

文章编号: 1003-0077(2023)02-0001-14

## 写作智能评测研究综述和发展前景

薛嗣媛<sup>1</sup>, 周建设<sup>1</sup>, 任福继<sup>1,2</sup>

(1. 首都师范大学 中国语言智能研究中心, 北京 100089;

2. 电子科技大学 计算机科学与工程学院, 四川 成都 611731)

**摘要:** 随着计算机技术的发展, 写作智能评测研究有了更加丰富的技术手段和应用场景。该文对写作智能评测的相关研究进行了梳理, 首先对写作智能评测系统的发展历程进行了阶段性梳理; 其次介绍了写作智能评测研究的任务模式、常用数据库、评估方式; 再次梳理了写作智能评测的主要技术方法; 再次以面向汉语母语者、面向汉语非母语者两个不同维度展开介绍中文写作智能评测研究的现状及面临的挑战; 最后对未来写作智能评测研究发展进行展望。

**关键词:** 写作智能评测; 汉语作文测评; 语言智能技术; 中文信息处理

**中图分类号:** TP391

**文献标识码:** A

## A Systematic Survey of Automated Essay Scoring

XUE Siyuan<sup>1</sup>, ZHOU Jianshe<sup>1</sup>, REN Fuji<sup>1,2</sup>

(1. Research Center for Language Intelligence of China, Capital Normal University, Beijing 100089, China;

2. School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, Sichuan 611731, China)

**Abstract:** This paper summarizes the researches on automated essay scoring, including the development of automated essay scoring system. It also examines the tasks, public datasets and popular metrics in of automated essay scoring. The main techniques and models for automated essay scoring are reviewed, as well as the challenges in terms of both native Chinese speakers and non-native Chinese speakers. Finally, the prospects for future automated essay scoring is discussed.

**Keywords:** automated essay scoring; Chinese automated essay scoring; language intelligence technology; Chinese information processing

## 0 引言

随着计算机技术的发展, 以及新冠疫情期间在线学习、移动学习等方式的普及, 计算机辅助写作教育的教育意义和商业价值吸引了学术界和工业界的广泛关注。写作智能评测系统 (Automated Essay Scoring, AES) 是一种教育应用程序, 能够通过机器评分来减轻教师评估学生作文的压力。写作智能评测系统的基本思路是通过各种自然语言处理技术, 从待评测文章中提取文本特征, 让机器对文本特征与分数之间的关系进行学习并评分。这种方式可在

一定程度上减少评估操作效率低、主观评价有偏差等问题, 可以减轻教师评阅负担。

写作是以书面语言文字为媒介的文化交流行为, 需要高度综合的语言能力。为了能够在机器中有效评估这种语言能力, 写作智能评测研究包含了计算机科学、语言学、心理学、神经科学等多学科的融合, 是典型的交叉应用。写作智能评测研究源于国外, 从 20 世纪 60 年代开始发展, 可以将其划分四个阶段: ① 1966—1990 年是智能评测技术的起步阶段, 评价模型主要基于线性回归模型。Page 教授开发了第一个自动写作评分系统 (Project Essay Grade, PEG), 标志着智能评测研究的开始<sup>[1]</sup>。其

收稿日期: 2022-09-02 定稿日期: 2022-10-09

基金项目: 科技创新 2030 重大项目 (2020AAA0109700); 国家语委项目 (YB145-16, YB135-163); 中国博士后科学基金 (2022M722231); 国家社会科学基金 (22CYY036)

中,PEG 通过语法、词汇、结构等语言特征进行作文评估。②1990—2000 年是智能测评技术的快速发展期。20 世纪 90 年代的 AES 研究基本采用基于统计学习的方法。大量智能评分系统相继出现在各类大型语言测试中,推动了智能测评系统对教育领域的赋能,如 E-rater、Intelligence Essay Assessor、IntelliMetric 等。③2000—2013 年是智能测评技术的广泛应用期,AES 模型逐渐从单一的评分模式过渡到全方位的综合评价,典型系统如 PEG Writing、MyAccess!、Bookette、SAGrader、LightSIDE、ETS 等。④2013—2022 年是基于深度学习的智能测评技术的发展期。受到深度学习的积极影响,AES 模型逐渐采用深度学习算法等技术,挖掘文本深层次特征,以实现端对端的智能评估。

近年来,研究者从不同维度对 AES 研究进行整理分析形成多篇综述。Burrows 等人<sup>[2]</sup>从数据集、技术方法、模型构建、模型分级、模型评估和模型有效性等六个方面对 AES 系统进行了总结,但此综述仅涵盖了基础机器学习方法,并不全面。Ke 和 Mohamed<sup>[3]</sup>主要介绍了自动测评系统的现状,两篇文章通过对比研究的方式对 AES 系统进行详细描述和比较,但并未列出其面临的困境。Hussein 等人<sup>[4]</sup>研究了两类 AES 系统,介绍了基于规则特征和基于神经网络方法的论文,但没有详细介绍特征提取技术、AES 模型的性能等方面的问题。Klebanov 等人<sup>[5]</sup>梳理了 50 年来 AES 研究的发展历程,重点阐述了基于特征工程方法和深度学习方法的 AES 研究。预训

练语言模型对 NLP 相关领域的颠覆正在悄然发生,然而这篇文章并没有涉及这部分的内容。Borade 和 Netak<sup>[6]</sup>重点关注与特征工程相关的 AES 研究。Uto 等人<sup>[7]</sup>主要介绍了基于深度学习的写作智能评测模型。综上所述,以上都是由计算机领域的专家撰写的,介绍以英文文本为主的 AES 系统、模型方法、特征工程等。面向英文的 AES 研究起步早,研究成果丰富,但面向中文的自动写作评测研究仍处于起步阶段。吴恩慈和田俊华<sup>[8]</sup>对汉语写作自动评分研究中的问题进行了详细分析,但对数据构建、技术方法、模型构建等方面的探讨不够全面。丁革建等人<sup>[9]</sup>聚焦于英文文本的 AES 研究,但并未涉及中文写作智能评测方面的相关调研。

本文结构组织如下:第 1 节对写作智能评测的任务定义、数据集、验证方法进行梳理;第 2 节系统综述此领域的代表性研究,详细介绍写作智能评测任务中所使用的技术方法;第 3 节从中文信息处理的角度,重点关注面向中文的写作智能评测研究;第 4 节提出写作智能评测未来的研究方向;第 5 节对全文进行总结。

## 1 背景

### 1.1 写作智能测评任务定义

本文将依据文献[10]的划分标准,将写作智能评测任务划分为四种任务模式,如图 1 所示。

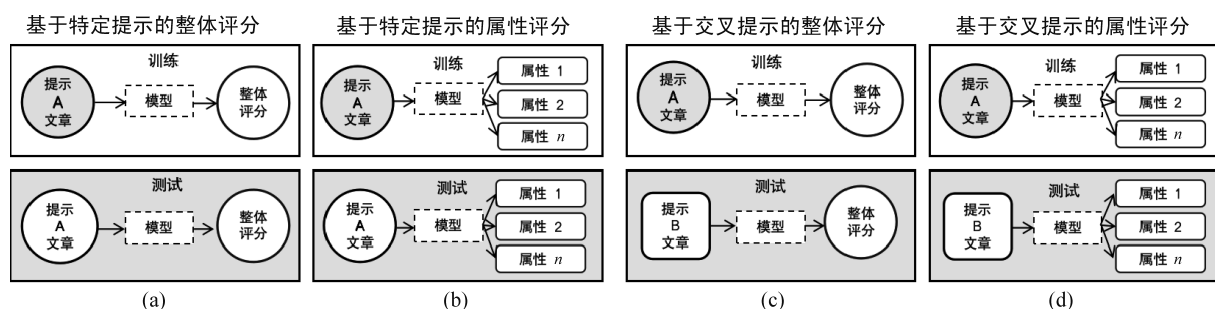


图 1 AES 任务的四种不同模式图示

#### 1.1.1 基于特定提示的整体评分

基于特定提示的整体评分是 AES 任务典型的解决方式。此处的提示(Prompt)可以理解为文章主题或者写作任务,通常由阅读材料和任务说明组成。目前,AES 评分模型大多是以特征值作为预测变量从拟合人工评分,其优势在于稳定的语言特征体系能更好地控制评价概念<sup>[11]</sup>,但特征采集过程容

易对结果造成影响。基于此,自动化提取语言特征的方法被提出,如浅层语义分析(Latent Semantic Analysis, LSA)以及基于深度学习的文本表征方法,它们都是通过构建端对端模型挖掘文本深层次语义特征以实现对其文本语义的深度理解。该任务的假设前提是,用于模型训练和测试的文章都来源于同一个提示主题,具体示例如图 1(a)所示。

