

语言·艺术

英语论文自动评分系统探索

张 梅

(重庆大学 外国语学院, 重庆 400030)

摘要:英语论文自动评分系统是由美国教育测试服务中心利用自然语言处理技术与信息抽取技术研究开发的一种英文在线测评英语写作能力的计算机程序。本文首先介绍了自动评分系统的结构及工作原理,然后探讨了其优点和存在的问题以及在大学英语四、六级考试作文网上阅卷的可行性。

关键词:英语论文自动评分系统;工作原理;可行性

中图分类号:H08

文献标识码:A

文章编号:1008-5831(2005)03-0095-03

A Research on English Automated Essay Scoring

ZHANG Mei

(College of Foreign Languages, Chongqing University, Chongqing 400030, China)

Abstract: E-rater is an automated essay scoring system developed at Educational Testing Services (ETS) by using natural language processing (NLP) and information retrieval (IR). This paper first presents the general structure and the performance of E-rater. And then, the advantages, potential problems and the feasibility of implementing online marking of College English Test compositions are also discussed in this paper.

Key words: E-rater; scoring engine; feasibility

一、简介

英语论文自动评分系统(E-rater, Electronic Essay Rater 的缩写),是由美国教育测试服务中心(Educational Testing Service)的 Jill Burstein 博士及其研究小组利用自然语言处理技术(NLP, natural language processing)与信息抽取技术(IR, information retrieval)研究开发的一种全美英文在线测评英语写作能力的计算机程序。“英语论文自动评分系统”的诞生,缘于美国的工商管理硕士(MBA)的入学考试(GMAT)计算机化之后,GMAT 考试的次数由每年 4 次增加到 216 次,考试次数的增加、参考人数的增多,评分工作量也就增大了很多。于是“英语论文自动评分系统”应运而生。据报道,自 1999 年起,用“英语论文自动评分系统”技术测评 GMAT 应试者的作文已逾 60 万篇,经与常规人工批阅作文的结果比较,“英语论文自动评分系统”评分的准确率超过 97%。现在,美国研究生管理专业委员会(GMAC)已经认为 E-Rater 电子软件判分相对于传统的由教师阅读判分更有效。

二、英语论文自动评分系统的工作原理

自动评分系统是用来评估论文写作的计算机化的技术系统,它是根据已被手工评判过的很多篇论文的结果总结经验而建立的计算机化评分系统。它以人工已完成的工作为基础,总结提高后作为计算机模式处理。美国教育考试委员会的成员就是要训练计算机习惯和熟悉判读文章,给出相应分数,越接近人工越好。

英语论文自动评分系统这个评分引擎(scoring engine)的评分流程其实就是人工评阅过程的复制。英语论文自动评分系统评分模式的建立,是在美国工商管理硕士入学考试中,通过大量分析经过两位(或两位以上)专业阅卷老师批改过的某一特定题目的作文(分数 1 到 6 分的文章都在英语论文自动评分系统的分析范围之内),根据阅卷老师批改学生作文时所应用的评比标准,例如句子结构、文章组织(即写作逻辑、修辞与论述条理)以及文章内容(词汇使用的能力)等因素,在英语论文自动评分系统里建立以大量

收稿日期:2005-02-20

基金项目:重庆大学语言认知及信息处理研究所 2004 年专项研究项目“网上作文评估研究”

作者简介:张梅(1964-),女,四川邻水人,重庆大学外国语学院副教授,主要从事英语教学法和语言教学研究。

万方数据

分析数据为依据的该特定题目的作文评分模式。

首先把人工已经判过分的论文,经过计算机处理,变成一个一个独立的单元。其次,给文章里每一个单元分成的小的元素加上标签,以便于计算机能够判读,比如语法、句子的主句等等。对文章的每一个元素加上标签后,给每部分一个分数,综合起来就是论文的总得分。实际使用时,被测者在用户终端输入根据某一特定题目所写的作文,通过互联网递交给美国教育测试服务中心的英语论文自动评分系统,系统将考生提交的作文与自动评分系统内的模板文章进行比较,即电脑评分器通过扫描考生的作文,寻找那些揭示逻辑推理的关键词汇并做出评分,再将分数通过互联网发送到考生的个人终端,从而完成一次作文评分工作。据报道,电脑评分器为每一题储存了成千上百的已计分的作文,包括1, 2, 3, 4, 5, 6分六种类型。电脑评分器通过扫描考生的作文,利用它的存储数据库决定考生的作文和这六种分数中的哪一种最相似,然后给出作文评分。作文的字数并不是得分高低的关键,因为自动评分系统是针对文章中的单词、词组、句子、文章的结构和写作内容与系统内的模式进行比较。若系统无法找出考生文章与系统内模板之间的关联性,自动评分系统就不会进行评分,而只是提供给你一些建议性的信息(Advisory Message)。

