基于词汇评分的汉语作文自动评分*

彭星源[1]，柯登峰[1]，赵 知[1]，陈振标[1]，徐 波[1,2]
[1]中国科学院 自动化研究所 数字内容技术研究中心，北京 100190
[2]中国科学院 自动化研究所 国家模式识别实验室，北京 100190
E-mail: xypeng, dfke, zzhao, zbchen, xubo @hitic.ia.ac.cn

摘 要：本文研究了通过作文词汇评分来实现汉语作文自动评分的新算法。在作文评分应与词汇评分高度相关的假设基础上，实现了这种关系的量化计算。本文从通用词表方法、常规方法，以及提出的三种改进算法上进行方法性能的比较，并对比了 e-rater 作文评分系统中同样采用基于词汇方法的性能。实验结果表明，基于新的词汇评分的作文评分方法相关度**达到接近 0.7 的水平，高于 e-rater 中采用的基于词汇的方法。同时，这一方法的结果已经接近于人工作文评分的相关度。

关键词：词汇评分；作文自动评分

Automated Chinese Essay Scoring Based on Word Scores

Peng Xingyuan[1], Ke Dengfeng[1], Zhao Zhi[1], Chen Zhenbiao[1], Xu Bo[1,2]
[1] Digital Content Technology Research Center, Institute of Automation, Chinese Academy of Sciences, Beijing 100190
[2] National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190
E-mail: xypeng, dfke, zzhao, zbchen, xubo @hitic.ia.ac.cn

Abstract: This paper studies new methods of automated Chinese essay scoring based on word scores. Under the hypothesis of the existence of a relation between the essay score and the scores of words which appear in the essay, the equation of the relation is defined. The conventional methods and three enhanced methods are implemented to estimate the parameters of the equation. Compared with the e-rater's methods, our new methods have a correlation close to 0.7, which demonstrates the performance of the latter is better. In addition, the performance of our methods is close to that of humans'.

Keywords: word scores; automated essay scoring

1 引言

作文自动评分已经成为写作评估发展的一个必然趋势[1]。按照 MHK(民族汉语水平等级考试)的评分准则，考生的作文将从语言、条理、内容三方面进行评价。其中，语言表现为句子流畅，用词恰当；条理表现为记叙、论述中，各部分衔接符合条理；内容表现为按规定的主题进行写作。最早的作文评分系统 PEG[1]，以及国内的李亚男[5]所侧重的研究是对语言形式的考察。而 IEA[2]，以及曹亦薇、杨晨[6]的方法则着重于比较相同内容的出现，以内容得分为主得到最终的作文评分。再有综合的系统，如 e-rater[3]、JESS[4]等，则是从语言、条理、内容三方面综合考虑评分的。

本文的研究思路在于通过对作文的用词进行评分，进而通过作文的词汇评分来对作文进行自动评分。首先，优美而富于变化的词汇，能够体现作文的语言水平；其次，词汇的特定指代意义，在很大程度上能够反映作文的内容。因此，词汇对作文评分的语言和内容两方面都有较大的意义。本文在作文的评分是由作文所使用的词汇评分叠加而得的假设下通过实现词汇评分估计，进而计算作文评分。本文提出行之有效的词汇评分估计方法，在准确估计词汇评分的基础上可以获得较高的作文评分性能，融合各种估计方法后，性能还可进一步提升。本文按如下方式组织：第二节

* 所属课题：国家自然科学基金委员会重大研究计划“视听觉信息的认知计算”重点项目(90820303)。
** 本文中相关度均采用皮尔逊相关度。

中介绍一般化的基于词汇评分的作文自动评分方法和用于对比的 e-rater 的方法；第三节介绍本文提出的改进的估计词汇评分方法；第四节分别介绍语料库、实验参数的影响和设定以及最终实验的结果；最后一节对基于词汇评分的作文自动评分进行工作的总结。

