DOI:10.13458/j.cnki.flatt.003854

语法自动分析与计算机辅助写作评分*

贺俊杰 王 昉
(陕西师范大学外国语学院,陕西西安 710062)

摘 要:本研究旨在利用多元线性回归统计模型,构建面向中国英语学习者的写作计算机辅助评分模型,并应用大规模语料验证其可靠性。研究发现:在考虑语法及基本写作风格等因素的条件下所构建的评分模块,辅以事后人工复查,可相当准确地预测写作的总体评分。因此在特定条件下,该评分模块可应用于计算机辅助写作评卷。

关键词:英语写作;自动评分;语法判断;辅助评分

中图分类号:H0 **文献标识码**:A **文章编号**:1004-6038(2012)06-0061-05

1. 写作机器评分的现状

写作测试在语言测试中属于主观测试。评分实践中可以通过多人独立评分,并制定具有可操作性的详细评分细则等手段来保证评分的客观性。但由于多人独立评卷成本较高,且难以保证评卷人员不受各种因素的影响而保持评分的稳定性,目前,教育评价界在努力探索写作机器自动评分的途径。这样的系统既可以充当多人评卷中的一名评卷者,从而减少人力投入,还可以发现评分环节的异常情况,以保证多人独立评卷的稳定性。

目前,国外比较成功的这类系统有:(1) Project Essay Grade (PEG)、(2) Latent Semantic Analysis (LSA)、(3) Electronic Essay Rater (*e-rater*) 和(4) Bayesian Essay Test Scoring System (BETSY) 等。PEG 依照文章的结构和形式评分,LSA 评估文章内容的相关性,而 *e-rater* 则综合二者,对文章的结构和主题内容两方面做出评判。具体细节,请参看梁茂成和文秋芳(2007:18-24) 以及葛诗利和陈潇潇(2007:25-29) 对这些系统所作的评介。

以上系统中,*e-rater*(由美国 Educational Testing Service 的 Jill Burstein 等研究人员所开发) 的表现比较突出。它通过撷取作文中的混合特征(hybrid feature),从语法(syntactic structure)、写作技巧(Rhetorical Structure) 和词汇与主题(lexical and topical content) 等三方面模拟真实评卷人对作文进行评分。据 Burstein 等(1998) 的报道,*e-rater* 的准确性达到 87%~94%[①],并且 *e-rater* 与评卷者之间评分的相关性也达到 0.73,非常接近两名评卷人之间的相关系数 0.75。如此可靠的评卷结果令 *e-rater* 很快地被应用到美国的 GMAT(the Graduate Management Admissions Test) 和 TWE(Test of Written English) 等大型考试的评卷实践中。

2. 本研究的意义

对于写作机器自动评分是否具有可以接受的信度和效度这一问题,很多人持谨慎的观望态度。其中重要的原因就是在目前条件下,机器并不能读懂文章的内容,对于文章是否具有新颖观点、是否运用了幽默手法或其他诗意的表达法,它都无法识别。此外,LSA 虽能就意义进行评价,但由于它的工作原理是计算词与词、句与句或篇与篇之间在意义上的距离,或相似程度,并不考虑与主题相关的词汇呈现和组合的顺序,因此只要文本中有相关词汇就可得高分,而不管这些词汇是否构成合语法的句子[②]。

Powers 等(2002) 开展了一项研究。他们邀请 27 位各方面人士提供了 63 篇文章,目的在于挑战 *e-rater* 系统,希望能通过各种手段"骗过"*e-rater*,让它能给出一个不可靠的分数,从而探讨写作等主观题型机器自动评卷的现实性和效度。结果表明,人们确实可以通过某些手段,通过提交给 *e-rater* 不太有意义的文本(人评为最低分) 而能获得最高分。因此,Powers 等(2002) 认为"若要将 *e-rater* 应用于教育测量实践,正如其开发者已让用户所明了的,写作者必须是在针对考题诚实而努力地做答。"有意义的文本首先是由合语法的语句所构成的,这从一个角度说明,自动评分之前对文本首先做出合语法性的评判是甄别真正有意义文本的前提。

