参赛队员姓名：邹桐　米子琪

中学：南京外国语学校

省份：江苏省

国家/地区：中国

指导教师姓名：史钭镭、蒋兴超

指导教师单位：南京外国语学校

论文题目：计算机智能辅助评分系统母语写作切题研究

计算机智能辅助评分系统母语写作切题研究

南京外国语学校　邹桐　米子琪

摘 要

本研究介绍了一种命题作文切题评价的原创算法。在适应大规模考试评卷工作的计算机智能辅助评分系统的实现中，该方法通过实践“对着命题提问题，带着问题读文章”的解题思路，提升母语写作难度水平下命题作文的切题评价任务的执行效率与精准程度。

本研究以南京外国语学校初中学生考试命题作文数据集作为研究对象，凭借预训练语言模型的机器阅读理解能力，实现了命题作文的切题评价任务的语义级别特征组合提取。基于深度学习技术的神经网络模型构建能力建立了回归预测模型，最终得以实现高效并精确地推理作文的切题程度分类。

在目前理想前提实验阶段中，我们建立的回归预测模型对于测试集的切题程度判断的正确率为 96.41%，验证了此母语写作切题程度分类算法的可行性。

关键词

自动作文评分系统；切题；问题生成；机器阅读理解；自然语言处理；神经网络。

Abstract

This study introduces an original algorithm for evaluating the relevancy between topic and content of compositions with assigned topics. In the implementation of computer intelligence-assisted scoring system adapted to large-scale examination marking work, this method improves the execution efficiency and accuracy of the task of relevancy evaluation of compositions with assigned topics under the difficulty level of native writing by practicing the problem solving idea of "asking questions accordingly and reading the composition with the question".

In this study, we take the data set of test written compositions of junior middle school students in Nanjing Foreign Languages School as the research object, and by relying on the machine reading comprehension ability of the pre-trained language model, we achieve the semantic level feature combination extraction of the topic related evaluation task of the test written compositions. Based on the neural network model construction ability of deep learning technology, the regression prediction model is established, which finally realizes the efficient and accurate inference of composition relevance degree classification.

In the current ideal premise experiment, the accuracy of the relevancy degree judgment of the test set is 96.41%, which verifies the feasibility of the relevance degree classification algorithm for native writing.

Key word

automatic composition scoring, relevancy, question generation, Machine Reading Comprehension, natural language processing, neural network

目　录

1. 背景

网上评卷系统[1]已普及应用于各类大规模考试评卷工作中。在网上评卷过程中，通过试卷扫描、任务分派、上机阅卷和仲裁处理的步骤完成评卷任务。通过一卷双改，差异超限后触发仲裁处理机制，有效的提升了评价标准的一致性以及阅卷和归档过程的安全性。

随着人工智能技术的发展，计算机智能辅助评分系统愈加广泛地运用在教育领域，在网上评卷系统之上进一步的减轻评卷人力负担，提升评价标准的一致性，实现快捷的考试评价体验。

自动作文评分系统 （Automated Essay Scoring）是语言类考试计算机智能辅助评分系统的一种。Ju-Lu, Yu 等[2]指出：英文写作评价部分（如在托福、雅思等上机考试中的写作题），已较为广泛地运用了自动作文评分系统，以计算机代替人工评卷的方式彻底解决人工参与评卷引起的效率和质量问题。

在面向中高考等以中文母语进行日常训练评价领域涌现了如 IN 课堂[3]、测文网[4] 等自动作文评分系统，实现了作文答卷 OCR，自动作文评分和自动修改建议。

何屹松等[5]指出：在 2017 年 6 月安徽省高考评卷期间，利用科大讯飞智能阅卷系统对考生的语文作文答题情况进行后台离线智能评分，并将评分结果应用于网评质量监控。其中语文字符的识别准确率为 97.6%。机器评分与报道分的评分一致率为 95.24%，证明了智能评分整体效果优良。

1.1 国外自动作文评分系统

国外英语较早实现机考，并随着机考的实施，逐步的开始了英文写作评价的研究和实现，早期的 PEG(Project Essay Grade)基于特征工程方案，通过提取学生作文的浅层语言学特征并结合命题人提供范文的浅层语言学特征进行比较，完成对文章的评分。

IEA(Intelligent Essay Assessor)系统通过潜在语义分析（LSA）将学生作文话转换为语义空间向量，并结合命题人提供范文的语义空间向量，完成对文章的评分。

E-rater[6]系统通过接近于阅卷教师的批改习惯，基于关键词的文章分类、在不同的文章分类中对于词语、句子和篇章结构提取特征，并根据回归模型给出最终得分。

1.2 国内自动作文评分系统体验

我们对 IN 课堂和测文网两种国内较为知名的自动作文评分系统进行了体验，得到了如表 1-1 的体验感受：

表 1-1　两种国内较为知名的自动作文评分系统体验

体验角度	体验感受
OCR 准确率	经测试正常卷面书写水平下正确率可达 95%以上
表达维度评价	可实现优美句子识别[7]、错别字检测、病句检测等功能
结构维度评价	可实现汉语篇章结构分析、句子通顺程度分析等功能
内容维度评价	两款自动评分系统都是按照关键词匹配机制进行的。对文章切题的判断较容易失误，对整体评分效果影响较大。

