自动作文评分研究综述

陈潇潇[1]，葛诗利[2]

（1. 广东金融学院外语系，广东 广州 510520；2. 华南理工大学外国语学院，广东 广州 510640）

摘 要：基于统计、自然语言处理以及人工智能技术的自动作文评分研究在国外，尤其是美国已颇具规模，并于 1999年付诸实用。当前的自动作文评分技术及应用方面有代表性的 6种系统各有其长处与不足。国内相关研究，尤其是该技术应用于中国英语学习者作文评分的前景，以及需要改进的方面值得关注。

关键词：英语写作；自动作文评分；自然语言处理；机助语言测试

中图分类号：H315 文献标识码：A 文章编号：1002-722X（2008）05-0078-06

A Review of Automated Essay Scoring

CHEN Xiao-xiao[1], GE Shi-li[2]

(1. Department of Foreign Languages, Guangdong University of Finance, Guangzhou, Guangdong Prov., 510520, China;
2. School of Foreign Languages, South China University of Technology, Guangzhou, Guangdong Prov., 510640, China)

Abstract: The research of automated essay scoring (AES) based on statistics, natural language processing (NLP) and artificial intelligence (AI) receives adequate attention abroad, especially in the U.S.A. One of the systems, E-rater, has been applied in GMAT essay scoring since 1999. Six current AES systems have their own advantages and disadvantages in respect of technology and application. The related research in China, especially the feasibility of the technique applied to Chinese EFL writing and the innovation needed, demand more efforts.

Key words: English essay writing; automated essay scoring; natural language processing; computer-aided language testing

0. 引言

计算机技术的发展为实现语言测试现代化提供了技术方面的支持，而机辅语言测试走向智能化也是发展的必然趋势（吴会芹，2006），自动作文评分（Automated Essay Scoring）系统的研究与开发就是这一趋势的具体体现。所谓自动作文评分就是利用计算机技术对作文进行评估与记分。（Shermis & Burstein, 2003）该方向的研究至今已历时 40年，在此过程中，采用了统计、自然语言处理以及人工智能等方面的最新成果，并于 1999年进入实际应用。（Kukich, 2000）

自动评分系统有很多优点。首先是可靠性。许多研究表明，计算机评分系统效果很好。（Shermis & Burstein, 2003）其次是客观性。电子判分系统很客观，评分标准定义清楚，评分不受人为因素影响。再次是经济性。自动评分系统的运作快捷而准确，节省大量人工。其他优点还包括即时性、互动性等。

当然，计算机评分也有缺点。Page（2003）强调，计算机并不能像人一样评判一篇作文，因为计算机只是“编程让它做什么”它就做什么，而并不能像人一样去“欣赏”一篇文章。另一种批评有关构念（construct）方面的缺陷。也就是说，计算机所计算的变量并不一定是作文评分中“真正”重要的方面，如关注文章的形式方面而不是组织方面。（Page, 2003）这使得有些学生以反工程（reverse-engineering）的方式，用一些毫无意义的文章来欺骗机器并获得高分成为了可能。（Lonsdale & Strong-Krause, 2003）

收稿日期：2007－11－06

基金项目：国家自然科学基金项目（60572159）

作者简介：1. 陈潇潇（1975－），女，辽宁大连人，广东金融学院外语系讲师，博士研究生，研究方向为英语语言学和文化传播学；2. 葛诗利（1969－），男，山东烟台人，华南理工大学外国语学院讲师，博士，研究方向为英语语言学和语言技术。

