大学英语作文自动评分方法比较研究

葛诗利
（广东外语外贸大学·广东·510420）

内容提要：自动作文评分的标准除内容外，更重要的是语言的运用。中国大学生英语作文在语言运用上尤具自身的特点。为了探索该类文本的有效的自动评分方法，本文采用中国英语学习者语料库中大学英语四级考试作文子库的部分作文为样本，针对语言使用提取特征，采用多元回归、KNN和SVM方法分别训练评分模型并进行评分实验，结果表明多元回归方法能最准确地模拟人工评分，但KNN方法更适合日常教学使用。

关键词：大学英语作文；自动作文评分；评分方法
中图分类号：H319 文献标识码：A 文章编号：1672-0962(2010)03-0087-04

一、引 言

自动作文评分（Automated Essay Scoring，简称为AES）就是利用计算机技术对作文进行评估与记分[1]，其实质就是基于人工准确评分的训练作文集的自动文本分类。但其分类标准除内容外，更要兼顾语言的运用。该方向的研究至今已历时近四十年，在此过程中，采用了统计、自然语言处理（NLP）及人工智能等方面的最新成果[2]，并于1999年进入实际应用阶段[3]。

自动作文评分有以下几个优点：1）实用性：可以提高工作效率。2）一致性：作文评分本质上存在着主观性，人工评分的一致性就会因此而受到一定的影响。3）反馈：给学生反馈是非常重要的，这种评分系统能够为作者提供具有针对性的修改建议[4]。多项研究证明，在写作评测方面，自动评分系统的准确性与可靠性，以及与人工评分的一致性方面都非常高[5-8]。

当然，计算机评分也有很多缺点。Page强调，计算机并不能像人一样评判一篇作文，因为计算机只是“编程让它做什么”它就做什么，而并不能像人一样去“欣赏”一篇文章。另外一种批评是构造方面的缺陷。也就是说，计算机所计算的变量并不一定是作文评分中“真正”重要的方面，比如，关注文章的形式方面而不是组织方面[9]。

二、当前AES系统在英语作为外语作文评价方面的表现

虽然近年来AES在国外已渐成为自然语言处理中的一个热点问题，成型的系统已有十余个，文章与著述也比较多，但大多是针对英语母语作文评分，涉及英语作为外语（EFL）作文评价的尚不多见。比较系统的研究只有E-rater[10]。

Burstein等人把母语为汉语、阿拉伯语和西班牙语的英语学习者的作文与母语为英语的人（包括美国本土出生与本土以外出生两类）的作文，在人工评分与E-rater评分的框架下做出了对比研究。他们收集了这五类作者两个题目的作文，分别是562篇和576篇。然后以整体评分的方式，从1分到6分，全部做出人工评分。在这两个作文集中各自随机挑出255篇作文（2分到6分各50篇，1分的5篇）作为训练集，训练评分模型，并对其余作文进行自动评分。

他们的研究表明，虽然人工评分的均值（4.16）与机器评分的均值（4.08）差别不大，但具有统计显著性（$F=5.469$，$p<.05$），而不同题目之间没有显著性差异。这说明E-rater在评价英语作为外语的学习者所写的英语作文方面与人工评分虽然差别不大，但还是存在一些影响机器评分准确性的因素。研究还表明英语作为母语的作者写的作文成绩要高于非母语作者，在这方面，机器与人的评分达到了一致。但是在非英语母语的语言组作文成绩的评价上，E-rater与人的评价出现了显著性差异（$F=12.397$，$p<.001$）。有的组（如：西班牙语）E-rater给的分值低于人给的评分，而有的组（如：汉语）E-rater给的分值高于人给的评分。回顾机器评分中的建模过程，训练集中75%的作文是由非英语母语者所写，而筛选出用于线性回归过程的特征与英语母语作文评分所选