2 传统词汇评分方法

基于作文词汇评分的作文自动评分方法的思想，最初是基于考试词汇大纲而来的思路。在语言学习中，词汇依据掌握的难度，在词汇大纲中分为不同的等级。这容易得到如下假设：词汇的难度与作文水平有密切关系。基于这个假设，可以得到一个等价的衡量作文评分的假设：作文的评分等于作文所包含词汇的评分的加权均值。用公式表示为：

$$Score_{essay} = \sum_j t_j w_j / \sum_j t_j + b \tag{1}$$

其中，t_j 表示为 j 词在作文中的出现次数，w_j 表示为 j 词的评分，b 为一个线性偏移量。这个公式为本文所提出方法的假设前提。在获得词汇的评分 w_j 后，便能够通过公式(1)对作文进行评分。

2.1 通用词汇表评分方法

通常词汇等级表是通过广泛的词频统计为基本依据，同时依靠资深的语言教师们进行经验判断得出的，MHK 考试词汇等级表也同样如此[7]。

本文采用了一份包含 3 个等级的词汇表作为通用词汇表，共计 7277 词。考虑到作文的人工评分范围为 1~6 分，本文将以上的词汇以一级词汇按 1 分统计，二级词汇按 3.5 分统计，三级词汇按 6 分统计。通过公式(1)计算出作文的最终得分，b 值取 0。通用词汇评分方法中，利用广泛的词频和人工建立的评分等级，进行作文评分的赋值，实现作文评分的自动评定。

2.2 常规的词汇评分估计方法

常规的词汇等级方法是按照公式(1)类比得到的。既然假定了作文的评分是由词汇的评分所决定，由对偶的原则，估计词汇的评分也同样可以由作文的评分入手，因此有如下假设：作文词汇的评分为词汇所出现作文的人工评分的均值。公式表示如下：

$$w_j = \sum_i t_{ij} score_i / \sum_i t_{ij} \tag{2}$$

其中 w_j 表示 j 词的评分，t_{ij} 表示 j 词出现在作文 i 中的次数，$score_i$ 表示作文 i 的人工评分。

2.3 e-rater 内容向量分析法

E-rater 的方法表上并不完全相似于以上的方法。它通过作文向量与各评分等级向量之间的相似度，对作文进行评分等级归属划分，得到作文评分。但是其本质上，仍然是通过纯粹的词汇统计，得到作文等级划分的向量，此向量上的词汇的特征类似于词汇评分。其方法介绍如下[3]：

每一篇作文都将由一个词汇向量表示，同样每一个作文评分等级也可以由一个词汇向量表示。每一个词在向量中都由一个权重表示。其中评分等级的词汇向量权重的计算公式为：

$$w_{js} = (F_{js} / Max(F_s)) * \log(N / N_j) \tag{3}$$

其中 F_{js} 为词汇 j 在 s 评分等级中出现的频数，$Max(F_s)$ 表示所有词在 s 评分等级中出现的最高频数的那个词的频数，N 是训练集的作文数，N_j 表示 j 词在 N 篇作文中出现的作文的数量。公式的前半部分为词汇在 s 评分等级的归一化频率；公式的 log 函数部分为一个倒排文档频率，是一个词语

普遍重要性的度量。作文的词汇向量权重的计算公式为：

$$w_j = (F_j / Max(F)) * \log(N / N_j) \tag{4}$$

其中 F_j 为词汇 j 在某篇作文中出现的频数，Max(F)表示作文中所有词中出现频率最高的那个词的频数，N 和 N_j 的含义同上。公式(4)含义与公式(3)一样，只不过针对的是单独的一篇作文。

E-rater V2 中，有两种方法由词汇向量得到最终的作文评分。其一是计算待评作文向量与各评分等级的词汇向量的相似度，作文评分为相似度最高的评分等级的评分；另一方法是计算待评作文向量与最高等级评分向量的相似度，最终的作文评分为相似度与最高分的乘积。