1. 国外成熟的自动作文评分系统

1.1 Project Essay Grade (PEG)

PEG是 Ellis Page于 1966年应美国大学委员会的请求而研发的，其目的就是为了使大规模作文评分更加实际而高效。(Page, 2003) PEG完全依靠对文章的浅层语言学特征的分析对作文进行评分，根本没有涉及内容。(Valenti et al., 2003) 它使用代理量度标准 (proxy measures) 来衡量作文的内在质量以模拟人对作文的评分。作文评分本应该直接针对作文的内在质量进行评判。但内在质量，如写作的流畅性、句子结构的复杂度、文章措辞的情况等难以用计算机直接测量。于是 PEG采取了间接测量写作构念分项指标的方法，即所谓的代理量度标准。比如：作文长度代表了写作的流畅性；介词、关系代词等表明了句子结构的复杂度；词长的变化表明了文章措辞的情况 (因为非常用词一般都较长)。(Kukich, 2000) 代理量度标准由计算机从人工评分的训练作文集统计得出，与训练集中作文的人工评分一起用于标准多元回归的计算，从而得出各项代理量度标准的回归系数。得出的回归系数代表了人对作文评分的最佳模拟。这些系数与代理量度标准一起用于待评阅作文的自动评分。(Valenti et al., 2003)

PEG由于其对语义方面的忽视和更多地注重表面结构而遭受指责。(Kukich, 2000) 由于对作文内容相关方面的忽视，该系统不能够给出对学生有指导意义的反馈。另外，该系统最大的问题，就是对写作技巧的间接测量很容易被写作者利用，如写出文理不通的长文以获取流畅性方面的高分，欺骗计算机。(Kukich, 2000) 在上世纪 90年代，PEG在很多方面得到改进，整合了多种分析器、词典及各种资源，Page最新实验结果与人工评分在多元回归相关性上达到了 0.87。(Valenti et al., 2003)

1.2 Intelligent Essay Assessor (IEA)

IEA是上世纪 90年代末由 Pearson Knowledge Analysis Technology公司在潜在语义分析 (latent semantic analysis) 技术的基础上开发的。(Hearst, 2000) 潜在语义分析本来是一个用于文本索引和信息提取的复杂统计技术，其定义为“一个单词用法的统计模型，该模型允许对片断文本包含的信息之间的语义相似性进行比较”。(Dikli, 2006) 其核心思想就是一个段落的意义，在很大程度上取决于该段落所包含的词汇的意义，即使只改动一个单词，也可能使这个段落的意义发生改变。该思想可以总结为“词汇 1的意义 +词汇 2的意义 +……词汇 n的意义 =段落的意义”。(Landauer et al., 2003) 另一方面，两段由不同词汇构成的段落，其意义也可能非常相似。通过大量文本的数学计算可以发现，当某些不同的单词以较高的频率出现于相同或相似的语境时，可以推算出这些词汇意义的相近。而由不相同但意义相近的单词构成的段落，其意义也可能非常相似。

在自动作文评分中，该技术能够将学生的作文按照它所包含的单词投射成为能够代表作文意义 (内容) 的数学形式，然后在概念相关度和相关内容的含量两个方面与已知写作质量的参考文本进行比较，从而得出学生作文的评分。(Landauer et al., 2003)

在做潜在语义分析时，将每一篇文章视为一个空间向量，所有的向量构成一个矩阵，其中的列对应文档向量，行对应文档特征。特征可以是词、句子或段落。通常是以词为特征，但不包括停用词。向量值大多数情况下取值为词频。于是每个用于评价的教师参考文本以及待评价的作文都被表示成如上所述的向量。然后对所有这些向量构成的这个矩阵采用线性代数的奇异值分解技术来降低向量维数。这种维数缩减方法增加了数据彼此之间的依赖，增加了词与上下文之间的联系。余弦相关性计算法被用来依次计算降维后的教师参考文本向量和待批改作文向量的相似度。选取前 n个与待批改作文最相似的参考文本后，根据各自的相似度对参考文本的分值加权平均，得到待批改文本的分数。

基于潜在语义分析的 IEA不仅能评判基于内容的作文，还能评判有创意的记叙文。这种作文评分系统虽然主要在于评价文章内容方面的质量，但也可包含对语法、文体以及写作机制方面的评价与反馈，并能发现抄袭现象。(Landauer et al., 2003)

IEA的主要特征包括：相对较低的单元耗费、快速客户定制的反馈以及抄袭的发现。另外，其开发者声称该系统非常适合科技、社会研究、历史、医药或者商业方面说明文的分析与评分。

