

国外作文自动评分系统评述及启示

梁茂成, 文秋芳

(北京外国语大学 中国外语教育研究中心, 北京 100089)

摘 要: 本文依据语言测试领域的作文评分要素, 对国外具有代表性的三种作文自动评分系统进行评介和比较, 指出这些评分系统在训练及作文的人工评分方法和机器评分效度等方面存在的问题, 并分析这些作文自动评分系统为我国自主开发作文自动评分系统所提供的借鉴作用。

关键词: 作文自动评分; 模型; 评分要素; 信度; 效度

中图分类号: H319.3

文献标识码: A

文章编号: 1001-5795(2007)10-0018-0007

作文是大规模语言考试(如 TOEFL, GRE, IELTS 等)中的一种必备题型。通过作文可以检测应试者综合运用语言的能力。然而, 大规模作文阅卷面临两大难题: 其一, 阅卷需要耗费大量人力、物力等资源; 其二, 评判作文质量具有很强的主观性, 阅卷的信度(reliability)和效度(validity)不强(Johnson et al, 1991)。近几十年来, 随着计算机硬件和软件性能快速提高, 自然语言处理等技术获得了长足的发展, 国外一批作文自动评分系统相继问世, 这两个长期困扰大规模作文阅卷的难题有望得到解决。

本文对国外最具代表性的三种作文自动评分系统进行述评。这三种系统是: PEG(Project Essay Grade)、IEA(Intelligent Essay Assessor)和 E-rater。PEG 重语言形式, IEA 重作文内容, E-rater 则既重形式又重内容。一般说来, 作文评分应形式和内容并重, 围绕作文的语言质量、内容质量和篇章结构质量三个主要方面进行(Blok & de Glopper, 1992; Purves, 1985; Weigle, 2002; 梁茂成, 2005), 然而这三种系统侧重各有不同, 在这三个方面的分析力度也存在很大差异。通过对比分析, 笔者力图揭示这些作文自动评分系统的优势与劣势, 以期对开发我国学生作文自动评分系统有所启示。

1 国外作文自动评分系统述评

1.1 PEG, 一个重语言形式的评分系统

PEG 于 1966 年由美国杜克大学(University of Duke)的 Ellis Page 等人开发(Daigon, 1966; Page, 1966)。PEG 的设计者们认为, 计算机程序没有必要理解作文内容, 大规模考试中尤其如此(Shemis et al, 2001)。因此, 他们在其网站上公开申明: “PEG 不能理解作文的内容”(http://134.68.49.185/PEG-DEMO/)。

在 PEG 的开发者看来, 作文质量的诸要素是作文的内在因素, 无法直接测量, 因此, 最为合乎逻辑的方法是从作文文本中提取一些能够间接反映作文质量的文本表层特征项(surface features)。Page 首先收集了一部分人工评分后的作文(训练集), 利用当时并不发达的自然语言处理技术, 从作文中提取若干个文本特征项(text features), 并在这些文本特征项与人工评分之间进行相关性分析。然后, Page 选择与人工评分相关的文本特征项作为自变量, 把人工评分作为因变量, 进行多元回归分析, 得到回归方程。回归方程为每一个变量确定了 beta 值, 这样, 在对新的作文进行评分时, PEG 只需要提取这些变量, 并把 beta 值代入回归

作者简介: 梁茂成, 男, 博士, 教授。研究方向: 应用语言学及计算语言学。

文秋芳, 女, 教授。研究方向: 语言学及应用语言学。

收稿日期: 2007-01-15

基金项目: 本研究得到教育部人文社科项目(编号 06JA740007)和中国外语教育研究中心重大研究项目的资助, 在此一并致谢。

方程之中,就可以预测出这些作文的得分。经过开发者的多年努力,PEG“目前的程序中使用很多复杂的变量”(Page & Peterson, 1995)。遗憾的是,对这些所谓“复杂的变量”,PEG的开发者们没有详细报告,展现给我们的仅是一个“黑匣子”(Kaplan et al, 1998)。比较 Page和 Peterson(1995)和 Page(1968)的变量列表,他们的确增加了几个冠以代码的新变量,其中部分变量可能的确比早期的更加复杂,但 PEG的理论基础和工作原理没有发生根本的变化(Page, 1994)。