收稿日期：2010-01-01 * 基金项目：广东外语外贸大学2009年度校级青年项目（面向大学英语教学的自动作文评分和反馈研究）。
作者简介：葛诗利(1969-)，男，山东烟台人，博士，广东外语外贸大学外国语言学与应用语言学研究中心副教授；研究方向：计算语言学，语言教学。

特征基本相同，而其中最重要的两个预测性因素都是关于词汇的使用。[11]这意味着汉语组的作者在词汇掌握上要优于其它英语作为外语的写作者，所以在机器评分时占有优势；但在写作的其他方面比如句法、篇章结构等可能不及其它英语作为外语的写作者，所以在人工评分时处于不利地位。

Burstein等人的研究最后得出的结论是，“虽然不同语言组中人工评分与E-rater评分存在显著性差异，但其差异的绝对值不大，所以与人工评分的一致率没有显著性差异。”[12]但这并不能说明E-rater可以应用于中国大学英语教学中的作文评分。首先，由于这项研究是基于托福考试作文，结果对大多中国英语作为外语学习者并不具有普遍性，因为参加托福考试的英语学习者，跟大多数的普通大学生，尤其是大学低年级英语写作水平亟待提高的学生相比，其英语水平相对较高。另外，就是E-rater的评分范围在1分到6分，所以其评分差异的绝对值不大，而中国大学英语四六级考试作文是15分制，平时的写作训练评分一般是百分制，这种评分的显著性差异就会导致评分差异的绝对值显著增大。

另外一个要考虑的重要因素是低水平的英语作文中高频率出现的词汇和句法方面的错误。在这方面，“传统的自然语言处理语法分析器在英语作为外语的教学应用上，尤其是作文自动评分上至今尚未取得广泛的成功。”[13]

这些研究说明，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。其最主要差异就在于句法方面。这在Wolfe-Quintero等人[14]的论述中也得到证明。在外语写作评价中，语言的使用，尤其是句法方面，所占比重相对较大。这就使得针对英语母语写作设计的自动作文评分系统不能够直接应用于英语作为外语写作的自动评分。

在对英语作为外语作文自动评分方面的努力首先当推Lonsdale和Strong-Krause[15]，他们采用了基于链语法（Link Grammar）的句法分析器来分析评判英语作为外语作文。链语法分析器能够跨越句子中不合语法的单词，找到后面的词汇，并连接构成有句法意义的词对，比如：主语+动词，动词+宾语，介词+宾语，形容词+状语修饰语，和助动词+动词。但是由于该分析器的机制，以及单独对句法的分析，机器评分的准确率较低。

三、中国大学英语作文自动评分系统模型

EFL作文自动评分的研究远远滞后于主流的AES研究，而针对中国学生英语作文的自动作文评分研究至今未见公开发表的成果。大学英语四级考试作文，基本能够代表中国大学低年级学生英语写作的水平。鉴于语料库中的作文都基本符合作文题意，遵循外语写作研究理论并吸收国外自动作文评分研究经验，我们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。

1. 语料

语料采用了中国英语学习者语料库的ST3子库，从总共1318篇作文中选取篇数较多的一个题目的作文：Getting to Know the World Outside the Campus， 共201篇。由于高分段作文数量太少，另外参考E-rater的六分制，选取了6至11分的作文共174篇作为本次研究的对象。从各分数档中依顺序从前往后大约选取四分之三的作文共127篇构成训练集，其余作文47篇构成测试集。各自的分数分布如表1所示。

表1 样本作文语料分数分布

分数	6	7	8	9	10	11	合计
训练集	14	22	29	29	24	9	127
测试集	5	8	11	10	9	4	47
合计	19	30	40	39	33	13	174

2. 研究方法

本研究基于作文语料库，提取语言使用方面的特征，包括从单词、句子到错误标注的31个方面。分别采用三种方法：多元回归、K近邻和支持向量机，建立评分模型，并用测试集评估模型性能，对三种方法做出比较。