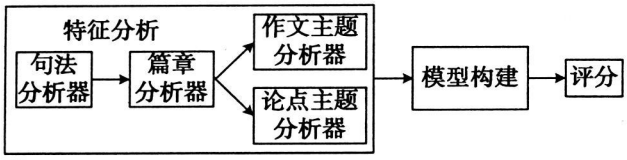
性能方面，在用 GMAT (Graduate Management Admissions Test) 作文所做的一次试验中，IEA与人工阅卷的基本一致性在 85%到 91%之间。(Valenti et al., 2003)

1.3 Electronic Essay Rater (E-rater)

E-rater是由 Educational Testing Service (ETS) 的 Burstein等人在上世纪 90年代末开发的。(Valenti et al., 2003) 目前 ETS正利用该系统对 GMAT中 Analytical Writing Assessment (AWA) 部

分进行评分，并于2005年开始应用于托福考试的作文评分。在E-rater付诸应用之前，GMAT的AWA由两名评卷员在6分的范围内做出整体评分，如果两名评卷员的评分差异超过1分，就需要第三名评卷员来处理。E-rater从1999年2月应用于AWA的评分。试卷的最终得分由E-rater和一名评卷员决定。同先前由两名评卷员共同阅卷的情况类似，如果E-rater跟评卷员的评分差异超过1分，第二名评卷员就参与解决这个问题。据Burstein讲，自从E-rater应用于GMAT的AWA的评分，E-rater与评卷员的分歧率一直低于3%，这并不高于两名评卷员的分歧，因此完全可以用于各种标准化考试的作文评分。（Burstein 2003）

图1. E-rater系统构建



E-rater系统构建如图1所示，它采用基于微软自然语言处理的工具包来分析文章，包括词性标注器为文本中每一个单词赋予词性；句法分析器分析文本中的句法结构；篇章分析器分析文本的篇章结构。采用词汇相似性度量器，以统计技术中的简单关键词分析法分析文本中的词汇使用。（Burstein 2003）另外，采用了基于语料库的方法建模。使用统计与自然语言处理技术来提取待评分文章的语言学特征，然后对照人工评分的标准作文集进行评分。评分过程主要由5个独立模块来进行。3个用来识别作为评分标准的特征，包括：句法模块、篇章模块和主题分析模块。这3个模块分别用来提取作文的句法多样性、思想的组织和词汇的使用方面的67个文本特征的特征值。第4个模块，即模型构建模块，用来选择和加权对作文评分具有预测力的特征。即把前3个模块提取的数据作为自变量，人工评分的分数作为因变量进行逐步线性回归，在67个变量中进行筛选，建立回归方程。第5个模块用来计算待评分文章的最后得分，即提取作文显著特征的特征值，代入回归方程计算得分。（Kukich 2000; Valenti et al, 2003; Dikli 2006）。

性能方面，在超过75万份经过评分的GMAT作文中，E-rater与专家评分的一致率大约是97%（Valenti et al, 2003）。Burstein & Chodorow（1999）的研究表明，此系统也可用于评判英语作为外语的作文，并已应用于托福考试的作文评分。该系统最大的问题是不能判别语法正确、但内容空洞的作文。

1.4 IntelliMetric™

IntelliMetric™是由Vantage Learning开发的，第一套基于人工智能（AI）的作文评分系统。它能够模仿人工评卷，在1到4或者1到6的分值范围内对作文的内容、形式、组织和写作习惯进行评分。它集中了人工智能、自然语言处理和统计技术的长处，是一种能够内化专家级评卷员集体智慧的学习机。（Elliot 2003）其核心技术是Vantage Learning的CogniSearch™和Quantum Reasoning™。前者是专门为IntelliMetric™开发，用来理解自然语言以支持作文的评分，如它能分析词性和句法关系，这使得IntelliMetric™能够依据英语标准书面语的主要特征来评判作文。二者结合使得IntelliMetric™能够内化作文中与某些特征相关的每一个得分点，并用于接下来的作文自动评分。（Elliot 2003; Dikli 2006）