由此看来,PEG实现作文自动评分有三个步骤:变量提取、多元回归分析和把多元回归得到的 beta值代入计算机程序换算出作文得分(Chung & O'Neil, 1997; Page, 1994)。

概括起来,PEG的技术大体包括两方面:其一,PEG使用的统计方法是多元线性回归,以此来确定各变量的 beta值,这样,基于训练集作文而构建的统计模型便可以用来为新的作文进行自动评分。这一技术合理而容易理解,后期出现的作文自动评分系统大多采用这一技术。其二,自然语言处理技术是 PEG提取变量的主要方法。基于这两种技术,PEG取得了很好的评分效果。有关 PEG的技术报告中申明,PEG的评分结果与人工评分结果十分一致。据 Page(1994)和 Page和 Peterson(1995)报告,PEG在 1994年的实验中取得了很好的结果,多元回归系数(multiple R)达到了 $R = 0.877$ 。

1.2 IEA,一个重内容的评分系统

IEA(Intelligent Essay Assessor)是一种基于潜伏语义分析(Latent Semantic Analysis)的作文自动评分系统,由美国科罗拉多大学的 Thomas Landauer等学者开发。与 PEG显著不同的是,IEA的设计者们在其网站上申明:“IEA是唯一能够测量语义和作文内容的程序”(http://lsa.colorado.edu)。据 IEA的设计者们报告,潜伏语义分析主要分析文本的内容和学生作文中所传达的知识,而不是作文的风格或语言(Foltz et al, 1998)。

根据 Landauer和 Dumais(1997)的描述,潜伏语义分析既是一种理论,也是一种技术。这种理论认为,在文本中隐藏着一个潜在的语义结构(semantic structure),这一潜在的语义结构正是所有词汇(潜伏语义分析称之为“词汇项”,即 terms)的语义之和(Dumais et al, 1982)。然而,由于自然语言中存在大量的多词同义(synonymy)和一词多义(polysemy)现象,常常使得语义结构带有大量的干扰信息(noise)。从技术

的角度看,潜伏语义分析是一种矢量空间模型(Vector Space Model, VSM)技术,但与一般的矢量空间模型相比,这种技术能够通过减少维数(dimensionality)的方法(Landauer & Dumais, 1997; 桂诗春, 2003),有效地过滤干扰信息,提取数据中的潜在语义结构(Chung & O'Neil, 1997)。潜在语义结构可以通过一个词汇项-文档矩阵(term-by-document matrix)来代表,矩阵中的每一行代表一个词汇项,每一列代表一个文档,而矩阵的每一个单元格中填入对应词汇项在对应文档中出现的频数。

使用矩阵代表潜在语义结构的好处,是对矩阵可以进行一种称之为奇异值分解(Singular Value Decomposition, SVD)的矩阵运算,通过该运算,原来的矩阵可以被分解成为三个不同的矩阵。减少维数后,对这三个矩阵进行进一步的运算,可以重建一个矩阵。重建后的矩阵因为使用了较少的维数,与原来的矩阵相比,可以更好地代表潜在语义结构。该矩阵保持了原来矩阵中最重要的语义联想关系(association patterns),同时又排除了大量的干扰信息(Deerwester et al, 1990; Dumais et al, 1982; Landauer et al, 1998)。

将潜伏语义分析用于学生作文自动评分时,待评分的作文与预先选定的范文(训练集)被视作为矢量,对矢量进行比较之后,可以得到每一篇待评分作文与范文在内容上的相似度得分(similarity score)。该得分被直接视为机器评分或经过转换后得到机器评分(Foltz et al, 1999)。根据 Landauer et al (2000)的报告,该系统所评出的作文得分与人工评分之间的相关性达到 $r = 0.85$ 。

