《新视野大学英语》作文自动评分系统的效度研究

王莺莺
(湖南科技大学 外国语学院,湖南 湘潭 411201)

摘 要:本实验检验了《新视野大学英语》作文自动评分系统的评分效度。实验数据表明:系统自动评分与人工评分之间的相关性较显著;系统评分中高分档的评分准确率较低,其他档的评分准确率较高;内容板块较语言板块、篇章结构板块对作文总体评分的影响显著。系统自动评分与适量的人工参与相结合,系统整体评分与细化的文本特征项相结合有助于提高系统的评分效度。

关键词:《新视野大学英语》;作文自动评分系统;效度
中图分类号:H319 文献标识码:A 文章编号:1674-5884(2012)12-0139-04

一 研究问题

(一)研究背景

AES(Automated Essay Scoring)研究自上世纪60年代以来,在国外取得了较大的进展,很多理论模型得以应用。从最初的作文自动评分系统PEG(Project Essay Grader)到1997年研发的IEA(Intelligent Essay Assessor),到1999年开始用于GMAT考试作文评分和2005年开始用于托福考试作文评分的E-Rater(Electronic Essay Rater),再到能够评阅多种语言文本的IntelliMetric和BETSY,作文自动评分系统始终在不断地更新和进步,力求更符合语言测试的要求。

从以上几种国外主流的作文自动评分系统来看,它们采用的样本都是美国学生的英语作文,能较准确地测试美国学生的英语写作水平,但对于非本族语学生,尤其是低水平英语学习者,"自动作文评分与人工评分会出现统计上的显著性差异"。因为以英语为母语的作文中,绝大多数句子都不存在严重的语法错误,而低水平英语学习者的作文中,有可能充斥着各种句法错误。此外,以上几种国外主流的作文自动评分系统都适用于大规模语言测试,针对任何一次测试,各系统都必须预先接受"训练集"的反复训练,此"训练集"通常"需要200甚至300篇以上已评分的作文作为训练语料"。由于"训练集"对样本作文的需求大,因此这些作文自动评分系统不适用于小规模语言测试,尤其不适用于自我测试。而使用效度较高的作文自动评分系统进行自我测试,根据系统提供的实时评分和反馈修改作文,是有效提高学习者英语写作水平的重要途径之一。同时,它能为大学英语低年级学习者提供基于网络的写作环境,在提高学习者英语写作水平的同时提高他们对大学英语4、6级网考的适应度,并能在一定程度上缓解因大学英语教师的严重短缺而引起的写作教学严重不足的现状。这就为AES系统在不断更新、完善大规模语言测试功能的同时提出了另一个应用目标,即提供即时的写作反馈以指导写作。

在这一研究领域,国内的外语教学与研究出版社做出了有益的尝试。2002年,它开发了《新视野大学英语》配套网络课程,为英语学习者提供了资源丰富的在线学习平台。"Write on"作文自动评分系统是新视野在线学习平台内的作文测评工具,它采用大学英语4、6级写作评分标准,将分值范围设定为1-15分,能够对任何题目的英语作文进行自动评分、计算单词总数并给出评语。2008年,美国著名的教育测评与研究机构CTB/McGraw-Hill开发了Writing Roadmap这一在线英语写作自动评分系统。它能从6个维度(思想内容、组织架构、文体、词汇选择、语言流畅程度和语言基本功)对作文进行分析、评分并给出评语。该系统的主要特色是它作为一种形成性评价工具,能够自动生成地区、学校和班级报告,便于教师和教学管理者及时了解写作教学效果,也便于他们利用此分析报告进行教学科研分析。2009年,浙江大学外语学院与杭州增慧网络