对于成功的写作行为来说,语言和思想两个要素缺一不可。母语写作者语言的合语法性基本不成问题,写作过程重点考虑主题内容。而外语写作者须同时考虑语言和内容两个方面。前述的机器自动评分系统都预设写作者在语言上没有太严重的问题。即使有的也考察文本的语法特征,但其目的仅在于寻找遣词造句方面的零星信息(诸如介词使用的频率、虚拟句的多寡等),从而确定文本的写作风格特点,而不是就语句进行合语法性判断,甚至就语法错误进行甄别。比如 *e-rater* 的句法模块就试图"识别包括不定式短语、补足语短语和从属子句在内的各种短语。*e-rater* 通过识别这些子句类型来捕捉写作文本在句法上呈现的差异"(Shermis & Burstein,2003:116)。因此,就写作测试而言,母语者和外语学习者(尤其是初、中级水平学习者) 有着本质上的差异。后者在写作过程中所用语言在语法上的准确性就是评分时需

* 基金项目:本文得到陕西省社会科学基金项目"中西部条件下面向 ESL/EFL 的计算机化任务型测试探索"(项目编号:07J009S) 的资助,为其阶段性研究成果。本文写作过程中曾得到桂诗春教授和曾用强教授精心的指导,孙晓惠老师和周石平老师无私的帮助,同时还特别受益于匿名审稿专家的宝贵意见,笔者一并致以诚挚的谢意。文中若有任何讹误,皆由笔者负责。

作者简介:贺俊杰,教授,博士,研究方向:音系学,计算语言学,语言测试学;王昉,讲师,硕士,研究方向:二语习得,语言教学

要重点考察的方面。对这样的测试文本进行评价，语言的合语法性判断也就占着很大的权重。

Burstein 和 Chodorow (1999) 利用非英语本族人的写作文本对 *e-rater* 进行了适用性测试。研究表明，针对中国学生英语作文，*e-rater* 的评分与人评分的一致性为 88.2%，相关系数为 0.54，明显低于其平均值 0.73。这表明，外国学生的作文在某些方面确实有别于英语本族人的作文。

在国内的一些大规模考试的写作评卷中，句子的合语法性判断在中国学生的作文评分中有着重要地位。虽然这些考试的写作评卷是综合语言和内容两方面给出总分，并不以统计语法错误数量来定档次，但总体评分与语法因素有着较高相关，这可以从本文后面的相关分析和在多元线性回归分析中所表现出的相关系数得到证实。

鉴于此，开发作自动评分的语法模块对于英语教育与测量，尤其是对于我国英语作为外语的测试实践有着重要的现实意义。

3. 语法模块的构建

3.1 基本思路

本文[③]将采用因子分析和多元回归分析方法(统计软件采用 SPSS11.0)。因子分析用于对诸多自变量之间关系的考察，并将自变量简化到少数几个在统计上相互独立的因子，以代表它们。通过探测性因子分析了解这些自变量相互联系的程度与方式及其数据结构，然后用因子分析产生的各个在统计上相互独立的因子来代表语法错误标注的聚合关系，最后将因子分析的结果用于语法错误判断的多元线性回归分析，考察合语法信息在写作评分中的作用。

首先，需要对文本进行语法分析，判断各句是否合语法。不合语法的句子，要大致确定其出错的位置，出错部分的词都是何种词性。这些特征在具体判分中应占多大权重可以通过因子分析和多元回归分析来确定。最后通过一定量的训练集得到回归公式，用此公式即可对其他写作样本进行判分。

3.2 模块的训练

本研究中用于训练的样本集共有 93 篇作文，它们全部随机取自 CLEC 语料库的 ST3(四级作文样本)子库。统计数据如表 1 所示。

表 1　训练集之统计数据

文本数	最低分	最高分	平均分	标准差
93	3.00	14.00	9.38	2.01

表 2 是训练集中样本作文得分在各档次上的分布。这表明训练集在各档次有合理的频数分布。

表 2　训练集各档次频数分布

档次	档次分	频数
1	2	1
2	5	3
3	8	47
4	11	36
5	14	6

首先将训练集作文由两名独立阅卷人[④]严格依照大学英语四、六级考试作文的评分办法评分，将得分与 CLEC 所提供的分数进行综合[⑤]，得到每篇的综合分——SCORE。

然后运用链语法(link grammar)的分析原理，将训练集作文经改进后的链语法分析器[⑥]进行语法分析。下面通过两个示例分析来说明如何通过语法分析发现错误。对于例(1)中的错误位置，链语法分析器显示无法正确链接，通过词性标注赋码，我们可以得到无法链接部分(即有错误的地方)的词性标注：wealth/NNP、deparment/NN、are/VBP、this/DT。

(1) Wealth deparment are build more rapidly, and this keep the people wealthy.