1.3 E-rater,一个模块结构的混合评分系统

E-rater是由美国教育考试处(Educational Testing Service, ETS)于 20世纪 90年代开发,其目的是评估 GMAT考试中的作文质量(Burstein et al, 1998a; 1998b; 1998c)。据 Burstein et al (2001)、Cohen et al (2003)和 Valenti et al (2003)的描述,E-rater自 1999年以来已经进入操作阶段,至 2003年,共评定作文 750,000篇。

E-rater的开发者们声称,他们的作文评分系统利用了多种技术,其中包括统计技术、矢量空间模型技术和自然语言处理技术(Valenti et al 2003)。凭借这些技术,E-rater不光能够像 PEG那样评判作文的语言质量,还能够像 IEA那样评判作文的内容质量。除此之外,E-rater还对作文的篇章结构进行分析。

与 PEG相类似,E-rater的评分方法基于线性回归

表 1 三种作文评分系统比较

系统	分析重点	内核技术	长处	不足
PEG	语言	1) 统计技术 2) 自然语言处理	语言质量分析	1) 对内容不作分析 2) 只对文本的表层特征进行分析,因而易于被考生识破
IEA	内容	信息检索技术	内容质量分析	1) 不分析作文的语言质量 2) 对篇章结构不作分析
E-rater	1) 语言 2) 内容 3) 篇章结构	1) 统计技术 2) 自然语言处理 3) 信息检索技术	1) 三个模块与人工评分要素更为吻合 2) 包含了篇章结构分析模块	1) 对作文的内容质量分析不力 2) 对篇章结构的分析拘泥于文本的表层特征 3) 对语言质量的分析不够全面

模型 (Powers et al 2000)。E-rater围绕三个主要方面对作文的质量进行分析和评判,Burstein等人把这三个方面称作为模块 (Burstein & Marcu, 2000)。E-rater的第一个模块为话语 (discourse)结构 (亦即篇章结构)分析模块,主要靠在文本中搜索“ In summary ”、“ In conclusion ”等提示词 (cue words)的方法得以实现 (Burstein et al, 1998b)。E-rater的第二个模块为句法多样性 (syntactic variety)分析模块,根据作文中句子结构的多样性来评判作文的质量。显然,该模块的目的是分析作文的语言质量。E-rater的第三个模块为内容 (content)分析模块。在这一模块中,E-rater通过矢量空间模型,观察作文中是否包含了足够的与作文题目高度相关的主题词。

E-rater的三个模块中分别包括哪些变量,有关报告一直含糊其辞 (或许是出于商业方面的考虑)。Kukich (2000)报告说 E-rater可以提取 100 多个变量,Powers et al (2000)声称 E-rater可以提取 50 多个变量,而 Attali和 Burstein (2004: 5)的报告则说最新版本的 E-rater“依据有关理论对写作能力的若干方面给予考虑”,只提取 12 个变量。

综合以上介绍不难看出,E-rater在自动评分过程中力求兼顾作文的内容和语言形式,是一种混合的自动评分系统。根据 Burstein et al (2001)和 Valenti et al (2003)的研究报告,E-rater与人工评分之间的一致性^① (agreement)一直高于 97%。

1.4 现有作文自动评分系统比较

表 1总结了以上三种系统的分析重点、内核技术、长处和不足。

PEG只测量作文中的语言质量,不顾及作文内容和篇章结构,其评分效度显然值得质疑。由于 PEG的开发者坚信文本的形式特征对作文质量的解释能力,而计算机无法也没必要分析作文的内容质量,因此他们只运用自然语言处理技术提取变量,并通过多元回

归的统计技术预测作文的得分。随着近几十年来自然语言处理技术的发展,PEG近来版本中的变量提取技术可能更为复杂,但总体上说,PEG的作文评分仍旧完全依赖作文形式特征 (即语言)的提取。

IEA注重对作文内容的分析,利用信息检索中的潜伏语义分析法进行作文的自动评分。客观地说,这是作文自动评分的一个重大突破,但由于该方法只能对作文内容进行评价,置作文的语言质量于不顾,难免令人对评分效度质疑。据称 IEA近来也在试图从作文中提取一些体现语言质量的变量 (Landauer et al, 2003),如拼写错误等,但 IEA对作文语言质量的测量效果远远比不上其对作文内容质量的测量效果。IEA原本设计为评价本族语作文所用,因此重内容、轻形式的做法后果可能不那么严重,但如果用于评测外语学习者的作文,其信度和效度就很难达到完美。