符耀章等[8]指出目前汉语作文智能评卷技术已经在部分考试中投入使用，文章介绍到将文字片段序列化处理之后，再使用 WORD2VEC 方法提取各维度特征，其中包括了离题检测特征，用以表征作文的客观情况，在此基础上利用建立回归模型的方式实现作文总体评分。但是，文章中并没有明确离题检测特征的详细实现情况。其他的数篇自动作文评分领域论文涉及到了语义离散度，优美句识别，嵌入语言深度感知等领域，也未明确的涉及到作文切题程度的自动评价方法。

本研究以南京外国语学校初中学生考试命题作文数据集[9]作为研究对象，凭借预训练语言模型的机器阅读理解能力[10]，实现了命题作文切题评价任务的语义级别特征组合提取。基于深度学习技术的神经网络模型[11]构建能力建立了回归预测模型，最终得以实现高效并精确地推理作文的切题程度分类。我们计划用此方法替代基于关键词匹配的上一代切题评测方法，并通过实验验证其具体效果。

2. 问题的提出

作文跑题是母语难度级别写作中内容部分的常见问题。以南京市中考作文评分标准[10] 为例，其内容部分评价标准由（1）文章中心明确且符合题意；（2）文章内容围绕中心详实的展开；这两点进行体现。满分 50 分，如果出现文不对题现象，则学生最高分不得超过 17 分，如果只是点缀了一下命题或是中心不明确，或是没有围绕符合命题的中心撰文，则最高分不得超过 26 分，可见切题与否对于母语写作考试评分影响之大。

在作文教学层面，保证对于命题要求的正确理解及在写作时保持文章内容围绕正确的中心展开确属难点。这也体现在学生答卷层面，虽然同一场考试中大部分文章是切题的，但总有走题的文章会被发现，从而在没有明显表达与结构问题的情境下获得偏低的分数。不仅如此，在阅卷教师阅卷层面，因为切题与否对于母语写作考试评分影响重大，所以误判切题程度是造成评卷质量明显下降的主要原因。

作文切题水平评价要求阅卷教师熟读全文，大规模考试评卷工作中实际的平均每卷批改时间一般不会超过 30 秒，有时甚至短于 15 秒。在阅卷工作开始后，随着阅卷教师长时间高强度的面对电脑评卷，逐渐会产生生理和精神疲劳现象。

考虑到设计良好的作文考试命题其得分应符合正态分布，也就是中分段得分作文数量较多，因此阅卷教师如果给出高分段或是给出低分段得分，相较于给出中分段会有更大的概率触发仲裁环节，若单纯以触发仲裁率考核阅卷教师的作文评价质量，则阅卷教师有可能会将有较高或较低评价质量尽可能按照中段位给分，最大可能的减少触发仲裁率，这种现象既埋没了好文章，同时也降低了问题文章的暴露机会。

在目前大规模考试评卷中，早先的研究者已经在作文的扫描、文字识别以及在评价作文的结构、表达维度上达到了较好效果。因此，本文聚焦于计算机智能辅助评分系统的母语写作切题评价流程，尝试解决如下问题：

（1）如何基于语义理解实现对于作文中心是否切题的准确判定？

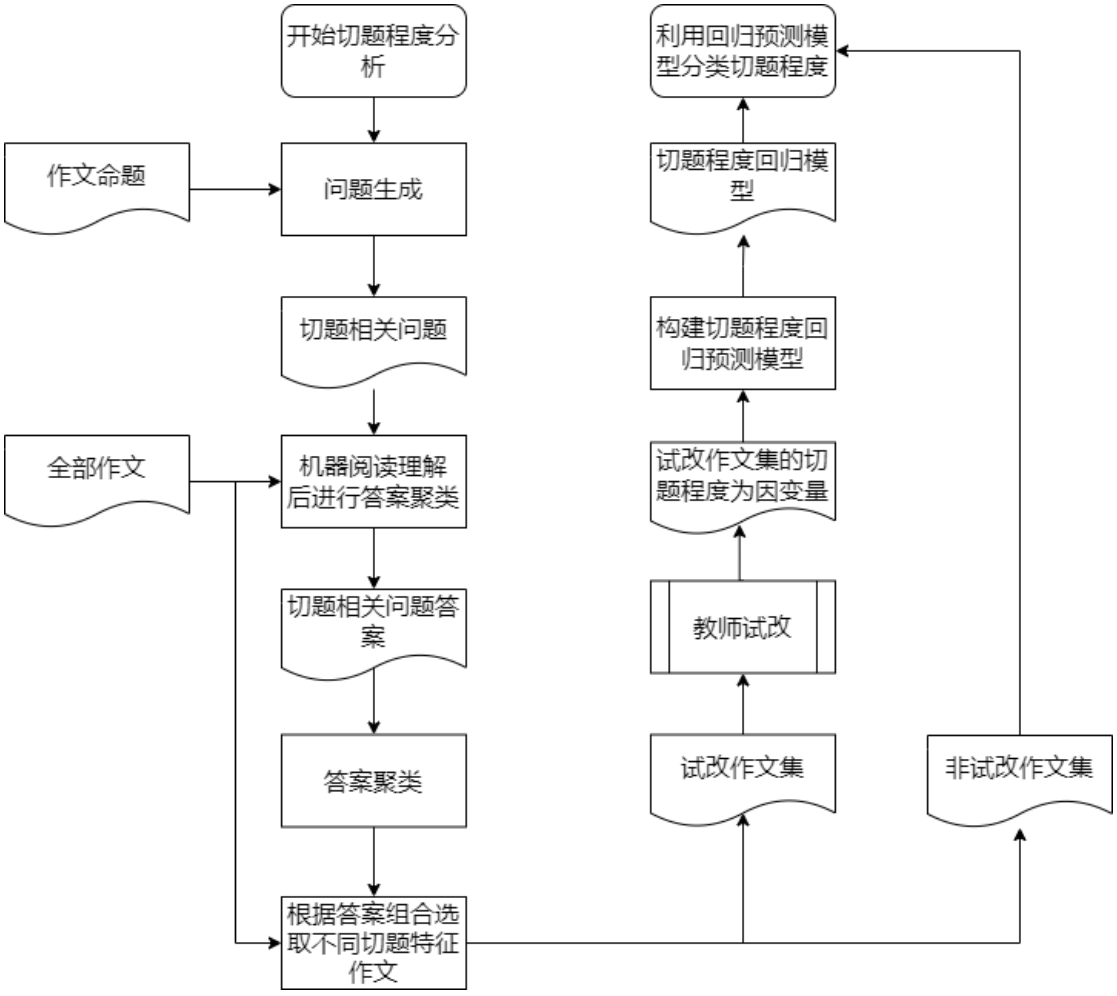
(2) 如何保证对于不同的命题都可以使用本写作评价流程?