链语法分析结果：[wealth][⑦] [deparment] [are] build. v more rapidly, and [this] keep. v the people. s wealthy. a .

再比如(2)，两个并列句之间以及第二句的两个并列成分之间都存在没有并列连词的错误，这导致链语法分析中逗点和动词 listen 得不到链接，从而被发现不合语法。其中无法链接部分的词性标注为：/COMMA、listen/VB。

(2) The most convinient is the public media, we can read the newspaper, listen to the radio, etc.

链语法分析结果：the most convinient is the public media [.] we can read the newspaper, [listen] to the radio, etc.

(1)中未成功链接的 NN 和 VBP 以及(2)的 COMMA 和 VB 暗示着这两句都存在错误，但这里不必明示错误类型，因为错误及其类型在评分中应占的权重可以通过后面的因子分析和多元线性回归分析得以确定。

接下来，将每句的总词数与句中出错部分词数之差和总词数的商作为每句的合语法词数比率。每篇各句的合语法词数比率之和再除以总句数，得到全篇的合语法词数比率，记作 LSCORE。此外，将合语法句数与全篇总句数的商作为全篇的合语法句数比率，记作 CRCTERR。

将训练集各篇的总得分 SCORE 与 CRCTERR 做散点图与相关检验。二者间的相关系数 $r = 0.486$，$r^2 = 0.236$，$p = 0.000$，有着中等程度的相关，在双尾检验和 0.01 的水平上具有统计意义。这说明 CRCTERR 能解释总得分中近 1/4 的变异。

然后，将各句中出错部分各词的词性分别标注[⑧]并以篇为单位按标注累加计算频数。以这些标注的频数和 LSCORE 作为自变量，以 SCORE 作为因变量进行多元线性回归分析。结果显示，$r = 0.781$，$r^2 = 0.609$。但 ANOVA 分析提示：$F = 0.826$，$p = 0.694$。也就是说虽然相关系数很高，但这个模型没有统计意义。接下来考虑从这些标注中提取能解释它们的若干因子，通过因子分析，观察有多少个因素可以解释这些错误信号，以期用较少的因子代替这些标注用于回归分析。

经过因子分析，一共提取了 9 个因子，它们的特征根(Eigenvalue)均在 1.0 以上。这 9 个因子一共能解释 85% 的变异。因子分析碎石图(图 1)也显示从因子 9 之后开始平缓，说明选取 9 个因子是比较合适的。

经旋转后，我们得到每个因子载荷较高的标注。在因子 1 中，DT(determiner)、VBP(verb, non-3rd person singular pres-

ent) 和 COLON(colon) 的载荷较大,可以将其称为“限定词、动词现在时、冒号因子”; 因子 2 中 NNP(proper noun,singular) 、VBD(verb,past tense) 的载荷较大,可以将其称为“专有名词、动词过去式因子”; 因子 3 中只有 COMMA(comma) 的载荷较大,可以将其称为“逗号因子”; 因子 4 中 FW(foreign word) 、MD(modal) 、PRP(personal pronoun) 的载荷较大,可以将其称为“外来词、情态动词、人称代词因子”; 因子 5 中 JJ(adjective) 和 VBZ(verb,present tense,3rd person singular) 的载荷较大,可以将其称为“形容词、动词三单现因子”; 因子 6 中 IN(preposition or subordinating conjunction) 和 NN(noun,singular or mass) 的载荷较大,可以将其称为“介词、名词因子”; 因子 7 中 CC(coordinating conjunction) 的载荷较大,可以将其称为“并列连词因子”; 因子 8 中 POS(possessive ending) 、RB(adverb) 的载荷较大,可以将其称为“所有格、副词因子”; 因子 9 中 VB(verb,base form) 的载荷较大,可以将其称为“动词基本型因子”。虽然这些因子中有一些覆盖了多个词类,不太容易得到解释,但从总体上看,它们考察了文本中绝大部分常见词类,因此能比较全面地捕捉这些词类上出现的失误。

