大学英语作文自动评分研究中的问题及对策

葛诗利[1]，陈潇潇[2]

(1.华南理工大学 外国语学院，广东 广州 510640； 2.广东金融学院 外语系，广东 广州 510520)

摘要：面向大学英语写作教学的自动作文评分研究存在四个难题：评分标准、针对性、通用性和人机界面的划分。自动评分要以人工评分为准，并结合写作教学理论；评分要考虑中国学生写作特点，使评价具有针对性；为了构建一次训练多次使用的通用评分模型，语言使用和内容需分别处理；作文评分必须有人参与，适当的人机界面能充分发挥机器和人的长处，使自动评分高效而准确。

关键词：大学英语；写作教学；自动作文评分

中图分类号：H319 **文献标识码**：A **文章编号**：1002-2643(2009)03-0021-06

40多年来，自动作文评分研究在国外，尤其是美国，得到了长足的发展，并已付诸应用。在这些研究中，充分利用了计算机统计技术、自然语言处理技术、信息检索技术(梁茂成等，2007：20)，甚至人工智能技术。(Elliot，2003：71)近年来，该方向的研究在国内也逐渐得到外语教学界和自然语言处理界的重视，相关研究如梁茂成(2005)，梁茂成等(2007：18－24)，和葛诗利、陈潇潇(2007：25－29；2007：43－50)等。但与国外相比，国内的相关研究具有明显的探索性和探讨性，研究的针对性和实用性较弱。

在我国大学英语写作教学中，由于教师作文批改负担过重，学生写作训练不足，直接导致了学生写作水平的提高缓慢。在这方面，自动作文评分不失为一个良好的解决方法。研究实用的、面向大学英语写作教学的自动作文评分方法，必须解决作文评分标准的形式化、作文评分的针对性、通用性和人机界面划分四个方面的关键问题。

1.0 自动作文评分的关键问题及对策

马希文指出，要想让计算机去解决某种问题，从理论上说，必须满足下列三个基本的前提条件：

第一，必须把待解的问题形式化。

第二，这种问题必须是可计算的。

第三，这种问题必须有一个合理的复杂度，也就是要避免指数爆炸。(袁毓林，2001：158)

满足这些条件的计算称为典型计算。但是，现实世界中有很多不符合或者不完全符合这三个条件的问题，也需要使用计算机来解决，这就是智能计算，如自然语言理解问题。

解决这类问题的关键方法，一个是限制原问题的范围，在该范围内该问题变成典型计算问题，超出该范围的部分由人另想办法处理；另一个是允许存在一定量的计算偏差。

计算机自动作文评分属于典型的智能计算，中国大学生所写的汉英中介语(inter-language)作文的自动评分同样是如此。在该问题的形式化、算法化和计算复杂度设计上都存在着巨大的难题，如何限制、缩小自动评分问题的范围，求得较好的近似解，即与真实水平差距不大的作文评分，就成了我们努力的方向。

1.1 作文评分标准问题

研究计算机自动作文评分，首先要面对的就是评分标准的形式化问题。当前主要有两种外语作文评分形式：整体评分和分项评分。(李志雪，2004：61)对比而言，分项评分的各项标准更加精细。这两种形式的标准都是面向人工评分设计的，很难或者根本不能够直接形式化并为计算机所用。但困难再大，自动作文评分中还是需要一个形式化的评分标准。当前的解决方法有两个，一个是从写作理论出

基金项目：本论文的工作得到国家自然科学基金课题(编号 60872121)的资助。

收稿时间：2009-04-05

作者简介：葛诗利(1969－)，男，汉族，山东烟台人，讲师，博士。研究方向：计算语言学。

陈潇潇(1975－)，女，汉族，辽宁大连人，副教授，硕士。研究方向：应用语言学。

万方数据

发,另一个是从人工评分出发。前者按照写作理论细化评分标准,提取内容和语言使用方面能够形式化的评分特征,同时分配各特征在作文评分中的权重,如曾用强(2002:26-31)的过程化作文评估;后者是采用人工评分,通常是多位专家共同评分而达到一致的各个分数档的多份试卷,利用统计和自然语言处理技术,提取数十至数百项文本特征,采用机器学习的方式,自动发现其中某些有效评分特征及其权重,用于未评分作文的评分。后一种方法被当前绝大多数自动评分研究所采用。