（1）多元回归方法

编制Perl程序，从训练集的127篇作文中提取作文分数作为因变量，31个方面的特征值作为自变量，形成一个127 x 32的矩阵，输入SPSS统计包，以两种方式进行多元回归。一种是以系统默认的采用所有的31个变量作为自变量进行回归，即强制进入法。另一种是采用逐步回归法，以决定31个备选变量的取舍，其基本思路是：对于所有的备选的自变量，如果一个自变量所对应的F统计量的值大于系统默认的“纳入标准”3.84，则这个自变量将被纳入模型；如果它所对应的F统计量的值小于系统默认的“剔出标准”2.71，则这个自变量将被剔出模型。经过多次叠代，得到最后结果。

（2）K近邻方法

K近邻方法是著名的模式识别非参数统计方法，也是最好的文本分类算法之一，已经有四十年历史。其算法思想简单，容易实现。

评分算法描述：给定一个待评分作文样本，系统在训练集中查找离它最近的（最相似）前K个邻居，并根据这些邻居的分数来给该样本评分，即前K个邻居作文的平均分值设定为该作文的得分。待评分作文与训练集作文的相似性采用了作文特征向量间距离来衡量。

在用于多元回归的Perl程序基础上，进一步编写用于K近邻统计的程序。根据训练集中127篇作文形成127个作文特征向量。测试集47篇作文也分别提取成为一个向量，并分别与训练集的127篇作文比较相似度，并找到最相似的前K篇作文，把其平均分赋予该测试作文。K值分别取5、15、和25进行了三次测试。

（3）支持向量机方法

支持向量机是近年来机器学习研究中的一项重大成果。主要用于解决二分类模式识别问题，也可以用于多维度的识别。支持向量机是统计学习理论（Statistical Learning Theory，简称SLT）的基础上发展出来的一种新的通用学习方法，在众多领域的成功应用表现了它很多优于现有各种方法的性能。

改编用于多元回归的Perl程序，分别从训练集和测试集作文中提取 32 个变量的值，以支持向量机训练器所要求的形式，构成一个 127 x 32 的矩阵作为训练文件，一个 47 x 32的矩阵作为测试文件。运行支持向量机训练工具得到支持向量机对测试集的评分结果。

3. 评分结果

评分结果经过汇总后，可以清楚地看出各模型的评分准确性。表 2 是对测试集中 47 篇作文评分的准确情况，强制进入法回归与 K=15 时的 K 近邻统计方法所获得的评分模型达到最高的准确率，25.5%；其次是另外两种 K 近邻方法以及支持向量机方法；最差的是逐步回归方法，只有 10.6%。即使是最高的准确率也只有25.5%，但像E-rater等商用自动评分系统通行的准确率衡量方法都是计入相邻分数，即上下只差一分的，也算准确评分。据此得到的评分情况如表3所示。

按照通行的计算方法，达到最高准确率的仍然是全变量回归，准确率达 76.6%，然后依次是逐步回归、K=15 的 K 近邻，另外两种 K 近邻方法和支持向量机方法准确率最低，只有 66%。

表2　各评分模型的准确评分情况

	enter	stepwise	Knn5	knn15	knn25	svm
6	0	0	0	0	0	0
7	1	0	0	0	0	0
8	5	2	2	3	1	2
9	5	3	4	7	7	5
10	1	0	2	2	0	0
11	0	0	0	0	0	0
正确评分的作文数量	12	5	8	12	8	7
准确率 (%)	25.5	10.6	17	25.5	17	14.9

表3　各评分模型的相邻准确评分情况

	Enter	stepwise	Knn5	knn15	knn25	svm
6	0	0	0	0	0	0
7	5	3	4	3	3	3
8	11	11	9	10	10	11
9	9	10	9	10	10	10
10	9	9	8	9	8	7
11	2	0	1	0	0	0
相邻正确评分作文数量	36	33	31	32	31	31
相邻准确率(%)	76.6	70.3	66	68.1	66	66