IntelliMetric™需要采用专家级评卷员已经评好分数的作文集进行训练。在评分过程中，系统采用了多个步骤。首先，根据已评分数的训练集进行内化训练，构建模型；然后用较小的测试集检测模型的效度和概括度。两项都得到确认后，便可用于待评分作文的评判了。一旦根据标准美式英语或者先前训练得到的标准，某些作文被评估为不正常，系统会自动做出标注。（Dikli 2006）

IntelliMetric™评估了作文中语义、句法、篇章3个层次的300多项特征。在性能方面据称能够跟专家级评卷员给出的分数一样准确，与评卷员的一致率达到了97%至99%。另外，IntelliMetric™能够评阅多种语言的作文，如英语、西班牙语、以色列语和印度尼西亚语。对荷兰语、法语、葡萄牙语、德语、意大利语、阿拉伯语以及日语等多种语言文本的评价现在也能够做到了。（Elliot 2003; Dikli 2006）

1.5 Bayesian Essay Test Scoring sYstem（BETSY）和Larkey的系统

BETSY是由美国教育部投资，由马里兰大学College Park的Lawrence M. Rudner开发的，以概率论为指导，基于训练语料对文本进行分类的程序（Valenti et al, 2003）。该系统使用了包括内容与形式等多方面的一个大型特征集，根据4点类型尺度（优、良、合格、不合格）把一篇作文划分到一个最合适的集合中去。（Rudner & Liang 2002）文本分类所采用的底层模型是多元伯努利模型（MBM）和伯努利模型（BM），两者都属于朴素贝叶斯模型，因为它们都以条件独立假设为前提。BETSY的计算

量非常大，但据其开发者声称，由于该系统使用的方法能够整合 PEG、LSA和 E-rater的最佳特征，“再加上本身所特有的长处，使它具有以下特点：能够用于短文评测，易于使用，适用的内容范围宽广，能够产生诊断性结果，能够调节以用于多种技能的分类，以及容易使非统计人员明白其中的道理”。（Ruder & Liang, 2002; Valenti et al., 2003）

在性能方面，根据 Ruder & Liang（2002），BETSY采用了 462篇作文的训练集，在 80篇作文的测试集上得到了 80%的准确率。值得一提的是，BETSY是作文自动评分领域唯一可免费下载使用的软件。

另外，最早把文本统计分类方法用于作文自动评分的 Larkey（1998）以及 Croft（Larkey & Croft 2003）在这个领域也做出了很大贡献。在他们的研究中，采用了贝叶斯独立分类方法和最近邻分类方法（k-nearest-neighbor，简称 kNN），并提取 11个文本复杂性特征用于线性的回归计算。在他们的实验中，单独的贝叶斯独立分类方法有着稳定而良好的表现。然而，加入文本复杂性特征和最近邻分类方法后，系统性能并没有得到显著的改善。在这种评分方法中，作文长度的重要性不像其他自动评分系统那样明显。（Larkey & Croft 2003）

除了以上介绍的 6种主要的作文评分技术以外，美、英、法、日、西班牙等国还有很多基于信息提取和文本聚类的对于作文和问答型主观题的自动评分系统，如 ATM、Automark、Jess和 IEMS等。鉴于国内强烈的需求，近年来也有研究者涉足自动作文评分领域，并做出了可贵的探索。

2 国内对自动作文评分的需求和研究

2.1 国内对自动作文评分的需求

随着国内各种大规模考试参与人数的增加，如汉语的中考、高考、对外汉语教学的 HSK，英语的大学英语四、六级、专业英语四、八级、PETS、高考、研究生入学考试等，各种语言考试中评卷教师的作文评分工作量越来越重，而且评卷信度饱受争议。至少在英语教学界“教师对学生作文质量的评定往往以主观印象为主，缺乏统一、标准、科学、客观的尺度”，而“如果不能对学生的作文做出客观、公正的评定，就会挫伤他们的学习积极性”，甚至“还会使学生对作文的质量标准产生错误的认识或把握不准，从而影响其写作能力的提高”。（李志雪、李绍山，2003）