这两个方法各有所长,前者可以说是透明的,作文分数比较容易解释,能提供更多更直接的写作反馈,符合外语学习和写作的研究理论,但问题是"使主观测试评分标准客观化的可行性办法尚未出现"。(Bachman,1990:70)比较容易提取的文本特征往往不能够直接反映写作质量,而写作理论中的各种要求很难转化为形式化的文本特征;并且,对于能够提取的文本特征在评分过程中其权重分配也是见仁见智,甚至在写作教学研究中也未能达成一致。后者类似一个"黑匣子",机器全自动地以训练集为基准,挑选有效特征并加权,得到评分模型,但特征挑选和加权的原因根本不为人所知。在这种方法中,为了追求机器评分与人工评分的一致性,投入很多的特征,但对于某一个作文题目所构建的评分模型,在机器筛选以后真正使用的特征并不多,而这些特征在不同题目作文的评阅中不尽相同。即便是同一个特征,在不同题目的评分模型中,其权重也不相同,在线性回归评分模型中甚至可能出现权重相反的情况。这些都很难使作文分数有一个合理的解释,而这些特征的使用也不能够提供真正有价值的写作反馈。

为了制定形式化的、尽可能客观的面向大学英语写作教学的作文评分标准,我们首先确立了语言和内容分别评价,语言为主、内容为辅的评价体系。这样做的理由有两个,一个是从大学英语四、六级考试作文来看,除了个别恶意欺骗的学生外,绝大多数作文都不会跑题;另一个是我们假定在日常教学的写作练习中,学生们都是积极参与,不会出现恶意欺骗的现象。这样我们对于作文的评价就可以把范围缩小到作文的语言使用上。在自动作文评分研究中,针对语言使用,我们采取了人工选取特征和机器统计加权二者结合的方法,即首先根据前期写作研究,人工确定需要提取的有限几个特征,然后通过与已评分作文的拟合确定各特征的权重。这样,每个特征对作文成绩都会具有较强的解释力,评分模型性能也达到了较高的水平。对于作文内容,我们也采取了相应的衡量方法。

1.2 自动作文评分的针对性问题

虽然多种国外自动作文评分研究取得了较好的评分效果,但这些研究主要是面向英语母语写作者,对外语写作者,即便是较高水平的外语写作者,如托福考生,自动作文评分与人工评分也会出现统计上的显著性差异。(Burstein & Chodorow,1999:68-75)对于我国大学英语写作者,尤其是低年级学习者这种充斥着高频率的词汇和句法方面错误的英语作文,尚没有较为成功的研究。要研究具有针对性的、面向大学英语写作教学的自动评分,首先必须研究中国大学生英语作文的写作特点,然后广泛参考自动作文评分研究中的各种技术手段,最后根据自然语言处理发展的状况确定最适合的自动评分方法。

1.2.1 中国大学生英语作文特点分析

我国非英语专业大学生英语作文中与作文成绩关系比较密切的语言使用特征包括词汇、短语、句法、搭配和错误等几个方面。

濮建忠(2005:130)基于 CLEC 语料库的研究指出,大学生英语写作能够熟练使用的词汇量,按照类符(type)(如,ask、asked、asking 计为 3 个单词)计算,也只是两千二三百个,如果按照词族(word family)(即 ask 的 3 种形式计为一个单词)计算,还会少得多。濮建忠(2005:131)的研究同时表明,四、六级作文的词汇密度和本族语者相比也还存在一定的差距。这项研究是关于大学英语四、六级考试作文的。尽管在考试中写作者词汇能力的表现受到一定的影响,但也基本能够反映写作者的词汇水平。

在短语或者说词块方面,国内近年来的很多研究都涉及到作文质量或者整体语言能力。刁琳琳(2004:37)的调查表明,英语专业本科生"低年级与高年级的词块能力呈现显著差异,且高年级的词块能力与语言整体能力显著相关"。丁言仁、戚焱(2005:52)的研究表明,对于英语专业四年级本科生,"与语法知识相比,词块知识对…写作成绩具有更强的预测力",这里的"词块知识"计算的是"以动词开头的词块数量"。王立非、张岩(2006:40)研究了英语专业学生的英语议论文语料,发现"中国学生写作中存在过度使用 3 词语块,语块的种类较少"并且"具有口语化倾向"。徐芳芳、应咏梅(2007:108)分析了英语专业大一和大四学生的作文,发现"高水平组学生使用名词性聚合短语的数量和种类较多,

