文章编号: 1009 - 2552(2020)09 - 0143 - 05 DOI: 10.13274/j.cnki.hdzj.2020.09.029

基于深度学习技术的语法纠错算法模型构建分析

景艳娥

(渭南师范学院,陕西 渭南 714000)

摘 要: 为了探究基于深度学习技术的语法纠错算法模型,文中从系统需求分析入手,首先介绍了模型构建基于 seq2seq 的深度学习技术模型和语料库的相关理论基础,然后对基于 seq2seq 的语法纠错模型进行了分析,最后对语法纠错算法模型的架构设计和核心模块的运行框架和主要原理进行了介绍。研究成果表明: 人工智能在语法纠错中的应用也逐渐受到相关研究者的关注,该技术的成型不仅能有效减少教师批卷工作量,更有助于学生的自主学习; 在基于 seq2seq 的深度学习技术模型的引入 Attention 机制,技能保证语法纠错的准确性,又能提高语法纠错模型的运算效率; 在模型中引入反馈建议模块,有助于及时发现并优化系统的不足之处。

关键词: 深度学习; 语法纠错; 语料库; seq2seq

中图分类号: TP311 **文献标识码:** A

Analysis of grammar error correction algorithm based on deep learning technology

JING Yan-e

(Weinan Normal University, Weinan 714000, Shaanxi Province, China)

Abstract: In order to explore the model of grammar error correction algorithm based on deep learning technology, starting with the analysis of system requirements, firstly, the model of deep learning technology based on seq2seq and the theoretical basis of corpus are introduced. Then, the model of grammatical error correction based on seq2seq is analyzed. Finally, the architecture design of the model of grammatical error correction algorithm and the operation framework and main principles of the core modules are introduced. The research results show that: The application of artificial intelligence in grammar error correction has gradually attracted the attention of relevant researchers. The formation of this technology can not only effectively reduce the workload of teachers' marking papers, but also help students' autonomous learning. The introduction of attention mechanism into the deep learning technology model based on seq2seq ensures the accuracy of grammatical error correction and improves the efficiency of the operation of the model. The introduction of feedback and suggestion module in the model is helpful to discover and optimize the shortcomings of the system in time.

Key words: deep learning; grammatical error correction; corpus; seq2seq

0 引言

信息化时代促进了全球化的发展速度,国家甚至个人的发展也越来越离不开其它国家的相关产品及知识体系,英语作为目前世界通用语言,在全球化发展中起到至关重要的作用[1-2]。我国人口基数大,英语学习者众多,每年参与各类大小型英语考试的人次更是不胜枚举,参与阅卷的教师数量增加,这让本来就人数不算多的教师,工作量更加繁重了[3-4]。计算机技术迅猛发展的今天,人工智能不

收稿日期: 2020 - 03 - 04

作者简介: 景艳娥(1982 -),女,硕士研究生,讲师,研究方向为英语教育,跨文化交际,翻译笔译。

仅取代了原有机械重复的工作，在其它工作中也不断取得了新的突破和进展[5]。鉴于语言的灵活性和不确定性，人工智能在语言教学方面的研究一直被关注，但进展却未能达到人们预期的效果[6]。随着近年来计算机软硬件技术的不断进步，各类本来应用于其他领域的算法也逐渐被引入并成功应用到语法纠错中[7-9]。本文在提出带反馈机制的英语语法纠错算法的基础上，结合序列到序列模型，对基于深度学习技术的语法纠错算法模型构建进行了分析和研究，旨于为人工智能在英语教学和学习中的推广应用提供理论及技术支撑。

1 需求分析

语法纠错算法模型主要为教师和学生提供语法纠错系统，并以网站的形式呈现给用户。模型主要模块及功能如下：

①用户管理：用户分为普通用户和高级用户。其中普通用户在登录后完成语法纠错；高级用户可以在登录后完成语法纠错、反馈意见修改和查看原始纠错结果等。

②语法纠错：用户输入待修改语句后，经过一系列计算，将纠错完成后的多条单独语句组合成与完整句子并返回至用户。

③反馈机制：用户拿到系统模型修改的语句后，对语句进行检查，如果用户对纠错后的结果不满意或者认为有改进的地方，可以通过该模块向系统管理者提出反馈。

④文本管理：主要提供文本储存、状态修改和导出功能，管理员及时查看用户反馈的修改建议并对修改建议进行过滤，并根据建议对系统的模型进行升级，并将用户建议纳入系统中，并提供状态查询功能，方便检查是否已经根据该条用户反馈文本对模型进行了升级。