与前两种系统相比,E-rater显然考虑到了更多的作文评分要素,因此更加符合写作测试的要求。从某种意义上来说,E-rater与 PEG十分相似,它们都使用自然语言处理技术从训练集作文文本中提取若干变量,并通过回归分析来确定所有变量对作文质量的预测能力。与此同时,E-rater与 IEA也存在一定的相似之处,两者都使用矢量空间模型技术对作文的内容质量进行分析。E-rater的独到之处在于,它没有像 PEG那样忽视对作文内容的分析,也没有像 IEA那样将作文的语言质量置之不顾。除此之外,E-rater对作文的篇章结构质量也给予了一定的考虑。由此可见,E-rater在评分过程中考虑到了作文质量的更多方面,更大程度地模拟了人工评分的过程,也更多地使用了其它学科的技术,因而与另外两种系统相比,评分信度和

① 现有的大部分机器评分系统的研究报告中一般都通过计算相关性的方法报告评分的信度,但 E-rater一直采用完全及相邻吻合百分率 (percent exact-plus-adjacent agreement)的方法报告评分的信度。

效度可能更高。

然而,E-rater的自动评分技术也并非无懈可击。首先,其语言质量分析模块对语言质量的若干方面的分析显然不够全面。众所周知,对学生作文中语言质量的分析应该包括词汇、句法、语言的准确性等多个方面,而E-rater对语言质量的分析主要考虑的只是作文中的句法多样性,这势必会影响机器评分的效度。其次,与IEA相比,E-rater的内容质量分析模块显然尚有提高的余地。E-rater与IEA所使用的矢量空间模型技术的不同之处,在于两者同是信息检索技术,但前者是一种基于主题词分析的技术,而后者使用的潜伏语义分析法则是一种降维(dimensionality reduction)技术(参见Chung & O'Neil, 1997)。若干研究(如Deerwester et al, 1990; Dumais et al, 1982; Foltz et al, 1998; Landauer et al, 1998等)表明,潜伏语义分析法的降维技术可以有效地去除文本中的干扰信息,对处理多词同义和一词多义具有良好的效果。据称,为了提高搜索效果,大名鼎鼎的搜索引擎Google都使用了潜伏语义分析法(<http://www.searchenginejournal.com/index.php?p=1296>),E-rater对作文内容的分析方法未免略显落伍。我们认为,IEA对作文内容的分析方法值得借鉴。第三,E-rater的篇章结构分析模块靠搜索作文中的In conclusion等话语标记语,容易被考生识破,导致不利的反拨作用(washback effect)。

综合以上分析,PEG的语言分析技术值得弘扬,IEA以内容分析技术见长,而E-rater的模块分析方法更符合语言测试的要求,但三种系统同时也存在各自的不足之处。

2 国外作文自动评分系统的启示

作文自动评分是一种十分复杂的技术,需要合理利用多学科的技术,结合语言测试理论,才能达到理想的效果。对以上三种作文自动评分系统的分析和比较为我们开发自主的作文自动评分系统提供了以下启示:

2.1 应最大限度地提高训练集作文人工评分的信度

作文自动评分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。因此,在对计算机评分模型进行训练时,训练集作文人工评分的信度至关重要。只有有效地模拟具有较高信度的人工评分,计算机评分才有意义。