(3) 如何让阅卷教师有能力引导和审核计算机智能辅助评分工作?

(4) 如何设计实验流程使得整个设计思路能够落地?

3. 问题的解决方案

3.1 基于机器阅读理解建模的母语写作切题评价流程



为了验证文章是否切题，阅卷教师会在大规模考试评卷工作的开端共同审题，预测考生可能的写作主题，尝试通过设置部分用于区分不同写作主题的问题。

例 1 作文命题的问题生成

作文命题：以“南京街头”为题写一篇文章。

命题解析：题目两个限定：南京，街头。要求有二：必须写街头发生的事，不是在乡村、小区、商场里、景区里等等；而街头又必须是南京的街头，不是其他地方的街头。否则跑题，直接不及格。那么我们就可以写出如下两个问题：

问题 1：是街头发生故事或风景吗？

问题 2：是写的南京吗？

阅卷教师会尝试带着问题小规模抽样试改作文，从而发现对于问题不同答案

的问题，再总结这些答案的不同排列组合形成初步的切题程度判定标准。

随着 Bert 等预训练语言模型类似训练方法的流行，通过神经网络驱动的预训练语言模型已经可以利用训练时网络学习到的知识进行基于文本生成任务的自由问答。而机器阅读理解是自由问答的一种特殊形式，同时提供一段参考文本和一个问题作为自然语言处理模型的输入，期望的输出是在参考文本上下文提供信息中基于语义的理解找到的对应问题的答案。

以中考作文评卷过程为例，阅卷教师经过提问、试改、形成标准，最后按照标准大量批改。我们基于这样的评卷过程提出了一种基于机器阅读理解建模的母语写作切题评价流程：

（1）根据作文命题要求提出能够区分切题程度的问题；

（2）对于全部作文文本执行切题问题的机器阅读理解，获得区分切题程度的问题答案；

（3）对于全部作文的每个问题合并其语义相近的答案；

（4）选取部分答案组合具有代表性的作文由阅卷教师试改得到切题程度；

（5）将试改结果中的作文切题程度作为预测因子（因变量），将各切题问题的答案的组合作为因果因素（自变量），设计切题程度的回归预测模型，通过深度学习的方法获得该模型的参数；

（6）利用回归预测模型计算其他待批改作文的切题程度，并形成结论。

若我们能够根据作文命题提出一些能够明确区分切题程度的问题，并有能力对于全部作文文本执行机器阅读理解获取这些问题的正确答案，最后利用合理的建模方法减少模型误差。当以上三个条件同时满足时，我们就可以认为基于语义理解实现了对于作文中心是否切题的准确判定。

3.2 切题程度判断问题的问题生成

语文老师判定作文的切题程度通常依照三种标准，这三种标准则依附于三种不同的题目类型。

第一种类型为给定主题，发散中心思想类，此种类型非常简单，达不到中考难度。作文题目会明确给出学生应该描写何种事件或事物，而可以从不同的写法和细节中提炼出不同的中心思想，进行升华。语文老师此时即可以直接通过学生文章所写事件是否符合规定事件评判是否切题。

第二种类型为给定中心思想，发散主题，此种类型的作文题目会提到若干关于学生作文内容的中心思想关键词，若偏离任意关键词则依照该关键词重要性判为部分切题或不切题。与第一种作文不同的是，学生先需从题目描述中提炼关键词，然后尽情发散想象力构思文章的事件内容，并将规定的情感中心体现在文章中。语文老师则将自己提炼的中心思想关键词与学生文章中体现的中心思想进行匹配，并评判是否切题。

第三种类型则为第一种和第二种的结合，既有给定事件，也有给定中心，但事件不会特别明确，比如只会给定“南京”，“漫步”或者其他限定词。评判标准则也是同时结合两种评判标准，规定事件和规定思想。

以上的三种类型其实都可以转化为提问回答的形式确认是否切题。切题程度判断问题的生成可以由阅卷教师人工进行，这有助于阅卷教师理解、引导和审计计算机智能辅助评分工作。阅卷教师可以编写并提供任何基于文本的问题，参考例 1。

切题程度判断问题的生成亦可由计算机使用问题生成技术完成。问题生成是文本生成技术的一种特殊情况，我们可以给定一段文本让计算机根据参考文本和答案生成一些问题，是机器阅读理解的逆向情况。ERINE3 模型已经提供了问题生成任务实现，但是因为目前我们研究尚处于早期，没有足够的命题问题数据集，因此没有进行问题生成技术相关的实验。