⑤升级触发：借助于阈值判定规则，当储存的文本信息达到某一数量时，便触发系统升级。

2 相关理论基础

2.1 基于 seq2seq 的深度学习技术模型

sequence to sequence 模型由编码器和解码器组成，而且编码器和解码器一般均为死循环，其中编码器主要用于对输入的序列进行特征提取并执行向量化，在序列适当位置输入 < EOS > 作为终止解码过程。Google 于 2014 年提出了该模型的框架图，后经大量学者不断完善和改进，目前最常使用（本文也采用）的框架为相比于最初框架具有解码效率更高、效果更好的基于内容的 seq2seq 框架，如图 1 所示。

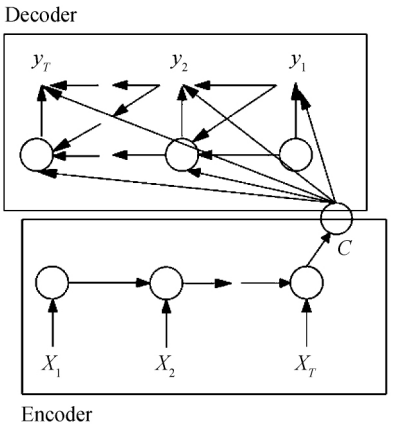

图 1 基于内容的 seq2seq 框架结构示意图

假设 $X = (x_1, x_2, \cdots, x_t, \cdots, x_n)$ 为输入序列；f 为转换规则，根据输入序列和转换规则便可以生成语义状态向量 C；假设 $h_t, t \in (1, 2, \cdots, n)$ 为隐藏状态；则 t 时刻的隐藏状态由 $t-1$ 时刻的隐藏状态和 t 时刻输入的序列共同决定：

$$h_t = f(h_{t-1}, x_t) \tag{1}$$

按照式（1）便可计算并获取任一时刻的隐藏状态，然后便可以计算得到包含输入 X 所有基本信息和并能够获取其重要特征：

$$C = q(h_1, h_2, \cdots, h_t, \cdots, h_n) \tag{2}$$

至此编码阶段结束，解码阶段开始。解码阶段在形式上与编码阶段正好相反，它是根据生成的语义向量 C 和经编码阶段生成的输出序列 $Y = (y_1, y_2, \cdots, y_{t-1})$ 去预测下一个输出的单词 y_t，计算过程如下：

$$\begin{aligned} y_t &= \operatorname{argmax} P(y_t) \\ &= \Pi_{t-1}^{T} p(y_t \mid \{y_1, y_2, \cdots, y_{t-1}\}, C) \end{aligned} \tag{3}$$

用 h_t' 表示解码阶段的隐藏状态，其计算如下式所示：

$$h_t' = f(h_{t-1}', y_{t-1}, C) \tag{4}$$

2.2 语料相关理论基础

许多基于人工智能的语言学习和教学系统都将语料库作为其基础资源之一，语料库通过对大量目标语言文本的筛选和标记等处理后得到的基础性数据库。根据不同阶段的学习，语料库中所含词汇数量、每一个词汇的注释数量均有所不同，目前常用的语料库主要有包含近 1 亿词汇量的，历经 17 年（1991 年～2007 年）建成的英国国家语料库（British National Corpus）和包含 100 余万词汇量且广泛包含了目前我国大中小学及各类英语考试所用词汇的中国学习者语料库（Chinese Learner English Corpus）。此外还有

新加坡国立大学英语学习者语料库 – release2.2(The National University of Singapore Corpus – release2.2) 和英语维基语料库也被广泛使用[10]。

在语料库支持下,面对一段完整的英语文本,在进行语法纠错时,一般需要经历断句、单词拆分和词性识别及句子分析几个步骤。

2.2.1 断句

断句就是先将一大段英语文本切割成一个一个的单独的句子,比如借助标点符号就可以完成断句,但这样的断句极容易出错,比如借助于英语中句号“.”断句时,会因为计算机无法区别句号是英文缩写的“.”还是实际用来断句的句号而出现错误。为此,宾夕法尼亚大学开发了自然语言工具包(Natural Language Toolkit,NLTK) 协助完成断句步骤[11]。该工具被广泛应于人工智能、信息检索和机器学习等领域。