万方数据

而低水平组学生虽然较多地使用约定俗成的语句和句子构造型短语，却在多样性上不及高水平组，且有较强的口语化倾向”。所有这些研究表面，作文中短语或词块的使用能够预测作文成绩，并且与学生的年级，即英语总体水平的高低有着显著的相关性。当然，这些研究的对象都是英语专业的学生，但同样对大学英语写作中短语使用的研究具有参考价值。

濮建忠(2005:135－139)的研究也指出，我国大学生四、六级英语作文在句法和搭配方面的言语失误种类和数量较多。这也是中介语写作的显著特点，若能充分挖掘，可显著提高自动评分的准确性。但目前自然语言处理中对于正常的自然语言句法和搭配处理准确率还比较低，对于这种充斥各种言语失误的中介语作文的自动处理更加困难。由于不正确的句法和搭配分析会严重影响自动评分的准确性，这些方面大部分特征目前尚不能够采用，在此不多做分析。

1.2.2 主流自动作文评分技术分析

当前，主流的自动作文评分技术以5类6种自动评分系统为代表，包括了注重语言形式评分的PEG、注重内容评分的IEA、二者兼顾的E-rater、发挥人工智能技术的IntelliMetric、基于文本分类技术的Larkey的系统和BETSY。(葛诗利等，2007:26－28)

作文的人工评分，不管是标准的制定还是评分的进行通常都是从语言形式和文章内容着手。计算机自动作文评分通常也是从这两个角度入手，但由于“语言范畴不能或很难形式化”(马希文，1987:18－21)，所以只能缩小问题的范围，设法取得文章语言和/或内容的较准确评价。如PEG就是统计文章长度、各种词类的数量、词长的变化等浅层文本特征来衡量文章的语言使用；IEA是用词汇统计，主要是实义词的统计得到文章内容的评价；E-rater的构造较为复杂，但其评分中权重较高的内容评价仍然是基于词汇的统计，而在语言使用的评价中，尽管相关研究宣称采用了自然语言处理的句法分析器，但鉴于目前自然语言句法分析的进展，仍然应该是浅层文本特征统计起到较大的作用；IntelliMetric的底层技术未见详细报道，但同样受制于当前自然语言处理技术的发展；最后，基于文本分类技术的自动作文评分系统实际上也是基于词汇的统计。

从以上简要的分析可以看出，在自动评分中，作为内容比对的词汇统计具有非常重要的地位，这是因为在英语母语，或者高水平的英语外语的作文写作中，“内容是衡量作文质量最重要的方面”。(文秋芳，2007:67)自动评分中通行的做法就是把内容评分转换为词汇向量的统计比较。当然，这不可避免地会损失大量的内容信息，因为词汇的堆积不可能等同于句子的意义，更何况词汇不仅多义，且有歧义。所以，自动作文评分中对内容的评分只能是对人工文章内容评分的一种近似的模仿。

对作文语言使用的评价通常包括词汇、句型、语法和操作细节几个方面，“词汇指作文所用单词的词频高低、单词的搭配以及恰当性；句型指作文所用句型的复杂度和多样性；语法指作文所用语言是否符合语法规则；操作细节指拼写、大小写、标点等使用情况”。(文秋芳，2007:67)在当今的自然语言处理研究中，词汇级的处理技术，如词频统计、操作细节中的拼写和大小写检查已经非常完善，对于二元词频统计的邻接词汇的搭配和采用模式匹配的某些特殊句型的识别准确率也较高，但稍高一些的要求，如远距离搭配、搭配的恰当性、标点使用和语法准确性方面目前尚没有达到实用的程度。所以，当前自动作文评分中对语言的评分主要也还是集中在词汇层次上和少部分较为成熟的语法分析，如一致性和动词构成上。由此可以看出，自动作文评分中对语言的评分也只是对文章中语言使用的一个大概的评估。

1.2.3 针对大学英语写作教学的自动评分方法

对于母语或者高水平外语作文，偏向内容的词汇统计就能较好地衡量作文的整体质量，若能辅以对作文语言使用的评价，便能取得更理想的结果。但由于我国大学英语学习者的写作属于外语作文，其评分标准有所不同，计算机在作文语言和内容上的处理策略也须有所改变。

