Research on Feature Selection in Automated Writing Assessment System

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

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

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

Yixin Zhong
Center for Intelligence Science and Technology
Beijing University of Posts and Telecommunications
Beijing, China
zyx@bupt.edu.cn

Abstract - Automated writing assessment system which uses artificial intelligence to evaluate essays and generate feedback. We designed and developed an automated writing assessment system. Not only words and phrases but also dependency triples are used to evaluate the content of the essays from long distance. Users can learn English from the feedbacks given by our system because it gives a detailed explanation of errors. Currently College English Test (CET) has been accepted by society and used as one of the assessment in employment. We did a lot of feature selection experiments on the official CET data and found 26 feature types were contributed to essays assessing. Experimental results provided our method was prominent.

Keywords-feature selection; writing assessment; Statistical Product and Service Solutions

I. 介绍

在英语的教学中，写作练习一直以来都被认为是重要的组成部分，但是由于工作负担的原因，很多时候老师没有足够的时间来批阅学生的作文并给出点评。使用计算机进行作文自动评分不但可以帮助老师减轻负担，而且可以即时给学生们作文批改的反馈信息，帮助学生提高写作水平。作文自动评分[1]是指使用计算机对学生的英语作文进行评分[2]，尤其对于一些简单的，重复性的错误进行批改，辅助英语老师的工作。

II. 相关工作

A. 国内外论文自动评价系统

国外的作文自动评分（AES）系统中比较代表性有：美国的 PEG（Project Essay Grader），IEA（Intelligent Essay Assessor），E-rater（Electronic Essay Rater），IntelliMetric 和最近的 Writing Road map 等。

PEG 是 1966 年开发出的作文自动评分系统 PEG（Project Essay Grader），为了能够大批量的进行作文的打分，并且打分的结果要接近人工的打分。PEG 的优点是机器打分的结果接近人工打分；缺点是，打分的规则较为单一，学生可以根据规则骗取高分。

IEA 是在 1997 年由美国科罗拉多大学开发的 AES 系统，与 PEG 的区别是增加了对进行语义分析的功能，并把语义作为打分的依据之一。IEA 的主要缺点是:不分析作文的语言质量和对篇章结构不作分析[2]。

E-rater 系统于 1999 年用于 GMAT 作文批阅，于 2005 年开始应用于托福考试的作文评分。E-rater 以句法特征结构为主要依据，因此，句法剖析的效果会影响评分的准确性[2]。

IntelliMetric 是本世纪初开发，采用人工智能（AI，Artificial Intelligence）和自然语言处理（NLP，Natural Language Processing）技术和统计方法，支持多语言的作文评分包括很多欧洲语言、希伯来语、阿拉伯语、日语等[2]。

Writing Roadmap 是美国教育测评与研究机构 CTB/McGraw-Hill 于 2008 年推出的在线英语写作评价工具，能及时对作文从六个方面（“思想与内容”、“逻辑与组织”、“论调”、“词汇选择”、“流畅度”和“语言基础”）进行分析、评分并给出评语[2]。

我国作文自动评分系统的发展历史较短，但是随着英语的学习越来越受到人们重视，作文自动评分系统也越来越受到关注。

梁茂成教授主持开发的“大规模考试英语作文自动评分系统”，该系统于 2008 年初步研制完成并通过了研究报告鉴定会[2][3]。

新视野大学英语在线学习作文自动评分，是新视野大学英语在线学习平台内提供的名为“write on”的在线作文自动评分系统，采用的是大学英语四、六级写作评分标准，分数范围为 1-15 分[2]。

冰果英语智能作文评阅系统，是由浙江大学外语学院与杭州增慧网络科技有限公司联合开发出的英语作文自动评分系统[2][4]。

B. SPSS (Statistical Product and Service Solutions)

SPSS(社会科学用统计软件包)统计分析工具，是著名的统计分析软件，适用于自然科学和社会科学各领域，主要

基金项目：北京邮电大学青年科研创新专项项目“基于本体的机器阅读背景知识库自动构建技术研究”(2011.5-2012.12)；国家自然科学基金项目(60873001)

978-1-61284-774-0/11/$26.00 ©2011 IEEE

功能包括：数据处理，统计分析等[5]。我们主要使用的SPSS 的相关性分析和回归分析这两个功能，以及 SPSS 的Python 插件，用于批量进行多组实验。

相关（Correlation）分析和回归（Regression）分析模型是统计学中最常用、最成熟、最基本的理论之一。无论在自然科学领域，还是在社会科学领域，一些现象与另外一些现象之间往往存在着相互依赖的关系。如果采用变量反映这些现象的特征，则表现为变量之间的函数关系或相关关系[5]。SPSS 中提供了 3 种相关模型，本文中我们使用的是 Kendall 秩相关模型。

计算 Kendall 秩相关系数，适用于有序变量或不满足正态分布假设的等间隔数据。Kendall 秩相关模型是一种比较直观的模型，该模型先按某一变量 X 进行排序，检查另一变量 Y 与变量 X 的排序差异，若无差异则变量 Y 与 X 完全相关，若差异大则变量 Y 与 X 为低度相关或无相关[5]。

