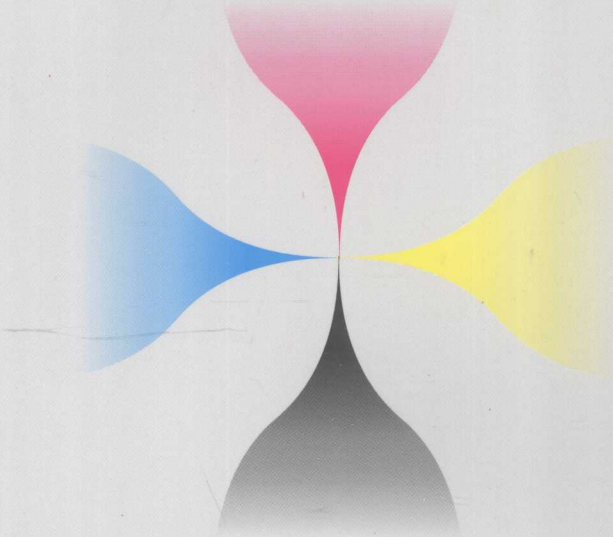
中国外语教育研究中心　外语考试自动评分研究系列丛书



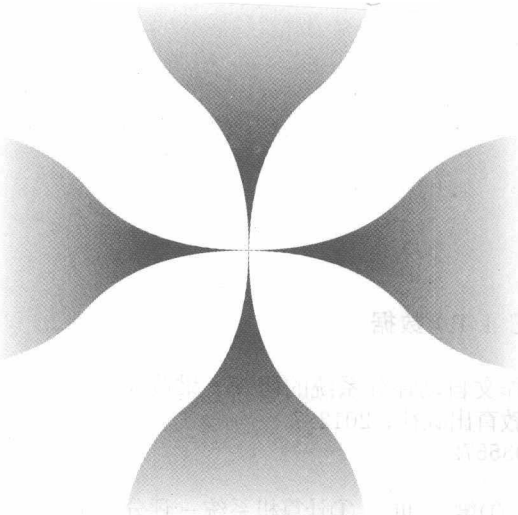
大规模考试英语作文自动评分系统的研制

梁茂成 著

高等教育出版社
HIGHER EDUCATION PRESS

中国外语教育研究中心　外语考试自动评分研究系列丛书

1538853



大规模考试英语作文自动评分系统的研制

daguimo kaoshi yingyu zuowen zidong ping
xitong de yan

梁茂成 著

高等教育出版社·北京
HIGHER EDUCATION PRESS BEIJING

图书在版编目（CIP）数据

大规模考试英语作文自动评分系统的研制／梁茂成著．-- 北京：高等教育出版社，2012.7
ISBN 978-7-04-035572-7

Ⅰ．①大… Ⅱ．①梁… Ⅲ．①计算机系统－评分－应用－英语－写作－考试－研究 Ⅳ．①H315-39

中国版本图书馆CIP数据核字（2012）第122281号

策划编辑 贾巍巍　　责任编辑 巩 婕　　封面设计 刘晓翔　　版式设计 刘 艳
责任校对 巩 婕　　责任印制 田 甜

出版发行 高等教育出版社
社　　址 北京市西城区德外大街4号
邮政编码 100120
印　　刷 北京市联华印刷厂
开　　本 787mm×1092mm 1/16
印　　张 9.75
字　　数 175千字
购书热线 010-58581118

咨询电话 400-810-0598
网　　址 http://www.hep.edu.cn
　　http://www.hep.com.cn
网上订购 http://www.landraco.com
　　http://www.landraco.com.cn
版　　次 2012年7月第1版
印　　次 2012年7月第1次印刷
定　　价 25.00元

本书如有缺页、倒页、脱页等质量问题，请到所购图书销售部门联系调换
版权所有 侵权必究
物 料 号 35572-00

序 言

美国未来学家Nicholas Negroponte在他的《数字化生存》里谈到，美国的《电视导报周刊》（TV Guide）的利润超过所有四家电视网利润的总和，这意味着关于信息的信息（the information about information），其价值可以高出信息本身。在因特网的风火轮推动下的信息社会里，每一个人都可以说是拥有一个全世界的数据库，淹没在海量的信息里。所以他提出“少就是多”的重要概念；这也正是Google这样的搜索引擎能够大行其道的原因。Google所依赖的技术是自动文本分类（text classification）或文本归档（text categorization），其目标就是从浩如烟海的信息中提取最有效的信息。上一个世纪90年代以来，建立以人工智能技术为基础的文本分类系统成为信息科学中的一个重要目标。用Google检索text classification这两个词，可以得到13,600,000个条目，光pdf格式的文章就差不多有1,240,000篇（2006年2月1日检索，可能有重复），就是一个明证。

