中国 EFL 学习者自动作文评分探索

葛诗利[1]，陈潇潇[2]

（1. 北京语言大学 2. 广东金融学院）

摘　要：本文首先回顾了自动作文评分研究在国外的发展，讨论了它们在二语写作评分方面的表现。接下来，分析了自动作文评分在 EFL 写作评分领域的探讨，以及针对中国英语学习者作文自动评分的研究。最后，在讨论自动评分应用目标的基础上，提出适合中国国情的研究方向及可能应用的方法。

关键词：中国英语学习者，英语作文，自动评分，自然语言处理

Automatic Essay Scoring for Chinese EFL Learners

Ge Shili[1]，Chen Xiaoxiao[2]

（1. Beijing Language and Culture University 2. Guangdong University of Finance）

Abstract: This paper first reviews the researches on automated essay scoring (AES) systems abroad and discusses the performance on ESL writing. Then, researches on AES of EFL writing are discussed including Chinese studies toward Chinese EFL learners. At last, on the basis of AES applications, the research areas needed in China and possible methods are suggested.

Keywords: Chinese EFL learners, English writing, automated essay scoring, NLP

1. 引言

“要学好英语，写的重要性怎么强调都不过分。”（王初明，2004）国外的“很多研究者也都认为写作是评价学习结果最有用的工具”。（Valenti, et. al，2003）但是作文评分与批改既费时又费力，给教师增加了沉重的负担。尤其是在我国高校扩招之后，据教指委统计，大学英语师生比例已达 1：130（张尧学，2003）。教师在写作的批改上变得越来越有心无力，从而也导致了学生在写作方面训练的不足。然而，自动作文评分（Automated Essay Scoring，简写为 AES）系统的研究与开发为彻底解决这一问题带来了希望。

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近四十年，在此过程中，采用了统计、自然语言处理（NLP）、及人工智能等方面的最新成果（Dikli, 2006），并于 1999 年进入实际应用阶段（Kukich, 2000）。

自动作文评分系统有以下几个优点：1）实用性：可以提高工作效率。2）一致性：作文评分本质上存在着主观性，人工评分的一致性就会因此而受到一定的影响。3）反馈：给学生反馈是非常重要的，这种评分系统能够为作者提供具有针对性的修改建议。（Lonsdale & Strong-Krause, 2003）

多项研究证明，在写作评测方面，自动评分系统的准确性与可靠性，以及与人工评分的一致性方面都是非常高的（Burstein & Chodorow, 1999; Keith, 2003; Landauer, Laham, & Foltz, 2003; Page, 2003）。

当然，计算机评分也有很多缺点。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。另外一种批评是构造方面的缺陷。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，比如，关注文章的形式方面而不是组织方面（Page, 2003）。

基金资助：本研究得到国家自然科学基金项目（60573184）的资助，是其前期理论研究的一部分。

本文首先回顾了自动作文评分研究在国外的发展，讨论了它们在二语写作评分方面的表现。接下来，分析了AES 在 EFL 写作评分领域的探讨，以及针对中国英语学习者作文自动评分的研究。最后，在讨论自动评分应用目标的基础上，提出适合中国国情的研究方向及可能应用的方法。

2. 自动作文评分的发展

2.1 早期研究

最早开始 AES 研究的是 Ellis Page，他在 1966 年应美国大学委员会的请求，开发了第一套 AES 系统：Project Essay Grader（PEG）。他是从已经评分的作文中自动抽取各种文本特征，用于多元线性回归，求出最能够预测人工评分的各加权特征的最佳组合。然后采用同一加权特征集给未评分作文评分。（Kukich, 2000）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的，如：作文长度，介词、关系代词、及其它词性的词汇数量，词长的变化等，根本没有涉及到内容（Valenti, Neri, & Cucchiarelli, 2003）。尽管当时 PEG 的评分准确率已经很高了，但并没能得到教育界的承认与接受。由于它对作文内容、组织、体裁等的忽视，该系统不能给出有指导意义的反馈。另外，该系统最大的问题，就是由于采用对写作技巧的间接测量而很容易被写作者欺骗，比如：写出更长的文章（Kukich, 2000）。

