[19] 中华人民共和国国家知识产权局

[51] Int. Cl[7]
G06F 17/00
G06F 19/00

[12] 发明专利申请公开说明书

[21] 申请号 200510040305.6

[43] 公开日 2005 年 11 月 23 日

[11] 公开号 CN 1700200A

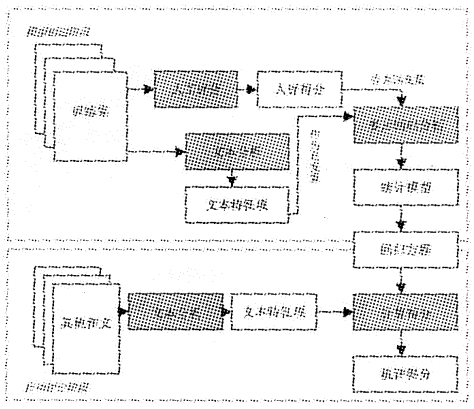
[22] 申请日 2005.5.30
[21] 申请号 200510040305.6
[71] 申请人 梁茂成
地址 221000 江苏省徐州市徐州师范大学二分部 2 号楼二单元 201 室
[72] 发明人 梁茂成
[74] 专利代理机构 南京苏高专利事务所
代理人 陈 扬

权利要求书 1 页 说明书 5 页 附图 1 页

[54] 发明名称 英语作文自动评分系统

[57] 摘要

本发明公开了一种英语作文自动评分系统，包括由一组英语作文集合而成的训练集、文本特征项、回归方程和带有输入和输出装置的电脑；训练集通过输入装置贮存在电脑中；文本特征项是通过对训练集中的作文进行文本分析而得到的信息，并将此信息作为自变量；回归方程是将训练集中的作文人评得分作为因变量与自变量进行多元回归分析并通过统计模型而建立的运行方式；对输入电脑中的待评分作文进行文本分析，将得到的文本特征项作为自变量，经过回归方程运算后得到评分结果，并通过电脑显示。 本发明可实现对中国学生英语作文的大规模机器评分，资源消耗低、评分信度可靠。 它广泛适用于各种大规模英语作文考试的评分中，具有极大的实用价值。



1、一种英语作文自动评分系统，其特征是：它包括由一组英语作文集合而成的训练集、文本特征项、回归方程和带有输入和输出装置的电脑；所述训练集通过输入装置贮存在电脑中；文本特征项是通过对训练集中的作文进行文本分析而得到的信息，并将此信息作为自变量；回归方程是将训练集中的作文人评得分作为因变量与自变量进行多元回归分析并通过统计模型而建立的运行方式；对输入电脑中的待评分作文进行文本分析，将得到的文本特征项作为自变量，经过回归方程运算后得到评分结果，并通过电脑的输出装置显示。

2、根据权利要求 1 所述的英语作文自动评分系统，其特征是：通过对训练集中的作文进行文本分析而得到的自变量包括能够体现作文特征的语言质量、内容质量和篇章结构质量。

3、根据权利要求 2 所述的英语作文自动评分系统，其特征是：所述语言质量包括流利性、词汇复杂性、句法复杂性和准确性；内容质量包括内容的相关性和内容的连贯性；篇章结构质量包括话语结构和段落安排。

4、根据权利要求 1 所述的英语作文自动评分系统，其特征是：所述自变量包括反映流利性的类符数；反映词汇复杂性的平均词长、词长标准差和名词化词汇比率；反映句法复杂性的平均句长和动名词数目；反映准确性的重现词丛数目、介词频率误差、定冠词频率误差、名词代词比；反映内容的相关性的内容相似度；反映内容连贯性的程序词汇数目；反映话语结构的语篇连接语数目；反映段落安排的段落数误差。

英语作文自动评分系统

一、技术领域

本发明涉及一种对试卷进行自动评分的系统，具体地说是一种英语作文自动评分系统。

二、背景技术

目前，国内还没有对英语作文进行自动评分的系统，国际上针对中国学生英语作文评分的技术也未见过任何报导。国际上对英语作文自动评分系统的研究主要有三种软件，都是利用人工评分培训机器评分模型，通过提取作文中的众多文本特征项，利用统计学的回归方法计算作文得分。这三种软件分别是 PEG(由 University of Duke 开发)，IEA（由 University of Colorado 开发）和 E-rater(由 Educational Testing Service 开发)。然而这三种软件并非针对中国学生的英语作文自动评分而设计，运行的总体原理基本相同，但提取的文本特征项各不相同并对外保密。从零星出版公开的研究报告来看，PEG 和 IEA 似乎主要为评阅以英语为母语的学生的作文而设计，E-rater 主要为评阅 GMAT 考试中的学生作文而设计。各软件分别提取哪些具体的文本特征项作为评分模型的变量，无从得知。