2.2.2 拆分单词

单词是英语的基本单元,在完成分句后,在每句话中分割出词汇是借助于机器完成英语学习至关重要的一步。虽然组成英语句子的词汇间一般由空格隔开,方便断句; 但诸如“New York”等由两个词组成的单词和“it's”等两个单词的缩写为以空格媒介拆分句子中单词的措施很难完美地完成分词的任务。为此,研究者们也借助于分词规则和正则表达式提出了一系列合理可用的分词模型。如: n – gram 语言模型等。同时,NLTK 支持使用者对数据集进行自定义,通过对分词场景的定制化,也具备良好的分词功能,本文研究中即采用 NLTK 进行分词。

2.2.3 词性识别及句子分析

因为语言环境的差异性,绝大多数英语单词翻译为中文后都具备多个词性,每隔词性下边也可能对应着多个词义。将句子中单词逐个分离出来后,为了对语句进行分析从而实现语法纠错目标,首先应该识别每一个单词在该句子存在时的词性为动词、名词还是其他词性。基于词性识别结果,结合单词在句子中所处位置及英语语法规则,便可以实现对句子的分析,其中词性分析的正误对语法纠错能否实现起着决定性作用。当前较为常用的词性分析工具有基于统计概率方法的词性标注系统(Stanford POS Tagger) ,Apache 开发的语言处理工具包(OpenNLP) ,基于 HMM 和 Viterbi 算法的 Trigram'n'Tags 及基于 Python 和 Cython 的 spaCy 等等[12-15]。

3 基于 seq2seq 的语法纠错模型

3.1 纠错评价标准

MaXMatch 算法是目前应用最为广泛的语法纠错评估算法,它可以借助于 MaXMatch scorer 得以实现[16]。该模型对语法纠错的评价主要从“纠错率 P”和“纠全率 R”两个方面,其中纠错率是指模型对一个错误句子修改后的正确程度; 纠全率是模型对句子中所有错误修正的比例。两个评价指标分别可以用式(5) 和式(6) 进行计算。

$$P = \frac{\sum_{i=1}^{n} | g_i \cap e_i |}{\sum_{i=1}^{n} | e_i |} \tag{5}$$

$$R = \frac{\sum_{i=1}^{n} | g_i \cap e_i |}{\sum_{i=1}^{n} | g_i |} \tag{6}$$

式中, $\sum_{i=1}^{n} | g_i |$ 和 $\sum_{i=1}^{n} | e_i |$ 分别为参考答案中需修正的句子总数和考生答案中修改了的句子总数。

定义 $F_{0.5}$ 为纠错评价的中央指标,其计算如下式所示。

$$F_{0.5} = \frac{(1 + 0.5^2) * R * P}{R + 0.5^2 * P} \tag{7}$$

3.2 Attention 机制

Attention 机制于 2014 年 Bahdanau 引入到基于 seq2seq 的英语语法纠错模型中,主要用于解决模型中当编码器输入的序列长度达到一定长度时,细节信息就会出现丢失[17]。而 Attention 机制受到人对物体的关注一般只需要关注主要部分,便可以对事物有相对完整认识的启发。在计算机计算中,引入权值因子,对一个英语句子不同部分进行不同的权重赋值,例如对于句子“This is a book.”主要关注的是“Book”一词,其权重便相对大一些,而对其它三个词的关注程度不高,其权重也可以相应小一些。根据系统对序列的计算范围,Attention 机制可以分为局部计算和全局计算,本文主要应用全局计算中的 Soft Attention 部分对语言错误进行识别和处理,以下对其进行详细介绍,如图 2 所示。

定义权重因子 α_{ij},计算公式如下:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} \tag{8}$$

$$e_{ij} = \alpha(s_{i-1}, h_j) \tag{9}$$

3.3 添加层规范化

基于深度神经网络的学习技术虽然计算功能比较强大,但其对计算机的要求较高,为此引入 Batch Normalization(简称 BN) 批规范化处理措施。它能够通过每层网络任意神经元输入激活参数并进行规范化操作,经过转化后的参数符合标准的正态分布,从而使得模型收敛速度大大提高,计算效率也得以

