

**本科毕业设计（论文）开题报告**

|  |  |  |  |
| --- | --- | --- | --- |
| 课题名称 | “剧本杀游戏”智能审稿系统的设计与实现 | | |
|
| 课题类型 | 系统设计 | 学院 | 软件与人工智能学院 |
| 班级 | 19大数据技1班 | 专业 | 数据科学与大数据技术 |
| 指导教师 | 林嘉诚 | 职称 |  |
| 学生姓名 | 黄鑫 | 学号 | 199030130 |

重庆工程学院教务处制

|  |  |
| --- | --- |
| 1.课题名称、来源、选题依据 | |
| **1.1课题名称**  “剧本杀游戏”智能审稿系统的设计与实现  **1.2来源**  课题来源于社会实践。  **1.3选题依据**  由于近几年疫情的影响，人们缺少线下、近距离、高效的互动交流[1]。然而“剧本杀”的出现刚好符合当前人们的需求，正成为年轻人线下社交、娱乐的重要方式。导致剧本杀行业高速发展。剧本杀作者可以将自己的剧本投递给平台、发行方赚取稿费，而收稿方则需要对剧本进行审稿从而保证剧本的质量。如今很多平台或发行方通过人工的方式进行审稿，往往耗费人力，因此开发出智能审稿系统有利于提升审稿的效率，从而节约成本。  本文主要是针对用户上传的“剧本杀”稿件进行智能审核，降低人工判稿的错误率，提升稿件的审核效率与质量。  通过智能化的审稿方式可以再短时间处理大量的稿件，与此同时能快速给予上传作者稿件的审核情况，促进作者与平台简单快速交互。不仅能够方便不同的投稿群体根据需求快速进行稿件审核，还能够有效满足对海量的稿件处理的效率问题。同时，对于某些互联网公司来说，快速有效的稿件审核也为他们提供了一定的便利，他们可以将不同内别的稿件进行审核分类。这样看来，有效的智能文本纠错不仅满足了用户的需求，而且也为某些公司带来了商业价值。  那么面对如此海量的数据，如何能够快速有效的对稿件进行纠错处理，以便给予作者快速、准确的反馈，同时节省人力物力，成为了一个值得研究的课题。 | |
| 2.课题国内外研究现状和发展趋势 | |
| **2.1文本纠错的国外研究现状**  计算机领域对自然语言处理的客观需求,最早产生于语言翻译领域。在20世纪90年代，自然语言处理逐渐进入繁荣期。在日本神户召开了第四届机器翻译高层会议,标志着自然语言处理进入一个崭新的纪元[2]。  英文文本纠错起源较早，国外研究者对英文文本纠错方法已经研究了多年,从最开始的拼写检查，到后期复杂的语法错误都有非常好的进展。英语是世界上第二常用的语言，超过6亿人将英语作为第二语言（ESL)或英语作为外语(EFL)使用，在互联网上大量信息是由英语展示的，同时由于欧美国家在经济上的长期领先地位，学习英语的人数保持着较高的增长趋势，因此英语文本的拼写检查和语法错误检测与纠正的研究很早就开始了。  目前最流行的英语纠错技术大多基于深度学习，随着使用大规模无监督语料库训练得到的预训练词向量和预训练语言模型的出现，机器学习文本纠错方法的性能也越来越强。近年来神经网络机器翻译（Neural Machine Translation，NMT）模型因为性能超越统计机器翻译模型占据了优势地位[3]。大多数NMT模型采用序列到序列(sequence-to-sequence，Seq2Seq）模型，通常使用大量的“正确-错误”句子对训练,此类模型可以纠正任何错误类型，而没有如序列编辑纠正类模型的限制。Chollampatt 等人提[4]出多层卷积 Seq2Seq模型，使用卷积神经网络代替循环神经网络,在GEC任务上取得了最好的效果。  **2.2文本纠错的国内研究现状**  中文文本纠错最开始的研究集中在中文拼写检查(Chinese spell check,CSC)，而对较复杂的语法错误研究较晚。由于早期缺乏平行语料训练数据，中文的拼写检查和错误检测大多采用基于规则和统计的方法[5]，由于中文语法灵活多变，这类模型的性能都难以满足应用需求。在常见的中文文本错误纠正任务中，通常只需要完成对错别字词改正。  由于中文语法错误复杂，常见的中文文本自动校对工具通常仅能进行中文拼写检查。早期中文错别字词纠错多基于语言学知识或混淆字典规则类方法，而目前性能较好的中文拼写检查模型大多是基于BERT等掩码语言模型(Masked Language Model)[6]。