英语自动作文评分方法及其可行性研究

● 杨有统

(云南财经大学, 云南 昆明 650221)

摘 要:为了检验中国学生英语作文自动评分模型的可行性,采用某高校英语专业高年级学生35篇作文为样本,探讨计算机提取语言使用特征和人工评分之间的关系。研究表明作文中语块的使用数量、类符数、作文长度四次方根和写作得分显著相关。

关键词:自动作文评分;二语写作;可行性

中图分类号:H319 **文献标识码**:A **文章编号**:1008-6765(2011)02-0145-03

一、引言

传统教学中写作被当作四项基本技能之一,一直在教学和测试中占据重要位置。Spolsky[1]回顾文献后发现,写作测试是近代科学建立以前语言测试的主要形式之一,另一项为翻译。近年来随着人们对交际能力以及交际语言测试的关注,写作在国内外各种考试中愈发占据重要地位:托福网考写作占总分25%[2],雅思考试写作共有两部分组成:看图说明或者书信和议论文[3]。国内大规模大学英语四、六级考试中,写作也一直是重要组成部分,英专四、八级和研究生英语入学考试中写作分别占到了25%、20%和30%。

通过作文可以检测应试者综合运用语言的能力,然而负担过重的英语教师已经无法抽身给每一位学生评分,再加上现今大规模考试耗费大量人力和物力,且阅卷信度和效度无法保证。应用基于人工智能开发出来的自动作文评分系统可以把教师解放出来,进而让学习者可以得到更多的训练,取得更快的进步。本文先回顾自动作文评分研究在国外的发展,接着介绍国内学者针对中国英语学习者所构建的自动作文评分模型,并对其可行性进行探讨。

二、自动作文评分系统研究现状

基于自然语言处理的自动作文评分系统已有40多年的历史,尤其在美国已付诸应用[4][5]。国内对自动作文评分系统的研究刚起步,多偏重于介绍[6][7][8][9][10]。也有学者对中国英语学习者作文自动评分系统研制做出了有益探索[11][12]。总的说来,国内研究和国外相比,还有一定差距。

自动作文评分研究的开拓者是美国杜克大学的Page,他在1966年设计出利用电脑基于统计技术提取文本特征的Project Essay Grader (PEG)系统。PEG采取了间接测量写作指标的方法,如:词长、文章长度、介词和生僻词使用情况等。1968年,Page发表了自动作文评分和人工评分的比较研究报告,多元回归系数R=0.78,取得了很好的结果[13]。

进入20世纪90年代,自然语言处理和信息检索技术的发展使得研究者可以寻找新的可以预测写作质量的手段。由美国科罗拉多大学的Thomas Landauer等学者开发的IEA (Intelligent Essay Assessor)是基于潜伏语义分析的自动作文评分系统的代表。IEA采取截然不同的评分方法,主要关注内容的评价,而不是诸如词数、句子数和关键词等语言特征。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。

这一时期的另一代表性系统是由Educational Testing Service (ETS)的Burstein等人开发的Electronic Essay Rater (E-rater),已经用于对GMAT中的Analytical Writing Assessment部分和托福考试的作文评分。该系统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。

1997年,Vantage Technologies开发出吸取认知处理、人工智能和计算语言学长处的IntelliMetric。它能分析词性和句法关系,主要评估了作文中语义、

收稿日期:2011-05-01

作者简介:杨有统(1978-),男,云南腾冲人,讲师,云南师范大学外国语学院2009级外国语言学及应用语言学硕士研究生。

句法和篇章三个层次的500多项特征,如话语、修辞、内容、句法和结构等。目前,IntelliMetric已被广泛应用于中学、大学以及其他一些机构。

三、中国学生英语自动作文评分模型的可行性研究

(一)中国学生英语自动作文评分模型

研究表明,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)都受到中国文化和汉语思维习惯的影响。英语作为母语的作文评分与英语作为外语的作文评分,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,其中最主要的差异就在于句法方面[8]。葛诗利[9]尝试开发了一种以语料库为基础,抽取浅层文本特征和语言错误为参数,采用多元线性回归、k近邻和支持向量机等方法建模,主要针对语言使用的大学英语作文自动评分系统模型。该研究表明,作为大规模考试自动评分的首选方法是线性回归法,因为该方法评分准确率最高。

梁茂成[12]在总结前人对写作质量测量指标的基础上,提出从语言、内容和组织三个方面考虑建立模型,利用计算机提取15项具体语言特征来预测作文质量。研究表明大多数指标能有效地测量中国学生英语写作质量。因为本族语和外语学习者作文写作最大差异在于语言的运用上而不是内容,因此本文将逐一介绍该模型中有关语言特征提取的方法。

1.语言流利性

语言流利性是指在规定的时间内写出或说出的形符数[14]。该模型采用作文长度的四次方根和类符数来计算。

2.词汇复杂性

词汇复杂性的计算:借助于Range软件自动计算出作文中基础词汇表1用词数和基础词汇表3用词数,然后用表3单词数除以表1单词数得出。

3.句法复杂性

句法复杂性由平均句长和作文中动名词出现次数来测量。为了计算出动名词用词数,需要借助Gotagger软件给语料标注,再用语料分析软件AntConc检索出。

4.语言使用得体性

语言使用得体性主要考察语块的熟练使用。语块指至少有2位高水平英语学习者作文中使用过的3-词或4-词组合[12]。借助于AntConc软件从人工阅卷得分较高的作文中提取出常用语块,进而到样本中检索语块的使用情况。

(二)中国学生英语自动作文评分模型的可行性研究

1.研究问题

本研究目的在于探讨中国学生英语自动作文评分模型是否可以用来预测二语作文得分。研究问题为:(1)该模型在多大程度上可以用来预测作文得分?(2)计算机提取语言特征和人工阅卷得分之间是否显著相关?