3 改进的词汇评分的作文评分方法

现在重新回到公式(1)这个假设上。可以看出，如果能够知道每一个词汇准确的评分，那么就能够计算出作文的得分。而估计词汇的评分 w_j 通常的做法就是用相关的训练集去估计。已知作文评分，可以通过(1)式建立起一个方程组，此方程组在最小二乘的方法下有一个全局最优解。因此，由一个已知人工评分的训练集，可以通过最小二乘法直接得到对 w_j 的全局最优估计，进而就可以通过(1)式完成对作文的自动评分。此方法在实际的操作中，会遇到两个问题：I. 实际 j 的取值范非常大，方程组中的未知变量个数过多，也即方程组矩阵过度庞大。在运用最小二乘法对方程组求解的时候，需要对矩阵求逆，过大的矩阵将导致求逆的难度增大。II. (1)式对每一个词汇都有一个单独的词汇评分，也即模型的参数变量过多，如果求得训练集中的最优解之后，会出现过拟合现象。为了解决存在的这两个问题，本文提出了一种解决的思路。

不再如公式(1)中那样对每一个词汇给予一个单独的评分变量，而将全部词汇评分划分为 c 个评分，每个词将属于其中一个评分。也即公式(1)变为：

$$Score_{essay} = \sum_j T_{ij} \cdot (\sum_c p_{jc} \cdot w_c) + b \tag{5}$$

其中 T_{ij} 表示(1)式中归一化后的词汇频数值，也即词汇频率。p_{jc} 表示当词汇 j 属于 c 评分类的概率。w_c 表示第 c 类词汇评分的确切评分值。b 为一个线性偏移量。这样，就将对词汇的评分 w_j 的估计，转化为对 c 个评分类的评分 w_c 的估计，和词汇 j 属于 c 类的概率分布的估计两个过程。现在本文提出三种方法来实现这两个过程的估计，同时在实现的过程中解决了以上的两个问题。

3.1 改进的估计方法一

分步求解 p_{jc} 和 w_c 来解决计算困难的问题，并且通过求得一个局部最优解替代全局最优解以防止过拟合的情况发生。在本方法中，p_{jc} 的取值固定为(0,1)。

算法流程：

I. 随机对初始 p_{jc} 赋值，实现分布初始化。

II. 固定 p_{jc} 值，这样待求解的方程组将化简为仅包含 C 个变量 w_c 的线性方程组。通过最小二乘法求得此分布情况下的最优解。

III. 固定 w_c 值，对 N 个词汇的 p_{jc} 分布按贪心算法进行逐词搜索，寻找能够让训练集作文按(5)式评分的方差最小的 p_{jc} 分布。

IV. 计算训练集作文按(5)式评分方差减小量 ε，如果 ε 小于某一预设值或者达到一定迭代次数 K，则进入步骤 V；否则回到步骤 II。

V. 按当前求得的 p_{jc} 和 w_c 值求得作文评分方程。按此方程得到作文的评分。

此方法中由于 w_c 的类别数 C 值较小，因此在用最小二乘计算的时候计算复杂度也在可操作之内。同时，求得的 p_{jc} 和 w_c 并非全局最优的，避免了过拟合现象的发生。具体的循环次数 K 以及变量个数 C 值如何确定将在后面的实验参数设置中讨论。

3.2 改进的估计方法二

此方法试图将词汇直接分为 c 类。将词汇在每一类人工评分中的分布概率通过训练集计算出来。以此作为特征对词汇进行聚类。同一类的词汇将获得同样的类别评分 w_c。在确定了词汇的类别分布 p_{jc} 和类别评分 w_c 后，也即确定了(5)式中的待估变量，实现了作文自动评分的方程。