4. 讨论与分析

（1）局限性和不足

首先是语料选择的局限性。由于中国学习者英语语料库（CLEC）建库时考虑到错误分析的可行性，没有收集 6 分以下的作文样本，以及 11 分以上作文样本数量太少，这也使得本次实验评分范围局限于 6 到 11 分。另外，由于语料库样本是按照总体采样收集，所以高分和低分作文，即6和11分的作文数量偏少，这就直接导致了高、低分段作文评分模型训练不足。

其次是作文人工评分的准确性问题。大学英语四级考试作文，包括 CLEC 中四级子库重的作文，通常只经过一名评分员评分，其评分信度存在一定的争议。训练集作文人工评分的准确性决定了评分模型的评分效果，而测试集作文人工评分的准确性直接关系到模型准确率的衡量。

最后是在研究方法上，主要是特征提取方面，不管是在词汇还是句法方面，都有很多的衡量标准，但鉴于本研究的探索性质，只是抽取了一些相对简单而容易计量的特征。在错误提取方面，为了克服数据稀疏问题，只计算大类别的错误，没有按照语料库的标注来详细分类。采用线性回归方法是基于这样一种假设：即作文得分是与其写作特征线性相关的。当然，这个假设本身也存在着一定的争议。使用的强制进入回归方法从理论上来说，也存在着一定的问题，因为有些变量之间的相关度较高。但出于实用目的，以及作为逐步回归方法的比较而采用。

尽管在语料、人工评分和研究方法上存在很多的问题，本次实验还是得到了比较有意义的结果。虽然最高准确率76.6%不是很理想，但这并不说明机器自动评分方法不可行，只是说明有问题的两个方面需要设法改善。首先是语料上，由于低分档的 6 分和高分档的 11 分用于训练的作文语料较少，这直接导致了评分模型训练的不足，由此引起这些分数段评分的不准确。除去这些分数段后可以看出，7到10分的38篇作文，机器评分与原分值正好相同的，强制进入回归和K=15的 K 近邻方法有 12 篇，占 31.6%，最差的支持向量机方法也有 7 篇，占 18.4%；包括与原分值相差一分的，强制进入回归的达 34 篇，占 89.5%，逐步回归、K=15 的 K 近邻和支持向量机方法正确评分的篇数分别为33、32、和31。这个最高的评分准确率甚至已经超过E-rater的下限87%。[16]

（2）三种方法的比较

从自动评分的结果看，不管是准确率还是相邻准确率，强制进入的多元线性回归都达到了最高值。在计算相邻准确率的时候，逐步回归方法也获得了较高的准确率。这说明在提取同样特征的情况下，线性回归方法能够最好地模拟人工评分。这也是大多数主流自动作文评分系统，如 PEG、E-rater和IntelliMetric，采用该方法的原因。但是这个方法也存在明显的弊端，即某一特征在线性回归方程中的权重是由训练集计算得出，很难人工作出解释。也就是说，一篇作文，

不管是被评为高分还是低分，很难说出该得分的原因，也因此不能够给学生以有益的反馈。

K近邻方法的评分准确率不是非常高，但如果参数K调整得当，也可以获得较高的评分准确率。该方法操作简单，更为可贵的是这个方法获得的作文评分可在一定程度上得到合理的解释，因为一篇作文的评分是与它最相似的K篇作文的平均分。这K篇作文可以列出，作文间哪些特征比较相似也可以计算得到。这样就可以给学生适当的反馈，使他们知道需要在哪些方面做进一步的修改。

支持向量机方法虽然在机器学习领域得到广泛研究并取得了非常好的分类效果，但从目前的研究看，该方法并不是很适合中国学生英语作文的自动评分。这也说明，相关领域的方法和工具可以充分利用，但必须经过实验，证明其有效性或者作出适当的改进。