### 1.1.2 基于特定提示的属性评分

在实际的教学场景中,整体评分提供的反馈信息非常有限,缺乏可解释性。在特定提示下的属性评分研究逐渐发展。属性评分可以辅助机器提供更详细的反馈,弥补整体评分的局限性。此任务可以对特定提示范围内的文章同时进行一个或者多个属性评分<sup>[4]</sup>,具体包括语法词汇、语篇组织、观点说服力、内容发展、衔接连贯、行文一致等。

Mathias 等人<sup>[12]</sup>首次提出属性评分的数据集,每个属性都被单独标注评分。此研究使用随机森林分类器(Random Forest Classifier)为各个属性分配分数以达到对具体属性的评估。在该任务设置中,用于模型训练的文章和用于测试的目标文章来源于同一个提示,具体示例如图 1(b)所示。

### 1.1.3 基于交叉提示的整体评分

在现实世界应用中,获得足量的标注文章成本较高。因此,如何在缺乏目标提示(Target-Prompt)训练数据的情况下进行整体评分预测成为热点问题。在该任务的设置中,用于模型训练的源域提示(Source-Prompt)文章和用于模型测试的目标提示(Target-Prompt)文章来源于不同的提示,具体示例如图 1(c)所示。

一些研究表明迁移学习方法能够有效处理低资源情况。迁移学习研究如何把源域的知识迁移到目标域上<sup>[13]</sup>。在标定数据缺乏的情况下,迁移学习可以利用相关领域有标定的数据完成未标注数据的标定。基于此,Jin 等人<sup>[14]</sup>构建两阶段训练方法 TDNN(Two-stage Deep Neural Network)实现交叉提示下的文章评分。第一阶段通过 RankSVM 生成伪数据,第二阶段借助伪数据作为训练集进行特征提炼,用于分数预测。在此方法的影响下,Li 等人<sup>[15]</sup>进一步扩展 TDNN 提出了 SEDNN(Shared and Enhanced

Deep Neural Network)模型,分为 Shared Model(S-model)和Enhanced Model(E-model)。S-model 是一个基于对抗网络原理的生成器,用于提取源域主题和目标域主题数据之间可用的共享信息,作为迁移特征。基于迁移特征生成目标主题文章的伪训练数据。E-model 利用伪训练数据提取文本特征实现分数预测。此外,Cao 等人<sup>[16]</sup>联合优化两个自监督任务对文本进行跨主题整体评分,分别是句子重新排序任务和噪声识别任务。

### 1.1.4 基于交叉提示的属性评分

基于以上研究,跨主题、交叉提示场景下的一个或多个属性评估逐渐被关注。Ridley 等人<sup>[11]</sup>首次提出了基于交叉提示的多维度评分任务,如图 1(d)所示。该任务要求模型在源域提示文章上进行训练获得预测模型,并迁移数据对目标提示文章进行多维度属性评估。在该任务设置中,用于训练的源域提示文章和用于测试的目标提示文章来源于不同提示。这项任务实际上综合了上述任务模式,对模型的概括和泛化能力要求极高。

## 1.2 数据集

写作自动评测大多是监督性任务,需要带有评分标记的作文数据库来训练模型。目前面向英文智能评测的常用数据集,如表 1 所示。国外的相关研究开始较早,至今已经形成了 CREE<sup>[17]</sup>、CS<sup>[18]</sup>、ICLE<sup>[19]</sup>、CLC-FCE<sup>[20]</sup>、BEETLE<sup>[21]</sup>、ASAP<sup>[22]</sup>、RTA<sup>[23]</sup>、SCI-ENTSBANK<sup>[24]</sup>、TOFEL11<sup>[25]</sup>、Power Grading<sup>[26]</sup>、AAE<sup>[27]</sup>、ASAP++<sup>[12]</sup>、FCE<sup>[20]</sup>等面向写作自动评测和短文本回答测试的通用型数据库。本文从数据集名称、数据来源对象、文章总数、数据库涵盖的主题数、所涉及的任务类型、评分范围等几个方面进行了梳理,如表 1 所示。

表 1 自动作文评测相关数据集汇总

| 数据集名称    | 数据来源对象       | 文章总数   | 主题数 | 任务类型    | 评分范围          |
|----------|--------------|--------|-----|---------|---------------|
| CREE     | 中级 ESL 学习者   | 556    | —   | 短文本理解评分 | [1-4]         |
| CLC-FCE  | ESL 学习者      | 1 244  | 10  | 整体评分    | [1-40]        |
| RTA      | 4~6 年级英语母语者  | 1 659  | 2   | 短文本回答评分 | [1-4]         |
| ASAP-AES | 7~10 年级英语母语者 | 17 450 | 8   | 整体评分    | [0-30];[0-60] |
| ASAP-SAS | 7~10 年级英语母语者 | 17 207 | 10  | 短文本回答评分 | [0-3]         |
| ASAP++   | 8 年级英语母语者    | 10 696 | 6   | 整体评分    | [0-30];[0-60] |

续表

| 数据集名称        | 数据来源对象        | 文章总数   | 主题数 | 任务类型     | 评分范围                        |
|--------------|---------------|--------|-----|----------|-----------------------------|
| ICLE         | ESL 学习者       | 1 003  | 12  | 文本逻辑结构评分 | [1-4]                       |
|              |               | 830    | 13  | 文本清晰度评分  |                             |
|              |               | 830    | 13  | 主题契合度评分  |                             |
|              |               | 1 000  | 10  | 论据说服力评分  |                             |
| TOFEL11      | ESL 学习者       | 1 100  | 8   | 语言流畅度    | 低、中、高                       |
| AAE          | 英语母语者         | 102    | 101 | 论据说服力分类  | [1-6]                       |
| BEETLE       | 电子与电气领域英语母语者  | 3 000  | 56  | 论据说服力分类  | 正确, 部分正确, 不完整, 矛盾, 不相关, 非领域 |
| SCIENSTBANK  | 科学领域英文母语者     | 10 000 | 197 |          |                             |
| PowerGrading | 美国入籍测试 ESL 人群 | 698    | 10  | 短文本回答评分  | 三个等级[-1, 0, -1]             |
| CS           | 计算机领域英文母语者    | 630    | —   | 等级评分     | [0-5]                       |
| FCE          | ESL 学习者信件随笔   | 1 238  | —   | 整体评分     | [1-40]                      |

### 1.3 验证方法

基于监督学习的整体评分 AES 模型较常使用以下几种常用指标进行验证:

(1) Quadratic Weighted Kappa(QWK)评估人类得分和系统得分之间的一致性,产生 0 至 1 的值。

(2) Mean Absolute Error(MAE)平均绝对误差是指人的评价分数与系统生成的分数之间的实际差异。

(3) Root Mean Squared Error(RMSE)均方根误差测量的是人的评价分数和系统生成的分数之间的平均平方差,反映了测量数据偏离真实值的程度,二者偏差越小,测量精度越高。

(4) Pearson's Correlation Coefficient (PCC)皮尔逊相关系数是指两个变量之间的相关系数。它将提供三个值,“0”代表人的评分和系统的评分没有关系。“1”说明人的评分和系统的评分关系是正向的,“-1”表示两者之间的关系是负向的。

## 2 自动作文评分技术方法分类

本节主要以特定提示下的整体评分为重点,简要梳理传统方法和深度学习方法两个部分的内容。

### 2.1 传统方法

#### 2.1.1 基于回归的训练方法

基于回归的训练方法指的是提取可计量语言特征构建多元线性回归模型预测文章质量。因此,基于回归方法需要挖掘对写作质量有影响力的文本特

征,以及探索各特征之间的函数关系等问题。基于回归方法的代表性工作有,Phand 等人<sup>[28]</sup>通过 EASE(Enhanced AI Scoring Engine)提取包括字符长度、词法错误、与主题的相关性等特征用于分数预测。

回归训练方法的优势在于,评测系统能够更好溯源,提供具体反馈意见。此种方法相较于深度学习等方法而言,数据量的需求小,在一定程度上降低了数据处理和标注的工作难度。但是,传统方法的局限性在于只关注了浅层文本语言符号,使得此类系统在处理复杂语义时非常脆弱。