三、自动评分系统的优点

第一,可靠性。目前许多数据表明,计算机判分系统效果很好。美国教育测试服务中心同时用一个人和一个E-Rater给作文判分,当有分歧时,找第三者来评判。为了测试E-Rater的表现,美国教育测试服务中心采用了一个可靠性的评估方法,它可以表明任意两个阅卷者判分一致次数的百分比。评估结果是,两个人之间评分一致性可达到87%至93%。把同样的文章给E-Rater,则它与人评分一致性达到88%至94%。和老师手工评判对比,结果可以达到97%—98%,而且非常稳定。根据美国教育测试服务中心的报道看,自动评分系统软件很可靠。

第二,即时性。该系统以互联网为媒介,实现信息的实时传递。学生通过该系统提交作文后几秒钟即可获得评分结果及相应的诊断信息。据介绍,E-rater作文测评系统的核心技术广泛应用于GMAT考试的作文评分,与常规人工评分相比,准确率在97%以上,使你对它的权威性和准确性毋庸置疑;另外,它的使用非常简便,完成作文后,30秒钟内就可得到分数和相应建议,大大节约了宝贵的备考时间,免去长时间等待之苦。

第三,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。实际上,对作文评分影响最大的是评分过程。同一篇作文,不同的阅卷员可能给出不同的分数,同一个阅卷员在不同时间也可能给出不同分数。只有保证评分的一致性,才能提高作文阅卷的信度。这个一致性包括阅卷员本人的一致性、阅卷员之间的一致性和阅卷点之间的一致性,很明显,阅卷员评分的主观性对分数的客观性有一定影响。

第四,经济性。自动评分系统给作文评分快捷而准确,省却大量人工。美国教育测试服务中心报告指出,美国全国各地的英语和传媒专业的教授现在可以在家通过互联网判卷,高科技帮阅卷者免除了旅行的费用,而且他们不用在会议上花费时间了。

第五,互动性。该系统对学生写作技巧进行整体测评的同时,还提供诊断分析和反馈意见,考生可按自己的安排,在无风险情况下进行写作练习,为学生提供更多练习写作的机会。在分析作文题目的基础上,从语法和逻辑推理等方面对学生的作文进行评估,根据普遍适用的标准提出可靠的写作建议。该系统注册流程简单,学生可轻松进入系统进行测评,并按照自己的计划通过练习提高写作水平。此外,考生还可将自己的作文与该分数线上的其他作文进行对比,并可看到其他考生就该题目写出的作文。同时,系统还可就如何提高写作水平为学生及其指导教师提供有价值的信息,教师可以及时获得学生写作技巧的相关信息,是对指导教师的有益补充。在线提交作文前,考生可按自己的安排进行写作,并可进行反复修改。

四、存在的问题

电子判分系统从一开始就受到一些同行专家及学者的批评。首先,在很多学校都没有足够的计算机。第二,有些权威人士认为,电子判分系统的准确性还有待进一步证实。第三,学生使用计算机的熟练程度在很大程度上要影响他们的得分。学生使用计算机写的作文用电子判分系统判分,而用笔和纸手写的作文是人工阅卷,这之间存在一定差异,对考生不公平。第四,学生可以采取很多方式取悦电脑评分器而得高分。比如,学生在参加GRE考试之前,会仔细阅读范文,找出最高分6分作文的感觉,让自己的作文在结构上看起来类似5分或者6分的作文;学生会利用诸如“for example”, “therefore”等词或短语清楚地界段落和区域。第五,有些东西电脑评分器发现不了,如幽默,它是通过一些段落过渡词、段落变化等分析文章结构,通过比较其他考生来评价作文的内容。所以,即使你用非凡的推理风格,电脑评分器也

不会发现。

五、大学英语考试作文网上阅卷的可行性

每年举行两次的全国大学英语四、六级考试是一项规模宏大的标准化考试。大学英语四、六级考试的目的是推动大学英语教学大纲的贯彻执行,对大学生的英语能力进行客观、准确测量,为提高我国大学英语课程的教学质量作出了重要贡献。大学英语四、六级考试经历了一个不断完善的过程,为提高考试的信度和信度,近年来采取了一系列改革措施。其中一项重要改革措施之一就是设作文最低分,使师生更加重视写作能力的培养。作文内容主要是情景提纲式写作,或图表作文或应用文,要求考生在半小时以内按照所规定的情景、提纲和所提供的信息,构思准确,合乎情景提纲要求,写出一篇简明流畅的短文。作文题评分原则是采用总体评分(Global Scoring)方法。阅卷人员就总的印象给出奖励分(Reward Scores),而不是按语言点的错误数目扣分。从内容和语言两方面对作文进行综合评判,采用一读定分的方式进行。评分标准是把15分分为5个等级:2分、5分、8分、11分、14分。各有标准样卷一份。阅卷人员经过一定培训,认真学习评分原则和标准,对专家已评阅过的试卷进行阅卷,然后将每位阅卷员给的分数与专家分进行对比。要求阅卷员根据阅卷标准,对照样卷评分。