算法流程：

I. 人工的作文评分有 11 类，计算每一个词汇 j 在此 11 类上的分布情况。

II. 用聚类方法对 N 个词汇在这 11 类上的分布情况进行聚类，得到每一个词汇分为某一类的判别。也即得到 p_{jc} 分布。

III. 在已知 p_{jc} 分布后，待求解的方程组将化简为仅包含 C 个变量 w_c 的线性方程组。通过最小二乘法求得最优解。

IV. 按当前求得的 p_{jc} 和 w_c 值求得作文评分方程。按此方程得到作文的评分。

此方法试图通过聚类的方法，直接求得词汇的评分类别所属。方法考虑了词汇在人工评分等级中的分布情况，一定程度上减少了数据带来的过拟合情况。但同时，引入了一个要判断的变量，即聚类的数目。具体的聚类数目 C 的确定将在后面的实验参数设置中讨论。

3.3 改进的估计方法三

训练集的作文有 11 等级人工评分，因此假设词汇的等级也分为 11 类。在训练集上计算词汇每一类的分布概率。词汇中出现频率低的词，并不具有良好的统计意义，因此可以作为噪声剔除；同样，人工评分分数段分布较为均匀的词汇，其对于作文评分起不到区分意义，这样的词汇也应剔除。这样就可以得到有效（被剔除的词汇不再参与计算当中）词汇的概率分布，再通过最小二乘法得到当前分布下的最优 w_c 值，也即确定了(5)式中待估变量，完成了作文自动评分方程。

算法流程：

I. 人工的作文评分有 11 类，计算每一个词汇 j 在此 11 类上的分布情况，即求得 p_{jc}。

II. 计算每个词出现的频率（tf）和每个词在 11 类上的分布方差（dv）。对于词汇频率低于特定频率 F 或者其分布方差小于某一特定值 DV 的词汇，删除其对作文分数的影响。即对于符合情况的词汇 j 有：$p_{jc}=0$ 对于任意的 $c \in C$。

III. 在确定了 p_{jc} 后，待求解的方程组将化简为仅包含 C 个变量 w_c 的线性方程组。通过最小二乘法求得最优解。

IV. 按当前求得的 p_{jc} 和 w_c 值求得作文评分方程。按此方程得到作文的评分。

此方法将人工评分等级等价于词汇等级。将词汇属于词汇评分等级的情况，用概率分布的方式描述，而不再是上面方法中的只属于某一评分等级。将统计特性不明显的词汇和分布较均匀没有区分作文评分意义的词汇去除，减少了噪声的引入。这里有两个需要得到的经验变量，一个是截断频率的取值，一个是分布方差的截断最小值。取值确定将在后面的参数设置中探讨。

总结起来，三种方法的基本思路一致，均是将公式(1)变化为公式(5)，通过分别求得词汇的分

布 p_{jc} 和评分类别的评分值 w_c 来解决以上提到的两个问题。区别在于其具体的实现方法上。方法一通过贪心算法求得词汇的划分，方法二则是通过对词汇的分布特征进行聚类来得到词汇的类别划分，方法三则是直接利用了词汇在人工评分中的分布结果。相比于直接去求每一个词的评分 w_j，求一个类别的评分值 w_c 则能够明显的减少模型的参数，从而避免了过拟合现象的发生。

4 实验设计与分析

4.1 语料库介绍

本文中的作文数据的人工作文分数评分设定为 1~6 分。每一篇作文由至少两个评分员进行评分。最终的人工作文评分分值为 1~6 分，间隔为 0.5 分一档，共分为 11 档。

本文实验的数据取自一作文集。此作文集中，最初的两个人工评分的相关度为约 0.54。考虑到作文评分中，人工评分较低的相关度，为了避免人工评分的不可靠性对实验带来的影响，本文实验的对象均选自两个人工评分中分差不大于 1 分的作文。本文共抽取 8000 篇作文作为实验的对象，其中 5000 篇作为训练集，3000 篇作为测试集，测试集分为 3 份，每份 1000 篇。每个数据集中两个老师人工评分相关度数据如下表：

