**项目编号：**

**创新训练项目**

**立项申请书**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 项目名称 | | | 基于自然语言处理的智能编辑辅助软件设计与实现 | | | | | | | | |
| 项目执行时间 | | | 2019年 | | 项目团队人数 | | | | | 3人 | |
| 项目  负责人 | 姓名 | | 胡冕琳 | 学号 | | | | 2017308130210 | | | |
| 院系班级 | | | | 信息与电气工程学院计算172 | | | |
| 手机 | | 15811252808 | 电子邮箱 | | | | [super\_humianlin@yahoo.com](mailto:super_humianlin@yahoo.com) | | | |
| 项目  成员1 | 姓名 | | 郭际虎 | 学号 | | | | 2017304010417 | | | |
| 院系班级 | | | | 信电计算172 | | | |
| 项目  成员2 | 姓名 | | 董瀚琪 | 学号 | | | | 2017301010313 | | | |
| 院系班级 | | | | 数学171 | | | |
| 导师  情况 | 姓名 | | 陈英义 | 职务职称 | | | | 教授 | | | |
| 电话 | | 13581712265 | 电子邮箱 | | | | 08019@cau.edu.cn | | | |
| 填表日期 | | | 2019年 1月 10日 | | | | | | | | |
| 项目简介（100字以内） | | | | | | | | | | | |
| 本项目基于自然语言处理技术，针对稿件格式的欠规范化现象，对投稿的文章进行格式审查，自动检测不符合要求的格式并自动生成针对论文格式的修改意见并将其反馈给审阅人和作者，以此提高审稿效率，实现论文审阅的部分智能化。 | | | | | | | | | | | |
| **一、申请理由** | | | | | | | | | | | |
| （内容应包括自身具备的知识条件、特长、兴趣和已有的知识基础）  **1.项目前提。**  本小组已取得《农业机械学报》的大力支持，并且学报已向本组提供研究所需的一系列文档和内部资料，例如：学报编辑部工作流程，论文格式要求明细以及写作要求，论文编改专项核对表，国内大学二级学院具体名称汇总信息等资料；除此之外我组与学报编辑部能够直接迅速的交流，已听取编辑社提出的详细要求，并包括纸质版、电子版在内的详细资料与需求，且学报的编辑都有丰富的工作经验，可以在第一时间给我们提供实现软件功能的建议，资源和软件开发后功能测试最直接的用户反馈。  以下为部分资料：  474096987.jpg  **2.知识储备。**  本小组为信息与电气工程学院和理学院的学生组成，包括计算机专业与数学专业，已经学习并掌握了高等数学、离散数学、高等代数、大学物理、概率论与数理统计等基础学科方面的知识，并且可以熟练运用C语言、C++、Java、python等编程语言，有使用Indesign等软件进行排版的经验，可以熟练使用word，LateX文档编辑功能，具有较强的编程能力。学习成绩良好，理论知识扎实，具备较强的学习能力和创新能力，且已经有在《农业机械学报》模拟投稿的经验，熟悉作者投稿和编辑审稿的流程。  **3.兴趣使然。**  本小组成员对于人工智能方向的自然语言处理技术具有浓厚的兴趣，且本项目所需要的理论基础与小组成员的兴趣方向相契合。  **4.导师带领。**  指导老师陈英义教授对于我组项目涉及的人工智能、智能计算领域有深刻的研究和丰富的经验。  **5.市场需求迫切，预期收益巨大。**  2016 年，全国共出版报纸1,894 种，总印数390.07亿份总印张1,267.27亿印张；全国共出版期刊 10,084 种，同比增长 0.70%；总印数 26.97 亿册；总印张 151.95 亿印张；数字出版业实现营业收入 5,720.85 亿元，实现利润总额 427.84 亿元，分别较 2015 年增长 29.91%和 27.89%。全国共出版电子出版物 9,836 种；出版数量 29,064.66万张，同比增长 35.57%。全国共出版图书 49.99 万种，较 2015 年增长 5.07%。其中，新版图书 26.24 万种，增长 0.76%；重版、重印图书 23.75 万种，增长 10.28%；总印数 90.37 亿册，增长 4.32%；总印张 777.21 亿印张，增长 4.58%；定价总金额 1,580.96亿元，增长 7.10%。据Futurism网站报道，根据学术界颇具权威的爱思唯尔（Elsevier）SCOPUS数据，2016年，中国共发表了42.6万份研究。我国期刊杂志等刊发量巨大，过去审阅稿件全靠编辑一职，一旦我组项目成功，将大大减少出版社的劳力支出和时间支出，大大的提高出版效率。 | | | | | | | | | | | |