四、进一步研究的方向

本次探索性研究表明了各种自动作文评分方法在中国EFL学习者写作评分上的有效性。作为大规模考试自动评分的首选方法应该是线性回归方法，因为该方法评分准确率最高，而且大规模考试作文评分并不需要反馈。但日常教学中的写作评分采用K近邻方法较好。虽然该方法评分准确率相比线性回归稍差，但日常教学对评分准确率的要求不像高利害关系考试（high-stake examination）那么严格，而且K近邻方法可向学生提供一定的反馈信息，有利于学生写作的改进。

当然，自动作文评分要在我国广泛投入使用，很多方面尚需深入研究与改进。首先是要有足够的具有权威评分的训练语料，然后在特征提取方面尚需加深研究，提取更能反映中国学生英语写作水平，尤其是语言应用方面的特征。本次研究主要指标还是依靠CLEC自带的错误标注，但通常要评分的作文是不带任何标注的，这就需要依靠现有的自然语言处理技术，发展相应的自动识别标注工具，这也是下一步重要的研究方向。

参考文献

[1] Shermis, M.D. & Burstein, J. Introduction [A]. Shermis, M.D. & Burstein, J. (eds.) *Automated Essay Scoring: A Cross-disciplinary Perspective* [C]. Mahwah, NJ: Lawrence Erlbaum Associates, 2003: xiii.

[2] Dikli, S. Automated Essay Scoring [J]. *Turkish Online Journal of Distance Education*, 2006, 7(1): 49-62.

[3] Kukich, K. Beyond automated essay scoring [A]. Hearst, M.A. The debate on automated essay grading [J]. *IEEE Intelligent systems*, 2000, (5): 27-31.

[4][13][15] Lonsdale, D. & Strong-Krause, D. Automated Rating of ESL Essays [EB/OL]. http://acl.ldc.upenn.edu/W/W03/W03-0209.pdf, 2003.

[5][10][11][12] Burstein, J. & Chodorow, M. Automated essay scoring for nonnative English speakers [EB/OL]. http://www.ets.org/Media/Research/pdf/erater_acl99rev.pdf,1999.

[6] Keith, T.Z. Validity of Automated Essay Scoring Systems [A]. Shermis, M. D. & Burstein, J.C. (eds.) *Automated Essay Scoring: A Cross Disciplinary Perspective* [C]. Mahwah, NJ:Lawrence Erlbaum Associates, 2003: 147-168.

[7] Landauer, T.K., Laham, D, & Foltz, P.W. Automated essay scoring and annotation of essays with the intelligent essay assessor [A]. Shermis, M. D. & Burstein, J.C. *Automated Essay Scoring: A Cross Disciplinary Perspective* [C]. Mahwah, NJ:Lawrence Erlbaum Associates, 2003: 87-112.

[8][9] Page, E.B. Project Essay Grade: PEG [A]. Shermis, M.D. & Burstein, J.C. *Automated essay scoring: A Cross-disciplinary Perspective* [C]. Mahwah, NJ: Lawrence Erlbaum Associates. 2003: 43-54.

[14] Wolfe-Quintero, K., Inagaki, S. & Kim, H.Y. Second Language *Development in Writing: Measures of Fluency, Accuracy, & Complexity* [M]. Honolulu: Second Language Teaching & Curriculum Center, University of Hawaii at Manoa, 1998.

[16] Valenti, S., Neri, F. & Cucchiarelli, A. An overview of current research on automated essay grading [J]. *Journal of Information Technology Education*, 2003, (2): 319-330.

A Comparative Study of Automated Essay Scoring Techniques for College Students' English Writing

Ge Shili

Abstract: The criteria of automated essay scoring include content and, more importantly, the use of language. Chinese college students' English writing has its unique attributes in language use. In order to find out an efficient automated scoring method for this type of text, in this study some essays from sub-corpus st3 of CLEC were used as the sample, attributes of language use were extracted, techniques of multi-regression, KNN, and SVM were adopted to train a scoring model respectively, and experiments of scoring were conducted. The result shows that the technique of multi-regression can reach the highest scoring precision while the technique of KNN is more suitable for daily teaching.

Key words: college English writing; automated essay scoring; scoring technique