不论是为了减轻教师工作量、降低考试费用，还是为了增加评分信度，促进学生的学习积极性，都应该加强计算机辅助作文测试，尤其是自动作文评分的研究。

2.2 国内自动作文评分研究

国内最早涉足自动作文评分领域的是梁茂成（2005），其研究方向是中国学生英语作文的自动评分。在他的研究中采用了 220篇已评分的作文样本，其中 120篇作为训练集，100篇作为验证集。在训练集的基础上得到评分模型后，用验证集交叉验证模型的可信度。并进而采用双重交叉验证，即交换训练集和测试集并重复以上步骤。梁茂成的建模方法兼顾了 PEG和 IEA的长处，在训练集中提取了大量的作文浅层文本特征，连同作文的内容得分作为自变量，人工评分作为因变量一起用于多元回归计算，得到作文评分的回归方程。提取待评分作文的相关特征值，代入回归方程即可得到该作文的得分。梁茂成的研究取得了较高的评分准确率，与人工评分相关系数 R最高达到 0.837。但由于作文样本来源范围较窄，数量较少，并且提取的特征主要是文本浅层特征，未能够涉及文章的深层结构，所得结果尚有待于进一步验证与加强。

国内另一位研究自动作文评分的是李亚男（2006），其研究方向是汉语作为第二语言测试的作文自动评分。这项研究以中国少数民族汉语水平考试三级作文为研究样本。研究中采用了两个样本，一个是包括 7个作文题目的 583篇作文，另一个是同一题目的 488篇作文，每个样本随机分成数量大致相等的两组。以多元线性回归为研究方法，以 45个可量化的评分要素作为自变量（其中第一个样本只利用了前 40个评分要素），阅卷员给出的作文分数作为因变量，利用逐步回归和强迫输入回归两种提取变量的方法，进行多元线性回归分析，并在样本内部两个随机组之间进行交叉验证。最后将得出的 8个回归方程进行比较，发现利用前 40个评分要素对多题目作文建立的回归方程虽然经方差检验都具有显著性，但有效性指标 R^2 都很低。用 4个方程对作文评分的结果与人工评分的结果相关度也很低。这说明对多题目作文自动评分模型的建立是不理想的。同时也发现，利用所有的 45个评分要素对特定题目作文建立的回归方程有效性都高得多，采用其中最好的方程对作文评分，与人工分数的相关度达到了 0.572，并且“自动评分与人工评分的完全一致性高出了人与人的一致性”。（李亚男，2006，45）

该项研究所得结果不甚理想，原因除了作者申明的样本抽取范围狭小、变量提取手段不完善和内

容变量缺失之外，更重要的是其所选变量绝大多数都是词汇层次的浅层文本特征，不能够真正反映作文的质量。而其中能够反映作文质量的特征：语法错误数和错误率首先要经过人工标注。另外，一个在所有回归方程中都起到显著作用的变量：书面表达第一部分的分数，完全是作文以外的变量，虽然可以用于该项考试作文的评分，但有碍于评分模型的推广使用。该研究虽存在诸多的问题，但还是在通用评分模型方面做出了有益的尝试。

最近发表的，也是第一个使用潜在语义分析技术对汉语作文进行自动评分研究的是曹亦薇和杨晨（2007）。他们的研究采用人工评分的202篇高中作文为样本，使用潜在语义分析技术评价作文得到内容分数，此分数与人工评价的内容分数的相关性达到0.47。其研究表明，潜在语义分析技术在汉语作文自动评分中起着重要作用，但仅采用该技术实行作文评分显然不够，尚需寻找更多的指标，并辅以其他方法提高自动评分效果。

3. EFL自动作文评分研究的发展与不足

虽然近年来自动作文评分在国外已逐渐成为自然语言处理中的热点问题，成型的系统已有10余个，文章与著述也比较多，但涉及EFL作文评分技术的研究尚不多见，目前只有E-rater（Burstein & Chodorow, 1999）、Lonsdale & Strong-Krause（2003）和梁茂成（2005）等。