3.3 切题程度判断问题的机器阅读理解

ERINE3 语言模型通过设计精巧的任务对于大量文本进行学习，从而学习语言的语义表达。当我们已经拥有了切题程度判断问题和全部待批改作文的文本内

容，我们就可以利用 ERINE3 预训练语言模型的机器阅读理解任务从语义的角度获得对于每一篇待批改作文的每一个切题程度判断问题的答案。

在本文对应后续实验中我们利用的 ERINE3 模型在 C3 数据集[13]上执行机器阅读理解任务的准确度在 2021 年中时达到了 SOTA 级别的水平，在验证集和测试集上正确率都在 86%以上。

为了验证机器阅读理解答案的正确性，我们用 100 篇南京街头的作文针对例 1 提出的两个区分切题程度的问题进行了机器阅读理解：

例 2：机器阅读理解结果-正面（ERINE3.0 公有模型 API 调用结果）

正文 1：城门城门几丈高，三十六丈高。骑白马，带把刀，城门底下走一遭。”新童谣年轻的歌手木小雅，用活泼的歌声将南京的传统文化与音乐结合，唱出了所有南京人对这座城市街头风光的感情南京一座旧与新碰撞交融的城市清晨，整个城市随着第一束阳光苏醒。踏上南京的街头，四面香气立刻将磕睡赶跑。老城区一条街道数不胜数的小吃店，匆忙却满足的食客，扑面而来的热气与菜香，组成了我对南京早晨的印象。东边一家粉丝店，老板娘一边招呼客人一边赶走偷吃的黄狗；两家锅贴店门口年轻伙计互相争着顾客，吆喝声一声高过一声；烧饼摊老板拿根小棍对着金黄烧饼翻翻捣捣，熟练地包起甜的或咸的烧饼，一边偷空喊楼上赖床的女儿起床上学。南京冬天并不温暖，今年更是冷风刺骨。但走进街头一家小吃店，捧着热乎乎的陶瓷大碗，拿起暖融触的纸袋，便被暖意包裹，吃着早餐像是细细咀嚼着春天到了傍晚，南京的街头就全换换了一副模样。刚从学校放学，走上街道，一片车水马龙。晚高峰来住的车辆行人演绎着属于城市的交响曲。边各式建筑闪着不同颜色的灯光橘黄的路灯从树枝中伸出，道旁的梧桐托着七彩霞光。商业区高耸的办公楼们是亮如白昼，街边商场人们鱼贯而入。抬起头，身边还是喧嚣，眼前还是灯光却不会感到渺小，因为这是南京的街道穿过几条小巷，又回到早晨的老街，黄狗在树下睡着，老板娘在洗第二天的菜，烧饼老板收了摊，把门外疯玩的女儿拎回家。“城墙根下寻常人家，我在这儿长大”耳机中循环着木小雅的歌声。每个南京人心中，都有独特的，属于他自己的南京街头。

正文 1 阅读理解结果：

问题 1 是街头发生故事或风景吗？ 答案：“是的”。

问题 2 发生在南京吗？ 答案：“是的”。

正文 2：紫金山上的植被，在初春时有的已是深绿，有的却仍枯槁，然而这

并不代表枯黄的植被已全无生机：在月余的等待与厚积薄发中，每一株植物都绽放了属于自己的一片绿。生活也是这样，面对一时的落后，不应气馁，只要不断努力，积蓄力量，耐心等待，终有追上差距的一天。常言道：“水滴石穿，绳锯木断。”英国著名科学家霍金，在大学期间被诊断出卢伽雷氏症，甚至被医生预言寿命不超过一年，然而，他没有放弃积蓄力量，努力学习天体物理学，把自己看作常人一样同同学与老师探讨学术，经过二十余年，终于等到了《时间简史》这部巨著的出版。西汉史学家司马迁，年轻时也遭酷刑，与别人相比条件有许多落后，但他能等待自己一步一步地寻访史料，能等待自己一处一处实地查考，能等待自己一点一滴为一部集大成的史书蓄力，终在暮年之时，写出“史家之绝唱，无韵之《离骚》，名贯古今。可见，既蓄力，又等待，一切落后的条件均不是阻碍。然而，仅力不等待，是无法有所成就的。“锲而舍之，朽木不折。“无论用多大的力，刻了一会儿就停止，连最软的朽木上也刻不出什么东西。同样，光等待不努力积聚力量，也一事无成。春秋时期一位著名的棋手弈秋教两个学生下棋，其中一个“惟弈秋之为听”，而另一个三心二意，“一心以为有鸿鹄将至”，结局不出所料，三心二意的学生过了三年仍旧棋艺平庸。《满江红》中亦云：“莫等闲，白了少年头，空悲切。”可见人不蓄力，从少年到耄耋的等待，也没有什么结果。所以，要想克服自身先天不足，实现超越，等待与蓄力是缺一不可的。物理学中，有一个公式于此十分适用，便是“功“功率时间”。功率由努力蓄力的程度决定，时间由等待决定，当两者兼备之时，其乘积在数量上是十分大的，足以让我们追上“初春”时的差距，迎来“春天”。正文 2 阅读理解结果：