在自然语言处理方面，对于正常的自然语言，对词形的研究已经非常准确，对词性的研究也比较准确，而句法自动分析的准确率却比较低，尚未到可实用的水平，对于更深入的实用语义处理更不可能。对于低水平的英语写作语言，即作文中高频率出现词汇和句法方面错误的汉英中介语，自动的句法和语义分析更不可能。传统的语法分析器效果不好，其他的尝试，如采用基于链语法(link grammar)的句法分析器来分析评判英语学习者的作文，得到的结果同样也不理想。(Lonsdale & Strong-Krause，2003:61－67)

这种结果应该是意料之中的，因为所有的句法分析都需要有一个形式化的语法模型，而过去人们研究自然语言的分析和生成的理论和技术时，对语

万方数据

法模型的要求不是很严格。比方说,机器翻译系统对源语言做语法分析时设定的形式语法可以相当宽泛,因为它可以认为凡是交给它的句子都是合乎语法的;生成目标语言时所采用的语法模型又可以相当缩减,因为它只要求生成的句子能被人看懂。而自动作文评分的语法分析,类似于计算机校对系统的要求。语法模型按最理想的目标来说必须完全同真实语言相吻合,既不宽泛也不缩减,即从理论上说,需要有这样一个形式化、算法化的语法系统:对于一切文稿,凡符合该语法的都是正确的,凡不符合该语法的都是错误的。(宋柔,2001:45 - 54)可惜的是,这个目标目前是达不到的。建立准确的语法语义分析系统对于母语文章尚达不到,对于错误百出的外语作文更不可能。

这是我国大学生汉英中介语作文自动评分的劣势,但这种文章的自动评分也有其优势,那就是词汇量相对较小,短语使用数量和变化形式有限,从理论上说,可以建立一个接近完备的列表供计算机使用。所以对这种作文语言的评价,采用目前自然语言处理中非常成熟的词汇级处理技术和模式匹配技术会更加有效,而更深层次的分析,如语法剖析或者语义分析都只能带来更多的分析错误。

当前自然语言处理研究中对内容的研究实际上也就是对文章中词汇,主要是实义词的比对研究,但对于大学英语写作这种充斥着各种词汇错用和误用的文章,这种比对方法与评卷人能够揣摩出的文章内容恐怕相去甚远,所以大学英语作文评价不能像本族语作文一样只是依靠内容评分,甚至内容比例较高的自动评分也会存在较大的问题。但如果内容评价只是作为参考,这种实义词汇比对的方法还是可以考虑的。

1.3 自动作文评分的通用性问题

当前的自动作文评分研究都是针对某一题目作文,几乎都需要采用数百篇人工评分的作文作为训练集(Dikli,2006:53),利用机器学习的方法建立评分模型,进而为该题目的其他未评分作文评分。这种方式对于几十万甚至数百万参与者的大规模英语考试作文评分效率较高。但是在大学英语写作教学中,每位大学英语教师教授的学生数量一般不会超过200人,如果每布置一个题目的作文就需要训练一个自动评分模型,所收集的学生作文数量还不足以供模型训练,所以面向大学英语写作教学的计算机自动作文评分需要一种"一次训练多次使用的、非特定题目的通用"评分模型。

在面向大学英语写作教学的自动作文评分研究中,把语言和内容分别处理。在训练集作文的收集上覆盖了尽可能多的作文题目,在训练集作文的人工评分上注意偏重于语言的使用,在构建评分模型的文本特征中剔除了题目相关的特征。由此构建成为非特定题目的、通用的、关于语言使用的自动作文评分模型,即该模型可用于大学英语低年级任何题目作文的自动评分。当然,由于该模型未涉及作文内容的比较,评分准确度难免会受到一定的影响。但是实验证明,该准确度在日常教学中也处于可以接受的范围,如果辅以适当的人机界面的划分,当可用于写作教学。