要克服 PEG 的缺点，就只能是对写作质量的更直接的评估。在这方面迈出第一步的是上世纪八十年代初由 MacDonald（MacDonald et al., 1982）等人开发的 Writer's Workbench（WWB）工具包。它并不是一个自动评分系统，而是为了给作者在拼写，措词及可读性方面提供有意义的反馈。它包括了拼写检查程序（世界上第一个），措词程序，及计算可读性的程序。虽然 WWB 的这些程序只是刚刚触及文章的表面，但也是向着文章质量自动分析的方向迈出了正确的一步。

2.2 最新研究

到了上世纪九十年代，NLP 与信息提取技术取得了长足的进步。依托于这些技术，开发出了数种有代表性的 AES 系统。包括：由 Pearson Knowledge Analysis Technology (PKT)在潜在语义分析（LSA）技术的基础上开发的 Intelligent Essay Assessor (IEA)；由美国教育考试中心（ETS）的 Burstein 等人开发的基于统计和 NLP 的 Electronic Essay Rater (E-Rater)；由 Vantage Learning 开发的，基于人工智能的作文评分系统 IntelliMetric；由马里兰大学 College Park 的 Lawrence M. Rudner 开发的，基于文本分类的评分系统 Bayesian Essay Test Scoring sYstem (BETSY)。

IEA 是基于 LSA 技术，而“LSA 的目的就是透过作文表面的词汇以量化的方式衡量其底层的语义内容”（Landauer & Dumais, 1997）。它把一篇作文看成是由词汇构成的向量，多篇文章的向量构成一个矩阵，然后采用歧异值分解降低维度的办法，归纳单词间的语义相似性。它的主要优势在于能够抓住词汇间语义传递关系和搭配效果。因此，尽管两篇作文的用词不尽相同，IEA 还是能够准确判断二者的语义相关性。（Kukich, 2000）IEA 的开发也证明了自动抽取更直接的衡量作文质量的标准，甚至语义标准，都是完全可能的。

E-Rater 采用了基于 NLP 的工具包，如：词性标注器、句法分析器、篇章分析器、及词汇相似性度量器，来分析文章中所有的句子，采用了基于语料库的方法建模。使用统计与 NLP 技术来提取待评分文章的语言学特征，然后对照人工阅卷的标准作文集进行评分。评分过程主要由五个独立模块来进行。三个用来识别作为评分标准的特征：句法模块、篇章模块、和主题分析模块。分别用来分析作文的句法多样性，思想的组织，和词汇的使用。第四个模块用来选择和加权对作文评分具有预测力的特征。第五个模块用来计算最后的得分。（Kukich, 2000; Valenti et al., 2003; Dikli, 2006）。

IntelliMetric 集中了人工智能、NLP、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。二者一起使得 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。IntelliMetric 评估了作文中语义、句法、篇章三个层次的 300 多项特征。（Dikli, 2006）

BETSY 是一个基于训练语料对文本进行分类的程序（Valenti, Neri, & Cucchiarelli. 2003）。该系统使用了包括内容与形式方面的一个大型特征集，根据四点类型尺度（如：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。文本分类所采用的朴素贝叶斯模型包括多元伯努利模型和伯努利模型。(Rudner & Liang, 2002)

PEG 在上世纪九十年代，很多方面也得到改进，整合了很多分析器、词典与各种资源（Dikli, 2006）

2.3 未来研究动向

几乎所有的 AES 系统都是以人工评分的作文作为训练集，建立模型后对其余作文进行总体评分。IntelliMetric 已经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%-99%（Dikli, 2006）从目前的研究趋势看，AES 的发展方向，一个是挖掘更直接的能够反映作文质量的特征，从而给出详细的反馈；另一个是对短文，甚至是问题回答的评分，以期能够对 NLP 中的问答系统进行评估。(Kukich, 2000; Dikli, 2006)

3. 当前 AES 系统在二语作文评价方面的表现

虽然近年来 AES 在国外已渐成为 NLP 中的一个热点问题，成型的系统已有十余个，文章与著述也比较多，但涉及 EFL 作文评价的尚不多见。比较系统的研究只有 E-Rater（Burstein & Chodorow, 1999）。

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（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 评分的框架下做出了对比研究。他们收集了这五类作者两个题目的作文，分别是 562 篇和 576 篇。然后以整体评分的方式，从 1 分到 6 分，全部做出人工评分。在这两个作文集中各自随机挑出 255 篇作文（2 分到 6 分各 50 篇，1 分的 5 篇）作为训练集，训练评分模型，并对其余作文进行自动评分。评分结果见表 1 到表 3（摘自 Burstein & Chodorow, 1999）。