对英语写作质量的评价，一般应从语言、内容和篇章结构三个方面入手，而对其语言质量的评价往往从流利度(fluency)、准确性(accuracy)和复杂性(complexity)三个方面入手，其中的复杂性又分别从词和句子两个方面加以观察。国外现有的作文评分系统因为没有遵循这样的第二语言写作评判原则，因而对中国学生英语作文的评分针对性不强，方法不力，要么只能适应对以英语为母语的学生的作文的评分，要么只能适应于对某种考试中作文的自动评分。

因此，上述三种英语作文自动评分系统都存在以下缺点：

1、中国学生的英语作文有其自身的特点，以上三种系统用于中国学生的自动评分针对性不强，不能客观地反映作文水平的高低。

2、这三种软件不能从全方位分析学生英语作文的特点。PEG 只分析作文中最基本的文本特征，如文本长度，平均词长等，其他变量却不加分析；IEA 利用信息检索中的 Latent Semantic Analysis 技术，主要分析作文的内容；而 E-rater 利用自然语言处理技术，分析作文的句法特点、切题度和修辞结构，其它具体变量也不加分析。

三、发明内容

本发明的目的正是要克服上述自动评分系统的缺点，提供一种适合中国学生的英语作文自动评分系统，该系统综合作文中诸方面的特点并以此为评判依据，对中国学

生的英语作文进行自动评分，可实现英语作文的大规模评分。

本发明的目的是通过以下技术方案来实现的：

一种英语作文自动评分系统，其特征是：它包括由一组英语作文集合而成的训练集、文本特征项、回归方程和带有输入和输出装置的电脑；所述训练集通过输入装置贮存在电脑中；文本特征项是通过对训练集中的作文进行文本分析而得到的信息，并将此信息作为自变量；回归方程是将训练集中的作文人评得分作为因变量与自变量进行多元回归分析并通过统计模型而建立的运行方式；对输入电脑中的待评分作文进行文本分析，将得到的文本特征项作为自变量，经过回归方程运算后得到评分结果，并通过电脑的输出装置显示。

本发明中，所述文本特征项包括能够体现作文特征的语言质量、内容质量和篇章结构质量。所述语言质量包括流利性、词汇复杂性、句法复杂性和准确性；内容质量包括内容的相关性和内容的连贯性；篇章结构质量包括话语结构和段落安排。

本发明中所述自变量包括以下 14 项：反映流利性的类符数；反映词汇复杂性的平均词长、词长标准偏差和名词化词汇比率；反映句法复杂性的平均句长和动名词数目；反映准确性的重现词丛数目、介词频率误差、定冠词频率误差、名词代词比；反映内容的相关性的内容相似度；反映内容连贯性的程序词汇数目；反映话语结构的语篇连接语数目；反映段落安排的段落数误差。

本发明中各自变量定义如下：

1) 类符数：指文本中所包含的类符（word types）数目。
2) 平均词长：指文本中所有词汇的平均长度（以单词中所包含的字母数计算）。
3) 词长标准偏差：指文本中所包含的词汇的长度（以单词中所包含的字母数计算）的标准偏差。
4) 名词化词汇比率：指文本中名词化词汇（-ion, -ment 等）与总词数之比率。
5) 平均句长：指文本中所有句子的平均长度（按句子中的单词数目计算）。
6) 动名词数目：指文本中以-ing 结尾的词数。
7) 重现词丛数目：指训练集中的最佳集（抽样样本中得分最高的 1/4）中出现 3 次以上的 3-4 词的词丛（word clusters）在文本中出现的次数。
8) 介词频率误差：指介词的比率（介词数与总词数之比）减去 13.21%后所得数

值的绝对值。

9) 定冠词频率误差：指定冠词的比率（定冠词数与总词数之比）减去6.5%后所得数值的绝对值。

10) 名词代词比：指文本中名词总数与人称代词总数之比率。

11) 内容相似度：指对词语—文档矩阵（term-document matrix）按照Okapi词语权重方案对词语进行权重后再经过奇异值分解（Singular Value Decomposition），重建矩阵后再按照点积数量积（dot product）求得的各文本与训练集中的最佳集在语义上的相似度（similarity）。Okapi词语权重方案为：

$$\text{词语权重（term weight）} = \frac{tf}{0.5+1.5*\frac{dl}{avg\_dl}+tf} * \log(\frac{N-df+0.5}{df+0.5})。$$