问题 1 是街头发生故事或风景吗？ 答案：“不是”。

问题 2 发生在南京吗？ 答案：“是的”。

自例 2 中我们可以观察到，这两篇文章机器阅读理解的结果还是较为理想的，特别正文 2 中只提到紫金山，并未提及南京，ERINE3 模型亦可正确回答问题 2。

例 3 机器阅读理解结果-负面（ERINE3.0 公有模型 API 调用结果）

正文 1：说到南京街头，想到的是单调的建筑、马路……但并不是这样的，城市的美，很大一部方体现在它的街头。只要你细心，就会发现，处处都隐藏着独特的蕴味。由朱赢椿老师设计的花迹酒店就隐藏在这城市的街头中在老门东的一个巷口，有一扇破旧的小木门，门边插着一束小花，边上是两块木牌，一块写着“欢迎入住谢绝参观。”另一块写着酒店的名字“花迹酒店”。木牌上一条淡淡的痕迹仅射着太阳光，我的嘴角不住的上扬，着细细的蜗牛痕带我走过了朱老师的各

种作品，沿着蜗牛的路，我发现了别样的美，而至今，我也没有看见痕迹尽头那只悠闲的蜗牛。进门，是木质的地板、天花板，白色的墙壁。一切看似简单，但如果你放慢脚步，便会看见，墙角上，门框上、花瓶上，一只只活灵活现的小昆虫，在朱老师笔下，这些昆虫有着各自的神态，看着这些虫子，我仿拂看见了坐在窗边聚精会神观察虫子的作者。除了虫子，还有花，带着我在小小的房子里走着。这着小花装点了白墙和一色的家具，它们生动了起来，散发着一种清香。房间中的家具有各种风格，组合在一起却毫无违和感。不完美，却像家一般温暖。院子四周是青砖砌成的矮墙，巧妙的挡住了周围的现代化建筑物，看着有限的天空，却又不觉得压抑。墙边的植物的叶片上挂着露珠，院子里是南京雨季特有的湿润。看着植物与青砖墙，我仿拂回到了民国时期；但转头看见玻璃推拉门，好像从梦中醒来却又不愿睁开眼睛，现实与想象现在与过去一起融入了梦境，这正是做为六朝古都的南东的独特蕴味。第二天，我又走出了那扇小木门回头看见那条蜗牛痕在阳光下闪烁。南京的街头既有车水马龙的马路，又有蜗牛慢慢爬过的美隐藏在南京的街头。

正文 1 阅读理解结果：

问题 1 是街头发生故事或风景吗？ 答案：“是拟人手法，把街头当作有生命的个体来写”。

问题 2 发生在南京吗？ 答案：“是的”。

正文 2：南京，是一座历史文化名城。可除开那些历史古迹，南京的街头有什么？我怀着这个疑问踏上了寻找的旅途那是一个元宵节。我和爸爸妈妈挤在去往老门东的路上——没错，我们要去看灯。一条本就不算太宽的道上挤满了慕名前来看灯的人，我们在这逼仄的空间里被挤不过气来。于是我们三人对视一眼，当即从人群中闪身而出，钻进了一条小巷子。巷口有一家糖芋苗店，我们走了进去。要了三碗糖芋苗，我们这才坐下来细细打量这家店：狭小的店面是敞开式的外面的丝丝冷气张牙舞爪地想冲进店铺，又被糖芋苗香甜的热气赶了出去墙上的小电视播放着万年不变的中央一套……嗯，这是一家很地道的街头小店。糖芋苗果真是又糯又香，我们三人吃完后，又继续向着老门东挪动南京街头，原来有不知名但美味的美食铺。走上北京西路，则会有与走在夫子庙完全不同的感觉。已是晚上六七点了，路灯微黄的光打在两旁斑驳的梧桐上，让人不禁去想这满城梧桐是否真的是当年蒋介石为宋美龄而种走到路的尽头，发现两边都是高楼大厦。定睛一看，楼下还有一个小贩在推车叫卖：“梅花糕，梅花糕”。原来南京街头还

有高楼与摊贩。南京的街头还有什么？有走街串巷卖糖葫芦的人，有会拿粉笔在地上画出蒙那丽莎的人，当然，大多还是高高的建筑我想，这就是南京这座古城的独特魅力。她不似上海那样现代，而是在高楼大厦间，“藏了些小摊、小店藏了些人间烟火气走在南京街头，你会发现，看似冷清的盏盏路灯，都在指引你走向回家的路。正文 2 阅读理解结果:

问题 1 是街头发生故事或风景吗？ 答案:“是一座历史文化名城。可除开那些历史古迹,南京的街头有什么?我怀着这个疑问踏上了寻找的旅途那是一个元宵节。我和爸爸妈妈挤在”。(生成式机器阅读理解输出如此，并非排版错误)

问题 2 发生在南京吗？ 答案:“是的”。

自例 3 中我们可以观察到正文 1 和正文 2 的问题 1 都答案的第一个“是”字都是符合期望的。接下来，正文 1 表述的内容或与文章实际并不符合；正文 2 出现对于原文的不完整摘抄。

目前实验结果表明，应用 ERINE3 模型对于南京街头的 100 篇作文样本阅读理解任务获取切题答案能获取到较为正确的答复。问题主要在于有一定概率会生成多余的文本影响判断，如例 3 正文 2 问题 2 所示。

本次需要实验的算法流程较长，在早期实验中，为验证整体流程可行性，我们通过人工审校了两个问题的答案。在保留原始输出用于后续实验的情况下，生成了人工标注答案数据集，并称利用该人工标注答案数据集的后续实验为在语言模型理想工况下进行的实验（简称理想工况实验），称利用原始输出答案数据集的后续实验为原有工况下进行的实验（简称原有工况实验）。