根据Barrett(2001)和Stemler(2004)的研究,评分员间的信度达到 $r=0.70$ 左右才是可以接受的,但现

有作文评分系统在对训练集作文进行人工评分时常常常达不到这样的信度要求,可能使得计算机评分模型很难模拟到人工评分的精髓。

根据Page(2003)的研究报告,在PEG最早的一次实验中,4名人工评分员的评分信度介于 $r=0.44$ 和 $r=0.57$ 之间,平均值仅为 $r=0.547$;在PEG 1994年的实验中,5名人工评分员的评分信度介于 $r=0.389$ 和 $r=0.581$ 之间,平均值仅为 $r=0.489$;即便是在信度最高的1995年的实验中,5名评分员的评分信度也只介于 $r=0.550$ 和 $r=0.748$ 之间,信度平均值也只达到 $r=0.647$ 。可见,尽管PEG的评分与人工评分之间具有较高的一致性,其训练集作文的人工评分信度明显偏低。在IEA所进行的几次实验中,所评分的作文并非来源于学生的语言测试,而主要是以英语为本族语的学生历史学、心理学等学科的论文(Landauer et al, 2003),人工评分也主要以论文的内容是否准确为依据,且研究者对评分过程并未作任何说明。E-rater的评分依据是ETS相关考试中的评分量表,但因其信度报告不采用传统的相关性分析,而使用容易夸大信度(Stemler, 2004)的完全及相邻吻合百分率,故而很难用统计学方法衡量其评分信度的优劣。但根据Page(1994)年的报告,ETS考试中,人工评分员间的信度一般介于 $r=0.50$ 和 $r=0.60$ 之间,显然也并不高。

为了使得计算机评分模型能够更好地模拟人工评分,我们有必要在训练集作文的人工评分方面多下功夫。语言测试领域的作文评分方法主要有整体评分(holistic scoring)和分析型评分(analytical scoring)两种,后者虽耗时费力,但更有利于提高评分信度(Weigle, 2002)。根据以上三种评分系统的研究报告,在对这些评分系统的评分模型进行训练的过程中,系统所模拟的人工评分大多并非出自于分析型评分。我们认为,分析型评分虽然耗时费力,但若组织为数不多的几名资深评分员采用分析型评分方法对训练集作文进行精细评分,以相对较小的投入对计算机评分模型加以训练,换取大规模考试中较高的评分信度,理所当然值得的。

2.2 机器评分模型的模块结构应与测试学理论相吻合,以提高机器评分的效度

评价对学生作文的评分是否合理,所需考察的另一个方面是评分的效度(Bachman, 1990; McNamera, 1996)。如上文所述,对作文进行评分一般至少需要从作文的语言质量、内容质量和篇章结构质量三个主要方面对作文的整体质量加以衡量。

以上三种作文自动评分系统在评分过程中并未能很好地兼顾这三个主要方面,因而评分的结构效度(construct validity)值得质疑。PEG虽然对作文的语言质量有着较强的分析能力,但忽略了作文的内容质量和篇章结构质量,因而其评分结果存在较大的效度问题。与此相类似,IEA突出了评分过程中作文内容的重要性,但忽略了作文的语言质量和篇章结构质量,显然也存在较大的效度问题。与这两种系统相比,E-rater虽然以其模块结构兼顾了作文质量的三个主要方面,但每个模块的分析能力尚可进一步提高。

使用计算机对学生作文进行自动评分,应该最大限度地模拟人工评分过程,考虑作文的语言质量、内容质量和篇章结构质量等评分要素,同时对这些要素进行周密的细化,从作文文本中提取最能够体现这些方面的文本特征项,以充分提高机器评分的效度。

2.3 评分模型中的内核技术问题

内核技术是机器评分模型能否有效预测作文整体质量的关键。上文评述的三种作文自动评分系统的设计都利用了多种现代技术,主要包括统计技术、自然语言处理技术和信息检索技术;但由于所使用的技术及其成熟程度不同,各系统的分析能力也因此存在较大差异,对作文评分结果必然产生较大影响。在构建我国自主的作文评分模型时,应该对以上三个系统的合理技术进行充分利用,同时摒弃其中不合理的部分。

多元回归的统计技术已经成为作文自动评分系统中的基本技术。从作文文本中提取多个文本特征项作为自变量,以人工评分作为因变量,通过多元回归分析的方法为待评分作文进行自动评分,这种方法直接、易于理解且便于操作(Chung & O'Neil, 1997)。PEG和E-rater从问世以来一直使用这种统计技术,且IEA的最新发展(Landauer et al, 2003)表明其设计者们也正在考虑在其产品的技术内核中融入多元回归的统计技术。可见,多元回归的统计技术已经成为作文自动评分中的基本技术,在构建我国自主的学生作文评分系统时应可以借鉴。