* 收稿日期:2012-09-15
作者简介:王莺莺(1979-),女,湖南益阳人,讲师,硕士,主要从事外语课程与教学论、计算语言学研究。

科技有限公司联合开发了“冰果英语智能作文评阅系统”。该系统利用最新的服务器处理芯片的大规模数据寻址及计算能力,结合文本语境处理、词法分析、句法分析、语义分析以及篇章分析等分析模块,能够对英语作文做出即时评分,还能从词汇、语法、文风、内容等方面给出反馈意见。该系统的主要特色为教师可以在机器评阅的基础上加以人工批改或进行班级点评。从上述几种适用于小规模语言测试和自我测试的作文自动评分系统来看,它们有着各自不同的特点,因而拥有各自的适用人群。《新视野大学英语》作文自动评分系统是专门为大学生开发的,它适合高等院校的大学英语学习者使用。Writing Roadmap 和“冰果英语智能作文评阅系统”的适用人群较广,包括中小学生、大学生和其他英语学习者。此外,上述作文自动评分系统在使用的准入方面存在差异。《新视野大学英语》作文自动评分系统属于《新视野大学英语》教材的配套网络课程,教材的使用者通过电子邮件获取账号和密码后即可免费使用该系统。Writing Roadmap 可以免费在线试用,长期使用则需付费购买。“冰果英语智能作文评阅系统”需要校方或使用者购买使用,且必须在局域网中运行。比较而言,《新视野大学英语》作文自动评分系统作为一种简单、便捷、经济的学习评估工具,更适合高等院校的大学英语学习者使用。

(二)研究问题

效度是语言测试关注的首要问题。《新视野大学英语》作文自动评分系统作为小规模语言测试和自我测试的适用模型,能否较准确地反映学习者的英语写作水平,关系到其能否取代传统的人工评阅,以实现计算机的工作效率最大化;同时也关系到它能否利用即时评分和反馈指导写作,成为学习者有效提高英语写作水平的学习辅助工具。影响作文自动评分系统效度的因素很多,如其工作原理和各分析模块的主要参数等等,限于篇幅,本文不作详述。检验作文自动评分系统效度的维度也有很多,如系统的自动评分是否与人工评分较为近似,它们之间的相关性是否显著,系统的效标关联效度如何,等等。本文主要从以下几个维度检验《新视野大学英语》作文自动评分系统(以下简称系统)的效度:

(1)系统自动评分与人工评分的相关性是否显著?

(2)系统自动评分中各分数档的精确率和误判率各是多少?

(3)系统所给评语的效标关联效度如何?

(4)系统所给评语中各版块是否对作文总体评分具有预测力?

问题(1)(2)侧重检验系统所给分值的效度。问题(3)(4)侧重检验系统所给评语的效度。

二 实验设计

(一)作文语料的提取与处理

从本校的大学英语第4册期末考试试卷库中随机抽取作文语料200份(其中文科试卷70份,理科、工科试卷各65份),编号并记录原始评分(分值范围为1-15分)。挑选有多年大学英语写作教学经验的教师4人,按照大学英语4级考试作文评分标准对上述200份作文进行重新评阅(分值范围为1-15分)。为消除原始评分对评阅人的心理暗示,我们隐去了200份作文语料的原始评分。重新评阅后的分值与原始评分相同的,作为该作文的最后得分。重新评阅后的分值与原始评分不同的,由其他3位教师复评,取4次评分的平均值(此平均值为小数点后一位四舍五入得到的整数)作为该作文的最后得分。按编号记录人工阅卷的最终评分。

由于部分单词拼写错误将严重影响系统对文章的理解,从而影响作文的总体评分,因此我们将作文语料输入自动评分系统后,利用系统配备的拼写检查工具对这些错误进行了人工改正,之后才提交给系统进行自动评分。按编号记录系统给出的评分和评语。

(二)数据统计与分析

1. 系统自动评分与人工评分的相关系数

表1显示了系统自动评分和人工评分的分数分布情况。由此表可知,人工评分较系统自动评分更集中在分数的中段(7、8、9分);系统自动评分的离散程度较人工评分的离散程度高;系统自动评分与人工评分的低段分一致,高段分明显多于人工评分。使用Pearson工具对系统自动评分与人工评分进行内部相关性检验,得到系统自动评分与人工评分之间的相关系数为0.62,表明系统自动评分与人工评分之间的相关性较显著,2种评分系统中的分数分布情况对二者的相关性具有一定的解释力。

表1 系统自动评分和人工评分的分数分布

作文分数	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
系统自动评分/%	1	5	6	6	8	10	9	8	9	7	8	9	7	5	2
人工评分/%	1	5	6	4	5	6	14	14	13	10	8	9	3	2	0

2. 系统自动评分中各分数档的精确率和误判率

将作文总分15分分为5个等距的等级(即2分、5分、8分、11分、14分)。按分数档统计系统自动评分的精确率和误判率。其计算公式如下:

各分数档的精确率 = 本为X档作文且被评为X档作文的数量 ÷ 所有被评为X档作文的数量 × 100%

各分数档的误判率 = 本为X档作文却未被评为X档作文的数量 ÷ 所有被评为X档作文的数量 × 100%

各分数档的评分精确率越高,说明作文被评为该分数档的可信度越高,系统自动评分的效度也越高。各分数档

的误判率越低，说明系统自动评分的误差越小，评分的效度越高。表2列出了系统自动评分中各分数档的精确率和误判率。

表2　系统自动评分中各分数档的精确率和误判率

分数档	2分档	5分档	8分档	11分档	14分档
精确率/%	100	72.2	70.6	73.3	39.1
误判率/%	100	27.8	29.4	26.7	60.9