12) 程序词汇项数目：指文本中所包含的程序词汇（procedural vocabulary）项的数目。程序词汇表由专利申请人自编。

13) 语篇连接语数目：指文本中所包含的语篇连接语（discourse conjuncts）的数目。语篇连接语列表由专利申请人自编。

14) 段落数误差：指训练集中的最佳集作文的平均段落数与文本实际段落之差的绝对值。

本发明中自动评分过程主要依赖评分模型的建立，而评分模型的核心部分是语言质量、内容质量、篇章结构质量这三大模块及各模块中的自变量。

首先，从大规模考试中收集批量的学生作文作为研究素材，并组织多个资深评分员对这批作文进行人工评分。评分后的作文作为训练集，用于创建评分模型。

在模型创建阶段，利用自然语言处理技术、语料库赋码及统计技术、信息检索技术对学生作文进行文本分析，提取大量的文本特征项，然后进行相关性分析，以确定模型中的自变量；同时以人工评分作为因变量，进行多元回归分析，建立回归模型，最终得到回归方程。这些自变量是一些能够体现作文的语言、内容和篇章结构的一些文本特征项。目前，基于已经进行的分析结果可知本发明的核心部分包括三大评分模块和已经确定的14个自变量，确定的三大评分模块为：语言质量、内容质量和篇章结构质量；自变量包括以下14项：类符数、平均词长、词长标准偏差、名词化词汇比率、平均句长、动名词数目、重现词丛数目、介词频率误差、定冠词频率误差、名词代词

比、内容相似度、程序词汇数目、语篇连接语数目、段落数误差。

在自动评分阶段，先对待评分作文进行文本分析，提取变量，然后将变量的数值代入回归方程之中，即可得到机器评分。

本发明一方面对训练集中的作文进行文本分析，提取大量的文本特征项，以确定模型中的自变量，另一方面以人工评分作为因变量，进行多元回归分析，得到回归方程，然后通过对待评分作文进行文本分析，提取变量，并将变量的数值代入回归方程之中，最终实现机器评分。本发明与现有的人工评分方法相比，资源消耗低、评分信度可靠，适合中国学生的英语作文自动评分。

四、附图说明

图 1 是本发明中英语作文自动评分流程图；

图 2 是本发明中英语作文质量分析图。

五、最佳实施方式

一种本发明所述的英语作文自动评分系统，首先收集一组英语作文的电子文本，可以是 50 篇，集合成训练集，并通过输入装置贮存在电脑中，电脑中应嵌入文本分析工具和统计分析工具，文本分析工具用于从英语作文的电子文本中提取变量，统计分析工具用于进行相关性分析和建立回归模型。然后从训练集中随机抽样，对抽样作文进行多人人工评分，得到因变量；对抽样作文进行计算机文本分析，提取文本特征项，共 14 种，如下表所示：

作文质量测评方面		变量
语言质量	流利性	类符数
	词汇复杂性	平均词长
		词长标准偏差
		名词化词汇比率
	句法复杂性	平均句长
		动名词数目
	准确性	重现词丛数目
		介词频率误差
		定冠词频率误差
		名词代词比
内容质量	内容的相关性	内容相似度
	内容的连贯性	程序词汇项数目
篇章结构质量	话语结构	语篇连接语数目
	段落安排	段落数误差

再分析各文本特征项与人工评分之间的相关性，将相关性显著的文本特征项作为自变量，人工评分的均值作为因变量，进行多元回归分析，得到回归方程；将待评分作文输入电脑中，提取待评分作文的电子文本中的变量，并将变量代人回归方程，得到待评分作文的机器评分。评分结果可通过电脑的输出装置显示。

