Features Selection of High Quality Essays in Automated Essay Scoring System

Mingtao Wang
Center for Intelligence Science and Technology Beijing University of Posts and Telecommunications
Beijing, China
mingtao0921@163.com

Yongmei Tan
Center for Intelligence Science and Technology Beijing University of Posts and Telecommunications
Beijing, China
ymtan@bupt.edu.cn

Chao Li
Center for Intelligence Science and Technology Beijing University of Posts and Telecommunications
Beijing, China
lichao@gmail.com

***Abstract*—In this paper, we analyze some essays of Chinese English learners, and text features such as words, phrases, paragraphs, and chapters. According to the correlation between the scores and these features, we do multiple regression analysis based on the feature recombination and extract the feature sets that high quality essays should be with. In the end, we build a reasonable model to distinguish high quality essays from others and effectively solve the problem of low differentiation scores in automated essay scoring system.**

Keywords-Automated Essay Scoring; Natural Language Processing; regression analysis; feature selection

I. 概述

英语写作对于很多英语学习者尤其是大学生来说是一个薄弱环节，因此提高英语写作教学的质量显得格外重要，但是传统意义的英语写作教学存在学生人数多，教师批改花费时间长，批改结果反馈不及时，批改结果容易受批阅人主观因素影响等弊端，而通过作文自动评分可以有效的克服这些弊端，使教学双方获益。

作文自动评分（AES, Automated Essay Scoring）是通过计算机技术对英语作文进行自动评阅和评分。最早出现于 20 世纪 60 年代，是一个名为 PEG 的系统，但是由于受当时计算机技术的限制，PEG 只提取了文章的浅层特征作为判别依据，尽管如此，这已为机器辅助教学在英语写作教学中的应用开创了新的思路。随着人工智能及自然语言处理技术等的发展，相继出现了 IEA、E-rater、BETSY 等系统，功能日臻完善，有些已在美国教育考试领域中被大规模应用，实际应用效果表明 AES 系统在英语作文批改的信度和效度方面都表现良好，其中信度是指对测试事先所作的判断与测试结果之间的一致性，效度是测试本身的可靠性，也就是测试与被测内容之间的联系[1]。

尽管国外英语智能评分系统性能良好，但是其并不适合中国的英语学习者，这是因为在以英语为母语的学习者的文章中，基本不会出现过多严重的语法、句法等错误，因此评判写作水平的高低不会取决于这些。相反地，对于英语为非母语的中低水平学习者来说，评价其写作质量高低的标准会更多地倾向于评判文章的语句是否正确、通顺等方面上，这是前后二者写作方面最大的不同。所以针对国内英语学习者的写作评价模型，会与母语为英语的英语学习者的写作评价模型有较大不同。然而在国内这项研究起步较晚，尚不成熟，因此提出针对国内英语学习者的英语作文智能评价模型仍然是十分重要的。

本文的内容结构是这样组织的，第一节介绍了作文智能评分系统的研究背景以及研究的重要意义，第二节主要介绍了国外在这一方面的研究成果和发展现状，并分析了实现作文智能评分需要解决的主要问题；第三节对高分作文如何进行特征项选择进行说明；接下来第四节对选定的特征项进行了多组回归实验，并对实验结果进行了分析，最后一节对本文的工作做了简要总结并提出后续的研究方向。

II. 相关工作

如今在国外，英语作文自动评分技术已经在大型考试中有较为成熟的应用，其实际评判结果与人工评分的契合度较高，达到可接受的范围，说明了智能作文评阅系统的可行性很高，本节会对已有的几个系统，包括 PEG、IEA 和 BETSY 做出简要概述。同时结合中国英语学习者写作的特点，探索如何找出能有效地反映作文质量的特征项，这也正是本文需要重点解决的问题。

A. 英语作文智能评价系统

PEG（Project Essay Grader）是 AES 的首次尝试，其主要方法是从已经过人工评分的作文中抽取出某些文本特征，如文章长度等，将这些特征项作为自变量，人工评分作为因变量，进行多元回归分析，得到回归方程，然后将待评分作文中对应特征项的值带入回归方程进行计算，结果值即可作为作文的得分参考值。然而据一些研究表明 PEG 系统仅提取了少量的浅层文本特征，导致文章的表层语言结构对评判结果影响较大，针对这些问题 PEG 在后期进行了多次改进。