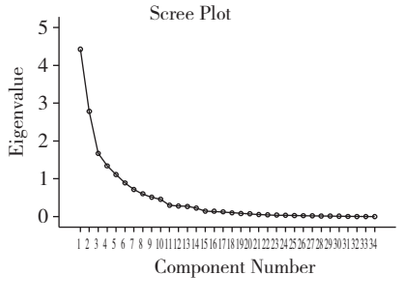

图 1 因子分析之碎石图

在因子得分系数矩阵中,找出那些能代表每个因子的标注,将其用于计算因子值的系数。在所有用于计算因子值的 17 个标注中,COLON 和 FW 的系数明显小于其他标注,也就是说它们对因子值的贡献有限。由于 LSCORE 在以上的因子分析中未被采用,因此在后面的回归分析中将其包括在自变量内。

最后,为了对写作风格因素有所考虑,我们将四个与写作风格有一定关系的变量也纳入自变量,它们分别是较为容易统计的 WORDLEN(平均词长) ,WDSTDEV(词长标准差) ,SENTLEN(平均句长) 和 STSTDEV(句长标准差) 。

这样,一共有了 16 个变量。它们分别是: SCORE、LSCORE、CRCTERR、FACTOR1 ~ FACTOR9(因子分析所提取的九个因子) 、WORDLEN、WDSTDEV、SENTLEN 和 STSTDEV。将 SCORE 作为因变量,其他变量为自变量,进行多元线性回归分析,结果显示: $r = 0.699$, $r^2 = 0.489$,调整后 $r^2 = 0.390$。且 ANOVA 分析显示: $F = 4.914$, $p = 0.000$。这说明模型具有统计意义,能解释因变量 SCORE 近 50% 的变异,即使是调整后的 $r^2 = 0.390$,也能解释 SCORE 1/3 强的变异。

接下来需要对回归模型进行诊断。根据图 2 中残差及预测值交叉分布情形,得知绝大多数的残差值都落在正负 2 个标准残差值之内,且均匀分布于 0 点上下,未成曲线分布,没有违反等分散性。图 2 中标出了分布于正负 2 个标准残差值之外的观测点的编号。此外,图 2 中常态概率分布情形所呈现的残差点,大致分布在由左下至右上 45 度的直线上,接近正态分布。

最后还须诊断自变量间的共线性。根据多元线性回归分析的结果,绝大多数变量的容忍度都比较接近 1,所有变量的 VIF 值都小于 10。因而排除了自变量间的共线性。经由以上检定之后,我们得出结论: 此多元线性回归模型的选取是合适的。其回归分析公式为: SCORE = -12.577 -0.416 × FACTOR1 +0.343 × FACTOR2 +0.300 × FACTOR3 -0.183 × FACTOR4 -0.132 × FACTOR5 +0.193 × FACTOR6 -0.651 × FACTOR7 +0.601 × FACTOR8 +0.544 × FACTOR9 -0.058 × WORDLEN +0.043 × SENTLEN +0.240 × WDSTEDV -0.003 × STSTDEV +5.222 × CRCTERR +18.793 × LSCORE