能否合理利用自然语言处理技术,从学生作文文本中挖掘对作文的语言质量和篇章结构质量具有解释力的变量,此项技术利用得好坏关系到作文自动评分系统的成败,需要做很多细致的工作。自然语言处理技术的最新发展为提高作文自动评分模型对作文质量的预测能力提供了有力的技术保障。正是由于自然语言处理技术对作文自动评分系统至关重要,PEG以自然语言处理技术作为立足之本,E-rater也以自然语言

处理技术作为其提取变量的重要途径,而且两种系统都取得了令人鼓舞的效果。然而,由于现有作文评分系统中的大部分变量对外保密,我们要开发自己的作文自动评分系统,就有必要从语言测试理论出发,利用自然语言处理技术反复尝试多种文本特征项,力争挖掘出能更直接地反映作文水平的变量(Kukich, 2000)。

作文自动评分中利用信息检索技术的主要目的是为了分析学生作文的内容质量。从以上三种评分系统看,PEG完全忽略了对作文内容的分析,因而其评分效度受到了学界的质疑(Chung & O'Neil, 1997)。IEA和E-rater在分析作文内容时都使用了信息检索技术,所不同的是E-rater的内容分析技术基于作文中的主题词,而IEA利用了潜伏语义分析法,有效地解决了同义词问题和一次多义问题,从而极大地提高了作文内容的分析效果(Chung & O'Neil, 1997)。从已有的研究看,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,但基于主题词的内容分析法有着显而易见的弱点。为了最大限度地提高作文内容分析的效果,作文评分系统中有必要融入最先进的信息检索技术。

因为自动分词的准确性是汉语自然语言处理中的瓶颈问题之一,对汉语作文进行自动评分,变量可能更难以挖掘,这需要国内自然语言处理界同仁的不断努力。由于英语不需要自动分词,而英语作文自动评分系统的开发对提高我国大规模英语考试的效率,对减少资源消耗意义重大,因此我国开发自主的英语作文评分系统既具有较大的可行性,又有必要性。

3 结论

作文自动评分系统在对评分模型进行训练时,应该使用分析型评分方法以提高作文评分的信度,同时应充分结合语言测试领域的理论,围绕作文的语言质量、内容质量和篇章结构质量对作文进行人工评分,并以所得到的评分对自动评分模型进行训练,以提高机器评分的效度。

由于所使用的核心技术不同,国外现有的作文自动评分系统对作文质量诸方面的分析能力存在较大差异。一个合理的作文自动评分系统应该充分利用统计技术、自然语言处理技术、信息检索技术及其它可能利用的技术,从作文文本中挖掘能够直接反映作文质量的文本特征项作为变量,有效地提高评分模型对作文质量的预测能力。

分析国外现有作文自动评分系统的得与失,对开发我国自主的作文自动评分系统具有十分重要的意义。通过计算机对学生作文进行自动评分是一个复杂的过程,需要总结前人的经验并不断汲取新的理念、开发新的技术。只有这样,才能找到对学生作文最具预测力的变量,保证机器评分的信度和效度。