#### 2.1.2 基于排序的训练方法

基于排序的训练方法将 AES 视为排序任务,即通过构建比较器对文本质量进行排序以实现评估。

Yannakoudakis 等人<sup>[20]</sup>提出支持向量机排序(Support Vector Machine for Ranking, SVMRank)方法为每一个数据点输出一个分数秩函数,以此构造数据的全局排序序列,提取单词、词性标记、N-gram、复杂语法特征等构建排序比较器。Chen 等人<sup>[29]</sup>构建排序比较器 LambdaMART (Lambda Multiple Additive Regression Tree)时采用了句法特征、语法特征、流畅性特征、内容提示等相关特征。Ronan 等人<sup>[30]</sup>通过转换排名算法来预测文章分数,提出计时聚合感知器(Timed Aggregate Perceptron, TAP)模型实现文章质量排序。

基于排序的方法利用对比的方式,而不是直接打分的方式,更符合人类实际判断时的认知行为。此种方法对于数据标注依赖较低,当无法明确给出

评价体系时,排序法可以减少数据标注阶段的难度。

### 2.1.3 基于分类的训练方法

基于分类的训练方法是指把 AES 视作分类任务,从不同等级的文本中挖掘一系列具有区别性的文本特征,并按照不同等级进行划分,以此构造文本等级评价模型。Persing 和 Ng<sup>[31]</sup> 使用支持向量机进行主题清晰度评估。Vajjala 等人<sup>[32]</sup> 使用支持向量机的变体 SMO,即序列最小优化算法(Sequential Minimal Optimization)在 TOEFL11 和 FCE 数据集进行训练,验证不同特征对分数预测的影响。Salim 等人<sup>[33]</sup> 运用集成算法框架 XGBoost 提取词汇数量、词性解析、句法树深度和连贯性等特征预测文章质量。

基于分类的训练方法受特征工程影响较大,优势在于有效特征纳入越多,其结果越好。其劣势在于模型对于标注数据的依赖度较大,训练数据的标注一般需要较多专家知识和经验,构建数据库的难度较大,时间周期较长。

综上,监督学习下的特征提取结果将极大影响模型预测的效果。经过文献梳理,表 2 从统计特征、主题特征、语法特征、语义特征、内容特征五个方面汇总 AES 研究中时常会选择的一系列特征。

表 2 AES 研究中的特征选择

| 特征类型 | 具体特征                                                          |
|------|---------------------------------------------------------------|
| 统计特征 | 文本长度、句子长度、平均句子长度、字符数、词汇数量、平均词汇长度等                             |
| 主题特征 | 词汇重叠、词汇话题性、提示重叠、主题相关度等                                        |
| 语法特征 | 词性标记,一元,二元, $N$ 元语言模型句法树特征,语法错误率,短语结构等                        |
| 语义特征 | 直方图(Histogram)特征、框架特征(FrameNet)、Glove 词向量、SSWE 词向量、神经网络编码词向量等 |
| 内容特征 | 实体网络、修辞结构树、连贯衔接、词汇链、语篇功能、论辩结构等                                |

## 2.2 深度学习方法

深度学习方法在各领域的成功逐渐拓展到了 AES 研究领域,构建出基于深度学习的 AES 模型(Deep Neural Network-Automated Essay Scoring, DNN-AES)。基于深度学习训练方法的 DNN-AES 主要包括:①基于循环神经网络的方法;②基于连贯性的深度学习建模方法;③基于预训练语言模型

的方法;④基于多模态深度学习的方法。

### 2.2.1 基于循环神经网络的方法

Taghipour 等人<sup>[34]</sup> 提出基于循环神经网络的 DNN-AES 模型进行文章评测。首先由独热(one-hot)表示的查找层(Look-Up layer)、卷积层(CNN layer)、循环层(RNN layer)、池化层(Pooling Layer)构成的编码器表征文章向量,然后经过 Sigmoid 激活的全连接层输出文章分数,其结构如图 2 所示。此模型开启了基于深度学习训练方法的 AES 研究。

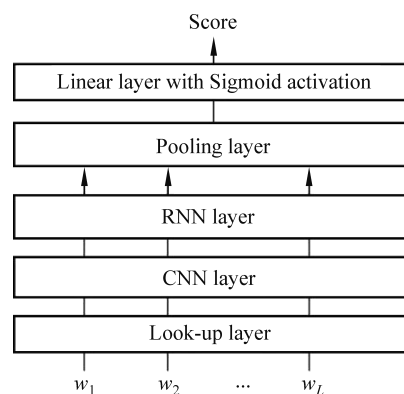


图 2 基于循环神经网络的 DNN-AES 模型

Alikaniotis 等人<sup>[35]</sup> 设计了与 Taghipour 和 Ng 相似的预测模型,该模型包含查找层、循环层、池化层。其特殊之处在于受 Collobert & Weston (C&W) 词语境嵌入的启发,这个模型创造性地使用特定分数的单词嵌入(Score-Specific Word Embeddings, SSWEs)表征文本语义,通过学习特定单词对文章分数的贡献程度来形成单词表征,使得文本表示向量能够有效融合写作质量信息,提升模型评分准确性。

Dong 等人<sup>[36]</sup> 考虑从文本中捕捉语篇结构特征,在模型中构建语义和结构的双重表征。此研究通过嵌套结构的卷积神经网络模型捕捉文本层次化结构特征,有效提升了 DNN-AES 模型的预测水平。Dong 等人<sup>[37]</sup> 在上述模型的基础上进一步融合注意力机制(Attention Mechanism),捕捉对评分起到关键作用的重要信息,提升了 DNN-AES 模型的预测准确率。

### 2.2.2 基于连贯性的深度学习建模方法

连贯性是评价文章质量的重要标准,尤其在文本长度较长的情况下。自然语言处理任务中普遍认为文章内各个句子之间的语义相似性与文本连贯性成正比,有部分研究通过衡量语块的语义相似性来表征文本连贯度以辅助 AES 模型优化。

Tay 等人<sup>[38]</sup> 探究文本连贯性与写作质量之间

的关系,发现文本连贯性语言特征影响写作质量。基于此提出融合语篇的连贯性语言特征的 SKIPFLOW 模型,具体如图 3 所示。在 SKIPFLOW 模型中,首先利用长短时记忆层(LSTM)表征文本语义特征,其次计算相邻时间步之间的相似性以获得其连贯性,然后将所有的相似度得分和文本特征拼接起来表征文章,最后映射到线性层进行分数预测。受以上研究启发,Liang 等人<sup>[39]</sup>提出孪生双向长短时

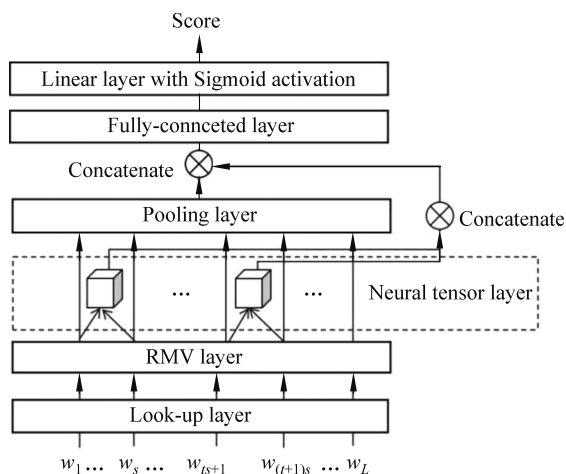


图 3 SKIPFLOW 模型结构图

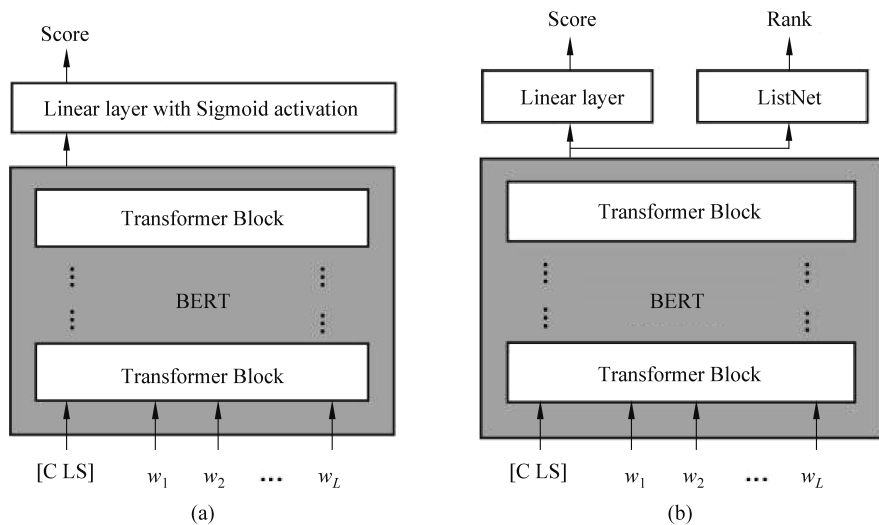


图 4 基于 BERT 预训练语言模型的 AES 模型

#### 2.2.4 基于多模态深度学习的方法

基于预训练模型的双模态智能写作评分模型如图 5 所示。现有评分模型只接受文本这一模态作为输入,没有考虑到光学字符识别引入的误差。为了能够解决这一问题,Gao 等人<sup>[43]</sup>提出双模态 AES 评分系统,引入文本和图片的双模态数据。首先,模型的输入为原始文本图片,图像特征由 FastRCNN