IEA（Intelligent Essay Assessor）采用了基于潜伏语义分析（Latent Semantic Analysis）的方法。这种理论认为在文本中存在一个潜在的语义结构，即一篇文章的意义取决于其包含的词语的意义。它把一篇作文看成是由词汇构成的向量，多篇文章的向量构成一个矩阵，然后采用奇异值分解来降低维度。该系统用于评分时，把待评分作文与预先选定的范文（即训练集）当作矢量，对这些矢量进行比较，得出每篇待评分作文与训练集在内容上的相似度，然后根据相似度得到系统评分[2]。

978-1-4244-8165-1/11/$26.00 ©2011 IEEE

BETSY（Bayesian Essay Test Scoring sYstem）采用的核心方法是贝叶斯方法，借鉴了文本分类的思想，即根据包括内容与形式的大量的文本特征把一篇文章划分到合适的类别中，类别是指文章的优劣（如优、合格、差）[3]。BETSY 主要通过训练集找到文章特征和文档类别之间的关系模型，再利用这种关系模型对新的文档进行类别判断，达到自动评分的目的[4]。

B. 特征选择的重要性

计算机在作文批阅中最大的不足是不能理解文章的内涵，无法像人一样真正地欣赏和评鉴文章，它所能处理的是量化的数据，将这些数据值的大小作为评判作文优劣的标准，所以实现英语作文智能评判需要解决两个问题。

第一是确立能够全面、客观地反映考生写作水平的作文分项评分标准及各项标准所占的比重，即根据作文评分标准的主要特征，把作文评分标准分解为若干不同的标准项，并按照一定的模式合成分数。

第二是计算机能否根据已经确立的作文分项评分标准自动而准确地从考生作文中提取出相关信息，这不仅依赖于英语本体研究的相关成果，同时也取决于自然语言处理技术的发展水平[5]。

本文对实验中的语料做了语法、句法等分析，并统计得到多个文本特征项，所以本文中对第二个问题不做进一步讨论。针对第一个问题，本文从文章的词，短语，句子，段落，篇章各个层次来进行特征选择的研究，力图能尽可能全面的包含反映文章质量的特征。

C. SPSS 介绍

本文中主要使用 SPSS 对数据进行分析，包括相关分析和回归分析等功能。SPSS（Statistical Product and Service Solutions）被称为“统计产品与服务解决方案”软件，其基本功能包括数据管理、统计分析、图表分析等。SPSS 统计分析过程包括描述性统计、相关分析、回归分析、聚类分析等几大类，每类又包括多个统计过程，比如回归分析中包括线性回归分析、曲线估计、Logistic 回归、加权估计、非线性回归等多个统计过程，而且每个过程中允许用户选择不同的方法及参数。

III. 高分作文特征选择研究

在本文中，我们主要从两个角度来对已有的特征项进行选择。

第一个是从专家评分出发，主要方法是对已经过专家评分的多篇作文，统计提取出其中有关的文本特征，通过相关性分析找出其中能有效反映作文质量的特征项。

第二个是从写作理论出发，即参考写作理论及评分标准，考虑一篇高质量作文应该具有的特征，提取对应这些特征的文本特征量值，作为候选的特征项[6]。

本文关注挖掘文章中能反映作文质量的文本特征项，从而构建一个能够判别高分作文的模型，即对于未评分文章，将对应特征项带入模型计算，便能判断出该文章是否属于高分。整个流程如图 III-1 所示。