3.4 文章特征生成与试改文章抽样

当获得了所有文章的切题程度判断问题的答案之后，文章在不同切题程度判断问题上的答案组合反映了文章的切题特征，故我们依据答案组合将其文章进行分类，最后在每个分类中随机抽取数篇文章用作阅卷教师试改。

例 4 南京街头 100 篇作文与其他 103 篇作文按例 1 问题的全部可能答案(原有工况实验)

问题 1 是街头发生故事或风景吗？ 可能答案集合:

{'是一座历史文化名城。可除开那些历史古迹,南京的街头有什么?我怀着这个疑问踏上了寻找的旅途那是一个元宵节。我和爸爸妈妈挤在', '是拟人', '是南京的现实生活写照', '是滴', '是方言', '是南京的大街小巷', '是街头发生故事或风景', '是南京街头', '是指一些在街头发生的故事和风景', '是南京写的一篇游记', '是南京著名的景点。', '是南京本地宝', '是泰安一带', '是的', '是真实故事。', '是南京话', '是南京的街头走一走', '是南京本最繁华的街头', '是拟人手法，把街头当作有生命的个体来写', '是南京街头的一种味道', '是高淳老街那条石板路', '是南京的各个街头', '是南京街头走着', '是拟人手法，生动形象地写出了南京街头的风景', '是南京街头一景。', '是的。', '是随机事件', '不是', '是南京街头最有古色古香的地方', '是南京老城区的街头', '是南京街头最让人感到温馨', '是南京街头最真实的样子。', '是对城市的记忆。', '是真实故事', '是实景', '是南京的街头', '是南京所独有的,能够温暖人心的夏风一般的温柔。', '是南京街头最靓丽的风景', '是啊', '是街头发生故事', '是指在街上、马路上发生的故事或风景。', '是南京一座古城，但南京同时也一座现代化的城市。', '是拟人手法，不是风景，是说明某一种现象', '是江苏省南京市的一条街道。'}

问题 2 发生在南京吗？可能答案集合:

{'是的', '不是', '是的。'}

因为我们使用的是基于文本生成技术的机器阅读理解任务，当我们没有限定输出的范围时，有可能会多生成一些内容或标点符号，例如问题 1 中的答案“是拟人”，例如上例中的”是的“和”是的。“。

如果尝试直接使用以上答案挑选特征文章，机会因为第一问答案有多达 44 种可能性，第二问答案有 3 种可能性，从而使得两个问题的答案组合最多达到多达 44x3 种。可是我们的数据集一共才 200 篇，每种可能性找个两篇，就相当于得把所有作文批改完成了。

因此我们必须合并语义类似的答案，这里又有两种方法来实现。

人工审核法，在早期实验中，为了同时验证步骤 3.3 的结果正确性，我们通过人工审核审查了两个问题的答案，这两个问题的答案恰好都可以是：“是”、“不确定”和“否”，用于在理想工况下实验。

基于句向量的聚类合并方法，我们可以利用 Sentence Bert 等方法，在问题

空间下将该问题全部答案先进行向量化得到其表征其语义的高维度向量，再利用将维度可视化人工聚类或是使用 K-Mean 等算法直接进行空间聚类，得到语义相近的各聚簇关系之后利用这些关系将语义相似的答案替换为各聚簇中心的答案，从而实现每个问题答案可能性的归并。因时间关系，尚未进行该类实验。

合并语义类似的区分切题程度的问题答案后，此时对于每篇作文，其答案的不同组合也就代表了其在写作内容上符合的不同特征。而对于过往阅卷教师发起的试改流程所不同的是，若基于计算机智能辅助评分系统的计算而不是人工抽样，我们完全可以在阅卷教师提出问题后，先行完成全部待批改作文对于全部问题的机器阅读理解工作。若切题程度问题具有区分度且每个问题答案机器阅读理解输出的答案完全正确，就能够完整地得到待批改作文的全部语义级别内容特征组合。此时从每种语义级别内容特征组合所对应的文章中抽取数篇结构和表达方面较为优秀的文章进行试改，就可以代表该类文章内容的切题程度。

表 3-1　南京街头 100 篇与其他主题作文 103 篇

理想工况下答案分布情况一览表

序号	问题 1 答案	问题 2 答案	作文数量
1	是	是	98
2	是	不确定	1
3	否	是	41
4	是	否	1
5	否	否	6
6	否	不确定	56
7	总计		203

根据表 3-1 可以得知本批次 203 片作文一共有 6 类特征组合，其中有 2 类恰好只有 1 篇作文，直接人工批改，其他 4 类每一类筛选 2 篇进行试改即可，共计 10 篇选入训练集，剩余 193 篇用于测试集。同时因为有两类答案分布恰好各仅有一篇作文，我们也把他们重复的放入了测试集，故测试集共计 195 篇作文。

3.5 文章切题程度试改

文章切题程度试改有两种方法实现：直接评价切题程度和自作文分数倒推。

直接评价切题程度可以考虑将切题程度划分为切题、一般和不切题三类，由老师直接阅读待试改文章，考虑文章中心是否切题且文章内容是否围绕中心展开进行判断即可。

如果阅卷教师不习惯直接评价方法，亦可参照评分标准，自分数逆向推出切题程度。但是需要注意的是，因作文总分内包含了对于内容、结构和表达三种维度的评价，由可能出现内容层面切题的文章，因为结构或表达问题被扣分，从而被错误的划分到一般档次的滑档现象。