记忆网络结构 (Siamese Bidirectional Long Short-Term Memory Architecture) 进一步拓展 SKIPFLOW 原始模型。此模型不仅可以捕捉文章中的语义特征,还可以捕捉文章背后的评分标准信息。实验表明,此方法能更好地表征冗长而复杂的文章内容,提升预测效果。

#### 2.2.3 基于预训练语言模型的方法

由于预训练语言模型在自然语言处理领域中的广泛使用,多位研究者将预训练语言模型应用于 AES 任务<sup>[40]</sup>。

Rodriguez 等人<sup>[41]</sup>首次使用 BERT 预训练模型进行整体分数预测,模型结构如图 4(a)所示。此研究沿袭“预训练+微调”的方式获得文本表征向量,然后将其输入由 Sigmoid 函数激活的全连接层中输出评分。然而,由于文章长度接近 BERT 模型输入的长度限制,很难直接对输入的文章进行处理。为解决上述问题,Yang 等人<sup>[42]</sup>提出了多损失的微调 BERT 模型,其结构如图 4(b)所示。该模型通过最小化 MSE 损失函数和排序错误损失函数进行参数调整,同时优化文章打分任务和文章质量排序任务,最终在 ASAP 语料库上达到最优结果。

抽取得到图像局部特征,文本由 OCR 技术提出文中字符内容,再由 Transformer 表征文本语义特征。之后,将图像特征和文本特征进行拼接得到双模态特征。然后,将双模态特征输入由 Visual BERT、Visual RoBERTa、Visual ALBERT 相融合的集成层进行深度表征,最终将表征输入全连接层进行分数预测。

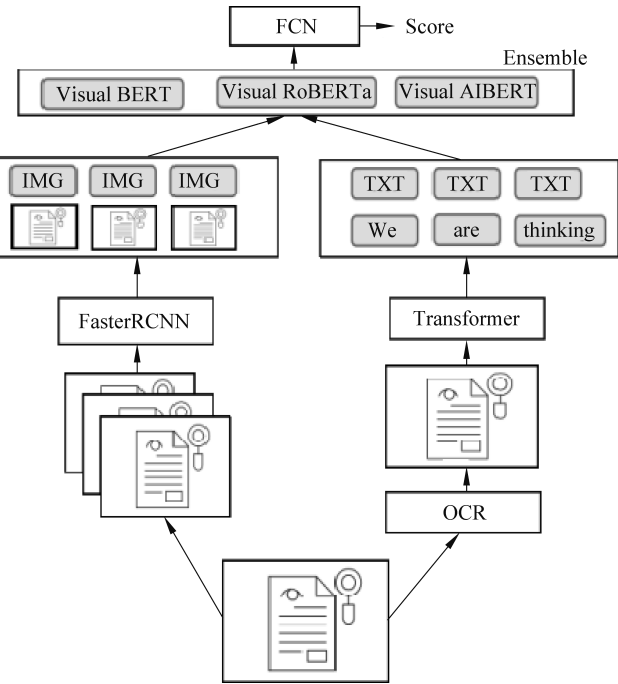


图 5 基于预训练模型的双模态智能写作评分模型

2.3 研究结果综合对比

表 3 综合比较了不同模型在 ASAP 数据集上的实验结果,以观测近十年此任务技术的发展程度和趋势。实验结果对比表明,基于深度学习算法的 AES 模型获得比传统学习方法更优异的预测结果,逐步成为 AES 研究方法的主流。面向交叉提示场景、多模态场景、小样本学习等的 AES 研究也逐步成为关注点,这与 NLP 本身的发展趋势相协同。

表 3 不同模型在 ASAP 数据集上的实验结果

| 任务          | 方法                              | 特征   | 平均    |
|-------------|---------------------------------|------|-------|
| 基于特定提示的整体评分 | EASE(SVR)                       | 统计特征 | 0.699 |
|             | EASE(BLRR)                      | 统计特征 | 0.705 |
|             | LSA                             | 统计特征 | 0.532 |
|             | CNN                             | 语义特征 | 0.734 |
|             | CNN+LSTM                        | 语义特征 | 0.761 |
|             | Regression                      | 统计特征 | 0.690 |
|             | Hierarchical LSTM-CNN-Attention | 语义特征 | 0.764 |
|             | LSTM+Memory Network             | 统计特征 | 0.780 |
|             | SKIPFLOW LSTM(Tensor)           | 语义特征 | 0.764 |
|             | Dilated LSTM With RL            | 语义特征 | 0.724 |
|             | CNN+Bi-LSTM                     | 内容特征 | 0.786 |

续表

| 任务          | 方法                         | 特征           | 平均           |
|-------------|----------------------------|--------------|--------------|
| 基于特定提示的整体评分 | HISK+BOSWE and v-SVR       | 语义特征         | 0.785        |
|             | BERT, XLNet                | 语义特征         | 0.760        |
|             | XGBoost                    | 统计特征<br>内容特征 | 0.681        |
|             | Multiple Linear Regression | 内容特征         | 0.770        |
|             | IRT+CNN-LSTM/ IRT+BERT     | 内容特征<br>语义特征 | 0.749        |
|             | R <sup>2</sup> BERT        | 语义特征         | <b>0.794</b> |
| 基于交叉提示的整体评分 | RANKSVM                    | 统计特征         | 0.546        |
|             | TDNN                       | 内容特征<br>语义特征 | 0.686        |
|             | SEDNN                      | 内容特征<br>语义特征 | 0.727        |
|             | CNN-LSTM-Attention         | 内容特征<br>语义特征 | 0.505        |
|             | BERT+SST+DAT               | 语义特征         | <b>0.791</b> |

3 面向中文的写作智能评测研究

从系统开发来看,国内最早涉足英语作文自动评分的专家系统是 Iwriter<sup>[44]</sup>,是包含语言、内容、篇章结构及技术规范四个维度的机器智能评阅系统,实现机评和人评深度结合,以机评促反馈。句酷批改网<sup>[45]</sup>是面向英语学习者的自动写作评分系统,其核心方法是通过算法映射作文文本和标准语料库之间的距离,再将距离转化成分数和评语。

国外的评估范式过渡到中文领域,并不能完全复用。原因在于中文属于汉藏语系,在语音、词汇和语法等方面与属于印欧语系的英语有着较大的差别。汉语本身的特点,如汉语词汇分词歧义、语义灵活、语法灵活等问题<sup>[8]</sup>,导致其无法为机器提供明确的计算规则。所以,面向中文的 AES 研究需要遵循中文的语言规则。科大讯飞研发的 IFlyEA<sup>[46]</sup>写作智能评测系统,该系统采用多种机器学习以及深度学习算法实现中文语言运用评估、语言表达评估、篇章异常检测、篇章质量评估等,实现多层次、多维度的语言分析功能。国家语委语言智能中心研发的 IN 课堂语文作文智能批改<sup>[47]</sup>,以人机交互为主,兼具评分和反馈功能,能够从内容、表达、发展三方面提供 13 个评估点,实现对汉语母语者的语言水平评估。

从研究来看,本文以中文母语者以及二语学习者两类评测为对象,展开叙述 AES 研究在中文领域中的进展。

### 3.1 面向中文母语者的写作智能评测研究

面向中文的智能评测研究已经从整体评分研究<sup>[48-49]</sup>逐渐发展到修辞技巧<sup>[50]</sup>、表现手法<sup>[51-52]</sup>、表达方式<sup>[53-54]</sup>、符合题意<sup>[55]</sup>、衔接连贯<sup>[56-57]</sup>等具体维度的

属性评分,与现有任务相配套的中文写作智能评测数据库,具体如表 4 所示。

#### 3.1.1 整体评分

受到国外评测方法的影响,中文智能评测研究的主流方法同样是筛选浅层文本符号特征并设计多元回归模型。这类研究的弊端在于只能考虑有限的文本特征,无法反映出文章的结构、逻辑、内容、立意等更深层次的信息和语言水平。