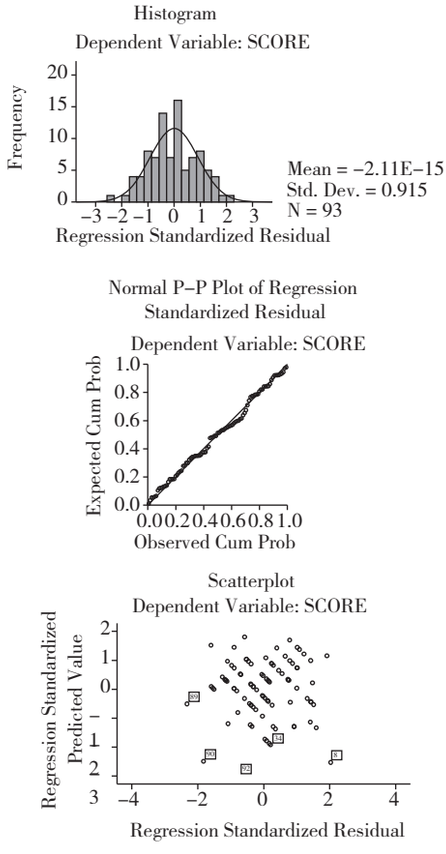

图 2 多元线性回归之残差统计图

此回归公式可通过语法因素对文本进行评判,预测其总体评分。下面,将使用三个测试集对此基于多元线性回归模型的语法模块进行测试,以观察其在实际评分中的效果。

3.3 对模块的测试

三个测试集分别由 CLEC 的 ST3 中随机抽取的 53 篇作文组成,共计 159 篇。它们也由评阅过训练集作文的两名评卷人独立评分⑨,并将评分结果与 ST3 的评分进行综合,得到测试对比的参照分。语法模块对测试集也分别进行评分,结果如表 3 所示。

表3　测试集之统计数据

测试集	篇目数	最低分	最高分	平均分	标准差
TEST1_H	53	7.00	14.00	9.59	1.71
TEST1_M	53	6.00	13.00	9.42	1.61
TEST2_H	53	7.00	14.00	9.55	1.65
TEST2_M	53	4.00	11.00	8.81	1.65
TEST3_H	53	6.00	13.00	9.60	1.57
TEST3_M	53	3.00	12.00	8.55	2.13
TEST_H	159	6.00	14.00	9.58	1.64
TEST_M	159	3.00	13.00	8.92	1.84

表3中，TEST1_H、TEST2_H、TEST3_H、TEST_H为人评结果，TEST1_M、TEST2_M、TEST3_M、TEST_M为机评结果。其中TEST_H和TEST_M为三个测试集的整体情况。整体上看，机评结果的平均分为8.92，比人工评分的平均分低0.66分，但两者的标准差比较接近，分别为1.64和1.84。

4. 评分准确性评价

我们将十五分制分数转换成五分制分数[⑩]，这有两方面的考虑。首先，Burstein等(1998)以及Burstein和Chodorow(1999)对*e-rater*系统所作的评价是基于六分制的。如果我们的分数体制与它相对接近，就有了可资比较的基础。其次，五分制对评分等次划分的粗细适当，便于人评掌握。基于五分制的评分实际相当于英语四六级考试写作评卷中的定档。这样，将十五分制转成五分制再作考查，可以很好地观察语法模块能否准确定档。本文在给出基于五分制比较的同时，也提供十五分制的比较结果以供参考。

对语法模块评分的准确性进行判断，须就其评分同时做出相关性评价和一致性评价。

4.1 评分相关性评价

将测试集全部样本的人评结果与机评结果作皮尔逊积差相关，十五分制的相关系数为0.52，五分制的相关系数为0.49，显示在自由度为157、p值小于0.01的双尾检验中具有统计意义。语法模块在对三个测试集所作的评分中，最低分为3分，最高分为13分，全距达到了10分。折合到五分制，最低分为1分，最高分为5分。这说明语法模块能将作文分别确定到五个档次中去，没有趋中效应。经过进一步观察，语法模块与人工评卷之间中等程度的相关，主要体现在人评高端分数的分歧上。对此，在4.3部分有深入讨论。

4.2 评分一致性评价

在对*e-rater*进行评分一致性评价时，Burstein和Marcu(2000)将评分一致确定为完全一致和临近一致(exact-plus-adjacent agreement)。完全一致是指参与比较的两名评卷人或一名评卷人与机器评卷系统之间所评分数差异为0分，临近一致是指两者所评分数差异不超过1分。需要说明的是，四六级评卷的档次分分别为2、5、8、11和14分，分别对应以上五分制中的各分数。五分制的临近一致应为不超过1分的评分差异。而十五分制的临近一致应为不超过3分的差异，即不超过一个档次的误差。

评分一致性的标准可以从两方面确定：(1)所有分数的一致性；(2)基准分的一致性。

从三个测试集分别的结果来看，十五分制中所有分数的评分一致(以不超过3分为准)比例为90.7%～92.5%之间。而五分制中评分一致(以不超过1分为准)的比例为92.4%～98.1%之间。完全一致的比例为49.1%～50.9%，这意味着语法模块定档完全准确的比例达到了近50%，这达到了*e-rater*的评分一致性水平。