Burstein等人把母语为汉语、阿拉伯语和西班牙语的英语学习者的作文与母语为英语的人（其中包括美国本土出生与本土以外出生两类）的作文，在人工评分与E-rater评分的框架下做了对比研究。他们的研究表明，虽然人工评分的均值（4.16）与机器评分的均值（4.08）（总分为6分）差别不大，但具有统计显著性（$F=5.469$, $P<.05$），而不同题目之间没有显著性差异。这说明E-rater在评价英语作为外语的学习者所写的英语作文方面与人工评分虽然差别不大，但还是存在一些影响机器评分准确性的因素。（Burstein & Chodorow, 1999）

Burstein等人的研究最后得出的结论是，“虽然不同语言组中人工评分与E-rater评分存在显著性差异，但其差异的绝对值不大（总分为6分），所以与人工评分的一致率没有显著性差异”。（Burstein & Chodorow, 1999）由于这项研究是基于托福考试作文，作文总分是6分，而我国的大规模英语考试，如大学英语四、六级，其作文是15分制。随着总分范围的扩大，由于人、机两者之间评分的显著性差异，评分差异的绝对值也会不断增大，从而导致自动评分与人工评分的一致率下降。

另外，国内低于托福测试水平的英语学习者为数众多，而低水平的英语作文对于自动评分有着不同的技术要求。其中一个重要的方面是低水平的英语作文中频繁出现的词汇和句法方面的错误。在这方面，“传统的NLP语法分析器在英语作为外语的教学应用上，尤其是作文自动评分上至今尚未取得广泛的成功”。（Lonsdale & Strong-Krause, 2003）

以上研究表明，英语作为母语的自动作文评分与EFL的自动作文评分，尤其是与低水平英语学习者的作文评分，存在着较大的差异。其最主要差异就在于句法方面。以英语为母语的作文中，无论其思想表达是否完整、顺畅，绝大多数句子都不存在严重的语法错误。但是英语学习者，尤其是低水平学习者的作文，充斥着各种句法错误，更遑论主题思想的表达。这在Wolfe-Quintero等人（1998）的论述中也得到证明，即在外语写作评价中，语言的使用，尤其是句法方面，所占比重相对较大。这就使得以表层文本特征为评分依据的PEG和以内容为主要评分依据的IEA并不适合英语作为外语的作文的评分。E-rater能够做句法，甚至篇章分析，但对于错误百出的初级英语学习者的作文，各种分析的准确性都会大打折扣，这也是E-rater对于EFL作文评分准确性降低的一个重要原因。基于人工智能的IntelliMetric™和基于文本分类技术的BETSY在EFL作文评分方面尚未有公开的资料。

在这方面进行改进的首先当推Lonsdale & Strong-Krause（2003），他们采用了基于Link Grammar（LG）的句法分析器来分析评判英语学习者的作文。LG分析器能够跨越句子中不合语法的单词，找到后面的词汇，从而能够连接上，构成有句法意义的词对，比如：主语+动词、动词+宾语、介词+宾语、形容词+状语修饰语和助动词+动词等。但是由于LG本身的局限和单独的句法分析，机器评分的准确率较低。

梁茂成（2005）对中国学生英语作文的自动评分做出了有益的探索，但未能涉及句法层次，而这应该正是中国学生英语写作自动评价的突破口。所谓的句法层次，包括作文中动词的各种用法以及句型结构。另外，与英语写作直接相关的一些项目，如各种搭配以及词块的使用，都是中国学生英语作文自动评分下一步的研究重点。

我国当前外语教学一方面有学生人数众多的压力，另外又有切实提高学生实际语言应用能力的较高要求，通过借助自动作文评分软件，将有望突破

写作批改量大、难度大的瓶颈，为教、学双方带来切实的帮助。

参考文献：

[1] 曹亦薇，杨晨. 使用潜在语义分析的汉语作文自动评分研究 [J]. 考试研究，2007，3（1）：63—71.