因此我们建议阅卷教师试改时对切题的进行直接评价。

本次实验中引用的南京街头 100 篇作文切题程度为自作文分数倒推法推出，划分的分数段如表 3-2 所示：

表 3-2 作文总分分布范围与切题程度对照关系一览表

分数范围	切题程度
[35,50]	切题
[18,34]	一般
[0,17]	走题

作为参考的其他作文 103 篇并不是按照南京街头命题的，此时它们的总分在南京街头命题下没有内容维度的参考意义。因此我们采用了基于区分切题程度问题答案的快速匹配的方法，解决间接判定其是否切题：若两个答案均为是，则人工阅读，确定的确切题，这种情况只有 1 篇。而对于其他文章而言任意问题答案为否则判别为走题，这种情况有 102 篇。

3.6 切题判断回归模型的建立和应用

切题判断回归模型的输入为每篇文章区分切题程度问题的答案，输出为这篇文章的切题程度。

我们神经网络的形式建立切题判断回归模型。在本文的南京街头主题两个区分切题程度问题的场景下网络的架构为：可接受 2 个输入信号的输入层（sigmoid 激活函数），全连接隐藏层（sigmoid 激活函数）和 1 个信号输出的输出层。

基于 Pytorch 框架提供的利用深度学习思想的训练工具构建了该神经网络的代码实现，并基于阅卷教师试改数据作为训练集构建了回归模型。

我们利用该模型对于测试集进行了验证，目前在语言模型理想工况下测试集的切题程度判断的正确率为 96.41%。

我们找到了 7 篇模型推测与实际评分有差异文章，我们发现造成这些差异的原因较为单一是在 3.5 文章切题程度试改中所提到的出现内容层面切题的文章，因为结构或表达问题被扣分，从而被我们错误的划分到一般档次的滑档现象。从而实际的正确率在语言模型理想工况下可以达到 100%，证明了从理论上说本文提出的切题程度分类算法是行之有效的。

需要注意的是，区分切题程度问题的提问和机器阅读理解在实际使用中的正确率并不总是理想的，所以未来还应当有进一步的工作进行实际工况下的总体切题判定正确率。

3.7 跨命题可用性

上文以南京街头为命题讲述母语写作过程中的切题程度判定流程，那么如果更换命题，我们还是可以按照此流程进行、其中机器阅读理解和语义向量化的模型可以直接复用，而答案到切题程度的回归模型需要进行针对命题进行重新建模。

市面上的语言模型训练集大多数都不是作文，通过大规模考试作文作为训练集有应当能够增加语言模型在作文领域的机器阅读理解任务和语义向量化任务的准确性。

4. 结论

4.1 结论

我们借鉴了语文老师在大规模作文答卷考试评卷中所使用的业务流程：通过试改发现部分有特征的内容主题分类，再通过将其他作文匹配到对应内容主题分类的方式进行高效切题程度判断。

在项目早期我们也尝试了基于统计特征、句向量、摘要生成等方法切题程度判断机制。但是效果始终不达预期。最终，我们观察到可以利用问答的形式，围绕命题提出部分主题相关的切题程度判断问题，根据这些问题的答案将文章主题进行分类。这种方法覆盖场景全面，并且易于阅卷教师理解评价过程的原理。

通过预训练语言模型的机器阅读理解能力，我们可以高效的推理全部作文的切题程度判断问题的答案，无盲区枚举全部内容主题特征分类，精准选出各分类具有代表性的作文进行试改，这就超越阅卷教师试改时抽样带来的主题特征分类不完整的局限性。

基于试改的结果，我们建立了切题问题答案到切题程度的回归模型，再以此模型预测其他待批改作文的切题程度。基于这样的机制，我们可以保证作文切题程度判定速度更快且结果更加稳定。

为了尽快将理念验证落地，我们目前已经完成的工作为语言模型理想工况下实验，这个阶段的主要目的是在语言模型理想工况下验证思路的可行性。此时由阅卷教师提出区分切题程度的问题，语言模型对该问题进行机器阅读理解得到各篇作文各自切题程度问题的答案，并由人工审校，再进行试改文章抽样和试改，最后建立切题判断回归模型。

目前在语言模型理想工况下测试集的切题程度判断的正确率为 96.41%。在未来的实验过程中，我们还将展开验证语言模型现状工况下切题程度实验以及通过作文文本训练集的针对性调优后工况下切题程度的实验，帮助计算机智能辅助评分系统在高效执行母语写作切题评价任务时保持可靠的评测质量。

4.2 后续工作规划

为了进一步了解到语言模型给出的阅读理解答案对于切题判断的影响，需要开展语言模型现状工况下实验，在此实验中不再进行对于各篇作文区分切题程度问题答案的人工审校环节，并正常进行后续环节，根据切题预测正确率的变化了解影响范围，并搜集不合理的机器阅读理解回答案例进行调整，生成语言模型调优训练集。再尝试利用语言模型调优训练集调优，帮助语言模型正确理解问题答案的回答方法。通过更多合理的切题问答答案，提升切题判断的准确度。

最后我们还将尝试更多命题的切题评价，并尝试由作文命题进行区分切题程度的问题生成，从而实现切题评价整体流程的自动化。

5. 参考文献

[1].何屹松，徐飞，刘惠等.新一代智能网上评卷系统的技术实现及在高考网评中的应用实例分析[J].《中国考试》.2019.1:57-65.http://www.ncpssd.org/Literature/articleinfo.aspx?id=NjEwMDI0NTU1MQ==&type=am91cm5hbEFydGljbGU=&datatype=am91cm5hbEFydGljbGU=&typename=5Lit5paH5pyf5YiK5paH56ug&nav=0&barcodenum=

[2].Ju-Lu, Yu，Bor-Chen Kuo，Kai-Chih Pa. Developing Chinese Automated Essay Scoring Model to Assess College Students' Essay Quality.[C]// P Xiangen Hu, Tiffany Barnes, Arnon Hershkovitz and Luc Paquette (eds.) Proceedings of the 10th International Conference on Educational Data Mining. Wuhan, Hubei, China. 2017:430-432. http://educationaldatamining.org/EDM2017/proc_files/papers/paper_151.pdf.