在内容方面,对于非特定题目的通用评分模型,不可能通过训练得到所有题目作文的内容评价标准,那么,就只能采取不经训练的内容评价方法。在面向教学的自动作文评分研究中,我们采用了文本自动聚类的方法。同一位教师所教授的一个班或者几个班的学生所写的同一个题目作文收集起来之后,采用实义词聚类的方法,判断某一篇或几篇作文可能是跑题作文。当然,未经训练的文本聚类方法比经过大量文本训练的文本分类方法精确性要差,但内容评价本身就处于从属的地位,再加上适当的人工参与,也可以对学生有一个较为满意的反馈。

1.4 自动作文评分的人机界面划分问题

所谓人机界面的划分,就是研究中必须决定哪些问题交给计算机去解决,哪些问题必须要有人的参与,从而达到以最少量的人工参与达到系统效率的最大化。由于计算机不可能真正理解人的语言,更别说充斥着各种错误的语言,而作文又是通过语言传达人的思想,所以作文不可能完全交给计算机去评判。但是计算机自动作文评分也有很多长处,首先,在大多数情况下,自动评分能够给出与人工评分较为近似的评分,另外,自动评分过程具有"再现性、一致性、客观性、可靠性和高效性"等特点。(Williamson,Bejar & Hone,1999:158 - 184)这就需要在方法研究和系统开发的时候深入分析计算机和人各自的长处,精确地划分人机界面,使得计算机最大限度地发挥高速运算的功能,以最少的人工参与,取得最高的工作效率。

国外研究,如 E-rater 在 GMAT 和托福等高利害关系(high-stakes)考试中是采取人机各半的划分方式,即计算机和一名评分员共同为一篇作文评分,而在基于网络的练习中完全交给计算机去评判。(Hearst,2000:25)我们认为这两种方式都不适合大

万方数据

学英语写作教学，因为前者丝毫未减少教师的负担，后者过于相信计算机，对学生会造成一定的负面影响。

我们在面向教学的通用自动作文评分研究中发现，针对语言使用的评分，低分段评分准确率非常高，中、高分段，尤其是高分段准确率较低。也就是说，自动评分评为低分的，基本都是低分的，评为高分的，不一定就是优秀作文。我们据此划分了人机界面，也就是教师要尽可能参与评为高分的作文的复核，如果确实时间和精力有限，至少也要向学生说明这种情况，让学生自我判断，确有疑问的，提交给教师复核。

在作文内容方面，通过同一题目作文的实义词自动聚类，可以发现某一篇或数篇作文与大多数作文用词不同，这些少量的作文需要提交给教师，人工判断是否确实跑题。

以这种分工方法，充分挖掘了计算机的潜力，教师的工作量可以降低到原来的 1/10 到 1/5，而学生对于作文的评分反应也类似于原来的人工评分。这样就可以在减轻教师负担的同时，促进学生的写作练习，使其更快地提高英语写作水平。

2.0 计算机自动作文评分的发展前景

同教师的人工评阅相比较，计算机自动作文评分有着极大的优势。首先，它可以提高阅卷速度，甚至达到网上即时阅卷。其次，它可以极大地降低教师的阅卷工作量，教师可以把充足的时间和精力放在教学组织和讲解上；最后，也是最重要的，就是学生可以写出更多的文章，写作能力会提高得更快。当然，对于写作教学而言，只有一个作文评分是远远不够的，详尽而准确的反馈对于学生写作能力的提高会起到更积极的推动作用。但限于篇幅，计算机自动作文反馈方面的研究将另文详述。

计算机作文评分的优势还在于它完全符合当今的《大学英语课程教学要求》，顺应了“以现代信息技术，特别是网络技术为支撑，使英语教学不受时间和地点的限制，朝着个性化学习、自主式学习方向发展”的潮流。目前，大多数高校的大学英语写作教学仍然停留在写作技巧的讲解，写作训练严重匮乏的阶段，使用计算机网上提交作文，教师做出评估反馈的少之又少。随着计算机的普及，尤其是网络教学的推广，计算机自动作文评分必将起到越来越大的作用。

写作教学与评测的深入研究，计算机性能的提高，计算语言学和人工智能研究的进展，将使计算机自动作文评分的准确率越来越高，从而写作评分的自动化程度也会得到相应的提升。写作教学与研究方面，首先要解决的就是评分标准的客观化问题，其次是发掘更能够反映写作质量的新特征，尤其是那些能够为学生提供有用的反馈、对写作改进具有指导意义的特征。计算语言学方面，直接相关的是中介语相关理论和技术，如中介语词性标注技术，基于巨大规模语料库的词语接续关系的理论和技术等。人工智能方面，最重要的是各类简单知识的表示方法。

随着计算机自动作文评分准确率的提高，该方法不仅可以用于教学测试这种低利害关系的作文评分中，也能够使大规模测试的作文评分更加客观，更加准确。此外，计算机自动作文评分也能够促进写作教学的研究，以及写作评价标准的客观化研究。

参考文献

[1] Bachman, L. F. *Fundamental Considerations in Language Testing* [M]. Oxford: Oxford University Press, 1990.