| **二、立项背景** | | | | | | | | | | | |
| （包括国内外研究现状、趋势、研究意义、参考文献和其他有关背景材料）  **1.1编辑文档处理现状：**  目前为国内编辑部所采用的采编系统：勤云采编系统，中国知网“腾云”采编系统，三才采编系统，玛格泰克采编系统。其中勤云采编系统具备校对补全错别字的能力。而同类型的采编系统则宣称可以优化结构的能力。  采编系统 网页上宣传的能力  勤云 自动注册DOI号、参考文献进行链接  知网腾云 参考文献校对比对、机器自动审校稿  （可定制），邮件单据模板  玛格泰克 元数据提取（增值服务）期刊校对(增值)  三才 参考文献辅助编校（另购），无  黑马校对等其他文字性校对工具：进行文字内容的校对进行错别字，政治性敏感词等的校对。虽然这些已经很大程度的提高了编辑们的审稿效率。但是文章中微小的格式错误，却并没有在他们的考虑范围之内。  微小的错误往往不容易察觉，例如：  错误 正确  中国农业大学信息电气工程学院 中国农业大学信息与电气工程学院  山东农业大学水利与土木工程学院 山东农业大学水利土木工程学院  中国农业大学水利土木工程学院 中国农业大学水利与土木工程学院  内蒙古农业大学水利与土木工程学院 内蒙古农业大学水利与土木建筑工程学院  一字之差，往往会使得一个单位得不到产出的成果。而针对诸如此类问题的校对功能是前面所提的四个采编系统目前不具备的。如：审查作者单位名称是否准确书写、审查论文中的序号是否符合格式规范、审查相关公式是否引用准确、审查引注参考文献方式是否准确等。上述功能正是我组项目着力实现的。  **1.2自然语言的应用现状：**  自然语言处理对于文档编辑帮助巨大，然而在当下并没有一种普遍的用于文档编辑的自然语言处理技术，我组在该方向的创新当属前列。即便如此，自然语言技术对于文档处理方面的成果对我组项目仍有很大的借鉴意义，以下是有关自然语言处理方面的参考文献。  《基于深度学习的文本自动摘要方案》一文，针对自然语言处理生成式摘要领域的语义理解不充分、摘要语句不通顺和摘要、准确度不够高的问题，提出了一种生成式自动摘要解决方案。该方案包括一种改进的词向量生成技术和一个生成式自动摘要模型。改进的词向量生成技术skip-gram 方法生成的词向量为基础，结合摘要的特点，引入词性、词频、逆文本频率三个词特征，有效地提高了词语的理解。其中作者提出的 Bi-MulRnn生成式自动摘要模型以序列映射与自编码器结构为基础，引入注意力机制、门控循环结构、双向循环神经网络、多层循环神经网络和集束搜索，提高了生成式摘要任务的摘要准确性与语句流畅度。基于大规模中文短文本摘要数据集，实验比较的结果表明，该方案能够有效的解决短文本生成式摘要问题，并在Rouge标准评价体系中表现良好，提高了摘要准确性与语句流畅度。该文章中的生成式自动摘要技术可应用于新闻行业、推荐系统、以及信息检索等领域，具有良好的应用价值。但是该模型在对一部分特有名词处理时无法识别这些信息，最终会导致摘要生成不准确。  《利用自然语言处理技术优化弹幕视频网站中纪录片的观看体验》一文中针对记录片在弹幕网站中呈现出的活力和与弹幕相关而引发的问题进行分析和研究，通过借助自然语言处理和优化交互操作两种手段，对弹幕内容进行分类展示和宏观把控，本项目也可以利用自然语言处理技术对文章中不规范格式进行分类处理和宏观把控，做到使文章规范化；《A Neural Approach to Automated Essay Scoring》一文提出一种基于递归神经网络的方法来学习论文与其制定分数之间的关系，且不需要任何的特征工程，在通过不断地实验后建立了一种记忆网络。通过该神经网络可以大大提高文章打分的准确率。