[2] 李亚男. 汉语作为第二语言测试的作文自动评分研究 [D]. 北京：北京语言大学博士论文，2006.

[3] 李志雪，李绍山. 对国内英语写作研究现状的思考——对八种外语类核心期刊十年（1993—2002）的统计分析 [J]. 外语界，2003（6）：55—60.

[4] 梁茂成. 中国学生英语作文自动评分模型的构建 [D]. 南京：南京大学博士论文，2005.

[5] 吴会芹. 用现代化手段辅助语言测试 [J]. 外语电化教学，2006（6）：49—53.

[6] Burstein J. The E-rater scoring engine: Automated essay scoring with natural language processing [C] // M. D. Shermis & J. Burstein. Automated Essay Scoring: A Cross-disciplinary Perspective. Mahwah, NJ: Lawrence Erlbaum Associates, 2003: 113—122.

[7] Burstein, J. & M. Chodorow. Automated essay scoring for nonnative English speakers [Z/OL]. (1999—06) [2006—03—20]. http://www.ets.org/Media/Research/pdf/erater_acl99rev.pdf.

[8] Dikli, S. Automated Essay Scoring [J]. Turkish Online Journal of Distance Education, 2006, 7 (1): 49—62.

[9] Elliot, S. IntelliMetric: from here to validity [C] // M. D. Shermis & J. Burstein. Automated Essay Scoring: A Cross Disciplinary Perspective. Mahwah, NJ: Lawrence Erlbaum Associates, 2003: 71—86.

[10] Hearst, M. The debate on automated essay grading [J]. IEEE Intelligent Systems, 2000, 15: 22—37.

[11] Kukich, K. Beyond Automated Essay Scoring [J]. IEEE Intelligent Systems, 2000 (5): 27—31.

[12] Landauer, T. K., D. Laham & P. W. Foltz. Automated Essay Scoring and Annotation of Essays with the Intelligent Essay Assessor [C] // M. D. Shermis & J. Burstein. Automated Essay Scoring: A Cross-disciplinary Perspective. Mahwah, NJ: Lawrence Erlbaum Associates, 2003: 87—112.

[13] Larkey, L. S. & W. B. Croft. A Text Categorization Approach to Automated Essay Grading [C] // M. D. Shermis & J. Burstein. Automated Essay Scoring: A Cross-disciplinary Perspective. Mahwah, NJ: Lawrence Erlbaum Associates, 2003: 55—70.

[14] Larkey, L. S. Automatic essay grading using text categorization techniques [C] // Proceedings of the 19th Annual International SIGIR Conference on Research and Development in Information Retrieval. Melbourne, Australia, 1998: 90—95.

[15] Lonsdale, D. & D. Strong-Krause. Automated Rating of ESL Essays [Z/OL]. (2003) [2006—03—20]. http://acl.ldc.upenn.edu/W/W03/W03—0209.pdf.

[16] Page, E. B. Project Essay Grade: PEG [C] // M. D. Shermis & J. Burstein. Automated Essay Scoring: A Cross-disciplinary Perspective. Mahwah, NJ: Lawrence Erlbaum Associates, 2003: 43—54.

[17] Rudner, L. M. & T. Liang. Automated essay scoring using Bayes' Theorem [J]. The Journal of Technology, Learning and Assessment, 2002, 1 (2): 3—21.

[18] Shermis, M. D. & J. Burstein. Automated Essay Scoring: A Cross-disciplinary Perspective [M]. Mahwah, NJ: Lawrence Erlbaum Associates, 2003.

[19] Valenti, S., F. Neri & A. Cucchiarelli. An overview of current research on automated essay grading [J]. Journal of Information Technology Education, 2003, 2 (2): 319—330.

[20] Wolfe-Quintero, K., S. Inagaki & H. Y. Kim. Second language development in writing: measures of fluency, accuracy & complexity [M]. Hawaii: University of Hawaii Press, 1998.

（责任编辑　周光磊）