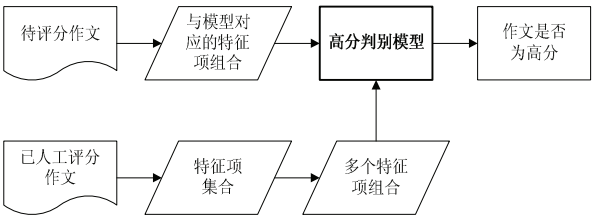

图 III-1 实验流程图

实验过程如下：

(1) 对语料进行词法、句法分析等处理，统计文章中量化的特征值，共得到包括文章长度、句长、各词性词语的数量等在内的共 140 多个特征项。

(2) 使用 SPSS 统计各个特征项和作文是否为高分之间的相关性，选取显著相关的特征加入候选特征集合。

(3) 根据写作理论或文献资料，将能反映高质量作文特征而在步骤 (2) 统计结果中未出现的特征项加入候选特征项集合。

(4) 将候选特征集合中反映作文相同特征的特征项分为一组，比如 word_type 和 token_type 均反映了文章的长度值，分到同一组，通过实验来确定采用哪个特征项。

(5) 从每组中依次选出一个特征项进行交叉组合，作为待判特征项序列，将待判特征项序列和文章是否为高分进行回归分析，并构建评判模型。

(6) 计算测试集中的文章通过上述模型对文章是否为高分的评判的准确率和召回率，评估模型性能。

本文采用三个不同的方法对特征项进行选择，分别进行了实验，其中实验一是在统计的基础上进行的特征项选择，实验二是基于写作理论及相关资料论述进行选择，实验三是将实验一和实验二中选出的特征项混合组合。

IV. 实验

A. 实验数据

本文用到的语料是近十几年来的大学生英语四、六级考试标准答案的参考范文，包括四级范文 60 篇，六级范文 50 篇。自 2005 年 6 月考试起，四、六级考试采用满分为 710 分的计分体制，其中写作部分占 15%，即 106 分。本文中用到的是百分制的计分方法，因此对非百分制的分数进行相应的换算，将作文分数转化成满分 15 分所对应的分数，然后按照分数将作文分为 2 分、5 分、8 分、11 分和 14 分共五个档次，其中我们规定 14 分档作为高分档，其他分数档作为低分档。语料在每个分数档中的文章分布情况如表 IV-1 所示。

表 IV-1 语料在各个分数档的篇章分布

	2 分档	5 分档	8 分档	11 分档	14 分档	总篇数
分数区间	0-3 分	4-6 分	7-9 分	10-12 分	13-15 分	
cet4	12 篇（20%）	12 篇（20%）	12 篇（20%）	12 篇（20%）	12 篇（20%）	60 篇
cet6	10 篇（20%）	10 篇（20%）	10 篇（20%）	10 篇（20%）	10 篇（20%）	50 篇

B. 实验及结果分析

本文实验是以 cet4 作为训练集，cet6 作为测试集，通过多种特征项组合构建模型，并考察其在测试集中对高分文章评判结果的准确率和召回率，评测模型性能。本文对特征项进行不同方法的选择组合实验。

首先，实验一采用了基于统计的方法对特征项进行选择。根据第III节中实验过程步骤 (2) 所述，选出显著相关及以上的特征项，并进行相似特征项的分组。相关程度是根据相关性系数 r 的值来确定：其中 $0 < |r| \le 0.3$ 为微弱相关；$0.3 < |r| \le 0.5$ 为低度相关；$0.5 < |r| \le 0.8$ 为显著相关；$0.8 < |r| < 1$ 为高度相关[7]。选定的特征项如表 IV-2 所示，其中 x_num, x_ratio, fitted_num, b0 与文章分数呈负相关[8]，即其值越大，文章的分数越低，其他特征项与分数均为正相关。

表 IV-2 特征项表示及含义说明

特征项	释义	备注
token_type lex_type lemma_type word_type word_diff_num	符号类数； 词类数； 原型类数； 类符数； 有难度值的词的类型数；	从不同方面反映文章长度
word_gt7 word_gt8 word_gt9	长度（类符）大于 7 的单词占比； 长度（类符）大于 8 的单词占比； 长度（类符）大于 9 的单词占比；	不同长词在文章中的占比
gramv gramv_vov gramv_vxp gramv_vxxp	动词短语数（3 元或 4 元）； 动词短语数（形如 plan to go）； 动词短语数（形如 make use of）； 动词短语数（形如 be in love with）；	动词短语数
v_ttr2	动词类符形符比；	
b2 b3	二级词汇比率； 三级词汇比率；	分级词汇占比
a_type	形容词类数；	
d_len	副词平均长度；	
word_diff_avg	单词平均难度；	
awl awl_sd atl atl_sd	average word length； awl 方差； average token length； atl 方差；	词长
ttr1	类符形符比；	
d_ttr2	副词类符形符比；	
x_ratio x_num	其他词性占比； 其他词性数量；	
fitted_num	分析失败的句子数；	
b0	零级词汇占比；	