参 考 文 献

- [1] Attali, Y. and Burstein, J. Automated essay scoring with E-rater V. 2.0 [A]. Paper presented at the Conference of the International Association for Educational Assessment (IAEA), Philadelphia, June 13 - 18, 2004.
- [2] Bachman, L. F. Fundamental considerations in language testing [M]. Oxford and New York: Oxford University Press, 1990.
- [3] Blok, H., and de Glopper, K. 1992. Large scale writing assessment [A]. In L. Verhoeven and J. H. A. L. De Jong (eds). The construct of language proficiency [C]. Amsterdam/Philadelphia: John Benjamins, 1992: 101 - 111.
- [4] Burstein, J. C., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L. & Harris, M. D. Automated scoring using a hybrid feature identification technique [A]. In The Proceedings of the annual meeting of the Association of Computation [C], 1998a.
- [5] Burstein, J. C., Kukich, K., Wolff, S. E., Lu, C., & Chodorow, M. Enriching automated scoring using discourse marking [A]. Paper presented at the Workshop on Discourse Relations and Discourse Marking at the annual meeting of the Association, 1998b.
- [6] Burstein, J., Kukich, K., Braden-Harder, L., Chodorow, M., Hua, S. & Kaplan, B. Computer analysis of essay content for automatic score prediction: A prototype automated scoring system for GMAT analytical writing assessment [R]. (Research Report RR-98-15). Princeton, NJ: Educational Testing Service, 1998c.
- [7] Burstein, J. C., & Marcu, D., Andreyev, S., & Chodorow, M. Towards automatic classification of discourse elements in essays [A]. In Proceedings of the 39th annual meeting of the Association for Computational Linguistics [C], France, 2001: 90 - 92.
- [8] Chung, G., & O'Neil, H. Jr. Methodological approaches to online scoring of essays [R] (Report No. CSE-TR-461). Los Angeles, CA: University of California, Los Angeles, Center for the Study of Evaluation, 1997.
- [9] Cohen, Y., Ben-Simon, A. & Hovav, M. The effect of specific language features on the complexity of systems for automated essay scoring [C]. Paper presented at the IAEA 29th Annual Conference Manchester, UK, 2003.
- [10] Daigon, A. Computer grading of English composition [J]. English Journal 55. 1, 1966: 46 - 52.
- [11] Deerwester, S., Dumais, S. T., Fumas, G. W., Landauer, T. K., & Harshman, R. Indexing by Latent Semantic Analysis [J]. Journal of the American Society for Information Science, 41, 391 - 407. 1990.
- [12] Dumais, S., Fumas, G., Landauer, T., Deerwester, S. & Harshman, R. Using Latent Semantic Analysis to Improve Access to Textual Information [J]. Machine Studies, 1982, 17, 87 - 107.
- [13] Foltz, P. W., Kintsch, W. & Landauer, T. K. The measurement of textual coherence with Latent Semantic Analysis [J]. Discourse Processes 1998, 25, 285 - 308.
- [14] Foltz, P. W., Laham, D., & Landauer, T. K. The Intelligent Essay Assessor: Applications to Educational Technology [J]. Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1999, 1 (2).
- [15] Kaplan, R. M., Wolff, S. E., Burstein, J., Lu, C., Rock, D. A., & Kaplan, B. A. Scoring essays automatically using surface features [R]. (GRE Board Report No. 94 - 21P). Princeton, NJ: Educational Testing Service, 1998.
- [16] Kukich, K. 2000. Beyond automated essay scoring [A]. In Hearst, K. (eds), The Debate on Automated Essay Scoring IEEE Intelligent Systems [C], September/October, 2000.
- [17] Landauer, T. K., Foltz, P. W. & Laham, D. An introduction to Latent Semantic Analysis [J]. Discourse Processes, 1998, 25, 2&3, 259 - 284.
- [18] Landauer, T. K., Laham, D., Rehder, B. M. E. Schreiner. How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans [A]. In Shafit, M. G. & Langley, P. (Eds), Proceedings of the 18th international ACM SIGIR conference on research and development in information retrieval [C]. 1997.
- [19] Landauer, T. & Dumais, S. A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge [J]. Psychological Review, 1997, 104. 211 - 140.
- [20] Landauer, T. K., Laham, D. and Foltz, P. W. 2000. The Intelligent Essay Assessor [A]. In Hearst, K. (eds), The Debate on Automated Essay Scoring IEEE Intelligent Systems [C], September/October, 2000.
- [21] Landauer, T. K., Laham, D. & Foltz, P. W. Automated