表 4 汉语自动作文评测研究中所使用的数据库

| 数据库      | 对象               | 文章数量/篇         | 测评对象    | 评分范围                              |
|----------|------------------|----------------|---------|-----------------------------------|
| 整体评分语料库  | 汉语水平考试(HSK)      | 7 174          | 整体评估    | 40~95                             |
| 整体评分语料库  | 7~12 年级作文        | 93 002         | 整体评估    | 1~4                               |
| 篇章结构语料库  | 高中作文             | 1 220          | 结构评分    | 分类(好、中、差)                         |
| 表现手法语料库  | 中小学记叙文           | 5 000          | 表现手法    | 识别(叙述、论述、描写、论证和情感)                |
|          | 小学作文记叙文          | 20 000         | 表现手法    | 识别(动作描写、直接抒情、外貌描写、心理描写、神态描写、语言描写) |
| 修辞语料库    | 学生作文             | 11 337         | 比喻修辞    | 比喻句识别(二分类)                        |
|          | 小学作文<br>新闻《人民日报》 | 1 022<br>4 000 | 排比修辞    | 排比句识别(二分类)                        |
| 表达质量数据库  | 中学生考试作文          | 21 053         | 语言表达    | 分类(好、坏)                           |
| 行文一致性语料库 | 中小学写人作文          | 401            | 文本主题一致性 | 6~9                               |
| 句间逻辑性语料库 | 中小学人物类作文         | 16 329         | 句间逻辑合理性 | 逻辑合理判别(二分类)                       |

为了更好地提炼文本语义信息,部分研究者尝试了其他模型构建方法。曹亦薇等人<sup>[49]</sup>使用潜语义分析技术(Latent Semantic Analysis, LSA)表征文章内容,并利用多元线性回归模型评估最终得分。Hao 等人<sup>[58]</sup>进一步提出正则化潜在语义分析算法(Regularized Latent Semantic Analysis, RLSA)来解决 LSA 矩阵分解时计算量大的问题,在保证分数预测准确性的同时也降低了计算复杂度。Xu 等人<sup>[59]</sup>提出语境模型(Contextualized Latent Semantic Indexing, CLSI)通过结合潜在语义信息和上下文语境信息来增强文本表征,实现文章内容和语言能力的评估。王耀华等人<sup>[60]</sup>提出了基于中心的文本语义离散度的表示方法,进一步优化了评分模型。钟启东等人<sup>[61]</sup>从语言感知为出发点,构建基于 AC 自动机的汉语作文评分算法。此方法通过文本分词和主干提取等技术实现了对句子流畅度的评价,并将评测作文和标准作文的框架进行相似性对比,建立了对作者语用能力的评估,增强了作文自动评分的可解释性和准确性。

#### 3.1.2 属性预测

##### (1) 离题检测研究

中心主旨在评分体系中具有重要比重。抽取文章主题特征并评估其是否偏题或者离题,能够有效提升智能评测系统的准确性。离题检测的核心问题是文本主题的提取和评估,主流解决方法是挖掘文本关键词以示主题。文本关键词的主流提取方法是潜在狄利克雷分配模型(Latent Dirichlet Allocation, LDA)。刘明杨等人<sup>[55]</sup>采用无监督的 LDA 主题模型挖掘文章的主题关键词表征主题。巩捷甫<sup>[50]</sup>通过 LDA 模型挖掘文本关键词并与扩展的主题词进行相似度比对,实现了作文的离题检测。蔡黎等人<sup>[62]</sup>提出三重分段模型,用 TextRank 结合词频的方法计算每篇文章的主题特征值,以此判断文章是否偏离主题。

##### (2) 修辞手法识别

修辞格是通过各种语言手段来修饰调整语句,从而优化表达效果的一种语言活动。常见的修辞手段有排比、拟人、比喻三种。在 AES 研究中考虑修辞语言特征能够在一定程度上优化评分模型的

性能。

比喻修辞是指以一种偏离字面意思的方式传达文章的复杂含义。比喻修辞识别可以考察学生使用比喻修辞的能力,为评估提供线索。明喻句使用明确的触发词,如“像、似、如”,识别较为容易,但是一个带有比较词的句子并不总是能引发比喻<sup>[63]</sup>。因此,Liu 等人<sup>[64]</sup>利用本体和喻体之间形成的跨领域映射挖掘比喻手法,利用多任务联合学习方法进行汉语明喻的识别研究,即明喻句子分类、明喻成分提取和语言建模。

拟人修辞是比喻性语言的另一种特殊情况,它借用人类的行为、表情或其他特征来赋予非人类对象的特定属性。IFlyEA 评分系统结合外部知识实现拟人手法识别。评分系统引入了一个外部知识同义词词林<sup>[65]</sup>,将单词根据词义聚类,并为每个词义簇分配一个可学习的嵌入向量。因此,每个字由其词嵌入向量和聚类嵌入向量联合进行表示。然后将向量输入编码器进行表征,再输入分类器进行识别。

排比句具有结构紧凑、句式整齐、富有表现力等鲜明的特点,但目前排比句自动识别方法的识别效率欠佳。Song 等人<sup>[66]</sup>通过结合广义词对齐策略和词序列之间的对齐度量方法,识别学生论文中的排比句。穆婉清等人<sup>[67]</sup>依据排比句结构相似、内容相关的特点,设计了融合卷积神经网络和结构相似度计算的排比句识别方法。朱晓亮等人<sup>[68]</sup>通过研究中文排比句的语法特征,设计融合预处理算法与 BERT 模型的排比句自动识别方法,在小学级别排比句数据库上达到 0.96 的  $F_1$  值。

### (3) 语言表现评估

文采是在语言通顺基础上,对创作者提出的更高层次的语言标准,是写作评估的重要考虑对象。冯胜利等人<sup>[69]</sup>以书面语体的“韵律语法”为基础,提出了“表达庄雅度”的自动测量方法。付瑞吉等人<sup>[53]</sup>提出中学生作文优美句识别任务,其中利用卷积神经网络和双向长短时记忆网络的混合神经网络模型对语句进行表征,最终实现表达是否优美的判断。

### (4) 表达方式识别

语篇模式,即表达方式,描述了以语言为基础的交流目的和惯例。研究者通过对篇章结构等语言特征进行统计分析,发现语篇模式对写作质量评估起到重要作用。基于此,Wei 等人<sup>[62]</sup>提出利用神经序列标注方法对记叙文中的叙述、论证、描写、论证和

情感表现五种句子级别的表达方式自动识别。实验结果验证了语篇模式识别可以改善作文评分效果。马晓丽等人<sup>[52]</sup>为提升表现手法分类性能,以中小学中文作文为对象,引入方差的  $TF \times IWF \times IWF$  算法和 Word2Vec 模型优化文本表征。结果显示此模型可以提炼重要语义信息,有效表征文本中特征词汇分布,提升表现手法识别的准确率。

### (5) 篇章结构评估

篇章结构的合理性是评判文章质量的一个隐性、深层因素,具有很大挑战。Song 等人<sup>[70]</sup>提出基于视觉布局的网格格式神经网络模型表征篇章结构,并进行篇章要素识别和篇章结构质量评估进行联合建模。此文将议论文结构划分为七类:引论、中心论点、分论点、事实论据、理论论据、结论和其他。结果表明篇章要素识别可以在多个方面辅助作文自动评分模型,例如把作文结构建模、观点识别等作为评分系统的特征。

### (6) 逻辑关系研究

篇章角度的逻辑分析与评测,如果只关注蕴含、矛盾、不相关三种逻辑关系,不足以对篇章整体的逻辑性进行精准评测,还需要结合逻辑学、语文学等学科对词、句、段的逻辑关系做进一步的范畴化分类。刘杰等人<sup>[57]</sup>利用两个任务对文本逻辑合理性进行判别,首先,构建位置判别模型来判断段落内部每句话是否在相应位置,其次,构建段落整体逻辑合理性判别模型来判断段落内句子的顺序是否合理。实验结果表明此种方法对段落句间逻辑合理性的判别是有效的。

刘杰等人<sup>[71]</sup>提出了一种通用的作文行文一致性评测模型。该模型借鉴孪生神经网络的思想,提取作文中核心人物性格特征、形象特征以及故事情节特征进行相似度度量,从而获取文本中心思想以及行文一致性的匹配分数。此方法有效解决了文本语义块的隐含语义特征捕捉问题及其一致性问题。