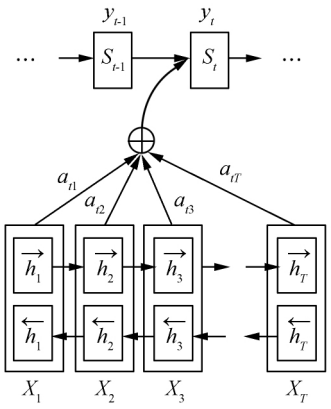

图2　Soft Attention 计算框架图

提升。但是 RNN 的输入长度是可变的，而 BN 运行时依赖于 mini-batch 一阶和二阶统计量，为了提升 BN 的适应性，提出层规范化（Layer Normalization，简称层规范化）[17-18]。

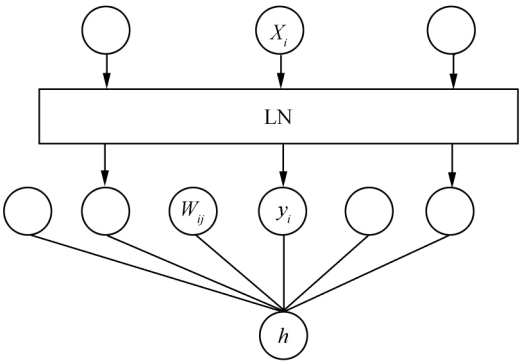

图3　Layer Normalization 工作原理图

LN 规范化的工作原理如图 3 所示，它运行时和其它样本数据并行运算，相比于 BN，对数据分布的要求明显降低，样本训练难度也较小。LN 先对 RNN 中整层神经元输入的信号求和并计算方差，然后将该层神经元输入的信号映射到同一分布中，所有隐藏元共享规范化 μ 和 σ [18]，计算公式如下：

$$\mu = \frac{1}{H}\sum_{i=1}^{H} x_i$$
$$\sigma = \sqrt{\frac{1}{H}\sum_{i=1}^{H}(x_i - \mu)^2} \tag{10}$$

3.4　词潜入

计算机无法直接识别和理解自然语言，需要对输入文本进行向量化操作，对文本进行向量化处理转为机器语言，才能让计算机去识别。以单词短句“This is a book”中“book”一词为例，常规做法如下：首先在句子中识别“book”出现的位置，然后将改位置标记为 1，其余位置则标记为 0，上述例句则可以转换为 [0,0,0,1]。常用的方法有 CBOW 和 Skip-gram 模型。前者是根据某个词前后 C 个连续的词查找该词出现的位置和呈现频率，后者则是根据某个词分别计算它前后出现另外几个词各自出现的概率。

4　基于深度学习技术的语法纠错算法模型架构设计

4.1　计算机软件技术概况

本模型构建过程中主要用到 Thrift RPC 框架、Nginx 反向代理与负载均衡、Python 与 Django 和 React 等相关技术[16-18]，该部分内容相关文献中均已提及，兹不赘述。

4.2　整体架构

英语语法纠错模型整体架构主要包含三个模块，“语法纠错”、“服务器接入”和“反馈过滤”，架构图如图 4 所示。

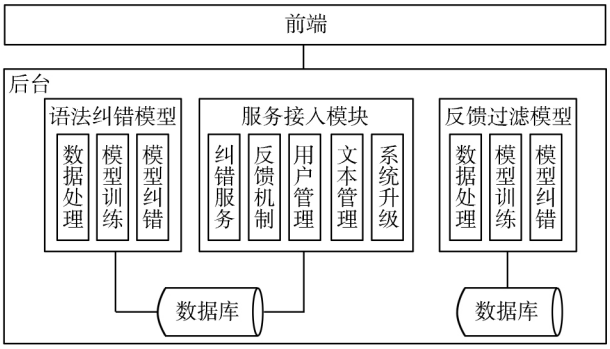

图4　英语语法纠错模型整体架构图

其中语法纠错主要包含数据处理、模型训练和模型纠错三个子模块，服务器接入则主要完成为用户的纠错服务、反馈机制、用户和文本管理及系统升级等模块；反馈建议模块中主要提供反馈过滤和语料处理等功能。以下主要对系统的语法纠错和反馈建议两个核心模块进行介绍。