全国大学英语四、六级考试阅卷工作量大,很难保证阅卷评分的准确性和客观性,随着参考人数的增多,这一问题日显突出。因此,迫切需要一种客观、准确、高效的作文自动评估系统来减轻传统的手工作文评判强度。于是,基于网络技术、计算机技术和人工智能技术的网上作文评估系统便应运而生。根据上海交通大学外国语学院王跃武老师的实验数据表明,在网上阅卷时阅卷员之间的评分一致性高于在传统阅卷中阅卷员之间的评分一致性。事实证明,大学英语四、六级考试作文网上阅卷可行。

六、结论

基于网络技术、计算机技术和人工智能技术的作文评估系统的研究和应用是国际学术界和教育界前沿的研究热点之一。英语论文自动评分系统的确有

很多好处,比如学生可按自己的安排,在无风险的情况下进行写作练习。网上作文评估还可应用于日常的外语教学中,增加评教作文的准确性,学生还可以借鉴该系统提供的修改建议,逐步提高写作水平,有助于学生产生积极的作文反馈从而喜欢上写作,同时,教学内容的系统性、新颖性、趣味性,教学手段的多样化能吸引学生的注意力,激发学生学习写作的动机,提高其积极性。教师可运用自动评分系统来批改学生的作业,了解、掌握学生的写作情况,这大大减轻了老师的负担,老师也会乐于布置写作练习。但自动评分系统评分模式的建立是通过大量分析经过两位(或以上)专业阅卷老师批改过的某一特定题目的作文,根据阅卷老师批改学生作文时所应用的评比标准,以大量分析数据为依据的该特定题目的作文评分模式。所以,笔者认为,虽然英语论文自动评分系统在大学英语四、六级考试作文网上阅卷可行,但英语论文自动评分系统不可能在短时间内完全取代教师阅卷。

参考文献:

- [1]王跃武. 大学英语四、六级考试作文网上阅卷实验研究[J]. 外语界, 2004, (5): 74-79.
- [2]LANDAUER T K, FOLTZ P W, LAHAM D. An introduction to latent semantic analysis [J]. Discourse Processes, 1998, (25): 259-284.
- [3]FOLTZ P W, KINTSCH W, LANDAUER T K. Analysis of Text Coherence Using Latent Semantic Analysis [J]. Discourse Processes, 1998, 25(2-3): 285-307.
- [4]ELLIOTT S. Intellimetric: From Here to Validity [A]. SH-ERMIS M, BURSTEIN J, eds. Automated essay scoring: A cross-disciplinary perspective [C]. Hillsdale, NJ: Lawrence Erlbaum Associates, 2003.
- [5]BURSTEIN J, KUKICH K, WOLFF S, LU C, CHODOROW M, BRADEN-HARDER L, HARRIS M D. Automated Scoring Using A Hybrid Feature Identification Technique [A]. Proceedings of 36th Annual Meeting of the Association for Computational Linguistics [C]. Montreal, Canada: 1998. 206-210.

作者：[张梅](#)，[ZHANG Mei](#)
作者单位：[重庆大学, 外国语学院, 重庆, 400030](#)
刊名：[重庆大学学报（社会科学版）](#) [PKU](#) [CSSCI](#)
英文刊名：[JOURNAL OF CHONGQING UNIVERSITY \(SOCIAL SCIENCES EDITION\)](#)
年，卷(期)：[2005, 11 \(3\)](#)
被引用次数：[2次](#)

参考文献(5条)

- 1.[Elliott S](#) [Intellimetric: From Here to Validity](#) 2003
- 2.[Foltz P W](#);[KINTSCH W](#);[LANDAUER T K](#) [Analysis of Text Coherence Using Latent Semantic Analysis](#) 1998(2-3)
- 3.[LANDAUER T K](#);[FOLTZ P W](#);[LAHAM D](#) [An introduction to latent semantic analysis](#)[外文期刊] 1998(25)
- 4.[BURSTEIN J](#);[KUKICH K](#);[WOLFF S](#);[LU C](#), [CHODOROW M](#), [BRADEN- HARDER L](#), [HARRIS M D](#) [Automated Scoring Using A Hybrid Feature Identification Technique](#)[外文会议] 1998
- 5.[王跃武](#) [大学英语四、六级考试作文网上阅卷实验研究](#)[期刊论文]-[外语界](#) 2004(05)

引证文献(2条)

- 1.[谭继红](#), [李卫](#), [王枫](#) [基于知识库的科技稿件质量综合评价模型](#)[期刊论文]-[北京邮电大学学报](#) 2008(6)
- 2.[汪泱](#) [主观文字试题评判相关技术研究](#)[学位论文]硕士 2006

本文链接：http://d.g.wanfangdata.com.cn/Periodical_cqdxxb-shkx200503026.aspx