综上,现有的 AES 研究点都具有针对性,但较为分散。此类研究需要搜集相应的数据,依据专家经验对特定维度进行标注,才能实现模型设计和评估预测。

## 3.2 面向非中文母语的写作智能评测研究

随着汉语作为第二语言的日益普及,面向汉语学习者的中文写作智能评测已成为一种重要的任务。面向二语学习者的写作智能评测研究主要集中于语言特征的有效性量化研究。

语言特征和写作质量相关性的定量研究方法已具有一定积累<sup>[72-73]</sup>,表5分别从汉字、词汇、句法、篇章内容、任务层面展示现有研究涉及的语言特征。

表5 面向汉语二语学习者的可读性特征体系

| 一级特征 | 二级特征    | 特征举例                                                  |
|------|---------|-------------------------------------------------------|
| 汉字层面 | 汉字熟悉度   | 汉字水平等级分布                                              |
|      | 字形复杂度   | 字形笔画复杂度                                               |
|      | 汉字正确性   | 字形错误、拼音代替汉字                                           |
| 词汇层面 | 词汇多样性   | 词种数、词型数、作文中相同的词 and 所有词的比例等                           |
|      | 词汇复杂性   | 词频、复杂名词短语比率、词汇等级数量占比等                                 |
|      | 词汇正确性   | 词形错误、词序错误、搭配错误、词语误用、词语多余、缺少词语等                        |
|      | 短语搭配复杂性 | 动宾、主谓、形名、状中、量词、框式介词、介动、述补、关联词搭配数量等                    |
|      | 词性复杂度   | 名词比例、动词比例、形容词比例、副词比例等                                 |
| 句法层面 | 句子复杂性   | 句子总数、T 单位长度、平均 T 单位分句数量等                              |
|      | 短语句法复杂度 | 句法树高度、短语句法结构复杂度等                                      |
|      | 语法正确性   | 语法错误小句与小句总数占比等                                        |
|      | 依存句法复杂度 | 平均句子依存距离、最大句子依存距离等                                    |
| 篇章层面 | 篇章正确性   | 标点错误、标点空缺和标点多余等                                       |
|      | 篇章长度    | 段落数量、句号标点句、逗号标点句、平均段落长度、最长段落数量等                       |
|      | 语篇衔接连贯  | 代词的使用、句间以及段落间的词语重复、介词的使用、连词使用等                        |
|      | 话题链数量   | 话题链数量、话题链分句综述、零形数量等                                   |
| 内容层面 | 内容正确性   | 语序错误、语义错误等                                            |
|      | 内容特征    | 讨论句、补充句、论点句数量统计等                                      |
| 任务层面 | 任务难度    | 是否此时此地、任务元素的多寡、是否需要时空推理、是否需要因果推理、是否需要目的推理,以及是否需要态度选择等 |

随着技术发展,一些研究者将利用机器学习方

法和深度学习方法展开面向汉语学习评分研究。Wang 等人<sup>[74]</sup>以汉语水平考试(HSK)作文为评测对象,提取语言复杂性和语言正确性等 90 个语言特征,利用线性回归模型、逻辑回归模型、随机森林算法、深度学习算法等多种模型进行分数评估。李琳等人<sup>[75]</sup>以汉语水平考试(HSK)作文为评测对象,使用“预训练+微调”的模式进行汉语作文自动评分,结果与人工评分结果具有较强一致性,QWK 可达到 0.726。

### 3.3 中文写作智能测评研究现阶段的挑战

近年来中文信息处理技术的发展为中文写作智能测评研究提供了更多突破的方向,但始终还存在严重不足:

① 面向中文母语者语言产出的指标研究不足 以统计方法为基础的二语测试研究成果较为丰富,但是其语言特征不能直接适配中文母语者的写作质量评测研究。其原因在于,二语学习者一般具有较高的认知水平,仅是缺乏语言表达的准确性,所以面向二语者的评测主要围绕表5中的统计特征对浅层文本符号进行计算。但是面向母语者的写作智能评测将会更倾向于关注其文本内容、立意等更高层次的语言能力,所以未来可以进一步研究和验证影响中文母语者写作质量的可读性指标,逐渐完善作文外部特征对内部质量的表征和验证。

② 面向中文母语者写作智能评测的开源语料库建设不足 深度学习模型的训练离不开高质量语料的支持。领域数据不足、加工深度不够、标注规范的准确度、标准是否开源等诸多问题始终未解决。相较于国外 AES 相关数据库,国内由于缺乏公开训练和测试数据集,研究者只能根据自己构建的规模较小的语料库进行研究。部分语料库仅标注了整体分数,缺乏详尽的规则,后人难以扩充迁移,影响中文的 AES 任务形成规模。此外,在语料标注阶段,其评价体系普遍都较为粗糙,亦或是借鉴国外标准,经常导致关键信息缺失、可解释性不强等问题。因此,未来亟需构建多维度的中文作文标准评测数据库并提供相关的基线实验用作参照。

## 4 写作智能评测研究的展望

从上面介绍的 AES 研究来看,未来可以向以下几个方向继续拓展:

#### 4.1 交叉提示任务研究

目前的 AES 研究都基于已标注数据进行训练,但是在真实的应用场景中,很难对每一个新主题都标注充足的训练语料。因此,基于交叉提示的 AES 研究逐渐受到关注,此任务旨在利用源域提示的训练语料对目标提示的测试语料进行分数预测,缓解数据稀缺问题。Song 等人<sup>[76]</sup>提出了分阶段训练方法,论文评分者首先在一个涵盖不同主题的大型论文数据集上进行预训练,对文章进行粗粒度的二分类评分,这可以被视为是一种弱监督。其次,在交叉提示的语料上对评分器进一步微调。最后,利用训练好的评分器对目标提示的测试语料进行分数预测。由此可见,深度学习多阶段微调等方式的目的都是为了充分挖掘目标提示和非目标提示之间可共享、可迁移的信息,提升跨主题模型的适应能力。因此,如何建模挖掘两者之间可迁移的关联信息是 AES 模型能否更好地适用于现实环境的核心问题。

#### 4.2 可解释模型研究

基于深度学习的神经网络模型或者预训练语言模型在 AES 研究中取得了优异的表现,由于神经网络模型无法实现对具体细节的分析,此类模型都被视为“黑盒操作”。此情况下,有研究提出机评与人评的结合可以有效解决神经网络的不可解释性。例如,Uto 等人<sup>[77]</sup>提出融合人工特征和嵌入特征的方式表征文本,以期通过控制特征参数变量透视模型的评估过程。这类混合模式训练方法迈出了 DNN-AES 可解释性研究的第一步,但未能显性揭示模型分析和决策过程。因此,如何解释 DNN-AES 模型决策过程,实现可解释性自动写作评估是一个值得被探索的研究方向。

#### 4.3 智能化反馈研究

反馈是评测系统的必要组件。评测系统能够生成多级评价,有效辅助学生进行写作内容的修改,帮助学生提升写作能力。目前有些评测系统可以根据一系列预定义模板生成相应的评论。评分模块从各个角度对文章评分,进而整合总体得分,基于总分的系统管理预定义模块并进行模板选择,随后生成连贯有效的评语来解释文章质量。然而,基于规则的方式不仅需要专家知识,还需要花费大量人力标注,知识迁移成本高、难度大。现有一些研究探索智能评测的自动化反馈技术,如 Miao 等人<sup>[78]</sup>扩展无监督的吉布斯采样方法,提出 CGMH (Constrained

Generation by Metropolis Hastings sampling)方法实现智能化反馈。因为 CGMH 允许对句子空间中的单词标记进行更灵活的操作,更容易生成不同长度的内容。整体而言,此类研究较少,如何在保证评分准确性的前提下实现智能化反馈需要进一步探索。

#### 4.4 模型评分的信度和效度

在机器评测语言的实际过程中,评分的信度和效度以及跨领域群体的公平性是比 AES 系统开发更重要的话题。Attali 等人<sup>[79]</sup>提出相对应的评价框架为 AES 模型的信度与效度研究打开思路。此框架包含三个维度,第一,特征的检验代表特征的可解释性。当一个特征被引入 AES 模型时需要清楚地反映出其测量意义,使得计算语言特征能够符合人类的基本假设。第二,AES 模型的准确性依赖于人工评分的质量,评分员间的信度大约在 0.7 以上才能够实现评分的基本“有信”。第三,跨样本稳定性的研究能够减轻评分工具产生的偏见。AES 在不同群体上进行评估具有一定差异,容易造成“机器偏见”。因此,AES 模型研究需要使用标准化“去偏”方法衡量不同群体之间的差异,筛选不公平案例,最大限度地保证数据和模型的公平性。Uto 等人<sup>[80]</sup>提出了一个整合了测试理论的写作自动评价框架。此框架利用人工评分对模型进行调控,为模型训练提供一个稳定且量化的参考分数,以保证模型评分的信度和效度。智能写作评测的目的是利用多学科技术有效地模拟人工评分,以达到快速评定作文质量的目的。于理想化的评分系统而言,仅仅以系统评分与人工评分的无限接近为目标是远不够的,需要更严谨地论证系统评分结果与文本之间的潜在关系,加强对模型评分信度、效度的关注。