再来看基准分一致的情况。GMAT六分制中的基准分是4分，也就是所有作文中得分最多的分数。在大学英语四六级的写作成绩中7、8、9分是得分最集中的分数段，这一分数段对应于五分制就是3分。通过考察五分制中基准分3分的一致性，就可以观察语法模块在基准分一致性上与*e-rater*[⑪]是否存在差异。

在三个测试集共同构成的样本集中，人评基准分3分占总数的46.5%。而其与机评分的一致情况为：完全一致46.0%，临近一致43.7%，总计89.7%。这说明，语法模块在基准分档次的确定上有90%是准确的，其中完全准确的近一半。

4.3 讨论

语法模块的评分在与人工评分仅具有中等程度相关的条件下能否具有辅助评卷的能力呢？让我们先看一看表4。这是测试集中人评分与语法模块评分差异在4分(十五分制)以上的所有作文的统计数据。

表4　测试集中4分以上差异作文统计表

人工评分	语法模块评分	分数差异	作文序号
13	6	7	2－3
9	3	6	3－36
11	6	5	1－3
12	7	5	1－30
11	6	5	2－13
14	9	5	2－44
8	3	5	3－14
7	12	5	3－49
12	8	4	1－8
14	10	4	1－14
11	7	4	2－21
12	8	4	3－3
12	8	4	3－25

由表4可以看出，在159篇测试作文中，一共有13篇作文的人机评分数差异在4分以上，占总数的8.2%。其中76.9%的作文是人评给了10分以上的高分，而语法模块仅给了相对较低的分数。通过进一步观察，可以看出它们的语法水平中等，而立意和内容则比较不错，因而人评分相对较高。若将语法模块用于辅助评卷，通过事后复查，这部分差异在很大程度上可以得到消除。比如说，当语法模块完成所有作文的评卷后，凡写作分在9分以下、而客观题部分得分很高[⑫]的作文由人工重评，即可消除约61.5%(即以上13篇中的8篇)的此类误差，使最终的人机评分相关系数达到0.70左右。此外，我们还观察到，几乎没有人工评分很低(7分以下)的作文与机评分相差2分以上，这也说明，只要我们控制了以上这部分差异，语法模块的评分总体上是可靠的。

通过以上分析，语法模块的信度在一定程度上可以接受。关于语法模块的评卷效度，有必要作以下说明。一般而言，在写作测试评价中，文章的内容和语言需要兼顾。因此，

e-rater 要从句法特征、写作技巧和主题等三个方面进行考察以模拟人工评卷；而大学英语四、六级写作评卷要综合考虑内容和语言进行总体印象评分。写作技巧包括篇章衔接、连贯手段和文体特征等，而语言则主要表现为句子合语法性和语法错误的多寡。对于一个具有高健壮性的写作评分系统而言，其效度检验需包括对以上各方面的考察。本文利用多元线性回归的语法模块对写作总分做出预测，并借助人工复查消除误差的辅助评分实践，自然不能以写作评分系统所应具有的效度标准来对此进行衡量。此外，本研究所采用的语法检查器自身的信度和效度均已得到检验（贺俊杰，2006）。若将语法模块与其他考察语义语用的模块进行整合，由此建立的完整评分系统将具有应有的效度。

5. 结论

语法模块在写作评分中仅考虑与语法有关的因素，所以与作文的综合评分仅表现出中等程度的相关。但作为写作评价过程中不可或缺的组成部分，它能够发挥应有的作用。比如，以机器评分极高的稳定性这一优势，语法模块可为作文评卷辅助定档。此外，还可作为初、中级学习者群体写作评分有力的辅助工具，应用于机助适应性语言学习。经过整合利用 LSA 等技术的评分模块，使语义和语法因素同时得以考察，最终的写作评分系统将更有应用价值。

注释：

①在 Burstein 等（1998）关于 *e-rater* 项目的报告中，评分准确性是根据两名评卷者或一名评卷者和 *e-rater* 之间在六分制里分数差异不大于 1 分的百分比。本文中，我们也采用这一为大家所接受的标准。