该文中利用自然语言处理对目标文章的分析以及文章内容提取方法对于本小组项目也有很强的借鉴意义。  **2.研究意义**  审稿的及时性与准确性关系到科研信息传播的时效和质量和学报的发展，所以学报编辑部面临审稿周期短，工作压力大的问题：据《农业机械学报》韩英编辑称：编辑平均工作量可以达到每天大约40-50页的审稿量，一个仅有5-6人的编辑部，一年的审稿量可达到600篇，国内约五百多家面临此问题的理工类出版社，但目前市场上被编辑部应用较多的几大平台还无法准确针对编辑部的需求，并且在短期这些平台并没有收入这些功能的计划。  据《农业机械学报》韩英编辑称其在与各大平台交涉后他了解到，上述平台对于针对这个需求开发相关功能并不感兴趣，因为成本超过了收益。这使有一些比较易于实现却可以节省编辑很多时间精力的辅助校对功能是编辑部所需求却不可求的。  综上所述，我们项目根据学报编辑提出的需求来进行设计针对学术期刊论文的智能校对，可为许多有类似需求的编辑部节省大量的人力，能够有很好的应用前景。  **3.项目基础材料：**  《论文编改专项核对表》—农业机械学报  《论文格式要求》—农业机械学报  《编辑部工作流程》—农业机械学报  《论文写作要求》—农业机械学报  《国内大学二级学院具体名称汇总》—农业机械学报  **4.参考文献：**  [1] 产品功能 - 国内首款参考文献编校系统[EB/OL]. 西安三才科技实业有限公司 [2019-01-09] http://proofread.notefirst.com/product/productour.aspx  [2] 北京玛格泰克科技发展有限公司----采编系统、采编软件、期刊采编软件、期刊采编系统编辑部办公系统、编辑部办公自动化、网刊发布系统、远程审稿系统、远程投稿系统、玛格泰克软件、期刊办公自动化、稿件管理系统、编辑部远程稿件处理系统、期刊群网、稿件处理、杂志社管理、采编、编辑部稿件管理、学术会议系统、网上投稿、网上审稿、专家审稿、网刊[EB/OL]. 北京玛格泰克科技发展有限公司 [2019-01-09] http://www.magtech.com.cn/CN/column/column290.shtml  [3] 腾云期刊协同采编系统平台优势[EB/OL].[2019-01-09] http://find.cb.cnki.net/advantage.html  [4] 北京勤云科技发展有限公司 官方网站[EB/OL].[2019-01-09] http://www.e-tiller.com/ch/index.aspx  [5]《基于深度学习的文本自动摘要方案》张克君、李伟男、钱榕、史泰猛、焦萌。  [6]《利用自然语言处理技术优化弹幕视频网站中纪录片的观看体验》张雪、何朝杰、彭韧。  [7]《A Neural Approach to Automated Essay Scoring》Kaveh Taghipour、Hwee Tou Ng。  [8]《科技期刊编辑如何处理修后重审稿件》张淑敏 段桂花 张维维 林 松 辛明红 向 政  [9]《学术期刊编辑应慎重对待再审稿件》 潘学燕 郭柏寿 杨继民  [10]《浅析如何提高高校学报的审稿效率》 朱夜明  [11]《高校学报可持续发展的战略思考》 韩维栋 党静萍 | | | | | | | | | | | |
| **三、特色与创新** | | | | | | | | | | | |
| **1.国内首家应用，创新领先。**  到目前为止，国内外尚无对接出版社的文档格式检查软件。市面上存在的诸如：EmEditor Professional文档编辑器、UltraEdit文本编辑器等都是基于作者本人的文档编辑器，且只能设置一般的文档格式，只具备修改格式功能而不具备检查格式功能，而多为编辑部所采用的勤云，知网腾云，玛格泰克，三才等平台也无相关针对性强的具体功能，忽略了编辑部的要求。而本项目属于国内中文杂志文档编辑的首次对接出版社需求的应用研究，软件功能直接而精准地针对理工类学报论文审阅校对过程中的具体需求来设计，是编辑部常用的各大系统中所不具有的功能。  **2.形式创新特色。**  利用python和java与word文档对接，实现文章的导入导出，并在此基础上利用自然语言处理技术，检索通篇文章，检测语法语义和格式，与编辑部的格式要求相匹配，发现需要修改的地方便记录并反馈给审阅人，自然语言技术使得更方便更智能的优化编辑软件成为了可能。 | | | | | | | | | | | |