本发明可实现对中国学生英语作文的大规模机器评分，资源消耗低、评分信度可靠。

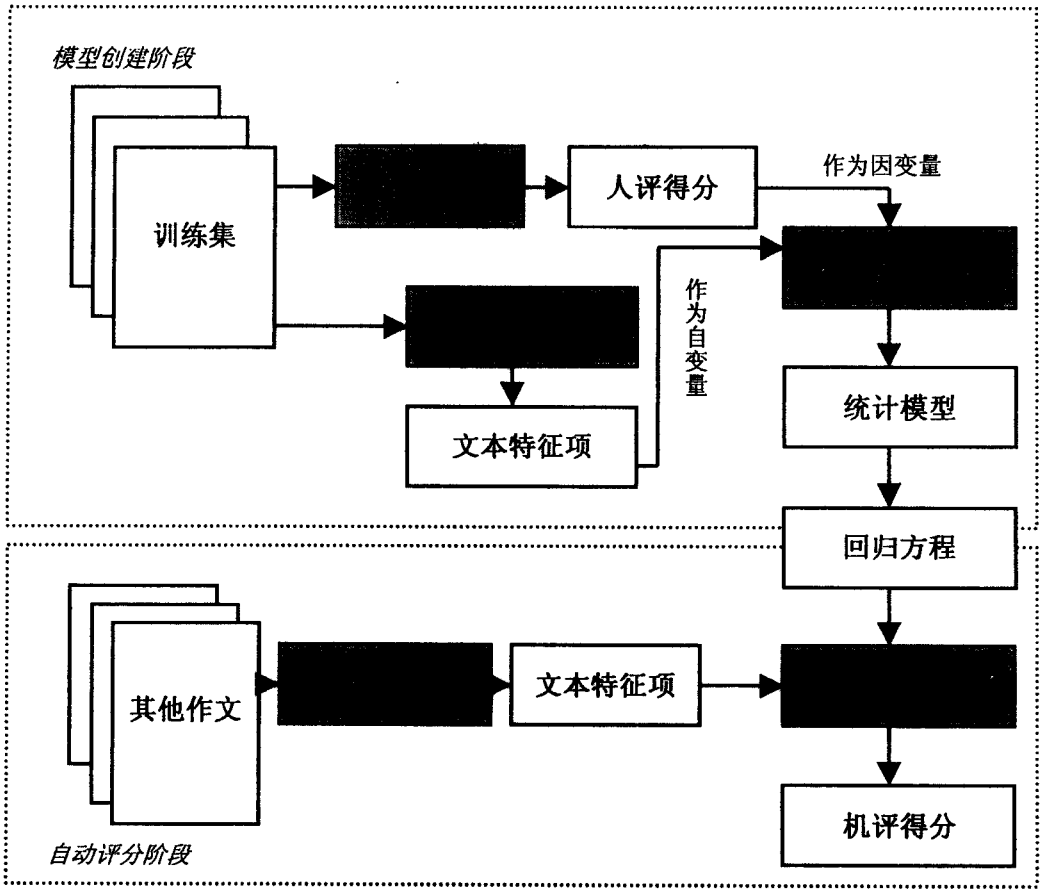

模型创建阶段
训练集
人评得分
作为因变量
作为自变量
文本特征项
统计模型
回归方程
其他作文
文本特征项
机评得分
自动评分阶段

图 1

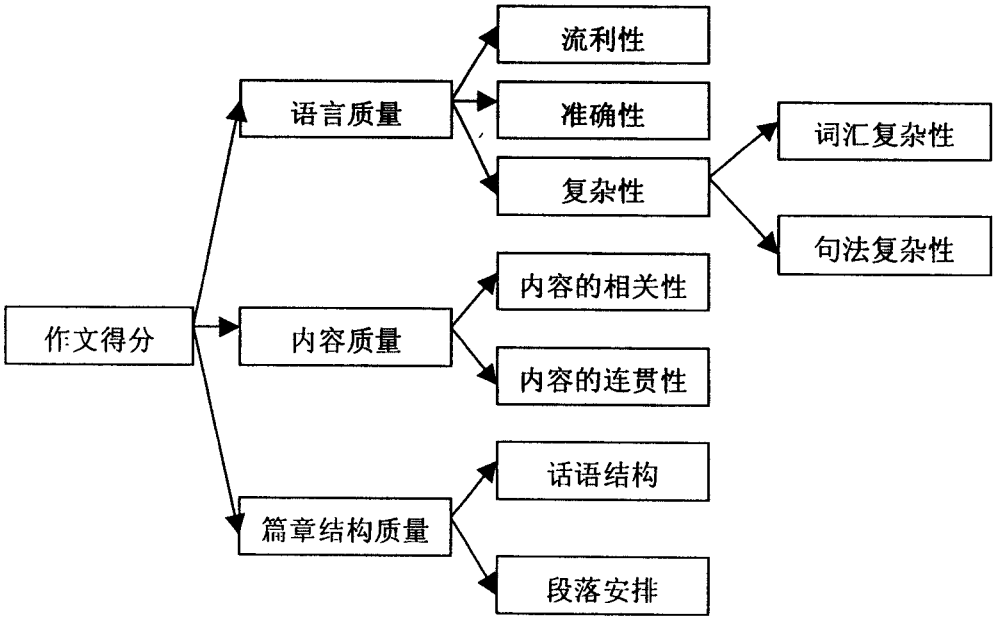

作文得分
语言质量
流利性
准确性
复杂性
词汇复杂性
句法复杂性
内容质量
内容的相关性
内容的连贯性
篇章结构质量
话语结构
段落安排

图 2