由表2可知，2分档的准确率最高，5分档、8分档、11分档的准确率较高，14分档的准确率最低。作者对2分档和14分档的作文语料分别进行了核查，发现系统自动评分为2分档的人工评分也均为2分档，而系统评分为14分档的有可能与人工评分相差一个乃至多个分数档。14分档误判的文章一般篇幅较长，而语句并非与文章主题紧密相关，系统因为文章篇幅的关系容易受到“欺骗”，这也印证了其他研究者已指出的自动评分系统的不足之处，如有学生“先写几个段落，然后简单地重复”以“骗取高分”。

3. 系统所给评语的效标关联效度

采用已经比较成熟的大学英语4级考试作文评分标准为效标，对系统所给的作文评语与4级考试作文评分标准进行相关性分析，得到它们的皮尔森相关系数，根据相关是否显著判断效度高低。大学英语4级考试作文评分标准将总分15分划分为5个等级，每一个等级从内容、语言和篇章结构3个方面都有具体的要求和描述。对这些具体的要求和描述(即评分细则)用表3的形式分别列出，并统计了系统所给评语与大学英语4级考试作文评分细则的相关系数。

表3　系统所给评语的效标关联效度

内容				语言				篇章结构	
文章切题	论点明确	文字连贯	阐述透彻	用词准确	词汇丰富	句式变化	语法正确	结构合理	条理清晰
0.78	0.57	0.72	0.51	0.52	0.42	0.37	0.52	0.45	0.32

由表3可知，系统所给评语在内容方面与大学英语4级考试作文评分细则相关较显著，在语言方面与大学英语4级考试作文评分细则相关较弱，在篇章结构方面与大学英语4级考试作文评分细则相关最不显著。因此从系统所给评语的效标关联效度来看，系统在自动评分过程中，较为关注文章的内容和语言，篇章结构不作为主要的评分依据。

4. 系统所给评语中各版块对作文总体评分的预测力

将200份作文语料的评语分3个板块(即内容、语言、篇章结构)与作文总体评分进行了比对分析，结果显示：系统评分为高段分(13－15分)的28篇作文语料中，有24篇作文评语含有“文章切题(to the point)”，有19篇作文评语含有“文字连贯(coherent)”，有12篇作文评语含有“用词准确(accurate wording)”，有7篇作文评语含有“结构合理(well－organized)”。系统评分为低段分(1－3分)的作文语料共有24篇，它们的评语基本一致，大多为“不符合四级写作要求(not meet CET requirements on writing)”或“字数不足(less than 100 words)”。从统计结果来看，系统评分为高段分的作文较低段分的作文评语更具体、更清晰地体现了各版块对作文总体评分的权重。“文章切题”和“文字连贯”均为衡量文章内容的标准，它们在高段分作文评语中出现的频率分别为85.7%和67.9%，因此，内容板块对作文总体评分的影响力最大。“用词准确”作为衡量文章语言的标准之一，在高段分作文评语中出现的频率为42.9%，因此，语言板块对作文总体评分的影响力较大。“结构合理”作为衡量文章篇章结构的标准之一，在高段分作文评语中出现的频率为25%，因此，篇章结构板块对作文总体评分的影响力较小。

(三)实验结果与讨论

1. 实验的局限性和不足

首先，作文语料的选取有一定的局限性。由于200份作文语料均取自同一所大学的大学英语期末考试试卷库，因此实验结果能较准确地反映《新视野大学英语》作文自动评分系统对某一地区或学校英语学习者英语作文的评分效度，但可能不具有广泛的代表性。

其次，作文语料的人工评分可能存在信度和效度问题。在本实验中，200份作文语料由4名有多年大学英语写作教学经验的教师评阅，最终的人工评分多为4名评阅人所给分值的平均值。此方法虽然较仅由一人评阅的方法更为科学、客观，但也不排除评阅人因受到“参与某种研究而非真实阅卷”的心理暗示而影响评阅结果的可能，因此，人工评分部分仍然可能存在信度和效度问题。

最后，实验选取的效标本身具有一定的争议。在验证系统所给评语的效标关联效度时，本实验采用的效标是大学英语4级考试作文评分标准。此评分标准自身亦处于不断完善之中，在某些方面仍存在一定的问题，如有学者指出大学英语4级作文评分标准“不够详细具体，对写作内容和结构的要求过于笼统”等等。