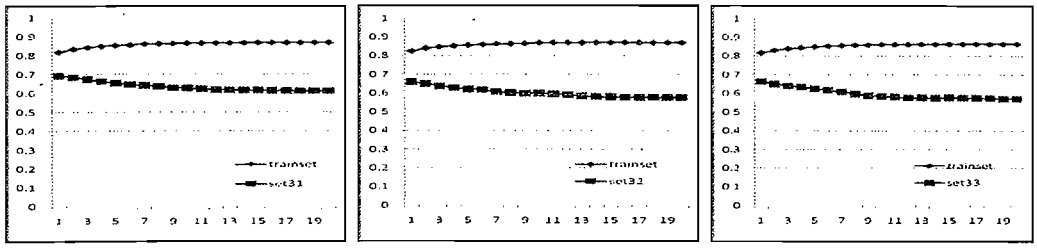
表 1　训练集与测试集人工相关度

	trainset	set31	set32	set33	测试集均值
人工评分相关度	0.7494	0.7567	0.7565	0.7499	0.7544

4.2 实验参数的选定以及影响

4.2.1 改进算法一中的迭代次数以及词汇评分等级数

对于改进方法一中，过大的迭代次数将导致评分公式出现过拟合，导致方程的泛化能力下降。因此，如何决定迭代次数，将是本小节所要解决的问题。图 1 所示，在一次迭代后，就出现过拟合现象。因此迭代次数选 1 次。由于迭代次数较少，这样也极大的减少了运算所耗的时间。



(a)　(b)　(c)

图 1　迭代次数对测试集相关度的影响曲线（方法一）

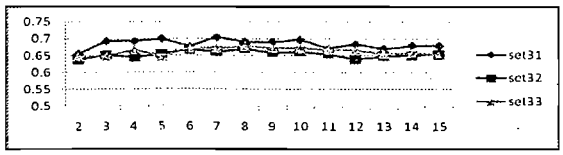
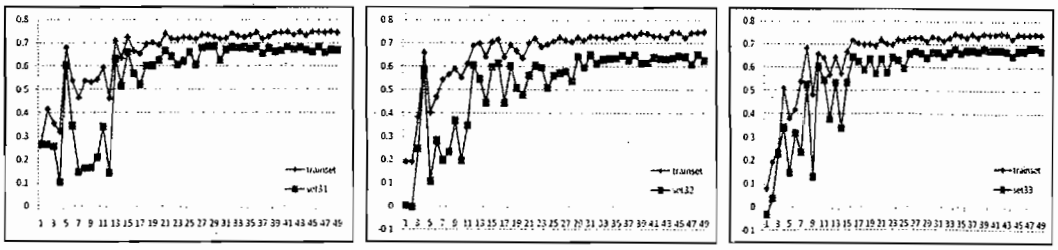


图 2　词汇评分种类数对测试集相关度的影响曲线（方法一）

在确定了迭代次数后，由图 2 可见，词汇评分种类数在此方法下对评分效果的影响有限而且并无显著规律，因此按人工对作文的评分分为 11 级评分，而选取词汇评分种类数 C = 11。

4.2.2 改进方法二中聚类类别数

在方法二中的聚类方法选用 K-means 方法。对于此聚类方法，聚类数目是一个预先需要确认的变量。为了获得合适的 K 值，本文在三个测试集上对不同 K 值下的测试集相关度进行统计。



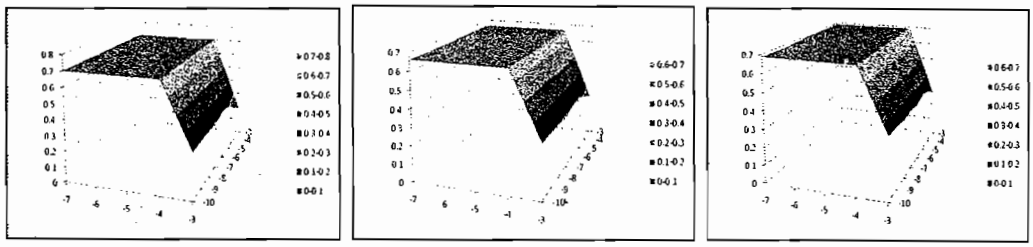
(a) (b) (c)

