用于 NLPCC2018 的 TASK2 提供的数据集。该方法将采用 jieba 分词预处理后的数据,运用 Skip-gram 模型得到词向量表示,通过 BI-LSTM-ATT 模型的 Softmax 分类器进行分类,进而采用 CRF 模型分类并定位。为验证模型的有效性,在复旦大学提供的语料集中抽样实验,结果表明,提出的模型在 Accuracy、精确率、召回率、F_meature 效果对比中,比传统深度学习模型均有提高,为中文语法错误诊断的相关研究提供了一些新思路。在未来的工作中,将获取更多的训练数据增强模型,争取更大限度地拟合出中文的固定规律,进一步优化和完善模型。

参考文献:

[1] 教育部语用司. 国家中长期语言文字事业改革和发展规划纲要(2012-2020 年)[J]. 语文建设,2013(28):163.

[2] YU L C,LEE L H,CHANG L P. Overview of grammatical error diagnosis for learning chinese as a foreign language[C]// 22nd international conference on computers in education (ICCE2014). Nara,Japan:[s. n.],2014:42-47.

[3] CHENG Shun-Man,YU Chi-Hsin,CHEN Hsin-His. Chinese word ordering errors detection and correction for non-native chinese language learners [C]//Proceedings of COLING 2014. Dublin,Iraland:[s. n.],2014:279-289.

[4] YEH J F,YEH C K,YU K H,et al. Condition random fields-based grammatical error detection for chinese as second language[C]//Proceedings of the 2nd workshop on natural language processing techniques for educational applications (ACL-IJCNLP 2015). Beijing:[s. n.],2015:105-110.

[5] 韩文颖. 面向问答的中文语法错误自动检测方法研究[D]. 哈尔滨:哈尔滨工业大学,2015.

[6] LEE L H,RAO G,YU L C,et al. Overview of the NLP-TEA 2016 shared task for chinese grammatical error diagnosis [C]//Proceedings of the 3rd workshop on natural language processing techniques for educational applications (NLPTEA2016). Osaka,Japan:[s. n.],2016:40-48.

[7] ZHENG B,CHE W,GUO J,et al. Chinese grammatical error diagnosis with long short-term memory networks[C]//Proceedings of the 3rd workshop on natural language processing techniques for educational applications (NLPTEA2016). Osaka,Japan:[s. n.],2016:49-56.

[8] YANG Y,XIE P. Alibaba at IJCNLP-2017 tack 1:embedding grammatical features into LSTMs for chinese grammatical error diagnosis task[C]//Proceeding of the IJCNLP 2017. Taipei:[s. n.],2017:41-46.

[9] 杨劲男. 基于神经网络的中文语法纠错关键技术研究[D]. 昆明:云南大学,2018.

[10] DEORAS A,MIKOLOV T,KOMBRINK S,et al. Approximate inference:a sampling based modeling technique to capture complex dependencies in a language model[J]. Speech Communication,2013,55(1):162-177.

[11] 张俊飞,毕志升,吴小玲. 基于词向量 Doc2vec 的双向 LSTM 情感分析[J]. 计算机与数字工程,2018,46(12):2385-2389.

[12] 林巧民,齐柱柱. 基于 HMM 和 ANN 混合模型的语音情感识别研究[J]. 计算机技术与发展,2018,28(10):74-78.

[13] 郎　波,樊一娜. 基于深度神经网络的个性化学习行为评价方法[J]. 计算机技术与发展,2019,29(7):6-10.

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

[15] ASHIQUR R S,DONALD A A. Deep learning using convolutional LSTM estimates biological age from physical activity [EB/OL]. 2019-08-06. https://www. nature. com/articles/s41598-019-46850-0.

[16] 刘　升. Bi-LSTM-CRF 模型在中文语法错误诊断中的应用研究[D]. 武汉:华中师范大学,2019.