2.研究设计

研究语料来自云南某高校英语专业高年级学生课堂限时作文35篇,题目为:Why should a future foreign language teacher have some basic knowledge of language testing theory? 3位评分员以分析性评分的方式,按照语言、内容和组织分别占作文总分50%、30%和20%的标准,参照英专八级作文总分20分做出人工评分,取其平均值为该作文得分。然后利用语料库软件提取出上述介绍语言特征,利用SPSS11.5计算出其和人工评分的相关系数,即假设作文得分与其写作特征线性相关。

3.结果与讨论

表1

		得分
语块	Pearson Correlation	.683(**)
	Sig. (2-tailed)	.000
类符数	Pearson Correlation	.289
	Sig. (2-tailed)	.092
作文长度四次方根	Pearson Correlation	.409(*)
	Sig. (2-tailed)	.015
词汇复杂性	Pearson Correlation	-.044
	Sig. (2-tailed)	.800
平均句长	Pearson Correlation	-.003
	Sig. (2-tailed)	.988
动名词	Pearson Correlation	.064
	Sig. (2-tailed)	.717

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

从表1相关分析可以看出,作文中语块的使用、类符数、作文长度四次方根和写作得分显著相关,可以作为二语写作质量有效的衡量标准。而词汇复杂

性、平均句长和动名词等普遍认为可以体现二语学习者水平的语言特征和得分之间没有显著关系，这可能是因为在限时命题作文中，写作者在紧张状态下为保险起见而采取的一种策略，应用简单的语言表达思想，而未能尝试复杂的表达方式，从而导致字数越多得分越高。

表1显示，语块的使用数量和作文得分相关系数高达0.683($p<0.01$)，验证了人们普遍认为的高水平学习者更多地使用固定表达方法的观念。此外，显著相关性亦表明高水平外语学习者在表达同样的话题时倾向于使用相似的语块。也就是说，语块的使用是和话题相关的。如对于给定题目 Why should a future foreign language teacher have some basic knowledge of language testing theory? 高水平学者在表达中有一些固定的选择：in language testing, language testing theory, to know something about, at the same time, in order to, that is to say, how to teach, is necessary for, pay more attention to, reliability and validity, with language testing theory 等。

四、结语

毫无疑问，计算机技术的快速发展正影响着人们对诸如写作过程、标准等本质性的认识，自动作文评分方法的产生也引起了人们的质疑。教育技术的引进不仅是技术层面的问题，它还意味着一场牵扯到使用者理念、认识、方法和行为等方面的系统变革[15]。教师只有创造性地把软件技术有机地融入到自己的教学中，不断探索能提供学生及时反馈的作文评估方法，才能满足学生需求和自身发展。

参考文献：

[1]Spolsky, B.. Measured Words[M]. Oxford: OUP, 1995.

[2]http://www.ets.org.

[3] http://www.ielts.org/test_takers_information/what_is_ielts/test_format.aspx.

[4] Wang, J. & M. Brown. 2007. Automated essay scoring versus human scoring: A comparative study [OL]. Http://escholarship.bc.edu/cgi/viewcontent.cgi? article = 1100&context = jtla (accessed 10/04/2011).

[5]Warschauera, M. & D. Grimes. Automated writing assessment in the classroom [J]. Pedagogies: An International Journal, 2008, (3): 22-36.

[6]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, 2005.

[7]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):18-24.

[8]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):43-50.

[9]葛诗利. 大学英语作文自动评分方法比较研究[J]. 广东外语外贸大学学报, 2010,(3):87-90.

[10]陈潇潇，葛诗利. 自动作文评分研究综述[J]. 解放军外国语学院学报, 2008,(5):78-83.

[11]李金辉. 使用潜伏语义分析理论研究计算机改中国学生英语作文[D]. 广州：广东外语外贸大学, 2009.

[12]梁茂成. 中国学生英语作文自动评分模型的构建[M]. 北京：外语教学与研究出版社, 2011.

[13] Page, E. B.. The use of the computer in analyzing student essay[J]. Intl Rev. Education, 1968, (14): 210-225.

[14]Wolfe-Quintero, et al. Second Language Development in Writing: Measures of Fluency, Accuracy & Complexity [M]. Hawaii: University of Hawaii Press, 1998.

[15]唐锦兰，吴一安. 在线英语写作自动评价系统应用研究述评[J]. 外语教学与研究, 2011,(2):273-282.

(责任编校：任　华)

Research into automatic scoring method of English composition and its feasibility

YANG You-tong

(Yunnan University of Finance and Economics, Kunming 650221 China)

Abstract: in order to test the feasibility of automatic scoring model of English compositions written the Chinese students, this paper adopts 35 positions written by senior English majors as sample and explores the relation between automatic extraction language use characteristics and man-scoring. The result shows that the use number of chunk of composition, type, the fourth root of composition length and writing score are markedly correlated.

Key words: automatic scoring of composition; second language writing; feasibility

英语自动作文评分方法及其可行性研究

作者：杨有统
作者单位：云南财经大学,云南昆明,650221
刊名：新余高专学报
英文刊名：Journal of XinYu College
年，卷(期)：2011,16(4)