4.3　模型训练

模型训练模块为自动运行模块，接到外部请求后，先对来源于语料库和用户建议文本里的语料句子进行统计，达到设定阈值后开始训练数据，过程中如果遇到异常则向管理员发送异常通知。模型训练结束后，借助于语法纠错模型评估本次纠错效果进行评估，如果对结果满意则结束运行，如果不满意则更新语法纠错模型，如图 5 所示。

4.4　语法纠错

语法纠错模块是本系统最为核心的模块，系统接收到输入的信号后，首先判断其参数是否有误，有误则直接结束运算，参数合法则进行下一步运算。首先判断是否断句，断句后采用前述训练完成的语

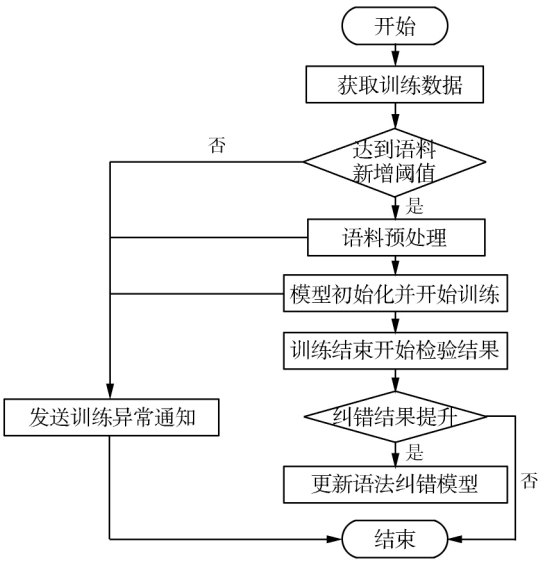

图5　模型训练模块作业流程图

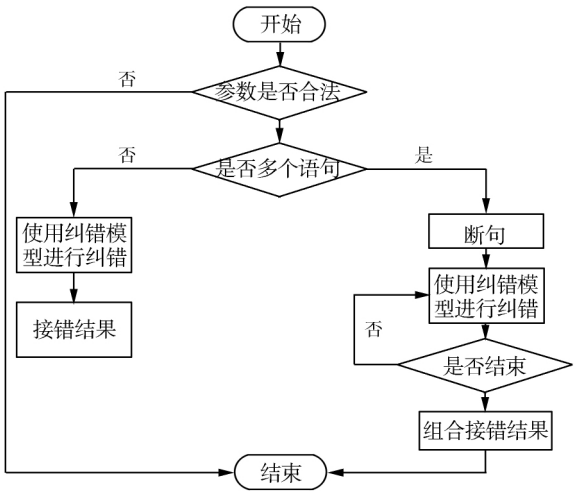

图6　语法纠错模块作业流程

法纠错模型进行语法纠错，完成全部句子的语法纠错后整合结果并返回，结束本次运算，如图6所示。

4.5　反馈建议

反馈建议模块也是系统的重要组成部分之一，用户完成语法纠错后，如果对纠错结果不满意则向系统反馈建议，该建议会集中在文本信息中，管理员按照阈值设定，定期对系统纠错模型进行更新和修改。与此同时，对反馈得到的文本信息的修改库也将作为语料库的资源之一，参与语法纠错运算，如图7所示。

5　结束语

本文从基于深度学习技术的语法纠错算法模型的系统需求分析入手，首先介绍了模型构建基于seq2seq的深度学习技术模型和语料库的相关理论基础，然后对基于seq2seq的语法纠错模型进行了