图 3 词汇评分种类数对测试集相关度的影响曲线（方法二）

图 3 可见，当聚类数目较小的时候，在测试集乃至训练集上的评分相关度较低，且极为不稳定，说明此时的类别数不足以反映实际的情况；当聚类数目达到 30 以后，测试集上的评分相关度逐步趋于稳定。因此选取聚类数目 K 值为 30。也即此情况下，词汇评分等级数 C 为 30。

4.2.3 改进方法三中词汇过滤条件

在方法三中，需要去掉统计特性不明显以及没有区分意义的词汇，以减少这部分词汇带来的噪声影响。实验采用网格搜索的方法对可能的参数进行逐一尝试，通过性能最优来决定参数。本文通过大致的参数尝试的方法初步得到截断频率的 tf 取值和截断分布方差 tdv 的初步值。随后，本文将 log(tf)的取值定在-10~-3 之间，而 log(tdv)的取值在-7~-3 之间。网格搜索的步长设为 1。



(a)set31 (b)set32 (c)set33

图 4 词汇过滤参数对测试集相关度的影响曲线（方法三）

图 4 可见，三个集上的最大评分相关度约为 0.68。其中，log(tf)在取值高于-4，相关度急剧下降；取值低于-4，相关度非常平缓的下降。log(tdv)的取值对相关度影响较低。约在-6 附近达到一个极值。本实验选取 tf 的最优取值为：0.015625(2^{-6})，tdv 的最优取值为：0.0625(2^{-4})。

4.3 实验结果与分析

本文将对各个基于词汇评分的作文自动评分方法进行对比。其结果见下表：

表 2 各方法下测试集自动评分与人工评分相关度

相关度	通用词表法	常规方法	E-rater_1	E-rater_2	改进方法一	改进方法二	改进方法三
Set31	0.0841	0.3244	0.6406	0.3658	0.6975	0.6820	0.7040
Set32	0.0102	0.2812	0.5960	0.3415	0.6528	0.6411	0.6642
Set33	0.0730	0.2832	0.6229	0.3512	0.6464	0.6697	0.6915

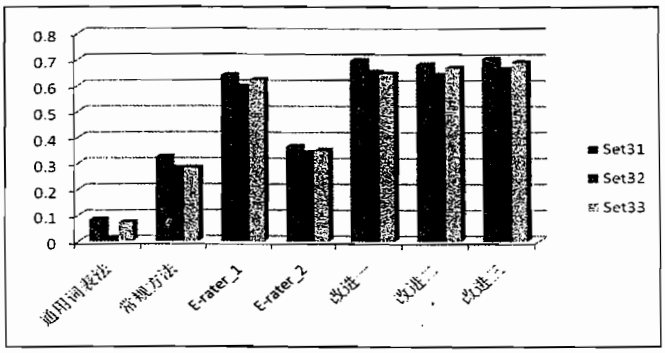


图5　各方法下三个测试集上的相关度

各方法的性能如表 2 所见。直观的从图 5 中，可以看到通用词表方法的相关度极低，这表明了一个广泛通用的词表等级对于作文评分并没有代表性的意义。而常规的估计方法，由于其估计的粗略性，因此准确度并不高，影响了其最终评分效果。本文将 e-rater 提出的两种方法作为改进方法的性能对比对象。可以看到，在本实验中的数据集上，e-rater_1 方法性能较好，自动评分与人工评分相关度达到了 0.6 左右，而相对而言，e-rater_2 方法性能则不够理想，分析其原因，应该是由于各个分数段作文并非完全是高分段作文的部分缩影，而可能具有每个分数段内自身的词汇分布特色。因此仅用与高分段作文的相似度来衡量，显得有些不足。从图中可以看到，本文所提出的三种改进方法上，性能均较平衡，平均相关度达到了 0.65 以上的水平，已经超过了 e-rater 的方法，因此，本文提出的对于公式(1)的假设是成立的。由于三种改进方法的本质一样的，差别在于具体的实现方法上，因此性能上也较为接近。考虑到由于各种方法上的实现差别，本文试图将各个方法进行线性融合，以期望获得基于词汇等级评分方法的一个综合性结果。