表 1：两个题目作文的人工评分与机器评分的对比

题目	n=	一致率（准确+临近）%	皮尔逊 r	人工评分		机器评分	
				均值	标准差	均值	标准差
题目 1	562	91.1	0.667	4.16	0.974	4.08	1.041
题目 2	576	93.4	0.718	4.16	0.936	4.07	0.989
均值		92.3	0.693	4.16	0.955	4.08	1.015

Burstein 等人的研究表明，虽然人工评分的均值（4.16）与机器评分的均值（4.08）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（Burstein & Chodorow, 1999）。这说明 E-Rater 在评价 EFL 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。

表 2：不同语言组题目 1 作文的人工评分与机器评分的对比

语言组	n=	一致率（准确+临近）%	皮尔逊 r	人工评分		机器评分	
				均值	标准差	均值	标准差
阿拉伯语	146	89	0.645	3.83	0.973	3.67	0.947
汉语	153	88.2	0.543	4.09	0.884	4.12	1
西班牙语	131	92.4	0.644	3.96	0.986	3.7	0.915
美国英语	97	96.9	0.632	4.96	0.624	4.93	0.814
非美国英语	35	91.4	0.544	4.31	0.9	4.51	0.981

表 3：不同语言组题目 2 作文的人工评分与机器评分的对比

语言组	n=	一致率（准确+临近）%	皮尔逊 r	人工评分		机器评分	
				均值	标准差	均值	标准差
阿拉伯语	151	96.4	0.783	3.85	0.959	3.7	0.909

汉语	139	91	0.707	3.92	0.957	4.04	1.03
西班牙语	138	93.5	0.616	4.07	0.845	3.69	0.733
美国英语	103	92	0.519	4.83	0.613	4.95	0.759
非美国英语	45	93.3	0.465	4.68	0.732	4.6	0.78

表 2、3 表明英语作为母语的作者写的作文成绩要高于非母语作者，在这方面，机器与人的评分达到了一致。但是在非英语母语的语言组作文成绩的评价上，E-Rater 与人的评价出现了显著性差异（F=12.397, p<.001）。有的组（如：西班牙语）E-Rater 给的分值低于人给的评分，而有的组（如：汉语）E-Rater 给的分值高于人给的评分。回顾机器评分中的建模过程，训练集中 75%的作文是由非英语母语者所写，而筛选出用于线性回归过程的特征与英语母语作文评分所选特征基本相同，而其中最重要的两个是文章层次上的词汇使用和论点层次上的词汇使用。（Burstein & Chodorow, 1999）这意味着汉语组的作者在词汇掌握上要优于其它 EFL 作者，所以在机器评分时占有优势；但在写作的其他方面比如句法、篇章结构等可能不及其它 EFL 作者，所以在人工评分时处于不利地位。

4. 当前外语自动作文评分研究

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，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，结果对大多 EFL 学习者并不具有普遍性，因为参加托福考试的英语学习者，跟大多数的普通大学生，尤其是大学低年级英语写作水平亟待提高的学生相比，其英语水平要么较高，要么已经过一定阶段的有针对性的考试培训。

有研究（梁茜，2002；方清，2003）表明，中国大学生，尤其是英语水平较低的学生，在英语写作的三个层次上（词汇、句法、篇章）都受到中国文化与汉语思维习惯的较大影响。随着英语水平的提高，对英语国家文化的了解，以及对英语思维习惯的熟悉，学生的写作水平在词汇和篇章层次上都会得到较大提高。但是在句法层次上，情况比较复杂。即使英语水平较高的学生，在句法的某些方面，仍然明显受到中国文化与汉语思维习惯的影响。另外一个要考虑的重要因素是低水平的英语作文中高频率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 EFL 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）

这些研究都说明，英语作为母语的作文评分与 EFL 的作文评分，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。这在 Wolfe-Quintero 等人（1998）的论述中也得到证明。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。这就使得以内容为主要评分依据的 IEA 不适合 EFL 作文的评分。E-Rater 在这方面的表现前面已有评价。基于人工智能的 IntelliMetric 和基于文本分类技术的 BETSY 尚没见到这方面的资料。

在这方面的努力首先当推 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 EFL 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，并连接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。但是由于只分析句法，机器评分的准确率较低。