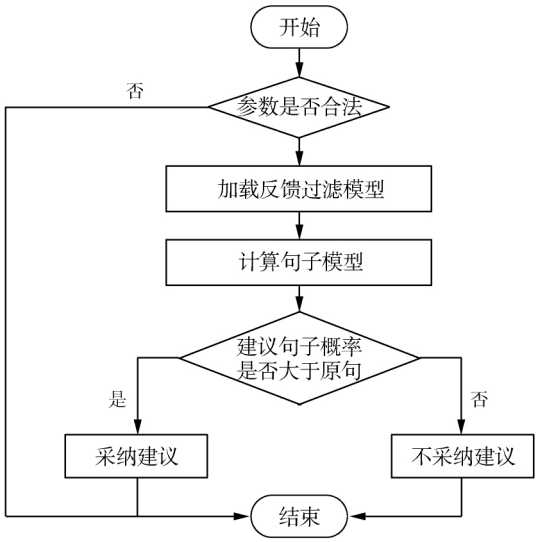

图7　反馈过滤模块作业流程图

分析，最后对语法纠错算法模型的架构设计和核心模块的运行框架和主要原理进行了介绍。主要研究成果如下：

①针对不同英语学习层析学习者和用户的语料库发展日趋完善，人工智能在语法纠错中的应用也逐渐受到相关研究者的关注，该技术的成型，不仅能有效减少教师批卷工作量，还能使得学习者随时随地对自己作文的语法正误得到指导和纠正。

②在基于seq2seq的深度学习技术模型的引入Attention机制，在不对原文有曲解的前提下实现了先抓主要因素对语句进行分析，从而实现提高语法纠错模型的运算效率。

③在模型中引入反馈建议模块，用于收集语法纠错过程中用户不满意的地方，并按照设定阈值及时对反馈文本信息进行筛选、整理和汇总，并依据反馈信息及时对系统模型进行优化。

参 考 文 献：

[1] 张玉清，董颖，柳彩云，等. 深度学习应用于网络空间安全的现状、趋势与展望 [J]. 计算机研究与发展，2018，55(6)：1117－1142.

[2] 金永红. 指向深度学习的小学生英语语篇阅读教学设计——以译林版小学英语6B Unit2 Cartoon time为例 [J]. 华夏教师，2018，102(18)：49－50.

[3] 李海龙，张维明，肖卫东，等. 通用标准SQL语法分析模型 [J]. 小型微型计算机系统，2003(11)：1969－1972.

[4] 周东祥，李群，王维平. 仿真模型的语法可组合问题及组合判定 [J]. 火力与指挥控制，2009，34(8)：4－9.

[5] 李波，高文君，邱锡鹏. 基于语法分析和统计方法的答案排序模型 [J]. 中文信息学报，2009，23(2)：23－27，41.

[6] 崔舒宁，吴宁，叶丹. 建立抽象语法树模型评测C++代码 [J]. 计算机应用，2015(S1)：183－185，191.

(下转第152页)

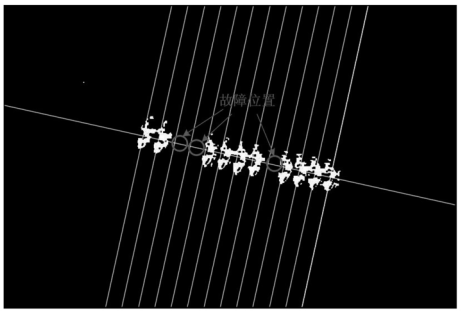

图7　输电杆塔实际故障位置

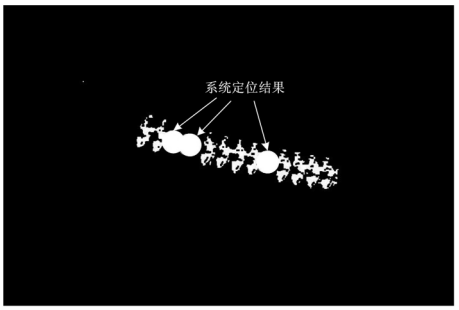

图8　基于视觉识别系统的定位结果

统具有较高的输电杆塔巡检准确性。

4　结束语

为了提高输电杆塔无人机巡检的准确性，提出基于视觉识别的输电杆塔无人机自动巡检系统。通过设计输电杆塔无人机自动巡检系统的硬件与软件，完成巡检系统的设计，实现输电杆塔无人机的自动巡检。仿真结果表明，基于视觉识别的巡检系统相比于传统巡检系统，输电杆塔无人机的巡检偏差小。

参考文献：

[1] 汪海，孙锋. 基于 RFID 的消防器材巡检系统设计 [J]. 电子设计工程，2017，25(23)：59－62.

[2] 陈益锋. 基于移动 GIS 技术的管道智能巡检系统开发与研究 [J]. 智能建筑与智慧城市，2017(4)：52－53.