| **四、实施计划及方案** | | | | | | | | | | | |
| （包括进度安排、研究方案、实验方案、技术路线等）  **4.1进度安排**   1. **2018年11月末-2018年12月15日**  * **与编辑进行沟通，了解编辑部诉求** * **查阅采编系统文献及自然语言处理技术的文献。** * **导师对选题的建议** * **咨询不同采编系统服务商的处理意见** * **准备国创文献**  1. **2018年12月底-2019年1月**  * **分配成员任务，选择软件编译环境** * **听取导师建议并制作技术路线，设计软件基本框架。** * **和编辑进行沟通，咨询编辑部以及同类型期刊的审稿流程** * **学习自然语言处理技术，查阅相关文献。**  1. **2019年2月----2019年3月**  * **采集编辑部格式数据，获取数据集并建立比对数据库。** * **做出最初不使用NLP技术的软件框架（可用库去做基本比对的文档），并联系编辑部进行测试。** * **根据老师意见改进NLP技术的路线设计。** * **进一步学习NLP技术，参加导师组会，听取师兄们的意见。**  1. **2019年4月----2019年7月**  * **做出测试版的使用NLP技术的检测软件，训练后满足学院名称校正的基本功能。** * **与编辑部进行沟通，改善使用功能。** * **参考行业内外的技术进展，确定下一步增加的功能。**  1. **2019年8月----2019年11月**  * **做出第二版测试版，增加序号筛选检测等功能** * **与编辑部及论文作者进行沟通，优化测试版使用体验** * **投放部分新版测试版，根据反馈改进功能**  1. **2019年12月---2020年2月**  * **增加自主选择模板的功能，适用于不同编辑部的模板** * **调试优化程序，改善编辑部及作者的使用体验。**  1. **2020年2月----2020年3月**  * **进行人机交互界面的优化，提供正式版软件** * **完成技术报告** * **申请专利** * **撰写结题论文**  1. **2020年3月----2020年5月**  * **借助农业机械学报进行产品投放** * **谋划商业化及后续产品维护**   **4.2研究方案**  **（1）使用IDE提取文章内容，建立数据库收录资料。**  **（2）利用NLP提取归类。检测序号序列是否正确。**  **（3）与数据库内容对比，文献引用查重。**  **（4）生成反馈与真实情况对照。**  **（5）由深度学习算法不断做出修改，直到准确率达到标准。**  **4.3实验方案**   1. **前期查阅相关文献，了解国内外研究现状与编辑部对于校对的需求，制定初步实施计划、研究方法与技术路线，采用已有采编平台资源进行初步实验，评判项目难点** 2. **中期（技术环节）** 3. **后期优化软件运行效率与人机交互友好界面，利用编辑部的数据资源进行实验，不断改进模型，成熟后分发给用户**   **4.4技术路线**  **第一部分：**  **建立论文审阅相关的数据库**  **第二部分：**  **实现软件的文本语义识别功能**  **第三部分：**  **实现软件判断文本是否符合要求的格式功能**  **第四部分：**  **实现软件自动生成反馈的功能**  **C:\Users\董瀚琪~1\AppData\Local\Temp\WeChat Files\5e1152adc34e2fca5a4ab28f2aa2ec1.jpg** | | | | | | | | | | | |
| **五、实施基础与条件** | | | | | | | | | | | |
| （包括项目实施应具备的专业知识、基本仪器、设备及实施条件）  **专业知识：**  **1.python、java、C等编程语言。**  **2.深度学习、自然语言处理方面知识。**  **基本仪器：**  **移动终端、个人电脑**  **实施条件：**  **1.获得农业机械学报的大力支持，已经得到相关的资料和具体详细的需求指导。**  **2.本小组已经制定详细的计划、查阅大量相关资料。**  **3.组内成员具有较好的专业知识背景。** | | | | | | | | | | | |