[2] Burstein, J. & M. Chodorow. Automated essay scoring for nonnative English speakers[A]. In *Proceedings of the Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing*, *Joint Symposium of the Association of Computational Linguistics and the International Association of Language Learning Technologies* [C]. College Park: Maryland, 1999.

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

[4] Elliot, S. IntelliMetric: From here to validity[A]. In M. D. Shermis & J. Burstein (eds.). *Automated Essay Scoring: A Cross Disciplinary Perspective* [C]. Mahwah, NJ: Lawrence Erlbaum Associates, 2003.

[5] Hearst, M. The debate on automated essay grading [J]. *IEEE Intelligent Systems*, 2000, 15(5): 22 - 37.

[6] Lonsdale, D. & D. Strong-Krause. Automated rating of ESL essays[A]. In Burstein J. & C. Leacock (eds.). *HLT-NAACL* 2003 *Workshop*: *Building Educational Applications Using Natural Language Processing*[C]. Edmonton, Alberta: Canada, 2003.

万方数据

[7] Williamson, D., I. Bejar & A. Hone. A mental model comparison of automated and human scoring [J]. *Journal of Educational Measurement*, 1999, 36:158 - 184.
[8] 刁琳琳. 英语本科生词块能力调查[J]. 解放军外国语学院学报,2004,(4):35 - 38.
[9] 丁言仁,戚焱. 词块运用于英语口语和写作水平的相关性研究[J]. 解放军外国语学院学报,2005,(3):49 - 53.
[10] 葛诗利,陈潇潇. 国外自动作文评分技术研究[J]. 外语电化教学,2007,(5):25 - 29.
[11] 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.
[12] 李志雪. 如何更加客观合理地给学生作文评分[J]. Sino-US English Teaching,2004,1(11):61 - 63.
[13] 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.
[14] 宋柔. 计算机辅助汉语校对系统[J]. 当代语言学, 2001,3(1):45 - 54.
[15] 王立非,张岩. 基于语料库的大学生英语议论文中的语块使用模式研究[J]. 外语电化教学,2006,(4):36 - 41.
[16] 文秋芳. "作文内容"的构念效度研究[J]. 外语研究,2007,(3):66 - 71.
[17] 徐芳芳,应咏梅. 高、低水平组学生英语写作中词汇短语的对比研究[J]. 绍兴文理学院学报,2007,27(11):107 - 109.
[18] 袁毓林. 计算语言学的理论方法和研究取向[J]. 中国社会科学,2001,(4):157 - 168.
[19] 曾用强. 过程化的写作评估模式[J]. 福建外语,2002,(3):26 - 31.
[20] 濮建忠. 基于学习者语料库的中国非英语专业大学生中间语状况调查[A]. 杨惠中、桂诗春、杨达复编. 基于 CLEC 语料库的中国学习者英语分析[C]. 上海:上海外语教育出版社,2005.

The Key Problems and Solutions in Automated Essay Scoring for College English Teaching in China

GE Shi-li[1], CHEN Xiao-xiao[2]

(1. School of Foreign Languages, South China University of Technology, Guangzhou 510640, China; 2. Foreign Language Department, Guangdong University of Finance, Guangzhou 510520, China)

Abstract: There are four difficulties in the research on automated essay scoring (AES) for college English teaching: scoring criteria, pertinency, generality, and human-machine interface. The scoring criteria of AES should be based on both the results of human scoring and writing theories; Chinese students' writing characteristics should be considered in AES researches so as to design a pertinent system; in order to construct a one-training-multi-using general AES system, language using and content in an essay should be evaluated separately; there must be human raters involved in essay scoring, and a suitable human-machine interface can give full play to their own advantages, which can make an AES system more efficient and accurate accordingly.

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

万方数据