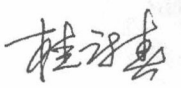
文本分类是一个应用范围甚广的领域，将学生在考试中的作文进行分类（即评分）就是一个很有前景的研究方向。Mark Shermis 和Jill Burnstein （2003）所主编的*Automated Essay Scoring: A Cross-Disciplinary Perspective*就介绍了5种评分系统。我国是一个考试大国，一个考试动辄几十万考生，而从语文（包括汉语和外语）试卷到其他科目的试卷都有许多开放性试题，改卷评分不但牵涉到很多阅卷员，而且很容易造成评分信度不高的问题，有失考试的公正。

在我国，也有不少仁人志士力图结合我国实际找出一条作文自动评分的路子，这是一个有很大现实意义的课题。但这也是一个复杂的攻坚课题，不可能一蹴而就，因为它牵涉到跨学科的研究和团队的共同努力。

我十分高兴地看到梁茂成教授的新著《大规模考试英语作文自动评分系统的研制》的问世。该研究不但对当前作文自动评分的各种探索进行评论和总结，而且结合我国的实际提出一个综合运用语料库语言学和统计学而构建的模型，并设计了可运行的程序。研究中所构建的模型包括了一些足以影响作文评分的变量，而且运用多元回归的方法建立了一个预测分数的公式，并进行验证。整个研究思路广阔、设计合理、论证充分，为我国日后建立更完善的自动作文评分系统做出了很有意义的尝试。

当然，作为一个模型，还有许多变量（如语言失误）有待摸索，而真正作

为一个自动化系统而建立和实施，也还有许多问题（如结合人工智能算法和建立语法分析器；除评估信度外，采用文本分类领域更常见的评价指标，如准确率（precision）和查全率（recall）等等）需要进一步考虑。但是梁茂成博士的研究的重要价值在于它在中国建立自动化评分系统的科学攀登中拉开了序幕，迈出了可贵的第一步。坚冰在融解，希望在前头！



2012年夏

试读结束：需要全本请在线购买： www.ertongbook.com

中文 摘要

本研究在梁茂成（2005）的基础上，挖掘新的文本特征变量，设计英语作文自动评分系统，并对该评分系统的评分信度、最低训练集样本量、适用文体类型、信度影响因素等问题进行探讨。

研究中收集了5个不同题目共1,067篇大学生英语命题作文，其中4个题目为议论文，1个为说明文。首先组织多名人工评分员对以上作文进行分析型人工评分，分析其评分信度，然后对这些作文进行多轮、多次抽样，组建训练集，对自行设计的作文评分系统的性能进行较大规模的验证。此外，研究中还对自动评分系统的构架、变量类型、汇编语言等进行了介绍，并就几个有代表性的变量进行了研究，对这些变量在自动评分系统中的应用进行了描述。

研究中主要回答以下问题：

1）以统计模型为基础设计而成的作文自动评分软件是否达到可操作水平？评分系统的评分信度是否能够达到语言测试的要求？

2）人工评分信度对自动评分信度和评分模型的稳定性有何影响？

3）训练集作文的最低样本量至少应该达到多少？

4）作文自动评分系统对不同文体学生作文进行自动评分时是否具有同等效果？

研究发现，由于我们在英语作文自动评分系统中设置了一些对作文质量具有较强预测能力的文本变量，使得系统在接受了足够的训练之后，自动评分的评分信度达到了r = 0.752或更高，可以满足统计学和测试学的要求。在训练集信度可靠的前提下，自动评分系统的评分信度最高达到r = 0.83以上，作文评分系统的评分结果与人工评分的结果之间的吻合率（在0 ~ 5的量表上）高于ETS的E-rater，表明当训练集样本信度可靠时，本研究中设计的英语作文自动评分系统的评分信度高于E-rater。因此，该系统已经达到了可操作水平。

本研究还发现，人工评分信度从两个方面对自动评分模型的稳定性产生影响。首先，当人工评分信度较低时，机器学习遵循“Garbage in, garbage out”的规则，自动评分信度也相应较低；当训练集人工评分信度较高时，自动评分信度也随之升高。其次，不同的人工评分员之间的评分信度差异对自动评分的信度也存在影响。由于训练集数据存在内在的不一致性，致使自动评分系统学习困难，评分信度降低。

对于第三个问题，研究中通过多轮多次、大小不同的训练集来对机器评分模型加以训练，并对验证集作文的自动评分信度加以比较和分析，发现当训练集样本量达到125时，自动评分系统的评分的信度开始趋于稳定并达到统计学和测试学要求。研究还表明，当训练集样本量达到125这个临界点后，大幅度增加训练集的样本量对提高机器评分的信度可能没有很大的作用。