| **六、预期成果** | | | | | | | | | | | |
| （包括发表论文、竞赛获奖、申请专利及其他）  **该项目的成功实施预计取得以下成果：**  **1.一款面向编辑社的智能格式规范检测软件，打开国内该方面的市场。**  **2.撰写相关论文并发表，推动国内有关该方向的研究。**  **3.申请软件系统专利，参加竞赛。** | | | | | | | | | | | |
| **七、经费预算** | | | | | | | | | | | |
| （包括大概支出科目、金额、计算根据及理由）  经费预算总额： **18328** 元   * 1. **书本费**      + **纸质版书籍可以随时查阅，通过阅读NLP技术书籍，提高组内成员的技术**      + **电子版, 阅读实时更新的技术，跟上技术的前沿。**   2. **采编系统试用**      + **通过使用4种不同的系统，感受可以改进的模块，感受开发者的需求**   3. **专家咨询**      + **涉及到我们不擅长的行业，需要通过咨询行业内专家的意见完成后期推广商业化的改进**   4. **旅费**      + **西安三才科技的参考文献程序对我们有借鉴价值，北京到西安的高铁二等座票价为515.6元**   5. **宣传手册**      + **制作宣传手册并分发宣传给必要的顾客**   6. **服务器**      + **搭建云平台，腾讯云两年期服务器官网最低价格为858.6元**   7. **云数据库**      + **搭建云平台，腾讯云6个月云数据库最低价格为626.4元**   8. **专利费用**      + **当前预算为6000元的实用新型专利**      + **如果是发明专利的话是8400元** | | | | | | | | | | | |
| 序号 | | 仪器、材料、药品名称 | | | | 单位 | 数量 | | 单价 | | 总价 |
| 1 | | 纸质书目 | | | | 本 | 15 | | 60 | | 900 |
| 2 | | 电子书目(网上各类资源) | | | | 份 | 2 | | 200 | | 400 |
| 3 | | 采编系统的试用(含全媒体) | | | | 种 | 4 | | 0-1000 | | 2000 |
| 4 | | 专家咨询 | | | | 次 | 2 | | 2000 | | 4000 |
| 5 | | 旅费(公司，报社拜访) | | | | 次 | 6 | | 515.5 | | 3093 |
| 6 | | 宣传手册 | | | | 本 | 30 | | 15 | | 450 |
| 7 | | 腾讯低配服务器2年使用费 | | | | 台 | 1 | | 858.6 | | 858.6 |
| 8 | | 发明专利申请费用 | | | | 次 | 1 | | 6000-8400 | | 6000-8400 |
| 9 | | 低配云数据库6个月 | | | | 台 | 1 | | 626.4 | | 626.4 |
| 10 | |  | | | |  |  | |  | |  |
| 11 | |  | | | |  |  | |  | |  |
| 12 | |  | | | |  |  | |  | |  |
| 13 | |  | | | |  |  | |  | |  |
| 14 | |  | | | |  |  | |  | |  |
| 15 | |  | | | |  |  | |  | |  |
| 合计 | | | | | |  |  | |  | | 18328 |
|  | | | | | | | | | | | |
| **八、申请人签名** | | | | | | | | | | | |
| **项目负责人：**  **项目团队成员：**  **年　　月　　日** | | | | | | | | | | | |
| **九、导师意见** | | | | | | | | | | | |
| **签名**    **年　　月　　日** | | | | | | | | | | | |
| **十、学院意见** | | | | | | | | | | | |
| **签名**  **学院公章**  **年　　月　　日** | | | | | | | | | | | |