[3]北京理琪教育科技有限公司.IN 课堂[OL].(2020)[2022-08-20].https://www.znpigai.com/

[4]集智量文教育科技.测文网[OL].(2021)[2022-08-20].http://www.cewenwang.com/

[5]何屹松，孙媛媛，汪张龙，等.人工智能评测技术在大规模中英文作文阅卷中的应用探索[J].《中国考试》.2018.314:63-71. http://www.cgl.org.cn/auto/db/detail.aspx?db=950001&rid=14464793[5]

[6].ETS. TOEFL iBT® Test Scores[OL].(2020)[2022-08-20].https://www.ets.org/toefl/ibt/scores/

[7]付瑞吉，王栋，王士进，等.面向作文自动评分的优美句识别[N].《中文信息学报》.2018.32:88-97. http://jcip.cipsc.org.cn/CN/abstract/abstract2586.shtml

[8]符耀章，厉浩，钱建良，等.人工智能网上评卷技术的应用探索[J].《考试研究》.2021.01:93-105. https://www.cnki.com.cn/Article/CJFDTotal-KSYA202101011.htm

[9]蒋兴超，史钋镭，等.人工智能视域下母语写作评价的构想[J].《教学月刊·中学版》.2019.33:53-57. http://www.ncpssd.org/Literature/articleinfo.aspx?id=SlhZS1pYQjIwMTkwMzMwMTY=&type=am91cm5hbEFydGljbGU=&datatype=am91cm5hbEFydGljbGU=&typename=5Lit5paH5pyf5YiK5paH56ug&nav=0&barcodenum=

[10] Yu Sun, Shuohuan Wang, Shikun Feng 等. ERNIE 3.0: Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation[J].arXiv preprint arXiv:2107.02137.2021.https://arxiv.org/abs/2107.02137

[11]LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. Nature 521, 436–444 (2015). https://doi.org/10.1038/nature14539

[12]佚名.南京市中考作文评分标准[OL].(2020)[2022-08-20]. https://wenku.baidu.com/view/aec9281943323968011c926d.html

[13] Kai Sun, Dian Yu, Dong Yu, 等. Investigating Prior Knowledge for Challenging Chinese Machine Reading Comprehension[J].arXiv preprint arXiv:1904.09679.2019.https://arxiv.org/abs/1904.09679

6. 致谢

本项目得到了南京外国语学校的史钭镭老师与蒋兴超老师的指导，在此衷心感谢两位老师的悉心指导。

6.1 论文的选题来源、研究背景

选题来源于中学作文批改环节，这是我们身边最触手可及的问题。

在高中信息技术必修教材中，提到了人工智能相关的知识；其实，人工智能已经出现在我们的生活中，有像“微软小冰”这样可以与人对话的人工智能，也有像 Siri 这样能接听指令并帮助人进行打电话之类操作的。人工智能仿佛能够理解提出的问题，并做出回答。带着对这类“问答型”人工智能的好奇，我们了解到百度文心的一个开放模型：ERINE3 文本理解与创作。这个模型包含“自由问答”模块，能够回答各种关于语言文字的提问，包括对段落主旨的理解。

在作文课堂上，老师总是会强调切题的重要性。我们发现，对于切题与否，人的判断方法其实就是回答几个关于内容的问题：以“南京街头”为例，老师在批改时判断是否切题，其实就是在问：“写的是南京吗？”“写的是街头吗？”而回答问题是我们可以直接调用强大的 ERINE 来实现。我们便萌生了做一个辅助判断作文是否走题的项目的想法。通过了解相关研究背景，我们发现同样的课题虽然也有人在关注，但均未提出过用问题回答来判断是否走题的方法，而这恰恰是最贴近现实批改场景的方法，具有创新性和高效性。

6.2 队员的各自贡献

全部内容均由两人共同探讨、撰写、修改定稿。其中米子琪同学负责了主要的背景调研，邹桐同学负责主要的代码实现，实验设计和实施中两人均有贡献。

6.3 指导老师所起的作用

史钭镭老师既是我们的信息启蒙老师之一，也指导我们完成本次课题。他在

选题方面给我们很多指导，包括如何判断一个想法的可行性，并推荐给我们一些阅读的文献。在课题进行阶段，他也时常和我们探讨想法。

蒋兴超老师曾作为南京市中考作文阅卷组组长，具有丰富的作文批改经验。他向我们传授了许多作文批改中最真实的规则与流程，让我们得以设计出符合实际需要的实验方法。同时，他还提供了作为数据的作文样本。

感谢两位老师的悉心指导和无私奉献。

6.4 他人协助完成的研究成果

论文与实验均在老师指导下独立完成。

特别感谢百度文心开放提供的 ERINE 3 文本理解与创作模型，给我们的实验提供了有力的支持。