有关自动评分系统能否适应对不同文体的作文进行自动评分的问题，研究发现，由于本研究中设计的作文自动评分系统中设置了一些可以侦测文体特征的变量，使得作文自动评分系统对学习者写作的英语议论文和说明文都可以实现自动评分，且评分信度可靠。在议论文和说明文评分模型中，权重变量既有相同的，也有不同的，表明模型中部分变量在对议论文进行自动评分时起较大作用，另一些变量在对说明文进行自动评分时起较大作用，还有一些变量在对两类文体的作文进行自动评分时都起作用。

关键词：作文自动评分，机器学习，统计模型，信度

English Abstract

On the basis of a previous research project (Liang, 2005), this study attempts to add new variables into the existing automated essay scoring (AES) model, and design an operational AES system. The study also reports results of an analysis of the reliability of the essay scores generated by the system designed herein. In addition, the research reported in this study also attempts to reveal the minimum requirement for the sample size of the training set needed for the system to yield reliable essay scores, the genre types of the essays the system can handle, as well as factors which may affect the reliability of the scores generated by the system.

Five sets of essays, totaling to 1,067, were collected, with each set addressing a different essay prompt. Among the 5 sets of essays, 4 are argumentative and 1 expository in nature. Mutltiple human raters were recruited to rate the essays with analytical essay scoring schemes. Reliability of the human-generated essay scores was then analyzed. The resulting data, composed of essays and their corresponding human-generated scores, were fed into the computer system to construct essay scoring models and to validate the model so constructed. During this process, multiple training sets were created by way of multiple sampling to avoid the possible hetereogeneity of the data, and the remaining essays in each set were then used as data for validating the models. Mean reliability estimates for the validation are reported. Besides, the report also introduces the architecture of the AES system, the types of variables in the scoring model, and the programming language employed to design the computer program. Further, a few representative variables in the scoring model are highlighted, with their research backgrounds and findings reported separately. Descriptions are also given as to how these variables are embedded in the architecture of the computer application.

The study attempts to address the following research questions:

1) Is the statistically-based AES system operational in terms of the reliability of the scores automatically generated? Can the reliability of the computer-generated scores meet the requirement described in the testing literature?
2) In what way does the reliability of the human-generated essay scores for the training set affect the reliability of the computer-generated scores?
3) In order for the system to generate reliable essay scores, what is the minimum requirement for the sample size of the training set?
4) Is the AES system equally effective in handling student essays of different genres?

Findings of the the study indicate that, thanks to the predictive power of the text variables in the essay scoring model, the AES system can yield essay scores with a reliability estimate of r = 0.752 or higher, so far as the minimum requirement for the sample size of the training set is satisfied. Such a reliability estimate is undoubtedly higher than what statisticians and language testers can expect in subjective testing items such as essay writing. It is also found that, as far as reliable scores have been assigned to the essays in the training set, the AES system can yield a maximum reliability of r = 0.83 or higher, and the consensus estimate (exact agreement and exact-plus-adjacent agreement on a 0-5 scale) of the essay grades generated by the computer runs higher than that reported of ETS' E-rater, an indication that the AES system is capable of generating more reliable scores than the E-rater. Therefore, the reliability anslysis shows that the AES system can be reliably put into operation.

Investigation into the second research question shows that human scores assigned to the essays in the training set can affect the reliability of computer-generated scores in two ways. First, the maxim in machine learning, "Garbage in, garbage out", turns out to hold true. When scores of low reliability are taken as input for learning, the AES system correspondingly outputs less reliable scores, and vice versa. Second, the magnitude of the differences between inter-rater reliability estimates in the training data also has a role to play in affecting the reliability of computer-generated essay scores. When all human raters think alike, the computer follows. When some human raters have consensus while others do not, expectedly, the computer is at a loss for what to learn.

To address the third research question, multiple training sets were created by way of multiple sampling, with results derived from each sampling. Results for such sampling are then compared. It is found that, when the sample size of the training set reaches 125, computer-generated scores become more stably reliable. Data analysis also seems to indicate that a large increase in the sample size of the training set beyond 125 may not bring about an expected growth of the reliability of computer-generated scores.

Whether or not the AES system can handle student essays of different genres is an issue related to the applicability of the computer program. A comparative study of the validation data shows that while some variables in the scoring model are equally predictive of the quality of both argumentative and expository student essays, some other variables in the model seem to be sensitive to different genres. The different weighting of the beta values for these variables in the scoring model

renders it possible for the AES system to use different marking schemes, an ability which enables the AES system to handle student essays of different genres.

Key Words: automated essay scoring, machine learning, statistical model, reliability

目 录

试读结束：需要全本请在线购买： www.ertongbook.com

表格目录

插图目录

第1部分
绪论及相关研究回顾

第1部分共分两章，第一章介绍研究背景，提出研究问题；第二章综述国外具有代表性的作文自动评分系统，并分析其得失，为本研究提供启示。