- scoring and annotation of essays with the Intelligent Essay Assessor[A]. In Shemis, M. D. & Burstein, J. (eds). Automated Essay Scoring: A Cross-Disciplinary Perspective [C]. Lawrence Erlbaum Associates, Mahwah, NJ, 2003: 87 - 112
- [22] McNamara, T. Measuring Second Language Performance [M]. Addison Wesley Longman Limited: New York, 1996
- [23] Page, E. B. Grading essays by computer: Progress report [A]. In Educational Testing Service (Ed), Proceedings of the Invitational Conference on Testing Problems [C]. New York City: Princeton, NJ: Educational Testing Service, 1966: 87 - 10.
- [24] Page, E. B. The Use of the Computer in Analyzing Student Essays[J]. Int'l Rev Education, Vol 14, 1968: 210 - 225
- [25] Page, E. B. New computer grading of student prose, using modern concepts and software[J]. Journal of Experimental Education, 1994, 62 (2): 127 - 142
- [26] Page, E. & Peterson, N. S. The Computer Moves into Essay Scoring: Updating the Ancient Text [J]. Phi Delta Kappan March, 1995: 561 - 565.
- [27] Powers, D. E , Burstein, J. C , Chodorow, M. , Fowles, M. E , & Kukich K Comparing the validity of automated and human essay scoring[R] (GRE Board Research Report 98 - 08aR). Princeton, NJ: Educational Testing Service, 2000
- [28] Purves, A. C. In search of an internationally valid scheme for scoring compositions [J]. College Composition and Communication 1985, 35, 426 - 438
- [29] Shemis, M. , Mzumara, H. R. , Olson, J. and Harrington, S. On-line Grading of Student Essays: PEG goes on the World Wide Web [J]. Assessment & Evaluation in Higher Education, 2001, 26(3):.
- [30] Stemler, S. E. A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability[J]. Practical Assessment, Research & Evaluation, 2004, 9(4).
- [31] Valenti, S. , Neri, F. and Cucchiarelli, A. An overview of current research on automated essay grading[J]. Journal of Information Technology Education Volume 2, 2003.
- [32] Weigle, S. C. Assessing writing[M]. Cambridge University Press: Cambridge, 2002
- [33] 桂诗春. 潜伏语义分析的理论及其应用 [J]. 现代外语, 2003, (1).
- [34] 梁茂成. 中国学生英语作文自动评分模型的构建 [D]. 南京大学博士学位论文, 2005.

A Critical Review and Implications of Some Automated Essay Scoring Systems

LANG Mao-cheng, WEN Qiu-fang

(National Research Centre for Foreign Language Education, Beijing Foreign Studies University, Beijing 100089, China)

Abstract: This paper evaluates and compares three representative automated essay scoring systems against the major assessment criteria in essay scoring in the field of language testing. On the basis of the comparison and the evaluation, the paper argues that these systems have reliability problems with the human-assigned scores used for training their models, as well as validity problems with the scores they assign. The paper also analyzes the implications that the three systems offer. The study sheds important light on the development of an automated essay scoring system in China.

Key words: Automated Essay Scoring; Model; Assessment Criteria; Reliability; Validity

NewClass 语言学习系统 系列产品二

**你一半，我一半，
“二享一” 数字化新生代**

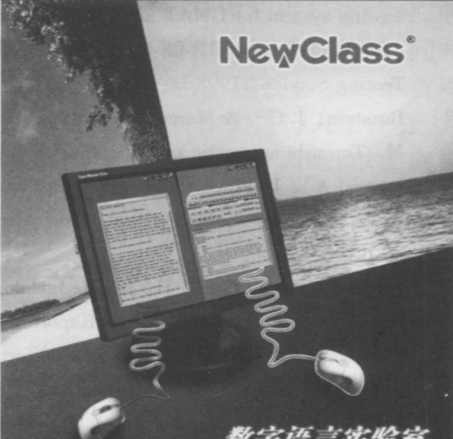
二享一终端型 DL500K

产品亮点：

二享一：两个学生共享一台显示器，全屏显示	音频点播，多频道广播
一分二：显示器屏幕一人一半，分屏使用	口语聊天室，数字双轨录音
一托二：一台学生终端联接两套鼠标、键盘	U盘下载，两人分屏AB卷考试

东方正龙 北京东方正龙数字技术有限公司 客服热线: (010)62969277
 电话: (010)51298899 传真: (010)62973888 E-mail: info@newclass.com.cn

NewClass®



数字语言实验室
当然 NewClass