#### 4.5 语言评测系统的构建

从以上研究梳理来看,目前的大多数 AES 研究均只能考虑单一评测维度进行针对性突破。实际上,写作智能评测场景十分复杂,在评估时往往需要同时考虑语言表达、语法错误、篇章结构、主题内容等不同评分属性。成熟的评测系统需要集成不同的单点模块,实现多维度的文本质量综合评价。在复杂场景下,综合性评测系统构建也会面临一些问题,例如,多个评分维度之间潜在复杂结构,多个维度在评分时不同的贡献度比重等等。因此,如何有效集成不同的研究结果,构建完整的写作评测系统,实现文本质量的综合评分,适应真实的教学场景,需要进

一步研究和探索。

## 5 结论

本文完成了写作智能评测研究的系统化调研。首先,从写作智能评测的任务模式、数据构建、技术方法等几个方面进行了归纳总结。然后,从中文信息处理的角度,分别从面向母语者和面向二语学习者两类人群梳理中文写作智能评测研究,并提出现阶段中文写作智能评测的挑战和局限。最后,本文综合分析写作智能评测研究现阶段的情况以及未来可发展的研究方向,以期对相关研究者提供一定参考。

写作评估的最终目的是通过语言使用情况判断个人的人文素养,即语言表达中的“人文基因”发现<sup>[80]</sup>。对比此目的,现阶段写作智能评测的研究还具有很大发展空间。伴随着大数据时代的到来,现阶段主流做法是依赖数据驱动方式的初级阶段,这类方式没有充分考察到语言运用手段背后的认知原理。文文章的逻辑结构、立意论证等方面,还未能实现有效评价。因此,如何持续提高机器的审美水平和预测能力,实现机器从文章浅层符号到人文素养的推断,不仅是 AES 领域同时也是人工智能领域的一个开放性问题。

## 参考文献

- [1] PAGE E B. The imminence of grading essays by computer[J]. *Phi Delta Kappan*, 1966, 47(5): 238-243.
- [2] BURROWS S, GUREVYCH I, STEIN B. The eras and trends of automatic short answer grading[J]. *International Journal of Artificial Intelligence in Education*, 2015, 25: 60-117.
- [3] KE Z, NG V. Automated essay scoring: A survey of the state of the art[C]//*Proceedings of International Joint Conference on Artificial Intelligence*, 2019: 6300-6308.
- [4] HUSSEIN A, HASSAN H, NASSEF M. Automated language essay scoring systems: A literature review[J]. *Computer Science*, 2019, 5: e208.
- [5] KLEBANOV B, MADNANI N. Automated evaluation of writing: 50 years and counting[C]//*Proceedings of the 58th Annual Meeting of Computational Linguistics*, 2020: 7796-7810.
- [6] BORADE G, NETAK D. Automated grading of essays: A review[C]//*Proceedings of the International Conference on Intelligent Human Computer Interaction*, 2020: 238-249.
- [7] UTO M. A review of deep-neural automated essay scoring models[J]. *Behavior Metrika*, 2021, 48(2): 459-484.
- [8] 吴恩慈,田俊华.汉语作文自动评价及其关键技术:来自作文自动评价(AEE)的经验[J].*教育测量与评价*, 2019(08): 45-54.
- [9] 丁革建,刘畅.作文自动评分技术综述[J].*计算机应用*, 2022, 42(S1): 386-390.
- [10] RIDLEY R, HE L, DAI X, et al. Automated cross-prompt scoring of essay traits[C]//*Proceedings of the AAAI*, 2021, 35(15): 13745-13753.
- [11] BENNETT D, PARRY G. The accuracy of reformulation in cognitive analytic therapy: A validation study[J]. *Psychotherapy Research*, 1998, 8(1): 84-103.
- [12] MATHIAS S, BHATTACHARYYA P. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores [C]//*Proceedings of LREC*, 2018: 1169-1176.
- [13] TORREY L, SHAVLIK J. Transfer learning[M]//*Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. IGI global, 2010: 242-264.
- [14] JIN C, HE B, HUI K, et al. TDNN: A two-stage deep neural network for prompt-independent automated essay scoring [C]//*Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2018: 1088-1097.
- [15] LI X, CHEN M, NIE J Y. SEDNN: Shared and enhanced deep neural network model for cross-prompt automated essay scoring[J]. *Knowledge Based System*, 2020: 210-219.
- [16] CAO Y, JIN H, WAN X, et al. Domain-adaptive neural automated essay scoring[C]//*Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020: 101-120.
- [17] BAILEY S, MEURERS D. Diagnosing meaning errors in short answers to reading comprehension questions. [C]//*Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications*, 2008: 107-115.
- [18] MOHLER M, MIHALCEA R. Text-to-text semantic similarity for automatic short answer grading[C]//*Proceedings of the 12th Annual Meeting of the Association for Computational Linguistics*, 2009: 567-575.
- [19] GRANGER S, DAGNEAUX E, MEUNIE F, et al. International corpus of learner English [M]. Presses Universitaires de Louvain, 2009.
- [20] YANNAKOUDAKIS H, BRISCOE T, MEDLOCK B. A new dataset and method for automatically grading ESOL texts[C]//*Proceedings of the 49th Annual Meeting of the Association for Computational Linguistic: Human Language Technologies*, 2011: 180-189.
- [21] DZIKOVSKA M O, NIELSEN R, BREW C. Towards effective tutorial feedback for explanation questions: a dataset and baselines[C]//*Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2012: 200-210.
- [22] Automated Student Assessment Prize (ASAP)[DB/OL]. <https://www.kaggle.com/c/asap-sas/>, 2012.