从表 IV-2 所列的每个分组中选出一个特征项进行交叉组合得到多个特征项序列，数据如下：

a_type, d_len, word_diff_avg, ttr2, d_ttr2, v_ttr2, fitted_num, b0, word_type, word_gt7, gramv_vov, x_ratio, b2, awl

a_type, d_len, word_diff_avg, ttr2, d_ttr2, v_ttr2, fitted_num, b0, word_type, word_gt7, gramv_vov, x_ratio, b2, awl_sd

……

对上述多组特征序列分别进行回归分析并获得模型，根据模型对测试集中高分作文判别的准确率和召回率来评估模型性能，选出性能最优的序列，结果如表 IV-3 所示。

表 IV-3 实验一的结果

准确率	召回率	特征项组合
80.00%	90.00%	a_type, d_len, word_diff_avg, ttr2, d_ttr2, v_ttr2, fitted_num, b0, word_gt8, gramv, b2, atl, x_ratio, word_diff_num
80.00%	90.00%	a_type, d_len, word_diff_avg, ttr2, d_ttr2, v_ttr2, fitted_num, b0, word_gt8, gramv, b2, atl, x_ratio, token_type

在实验一结果的特征项组合中，word_diff_num 和 token_type 属于相同分组，均反映了作文的单词数量，表明二者对作文质量的影响相同，这也说明了我们抽取的 140 多个特征项中，存在多个特征项反映作文相同特征的情况。

依据写作理论进行特征选择，进行了实验二。首先从写作理论的角度出发，在 140 多个特征项中找出高分作文的应具有的特征项，再结合已有的研究[[9][10][11][12]选出多个特征项，将多次出现的特征项加入确定特征项集合（定义为 DefSet），在接下来的实验中会对其进行评估。

DefSet = {word_type, word_num, snt_num, ttr2, awl, awl_sd, asl, mwe_disconj（篇章连词数）, b2, gramv, word_gt7, word_gt8, word_gt9 }

同时把与 DefSet 中特征项类似的、且与文章是高分呈现显著相关的特征项选出来，这些特征项包括：atl, atl_sd, lemma_type, lex_type, token_type, word_diff_num, word_diff_avg, gramv_vov, gramv_vxp, gramv_vxxp。

将上述的特征项分别与 DefSet 中对应的特征项进行替换，形成不同的组合，然后考察哪个能更好的反映文章质量，如 awl 和 atl 均反映了文章的平均词长，就需要通过实验来确定二者哪个能更好的反映平均词长的特征。

对这些特征项序列进行回归分析，选择性能最优的序列，如表 IV-4 所示，发现性能有所提升，召回率保持在 90.00%，而准确率由 80.00%提升到 82.00%。

表 IV-4 实验二的结果

准确率	召回率	特征项组合
82.00%	90.00%	snt_num, asl, mwe_disconj, word_gt7, word_gt8, word_gt9, atl, b3, gramv, ttr1, word_type, word_num
82.00%	90.00%	snt_num, asl, mwe_disconj, word_gt7, word_gt8, word_gt9, atl, b3, gramv, ttr2, word_type, word_diff_num
82.00%	90.00%	snt_num, asl, mwe_disconj, word_gt7, word_gt8, word_gt9, atl, b3, gramv, ttr2, lex_type, word_num
82.00%	90.00%	snt_num, asl, mwe_disconj, word_gt7, word_gt8, word_gt9, atl, b3, gramv, ttr1, lex_type, word_diff_num

从表 IV-4 的数据可以看出，结果跟实验二的类似，即部分同组的特征项，对作文质量高低的影响相同。

因此在后续的实验三中，对于多个特征项序列，我们选在 DefSet 中出现的特征项，例如在表 IV-4 中，不同特征项序列出现了 word_type 和 lex_type，而 word_type 包含在 DefSet 中，所以选择 word_type。根据这个原则从五个特征项组合中最终确定的如下组合，定义为 UseSet。