②其中部分原因可能是由于这些系统设计时所针对的文本主要是由母语者所提供的。

③作为示例，本文以大学英语四级考试写作评分来讨论语法模块的构建。若要推广到其他层次和类型考试的作文评分，只需针对特定的样本集进行训练即可。

④两位评分人都具有四六级评分资格，都曾参与四六级评卷工作。

⑤之所以这么做，是为了确保评分的信度。这两位评分人的评分信度为 0.72，他们的综合分与 CLEC 中原有分数的相关系数为 0.62。

⑥主要对链语法分析器的候选链排序进行了优化，具体改进详见贺俊杰（2006）。链语法的分析原理可参看有关文献，如 Sleator 和 Temperley（1993）等。

⑦方括号内的词为链语法分析器所无法成功链接的词。

⑧本研究使用的标注集为 PENN TREEBANK 语料库的标注集。详见 http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html。在训练集中有如下标注没有出现：JJS、LS、NNPS、RP、SYM、WP $。

⑨在测试集中，两名独立评分人的评分者信度为 0.70，他们的综合分与 CLEC 中原有分数的相关系数为 0.64。

⑩十五分制分数转成五分制分数公式：$SCORE_5 = ROUND\left(\frac{(SCORE_{15}+1)}{3}\right)$ 公式中 ROUND( ) 函数为取整函数；$SCORE_{(15)}$ 为十五分制分数；$SCORE_5$ 为五分制分数。

⑪据 Burstein 和 Chodorow（1999）报告，TWE 中基准分的评分一致性（包括完全一致和临近一致，下同）为 84%，GMAT 中基准分的评分一致性为 83%。所谓基准分一致性，是指两个评分人一致给出基准分 4 分（在六分制中最常见的分数）所占的比例。在此，我们设定五分制中的基准分为 3 分。

⑫诸如写作等的综合语言能力与由客观题反映出来的各单项语言能力有较高相关。但客观题分高到何种程度则需进一步研究确定。

参考文献：

[1] Burstein, J. & M. Chodorow. 1999. Automated essay scoring for non-native English speakers [R]. Proceedings of a Symposium on Computer-Mediated Language Assessment and Evaluation in Natural Language Processing. College Park: University of Maryland.

[2] Burstein, J., K. Kukich, S. Wolff, C. Lu, M. Chodorow, L. Braden-Harder & M. Harris. 1998. Automated scoring using a hybrid feature identification technique [R]. Proceedings of the 17th International Conference on Computational Linguistics. Montréal: University of Montréal.

[3] Burstein, J. & D. Marcu. 2000. Toward using text summarization for essay-based feedback [R]. Proceedings of Le 7e Conference Annuelle sur Le Traitement Automatique des Langues Naturelles TALN'. Lausanne: Suiss Federal Institute of Technology.

[4] Powers, D., J. Burstein, M. Chodorow, M. Fowles & K. Kukich. 2002. Stumping E-rater: Challenging the validity of automated essay scoring [J]. *Computer in Human Behavior*, 18(2): 103 – 134.

[5] Shermis, M. & J. Burstein. 2003. *Automated Essay Scoring: A Cross-Disciplinary Perspective* [M]. Mahwah: Lawrence Erlbaum Associates.

[6] Sleator, D. & D. Temperley. 1993. Parsing English with a link grammar [R]. Proceedings of the Third International Workshop on Parsing Technologies (IWPT'93). Tilburg: Tilburg University.

[7] 葛诗利 陈潇潇. 2007. 国外自动作文评分技术研究 [J]. 外语电化教学，(5): 25 – 29.

[8] 贺俊杰. 2006. 基于组合模式的语法检查 [D]. 广州：广东外语外贸大学.

[9] 梁茂成 文秋芳. 2007. 国外作文自动评分系统评述及启示 [J]. 外语电化教学，(5): 18 – 24.

Abstract: This research has constructed a computer – assisted essay scoring system for Chinese learners of English through the statistical model of multiple linear regressions, and tested the reliability of the system with a large-scale corpus. The conclusion is drawn that the essay scoring module thus derived, examining grammar and basic writing styles, is capable of making reliable prediction of the overall scores for writing samples, when complemented with post-hoc human inspection. Thus the scoring module is applicable to computed-assisted essay scoring practices under designated circumstances.

Key Words: English writing; automated scoring; grammaticality evaluation; computer-assisted scoring