- [23] CORRENTI R, MATSUMURA L C, HAMILTON L, et al. Assessing students skills at writing analytically in response to texts[J]. *Elem Sch J*, 2013, 114(2): 142-177.
- [24] SAKAGUCHI K, HEILMAN M, MADNANI N. Effective feature integration for automated short answer scoring[C]//*Proceedings of the Conference of the North American Chapter of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2015: 1049-1054.
- [25] BLANCHARD D, TETREAULT J, HIGGINS D, et al. TOEFL11: A corpus of non-native English[J]. *ETS Research Report Series*, 2013(2): 1-15.
- [26] BASU S, JACOBS C, VANDERWENDE L. Power-grading: A clustering approach to amplify human effort for short answer grading[J]. *TACL*, 2013, 1: 391-402.
- [27] STAB C, GUREVYCH I. Identifying argumentative discourse structures in persuasive essays[C]//*Proceedings of Empirical Methods in Natural Language Processing*, 2014: 46-56.
- [28] PHANDI P, CHAI K, NG T. Flexible domain adaptation for automated essay scoring using correlated linear regression[C]//*Proceedings of Empirical Methods in Natural Language Processing*, 2015: 431-439.
- [29] HONGBO C, BEN H. Automated essay scoring by maximizing human-machine agreement[C]//*Proceedings of Empirical Methods in Natural Language Processing*, 2013: 1741-1752.
- [30] RONAN C, MENG Z, TED B. Constrained multi-task learning for automated essay scoring[C]//*Proceedings of the 54th Annual Meeting of the Association for Computational Linguistic*, 2016: 789-799.
- [31] ISAAC P, VINCENT NG. Modeling thesis clarity in student essays[C]//*Proceedings of the 51st Annual Meeting of the Association for Computational Linguistic*, 2013: 260-269.
- [32] VAJIALA S. Automated assessment of non-native learner essays: Investigating the role of linguistic feature[J]. *International Journal of Artificial Intelligence in Education*, 2018, 28(1): 79-105.
- [33] SALIM Y, STEVANUS V, BARLIAN E, et al. Automated English digital essay grader using machine learning[C]//*Proceedings of the IEEE International Conference on Engineering TALE*, 2019: 1-6.
- [34] TAGHIPOUR K, NG H T. A neural approach to automated essay scoring[C]//*Proceedings of the Empirical Methods in Natural Language Processing*, 2016: 1882-1891.
- [35] Alikaniotis D, HELEN Y, MAREK R. Automatic text scoring using neural networks[C]//*Proceedings of the 54th Annual Meeting of the Association for Computational Linguistic*, 2016: 715-725.
- [36] DONG F, ZHANG Y. Automatic features for essay scoring: an empirical study[C]//*Proceedings of the Conference on EMNLP*, 2016: 1072-1077.
- [37] DONG F, ZHANG Y, YANG J. Attention-based recurrent convolutional neural network for automatic essay scoring[C]//*Proceedings of the Conference on Computational Natural Language Learning*, 2017: 153-162.
- [38] TAY Y, PHAN M C, Tuan L A, et al. SKIP-FLOW: Incorporating neural coherence features for end-to-end automatic text scoring[C]//*Proceedings of the Association for the Advancement of Artificial Intelligence*, 2018: 5948-5955.
- [39] LIANG G, ON B W, JEONG D, et al. Automated essay scoring: A siamese bidirectional LSTM neural network architecture[J]. *Symmetry*, 2018, 10(12): 682.
- [40] MAYFELD E, BLACK A W. Should you fine-tune BERT for automated essay scoring? [C]//*Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, 2020: 151-162.
- [41] RODRIGUEZ P U, JAFARI A, ORMEROD C M. Language models and automated essay scoring[J]. *arXiv*: 1909.09482, 2019.
- [42] YANG R, CAO J, WEN Z, et al. Enhancing automated essay scoring performance via finetuning pre-trained language models with combination of regression and ranking [C]//*Proceedings of Empirical Methods in Natural Language Processing*, 2020: 1560-1569.
- [43] GAO J, YANG Q, ZHANG Y, et al. A bi-modal automated essay scoring system for handwritten essays [C]//*Proceedings of the Conference of International Joint Conference on Neural Networks*, 2021: 1-8.
- [44] 梁茂成. 中国学生英语作文自动评分模型的构建[D]. 南京: 南京大学博士学位论文, 2005.
- [45] 唐芳, 庄翠娟, 巩艺超. 作文自动评分系统在大学英语写作教学中的应用: 以句酷批改网为例[J]. *海外英语*, 2017, 20: 48-49.
- [46] 魏思, 巩捷甫, 王士进, 等. 利用深层语言分析改进中文作文自动评分方法[J]. *中文信息学报*, 2022, 36(04): 111-123.
- [47] IN 课堂智能教育平台. IN 课堂: 语文作文智能批改教育迈向智能化阶段[N]. [http://inketang.com/v8/news\\_detail\\_00.html](http://inketang.com/v8/news_detail_00.html), 2018.
- [48] 陈一乐. 基于回归分析的中文作文自动评分技术研究[D]. 哈尔滨: 哈尔滨工业大学硕士学位论文, 2016.
- [49] 曹亦薇, 杨晨. 使用潜语义分析的汉语作文自动评分研究[J]. *考试研究*, 2007, 1(2): 65-73.
- [50] 巩捷甫. 面向语文作文自动评阅的修辞手法识别系统的设计与实现[D]. 哈尔滨: 哈尔滨工业大学硕士学位论文, 2016.
- [51] WEI S, DONG W, RUIJI F, et al. Discourse mode identification in essays[C]//*Proceedings of the 55th Annual Meeting of the Association for Computational Linguistic*, 2017: 112-122.
- [52] 马晓丽, 刘杰, 周建设, 等. 一种中小学汉语作文表现手法分类方法[J]. *计算机应用与软件*, 2018, 35(10): 49-54.
- [53] 付瑞吉, 王栋, 王士进, 等. 面向作文自动评分的优美句识别[J]. *中文信息学报*, 2018, 32(06): 88-97.

- [54] 刘明杨,秦兵,刘挺.基于文采特征的高考作文自动评分[J].智能计算机与应用,2016,6(01):1-4.
- [55] 刘明杨.高考作文自动评分关键技术研究[D].哈尔滨:哈尔滨工业大学硕士学位论文,2015.
- [56] 吴恩慈,田俊华.基于语言学特征的小学生作文流畅性自动评价[J].教育测量与评价,2020(03):41-50.
- [57] 刘杰,孙娜,袁克柔,等.中文作文句间逻辑合理性智能判别方法研究[J].计算机应用与软件,2019,36(01):71-77.
- [58] HAO S, XU Y, PENG H, et al. Automated Chinese essay scoring from topic perspective using regularized latent semantic indexing[C]//Proceedings of the International Conference on Pattern Recognition, 2014: 3092-3097.
- [59] XU Y, KE D, SU K. Contextualized latent semantic indexing: A new approach to automated Chinese essay scoring[J]. Journal of Intelligent Systems, 2017, 26(2): 263-285.
- [60] 王耀华,李舟军,何跃鹰,等.基于文本语义离散度的自动作文评分关键技术研究[J].中文信息学报,2016,30(6):173-181.
- [61] 钟启东,张景祥.嵌入语言深度感知的汉语作文评分算法[J].计算机工程与应用,2020,56(08):124-129.
- [62] 蔡黎,彭星源,赵军.少数民族汉语考试的作文辅助评分系统研究[J].中文信息学报,2011,25(05):120-126.
- [63] GEORGE L, MARK J. Metaphors we live by[M]. University of Chicago Press, 2008.
- [64] LIU L Z, HU X, SONG W, et al. Neural multitask learning for simile recognition[C]//Proceedings of Empirical Methods in Natural Language Processing, 2018: 1543-1553.
- [65] 梅家驹.同义词词林[M].上海:商务印书馆,1984.
- [66] SONG W, LIU T, FU R J, et al. Learning to identify sentence parallelism in student essays[C]//Proceedings of the Conference of International Conference on Computational Linguistics, 2016: 794-803.
- [67] 穆婉青,廖健,王素格.融合 CNN 和结构相似度计算的排比句识别及应用[J].中文信息学报,2018,32(02):139-146.
- [68] 朱晓亮,谯宇同.基于 BERT 模型的排比句自动识别方法[J].计算机应用与软件,2021,38(07):153-158.
- [69] 冯胜利,王洁,黄梅.汉语书面语体庄雅度的自动测量[J].语言科学,2008,7(2):113-126.
- [70] SONG W, SONG Z Y, LIU L Z, et al. Hierarchical multi-task learning for organization evaluation of argumentative student essays[C]//Proceedings of the conference of International Joint Conference on Artificial Intelligence, 2020: 3875-3881.
- [71] 刘杰,张文轩,李亚光,等.基于孪生神经网络的行文一致性测评研究[J].北京理工大学学报,2022,42(06):649-657.
- [72] 黄志娥,谢佳莉,荀恩东. HSK 自动作文评分的特征选取研究[J].计算机工程与应用,2014,50(6):118-122.
- [73] 徐昌火,陈东,吴倩,等.汉语作为第二语言作文自动评分研究初探[J].国际汉语教学研究,2015,1(1):83-89.
- [74] WANG Y, HU R. A Prompt-independent and interpretable automated essay scoring method for Chinese second language writing[C]//Proceedings of China National Conference on CCL, 2021: 450-470.
- [75] 李琳,董璐璐,马洪超.基于 BERT 的汉语作文自动评分研究[J].中国考试,2022(05):73-80.
- [76] SONG W, ZHANG K, FU R, et al. Multi-seaey pre-training for automated Chinese essay scoring[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2020: 6723-6733.
- [77] UTO M, XIE Y, UENO M. Neural automated essay scoring incorporating handcrafted features[C]//Proceedings of the International Conference on Computational Linguistics, 2020: 6077-6088.
- [78] MIAO N, ZHOU H, MOU L, et al. CGMH: constrained sentence generation by metropolis-hastings sampling[C]//Proceedings of Association for the Advancement of Artificial Intelligence, 2019, 33: 6834-6842.
- [79] ATTALI Y. Reliability-based feature weighting for automated essay scoring[J]. Applied Psychological Measurement, 2015, 39(4): 303-313.
- [80] UTO M, UCHIDA Y. Automated short-answer grading using deep neural networks and item response theory[C]//Proceedings of the Artificial Intelligence in Education, 2020: 334-339.
- [81] 周建设.人文基因智能计算将成为语言文字资源建设的新途径[J].语言战略研究,2022,7(05):10.



薛嗣媛(1993—),通信作者,博士,主要研究领域为自然语言处理、语言智能。  
E-mail: xuesiyuan987@163.com



任福继(1959—),博士,教授,主要研究领域为情感计算、自然语言处理。  
E-mail: ren2fuji@gmail.com



周建设(1957—),博士,教授,主要研究领域为语言智能。  
E-mail: zhoujianshe@solcnu.net