UseSet = {snt_num, asl, mwe_disconj, atl, b3, word_gt7, word_gt8, word_gt9, word_type, word_num, gramv, ttr1}

进一步，我们再选出实验（一）和实验（二）用到但未能选入 DefSet 的特征项，包括如下 13 项：a_type, d_len, d_ttr2, para_num, p_num, p_ratio, p_type, v_len, v_ttr2, fitted_num, b0, x_num, x_ratio。

然后从中分别取出 1 个、2 个……直到 13 个特征项的所有组合，与 UseSet 组合形成新的特征项序列，进行回归分析，选出最优的实验结果，如表 IV-5 所示，可以看出实验性能得到进一步提升，召回率保持在 90.00%，而准确率提高到 92.00%。

表 IV-5 实验三的结果

准确率	召回率	特征项组合
92.00%	90.00%	snt_num, asl, b3, word_gt7, word_gt8, word_gt9, word_type, atl, word_num b0, gramv, ttr1, v_ttr2, p_type, d_ttr2, mwe_disconj

从实验三的结果来看，句子数、单词数、篇章连词数、长词数量、平均词长、三级词汇数量、动词短语的数量、类符形符比、动词的类符形符比、副词的类符形符比、以及介词类数等数值的大小能在很大程度上反映文章质量的高低。这些特征项从不同方面反映了作文的复杂句式，搭配、复杂词汇，词汇多样性等方面的运用情况，基本上包括了写作评判所关注的特征。

V. 结语

本文通过对英语文章的多个文本特征项进行分析，选出了部分特征项，将这些特征项进行了交叉组合，逐步尝试了多个不同的特征项序列，使得模型的性能不断得到提高，得到了一个能较好反映文章质量的特征项组合。

实验结果表明，在已抽取的特征项中，不同的特征项组合对高分作文的判别性能差别较大，但是有些特征项可以稳定地反映作文质量。在后续的研究中，我们考虑扩大实验所用的语料库规模，并尝试更多的特征项组合。

英语作文智能评价系统与传统的英语教学模式有很大不同，具有可靠性、客观性和经济性等的诸多优点，会给英语写作的教学双方带来切实的利益。可以预见，随着有关技术的日益完善，这种英语写作教学方式会被越来越多的英语学习者接受，对于英语作文智能评价系统进一步的研究和探索具有重大意义。

致谢

本论文获得中央高校基本科研业务费专项资金资助。在论文的完成过程中，得到了很多人的帮助，感谢王小捷教授、张跃老师和郭永生老师，提出了宝贵意见并给予了悉心的指导。

REFERENCES

[1] Alan Davies, Principles of Language Testing, 1997

[2] Maocheng Liang, Review and Implications of Foreign Automatic Essay Scoring, 2007 (In Chinese)

[3] Lawrence M. Rudner, Automated Essay Scoring Using Bayes Theorem, 2002

[4] Jin Tang, The Theory and Application of BETSY in Automated Essay Scoring, 2011 (In Chinese)

[5] Jill Burstein and Martin Chodorow, Automated Essay Scoring for Nonnative English Speakers, 1999

[6] Shili Ge, Common Automated Essay Scoring and Feedback for College English Teaching, 2008 (In Chinese)

[7] Ye Lu and Zhiming Xiao, Theory of Statistics, 2006 (In Chinese)

[8] Chao Li, Yongmei Tan, Mingtao Wang and Yixin Zhong. " Research on Feature Selection in Automated Writing Assessment System", The Chinese Symposium on Multimedia Technology and Information Science (MTIS 2011), 2011 (In Chinese)

[9] Leah S. Larkey, Automatic Essay Grading Using Text Categorization Techniques, 1998

[10] Maocheng Liang, Model Construction of Automated Essay Scoring for Chinese Students, 2005 (In Chinese)

[11] Aiguo Cui, Research of Feature Selection Impact on the Automated Essay Scoring, 2009 (In Chinese)

[12] Yali Li and Yonghong Yan, Automated Essay Scoring System for CET4, 2010 (In Chinese)