尽管在上述方面存在一定的局限性和不足，本实验对《新视野大学英语》作文自动评分系统进行了有效的验证，并得到了以下较有意义的结论：系统自动评分与人工评分之间的相关性较显著，2种评分系统中的分数分布情况对二者的相关性具有一定的解释力；系统评分中高分档的评分准确率较低，其他档的评分准确率较高；系统在自动评分过程中，较为关注文章的内容和语言，篇章结构不作为主要的评分依据；内容板块对作文总体评分的影响力最大，语言板块对作文总体评分的影响力较大，篇章结构板块对作文总体评分的影响力较小。

2. 提高系统评分效度的途径

系统自动评分与适量的人工评分相结合。实验数据表明：系统评分中2分档的评分准确率最高，5分档、8分档、11分档的评分准确率较高，14分档的评分准确率最低。也就是说，系统评为低分的作文一定是低分作文，系统评为高分的作文则不一定是高分作文。因此，高分档作文的

评阅需要一定的人工参与。这一点系统应在使用指南中明确提示使用者。这样，使用者提交作文，得到反馈，经过自我判断之后，就可以根据系统的提示将“疑似高分”的作文提交给教师或系统管理员复核。系统自动评分一旦有了适量的人工参与，就像流水线上又多了一位质检员，给评分的效度增加一份保障。

系统整体评分与细化的语言特征项评分相结合。系统评分应该是一个对作文文本多次扫描分析的过程，既包含对作文语言质量、内容质量和篇章结构质量的整体把握，也包含对细化的各文本特征项的统计处理，因此所取的分值应为二者的均值。在整体把握文本语言质量、内容质量和篇章结构质量方面，系统可以采用“文本聚类方法”，以识别跑题作文，实现对文本内容的基本测量。另外，“将信息检索技术应用于作文内容的自动分析是一种可行的方法”。系统还需要配备一个功能强大的句法、词法和语义规则库，以实现对文本语言质量和篇章结构质量的总体评估。在细化文本特征项方面，必须借鉴自然语言处理技术，挖掘对文本的语言质量和篇章结构质量具有解释力的变量，并将这些文本特征项的统计学特征列入系统的主要工作参数。以文本的语言质量评估为例，系统的主要工作参数中应包含词汇、句型、语法、拼写和标点等一级指标的数据。将一级指标进一步细化，如词汇可细化为词频高低、词的搭配及恰当性、文章总词数、词的平均音节数、唯一词数等二级指标。这样，系统将作文文本的“总体印象得分”和“分项得分”综合起来，得到一个较为准确、客观的分值。

三　结　语

实验证明，《新视野大学英语》作文自动评分系统作为一种适用于小规模语言测试和自我测试的通用评分模型，具有较高的测试效度。对照人工评分，对系统的评分模型进行反复训练，能有效提高系统的评分效度。统计技术、自然语言处理技术和信息检索技术的进一步发展，将提高自动评分系统的评分效度，同时推进写作评分的自动化进程。

参考文献：

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

[2] Hearst M. The debate on automated essay grading [J]. IEEE Intelligent Systems, 2000, 15(5).

[3] Kukich K. Beyond Automated Essay Scoring [J]. IEEE Intelligent Systems, 2000(5).

[4] Weigle S C. Assessing Writing [M]. Cambridge: CUP, 2002.

[5] 葛诗利，陈潇潇. 国外自动作文评分技术研究 [J]. 外语电化教学，2007(5).

[6] 葛诗利，陈潇潇. 中国 EFL 学习者自动作文评分探索 [J]. 外语界，2007(5).

[7] 谢贤春. 英语作文自动评分及其效度、信度与可操作性探讨 [J]. 江西师范大学学报(哲学社会科学版)，2010(2).

[8] 蒋春丽，张青妹. 基于语料库软件的大学英语写作评估量表的设计 [J]. 语文学刊，2010(1).

[9] 谢贤春. 英语作文自动评分及其效度、信度与可操作性探讨 [J]. 江西师范大学学报(哲学社会科学版)，2010(2).

[10] 葛诗利，陈潇潇. 文本聚类在大学英语作文自动评分中应用 [J]. 计算机工程与应用，2009，45(6).

[11] 梁茂成，文秋芳. 国外作文自动评分系统评述及启示 [J]. 外语电化教学，2007(5).

(责任编校　杨凤娥)