国内学者近年也有涉足这一领域。梁茂成（2005）对中国学生英语作文的自动评分做出了有益的尝试。他以提取浅层文本特征为主，进行线性回归，得到了较高的评分准确率。

5. AES 系统的应用目标

AES 系统有两个应用目标：1）用于大规模考试的自动评分，2）用于写作教学，作为一个提供反馈的工具。前面讨论到的几个系统，基本都是以第一个目标为主。也有在此基础上兼顾第二个目标的，比如：E-Rater 和 IntelliMetric。

Criterion 就是为后一目标开发的作文评测系统。该系统内嵌了 E-Rater 及 Critique 写作分析工具。该工具包

含了一组程序，用来识别语法、用词、及写作机制（比如大小写、标点符号等）等方面的错误。Criterion 在给出作文实时评分的同时，还能对作文给出个性化的诊断性反馈（Burstein, Chodorow, & Leacock, 2003）。该系统只能评判与分析事先训练过的作文题目。而训练该系统，一个题目至少需要 465 份经过专家评分的作文作为训练语料（Dikli, 2006）。

My Access 是另外一套网上写作评测工具。它是用 IntelliMetric 技术构建的。其目的就是为学生提供一个写作环境，能够给出实时评分与诊断性反馈，使学生据此修改作文，刺激他们在该题目上不断改善以提高写作能力。该系统储备了 200 个题目，对这些题目的作文都能够给出即时的评判与分析。对于新题目，需要大约 300 篇评分过的作文作为训练集。（Dikli, 2006）

另外也有基于 LSA 技术的作文自动评分并提供反馈的研究（Foltz, Gilliam, & Kendall, 2000），但这里所提供的主要是内容方面的反馈。

6. 适合中国国情的研究方向

Page 在 1996 年把作文评分分为对内容评分与对文体评分，前者指文章讲了什么；而后者是指句法、写作机制、用词以及文章如何表达的其他方面（Valenti, Neri, & Cucchiarelli, 2003）。有的系统偏重于分析文体（如 PEG），有的系统偏重于分析内容（如 IEA）,有的二者兼收并蓄（如 E-Rater，IntelliMetric）。对于中国的 EFL 作文，只分析内容显然不切合实际。但兼顾内容与文体的评分系统也往往偏重于内容，像前面对 E-Rater 的讨论。在 6 分档的范围内，与人工阅卷的一致率差别不大，但扩大到 15 分或者 100 档，由于其与本族语者英语作文的评分存在显著性差异，与人工阅卷的一致率差异必定会增大。而梁茂成所做的，类似于 PEG 的对文本浅层特征的提取与计算，却较好的模拟了人工评分。

英语考试，尤其像我国的 TEM、CET、和 PETS，最终目的还是为了促进英语学习。所以我国 AES 研究应着重于第二个目标，即为学生提供一个基于网络的写作环境，能够为学生的作文给出即时的评分与反馈，指导学生写作，以提高他们的写作水平，缓解中国大学英语教师的严重短缺。在这个应用上，准确而详尽的反馈至为重要。而提供反馈，只是浅层的文本特征提取与分析是达不到目的的，必须结合中国学生英语作文的实际情况，采用各种 NLP 工具，对文章做出细致的深层次分析。

Criterion 能够给出的反馈信息类型有以下几种：

1） The text is too brief to be a complete essay.（建议学生写得再多一点）

2） The essay text does not resemble other essays written about the topic.（暗示可能跑题了）

3） The essay response is overly repetitive.（暗示学生使用了过多的同义词）（Burstein, 2003）

对于中国学生英语作文的反馈，仅此几项是远远不够的。当前的词汇、语法的分析技术已经比较成熟，只要做出适当的改进，当可适应中国 EFL 作文评分，并给出这两方面的详细反馈。EFL 作文内容方面的重要性虽不像英语母语作文那么突出，但也不可忽视，可参考 LSA 的方法，给出内容方面的提示。在篇章结构方面，可参考 E-Rater 的方法。更重要的是，根据英语写作教学理论，在反馈中要给出正面的表扬、鼓励性提示。（王初明，2004）。这方面可以考虑模式语法与语料库的结合运用。