将以上方法结果中，性能较优的 e-rater_1 方法和三种改进方法进行线性融合。在 3 个测试集中抽取 1-2 个作为拟合方法的训练集，剩余的一个作为测试集。实验结果如下表：

表3　基于词汇评分等级的作文评分性能

测试集 \ 训练集	Set31	Set32	Set33	Set31 +Set32	Set31 +Set33	Set32 +Set33
Set31	——	0.7187	0.7156	——	——	0.7180
Set32	0.6863	——	0.6816	——	0.6849	——
Set33	0.6944	0.6947	——	0.6947	——	——

从表 3 中可以看到，融合后的结果与单一的方法比较均有一定的提升，在测试集上的相关度均值达到了 0.6988，而此三个测试集上的人工评分相关度均值为 0.7544（本实验中数据集是经过人工挑选的初始两人工评分不大于 1 分的作文，实际人工评分相关度约为 0.54）。虽然自动评分的相关度低于人工评分相关度，但差距不大；另一方面，本文仅仅考虑词汇评分等级上对作文进行评分，并没有考虑其他许多能够表现作文水平的特征，能够取得如此接近人工评分相关度的性能已经表明本文提出的方法的可行性。如果进一步融合其他作文评分的方法与特征，作文自动评分的性能还将再进一步提高，但这已经超出了本文所讨论的范围。

5　总结与将来的工作

本文从词汇评分和作文评分之间的关系入手，通过建立合理的关系假设，从方法上讨论了如何通过词汇的评分得到作文的评分，并通过实验验证了假设的正确性，实现了基于词汇评分的作

文评分。实验结果表明，如何通过相关的训练数据获得准确的词汇评分是进行基于词汇评分的作文评分的关键。基于词汇评分的作文评分在相关度性能上高于 e-rater 的同样基于词汇的方法，并且在融合了各种方法之后，最终的评分相关度可以达到接近 0.7，说明了方法的有效性。

词汇仅是体现作文水平的一个重要特征。虽然基于词汇评分的作文自动评分方法在性能上已经达到不错的地步，但是相对于作文自动评分研究而言还仅仅是冰山一角。将来，可以继续从作文的语言、条理、内容三方面进行探索，从更加丰富而综合的层面进行作文自动评分的研究。

参 考 文 献

[1] S. Dikli. An overview of automated scoring of essays[J]. Journal of Technology, Learning, and Assessment, 2006, 5(1): 1-35.

[2] T. Landauer, D. Laham, P. Foltz. Automatic essay assessment[J]. Assessment in Education: Principles, Policy and Practice, 2003, 10(3): 295-309.

[3] Y. Attali, J. Burstein. Automated essay scoring with e-rater v.2[J]. Journal of Technology, Learning, and Assessment, 2006, 4(3): 1-30.

[4] T. Ishioka, M. Kameda. Automated japanese essay scoring system based on articles written by experts[C]. In Proc. ACL. Sydney, Australia, 2006: 233-240.

[5] 李亚男. 汉语作为第二语言测试的作文自动评分研究[D]. 北京：北京语言大学, 2006.

[6] 曹亦薇, 杨晨. 使用潜在语义分析的汉语作文自动评分研究[J]. 考试研究, 2007, 3 (1) : 63-71.

[7] 彭恒利. 中国少数民族汉语水平等级考试[J]. 中国考试, 2005, 10: 57-59.