2019年，Hong 等人[7]出 FASPell，一种使用BERT作为去噪自编码器、CSD(确信度-字符相似度解码器）作为解码器的无监督训练方式的模型。解码器通过量化的确信度和字符相似度在候选项矩阵中寻找最佳候选项路径作为输出。2020年，Cheng 等人[8]提出了SpellGCN方法，使用图卷积对音近形近字进行图嵌入编码，然后与 BERT 编码的字向量计算即可得到修改结果，其优点是使用图卷积神经网络约束了输出的范围，误报率较低。目前很多学者选择以Seq2Seq模型框架为基础进行中文文本纠错，主要实现方法是基于注意力机制的Encoder-Decoder模型。王辰成等人[9]提出使用动态残差结构增强Transformer机器翻译模型捕获语义信息的能力，并通过腐化语料的方法进行数据增强，在 NLPCC2018数据集上取得较好的性能。  **2.3发展趋势**  综上所述，文本纠错现在正处在一个蓬勃发展的科技时代中，它还具有很大的发展空间。其次可以将其与社会生活相结合做到1+1大于2的效果。例如将文本纠错与文本识别相结合可以实现自动化的文本信息采集，将文本纠错与语音识别想结合可以弥补识别错误的缺陷。此外，纠错任务更加依赖于预训练模型，随着预训练模型技术的发展，预训练模型的规模越来越大，参数量越来越多，不仅可以学习到文本的语义关系、衔接关系等，还可以学习其他形式的字符，例如数学公式、古文以及其他非常用的文字符号等，模型在学习过后，会对文本中可能的错误位置做出提示，并给出相应的修改建议[10]。 | |
| 3.本课题的目的及意义 | |
| 本课题是以网络“剧本杀”稿件以及其他相关文章数据为基础，根据文章内容的语句分析，对“剧本杀”稿件进行纠错。这样不仅能够用于审核“剧本杀”相关方面的文章，对于相关领域的文章也能起到纠错的作用，扩大了模型的运用面。同时，对于某些互联网公司来说，公司可能不只是拥有“剧本杀”业务板块，一个高效、易扩展的模型极大的降低公司开发新项目的成本，他们可以在当前模型上做简要的相关领域训练即可投入使用。这样看来，有效的智能文本纠错不仅满足了用户的需求，而且也为某些公司带来了商业价值。 | |
| 4. 本课题的任务、重点内容、研究方法、实现途径 | |
| **4.1本课题的任务**  以剧本杀文章数据为基础，首先对模型进行微调使得在剧本杀审稿获得更好的结果。完成智能审稿系统的搭建，用户可以上传文章到后台处理，之后再将处理结果展示在用户界面。  **4.2重点内容**  ①数据处理模块实现。首先可以利用爬虫相关技术爬取如微信等平台上的数据。然后再对爬取到的数据进行预处理。以及数据的清洗，有对分词后的处理、去重的处理等。  ②中文文本纠错模块实现。通过使用bert模型来进行数据纠错，同时将纠错结果展示在用户界面上。  ③系统搭建。使用Django开发Web系统实现，实现前后端的数据处理，用户上传数据，后端处理完成后，将处理结果展示给用户。  a.用户登录与注册模块；  b.文件上传模块；  c.文件评分与审核模块。  **4.3研究方法**  ①文献研究法。主要阅读国内外的相关文献，经过系统全面的研究，了解该课题的现状以及课题所涉及到的相关技术，为课题的研究提供理论依据。  ②数据分析法。通过爬取新浪、网易等新闻网站上的数据，清洗去重等预处理之后，分析数据的特点，根据其特点完成模型的搭建。  ③实验分析法。在完成模型搭建的基础之下，可视化展示该课题的最终需求。  **4.4实现途径**  ①本课题基本实现途径  a.“剧本杀”数据采集与处理；  b.文本纠错模型搭建；  c.审稿系统搭建；  d.系统运行测试。  ②数据采集与处理：首先从剧本杀Times、推理大师等各个网络剧本杀平台公开的剧本数据上进行数据的爬取；其次可以整合目前开源文本纠错数据集，主要选用SIGHAN 2013 Bake-off: Chinese Spelling Check Task等。根据本课题的最终需求，将采集到的文本数据进行数据的抽取、清洗后合并到训练数据集中。同时在数据处理模板需要根据特征词进行分词处理。  ③文本纠错模型搭建：该模型是基于pycorrector 搭建的中文文本纠错模型。Pycorrector主要用于音似、形似错字纠正，可用于输入法、OCR、ASR的文本错误纠正，兼容Kenlm语言模型纠错[11]，和深度模型纠错，包括：Seq2Seq，Bert，MacBert，Electra，Ernie等，该智能审稿系统采用pycorrector提供的bert模型作为主要处理模型。  ④模型评估：召回率与准确率来评估模型好坏。  ⑤智能审稿系统的搭建：在文本纠错模型搭建的成功的基础上，使用Django搭建Web系统，其中运用到的主要技术有以下几项：django model、django form表单、django 模板的继承、django view。用户上传数据、用户登录与注册、文件上传模块、文件评分与审核、结果展示。前端软件架构页面由HTML5进行页面布局，CSS3对HTML元素进行样式的制定及部分动画的制作，由JavaScript完成用户行为的交互。后端软件架构主要由Django框架。系统运维以Shell脚本和Python脚本为工具，实现日志切分、数据库备份、文件上传云端、IP检测、DDNS解析等功能。  **4.5时间安排**  ① 2022.11.08-2022.12.03 ：完成开题报告  ② 2022.11.27-2023.01.10 ：完成毕业设计项目设计与实现【文本纠错模板】  ③ 2023.01.10-2023.02.10 ：完成毕业设计项目设计与实现【审稿系统搭建】  ④ 2023.02.10-2023.03.10 ：完成毕业论文初稿  ⑤ 2023.03.10-2023.04.10 ：完成毕业论文定稿和查看  ⑥ 2023.04.10-2023.05.17 ：毕业答辩【ppt制作、答辩展示视频及毕业答辩】  ⑦ 2023.05.17-2023.05.25 ：完成毕业论文最终稿与查重 | |
| 5.完成本课题所需工作条件（如工具书、计算机、实验、调研等）及解决办法 | |
| 需要如下工作条件：  ① 电脑1台：本人已有  ② Python语言及其他软硬件环境：已从相关官方网站下载  ③ 实验数据集：通过编写爬虫代码收集以及下载开源数据集  ④ 工具书及相关资料：本人已通过中国知网等网站收集多篇论文，其他编程工具书可在图书馆借阅或自购。 | |
| 参考文献 | |
| [1] 洪艺涵,赵雨涵,王盈心,周雨凝,熊超颖.体验经济下“剧本杀”的商业模式分析[J].商展经济,2022(20):76-78.  [2] 宋一凡.自然语言处理的发展历史与现状[J].中国高新科技,2019(03):64-66.  [3] 王天极. 中文文本纠错方法研究与应用[D].贵阳:贵州大学,2022.  [4] Chollampatt S, Ng H T . A Multilayer Convolutional Encoder-Decoder Neural Network for Grammatical Error Correction[J]. 2018.  [5] 赵国红.中文语法纠错方法的研究综述[J].现代计算机,2021,27(28):65-69.  [6] 张仰森,唐安杰.面向政治新闻领域的中文文本校对方法研究[J].中文信息学报, 2014,28(6): 79-84.  [7] Hong Y , Yu X , He N , et al. FASPell: A Fast, Adaptable, Simple, Powerful Chinese Spell Checker Based On DAE-Decoder Paradigm[C]// Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2019.  [8] Cheng X, Xu W, Chen K, et al..Spellgcn:Incorporating phonological and visual similarities into language models for chinese spelling check[J]. arXivpreprint arXiv, 2020: 2004.14166.  [9] 王辰成,杨麟儿.基于Transformer增强架构的中文语法纠错方法[J].中文信息学报,2020.34(6):106-114.  [10] 袁阳. 基于深度学习的中文文本纠错方法研究[D].北京:北方工业大学,2021.  [11] 吕飞洋. 基于分布式视频流文本提取的文本纠错系统设计与实现[D].北京:北京邮电大学,2020. | |
| 指导教师意见 | 签字： 年 月 日 |
| 系（教研室）意见 | 系（教研室）主任签字： 年 月 日 |