另外就是以上几种系统，除了 IEA 之外都需要 200 甚至 300 篇以上已评分的作文作为训练语料。这已超过一个大学英语教师教授的学生数目。而 IEA 评分的前提条件可以是三种：100 篇已评分的学生作文；样板作文和知识源材料；未评分作文集的内部比较。（Landauer et al., 2003）可以考虑以第二种方法来训练日常教学用的中国学生英语作文评分系统。

E-Rater 的开发人员超过 15 个，投资超过了 100 万美元（Burstein, Chodorow, & Leacock, 2003）。一个理想的，适合中国英语学习者的自动评分、反馈系统的开发，需要广大的英语教学工作者与计算语言学工作者齐心协力，才有望在不远的将来达到目标。

参考文献：

[1] 王初明. 论外语“写长法”的教学理念[A]. 郑超编. 以写促学：英语“写长法”的理念与操作[C]. 北京：科学出版社，2004.

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

[3] 张尧学. 大力推进大学英语教学改革[EB/OL]. http://www.englishvod.net/Article/hdsvs/200410/2070.html, 2003-10-09/2006-03-20

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

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

[6] Kukich, K. Beyond Automated Essay Scoring [A]. In Marti A. Hearst (ed), The debate on automated essay grading [J]. *IEEE Intelligent systems*, 2000, (5): 27-31.

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

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

[9] Keith, T. Z. Validity of Automated Essay Scoring Systems [A]. In Mark D. Shermis and J. C. Burstein (ed.), Automated Essay Scoring: A Cross Disciplinary Perspective [C]. Mahwah, NJ:Lawrence Erlbaum Associates, 2003. 147-168.

[10] Landauer, T. K., Laham, D., & Foltz, P. W. Automated Essay Scoring and Annotation of Essays with the Intelligent Essay Assessor [A]. In Mark D. Shermis and J. C. Burstein (ed.), Automated Essay Scoring: A Cross Disciplinary Perspective [C]. Mahwah, NJ:Lawrence Erlbaum Associates, 2003. 87-112.

[11] Page, E. B. Project Essay Grade: PEG [A]. In M. D. Shermis & J. Burstein (ed.). Automated essay scoring: A cross-disciplinary perspective [C]. Mahwah, NJ: Lawrence Erlbaum Associates. 2003. 43-54.

[12] MacDonald, N. H., Frase, L. T., Gingrich, P. S., & Keenan, S. A. The Writer's Workbench: Computer Aids for Text Analysis [J]. *IEEE Transactions on Communications*, 1982, (1): 105-110.

[13] Landauer, T. & Dumais, S. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge [J]. *Psychological Review*, 1997, (104): 211-240.

[14] Elliot, S. IntelliMetric: from here to validity [A]. In Mark D. Shermis and J. C. Burstein (ed.). Automated essay scoring: a cross disciplinary perspective [C]. Mahwah, NJ: Lawrence Erlbaum Associates, 2003. 71-86

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

[16] 梁茴. 中国学生英语作文中的文化迁移及其教学对策[D]. 西南师范大学, 2002.

[17] 方清. 中西方思维模式的不同及其对中国学生英语作文的影响[D]. 中山大学, 2003.

[18] Wolfe-Quintero, K., Inagaki, S. & Kim, H. Y. Second language development in writing : measures of fluency, accuracy, & complexity [M]. Honolulu: Second Language Teaching & Curriculum Center, University of Hawaii at Manoa, 1998.

[17]梁茂成. 中国学生英语作文自动评分模型的构建[D]. 南京大学, 2005.

[18] Burstein, J., Chodorow, M. & Leacock, C. Criterion Online Essay Evaluation: An Application for Automated Evaluation of Student Essays [EB/OL]. http://citeseer.ifi.unizh.ch/cache/papers/cs/28192/http:zSzzSzwww.ets.orgzSzresearchzSzdloadzSziaai03bursteinj.pdf/burstein03criterion.pdf, 2003/2006-03-20.

[19] Foltz, P. W., Gilliam, S & Kendall, S. A. Supporting content-based feedback in online writing evaluation with LSA [J]. *Interactive Learning Environments*, 2000, 8(2): 111-129.

[20] Burstein, J. The e-rater scoring engine: Automated essay scoring with natural language processing [A]. In M. D. Shermis & J. Burstein (ed.). Automated essay scoring: A cross-disciplinary perspective[C]. Mahwah, NJ: Lawrence Erlbaum Associates, 2003. 113-122.