[3] 李晓琳，史国振，杨孟，等. 基于 Android 的地下管线巡检系统的设计与实现 [J]. 测绘与空间地理信息，2017，40(12)：39－41.

[4] 林俊国，丛强，许晨. MEMS 激光扫描视网膜投影显示系统设计 [J]. 光学学报，2017，37(12)：314－320.

[5] 黄耀林，王敏，林正. 大孔径大视场变焦投影镜头设计 [J]. 应用光学，2018，39(3)：412－417.

[6] 侯国柱，吕丽军. 大孔径变焦投影镜头设计 [J]. 应用光学，2018，39(3)：405－411.

[7] 伍坪，秦彩杰. 基于高速 DSP 深度视频帧内编码的农机导航系统设计 [J]. 农机化研究，2018，40(3)：84－87.

[8] 韩昊旻，张崇明，陈志红. 基于 ROS 和激光雷达的 AGV 导航系统设计与实现 [J]. 电子测量技术，2018，41(8)：112－117.

[9] 李海平，张二剑. 卫星导航信号转发器的设计 [J]. 电子世界，2017(9)：184－185.

[10] 郭凡，李东，许犇. 易燃品仓库群三维移动智慧巡检路径优化 [J]. 西安科技大学学报，2019(1)：160－167.

[11] 赵燕东，涂佳炎，刘圣波. 基于北斗卫星导航系统的林区智能巡检测绘系统研究 [J]. 农业机械学报，2018，49(7)：177－185.

[12] 薛阳，张晓宇，江天博. 基于视觉导航的巡检机器人双模控制研究 [J]. 控制工程，2018，25(11)：1982－1987.

[13] 王鹏，李军锋，熊山. 基于无人机的大范围内长距离配网线路智能巡检研究 [J]. 科技通报，2018，34(10)：149－153.

[14] 姜涛，葛少伟，李德泉. 济南市隧道巡检机器人电缆故障监测方法分析 [J]. 科技通报，2019，35(4)：69－73.

[15] 邓林峰，邓剑，朱磊森. 聂耳公园海绵城市建设中质量安全巡检系统的应用 [J]. 中国给水排水，2019(12)：96－99.

责任编辑：丁玥

(上接第 147 页)

[7] 杨永林. 大学英语在线课程及其建设 [J]. 外语教学，2019，40(1)：53－58.

[8] 胡佩，李小青. “人工智能＋校对”的应用前景分析 [J]. 现代出版，2019(2)：59－61.

[9] 刘芳，李戈，胡星，等. 基于深度学习的程序理解研究进展 [J]. 计算机研究与发展，2019，56(8)：1605－1620.

[10] 李志凌. 基于反馈信息论的作文评改纯理依据及原则 [J]. 海外英语，2018(5)：6－10.

[11] 程广兵. 运营商问答系统融入 AI 打造智能客服 [J]. 通信世界，2018(14)：45－47.

[12] 尚轩轩，陈体忠，杨柳青，等. 运用 ICT 辅助提升中学生英语单词记忆效率的应用研究 [J]. 中国教育信息化，2017(3)：66－70.

[13] 雷晓东. 英语作文自动评价系统技术的国内研究与应用 [J]. 科技视界，2015(35)：43－45.

[14] Sun Zhi-jun，Xue Lei，Xu Yang-ming. Marginal fisher feature extraction algorithm based on deep learning [J]. Dianzi Yu Xinxi Xuebao/journal of Electronics & Information Technology，2014，35(4)：805－811.

[15] 朱彦. 透过“反馈”之镜，倾听课堂之音——大学英语学习者对口头纠错反馈的信念探究 [J]. 外语与外语教学，2016(1)：33－40，147.

[16] 桂花，杨征权. 微课程教学法在高职英语语法教学中的运用 [J]. 高教学刊，2016(7)：114－116.

[17] Wang Long-fei，Li Xu，Zhang Li-yan，et al. Analysis of the positioning error of industrial robots and accuracy compensation based on ELM algorithm [J]. Jiqiren/Robot，2018，40(6)：843－859.

[18] 靖纯. 基于新闻语料库的中文自动校对改进方案探讨 [J]. 中国传媒科技，2016(6)：15－17.

责任编辑：丁玥