Logistic 回归分析是十分有用的，它能更细微刻画自变量对因变量的影响，尤其是对于二值（如 0、1 编码数据）因变量的预测和解释。Logistic 回归系数能够用于预测模型中的自变量之间的比值，在很多领域都有重要的应用。二值回归分析也称二分回归或二项回归分析，即因变量仅能取两个值的回归分析[5]。在本文中，我们对作文是否为低分（0 或 1）进行二值回归分析。

(1)因变量 y 必须是二值变量[5]。

如果 y 为 1，则表示出现结果为真，如死亡、感染、成功；反之，y 为 0 表示出现的结果为假，如生存、未感染、不成功[5]。

(2)假设 P 表示出现“真”的概率，1-P 表示出现“假”的概率，其比例也称优势比、比数、比值等，OR，即 odds=P/(1-P)[5]。

此处概率 P 的数值总是在 0～1 之间，当 P→1 时，odds→∞；当 P→0.5 时，odds→1，当 P→0 时，odds→0[5]。

(3)如果取 y=logit(P)=ln(odds)=ln(P/(1-P))，则当 P→1 时，y→∞；当 P→0.5 时，y→0；当 P→0 时，y→－∞[5]。

由于 y=ln(P/(1-P))，则 e^y = P/(1-P)，P= e^y /(1+ e^y)[5]。

由此可见，P 值在区间(0,1)时，对应 y 值在区间$(-\infty,\infty)$[5]。

(4) 如果令 y=logit(P)=ln(odds)=ln(P/(1-P))= $b_0 + b_1x_1 + b_2x_2 + \cdots + b_ix_i$ ，y 可代入：P= e^y /(1+ e^y)[5]。

III. 英语作文智能评价系统中的低分作文特征选择研究

我们借鉴了从事英语教学工作的老师们的经验及英语研究工作者的研究建议，本论文中我们从篇章，段落，句子，词语等各个不同的层次来进行特征的抽取，得到 141 个特征模版。

我们根据官方提供的历年大学英语四、六级考试范文来对英语作文智能评价系统中的低分作文特征进行研究，其中的低分作文特征选取框图如下所示：

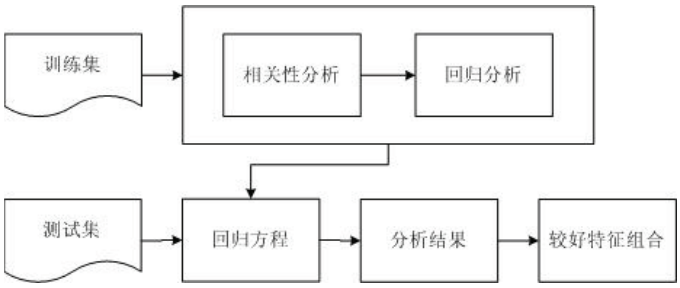

Figure 1. 低分作文特征选取框图

首先，我们使用 SPSS 统计分析工具中提供的相关性分析功能，找出与低分作文显著相关的特征(f1, f2, …, fi)；对所获得的显著性特征进行各种组合，使用 SPSS 提供的回归分析功能，进行回归分析，得到相应的回归方程。

接下来，我们对前面获得模型进行测试。我们把测试数据中相应特征(f1, f2, …, fi)的值代入到回归方程中做回归运算；分析运算结果，计算准确率和召回率；以准确率和召回率为基础，找出最符合低分作文的特征组合。

A. 实验数据

实验中使用的作文是历年大学英语四、六级考试官方提供的范文。总共有作文 110 篇，其中低分作文 22 篇，非低分作文 88 篇，四级作文 60 篇，其中低分作文 12 篇，非低分 48 篇，六级作文 50 篇，其中低分作文 10 篇，非低分 40 篇，在后续的方案一和方案二中分别把这些作文作为训练集和测试集。

在进行实验之前，我们把语料中的英文作文分为五档：2 分、5 分、8 分、11 分、14 分。其中 2 分的作文在实验过程中被认为是低分作文，标记为 1；其它的作文标记为 0。

TABLE I. 实验数据统计信息

	四级作文	六级作文
篇	60	50
段落数/篇(平均)	4	4
句子数/篇(平均)	13	14
单词数/篇(平均)	160	188
句子个数/段(平均)	3.71	4.05
单词数/段(平均)	40	47
单词长度/段(平均)	53.24	63.73
单词长度/句子(平均)	4.35	4.44

B. 特征选择实验

特征选择方案一以四级作文作为训练集，六级作文作为测试集来进行特征选择的研究，分别对单特征，双特

征，三特征和多特征的组合进行实验，找出较好的特征组合。

特征选择方案二以六级作文作为训练集，四级作文作为测试集来进行特征选择的研究，对方案一中选择出来的较好的特征组合进行实验，找出较为稳定的特征组合。

1) 方案一

首先，我们对训练数据使用 SPSS 进行相关性分析，一共找到 26 个特征与低分作文显著相关。

TABLE II. 低分相关特征

文章中所有词的总数与文章中所有词的总数（去重后）比值的平方根
文章中所有句子的平均长度，以单词为单位（单词/句）
错误单词比率(零级词汇占比)
三级一级词比
字符个数大于 7 的词占比
错误句子数(分析失败句子个数)
篇章连词数（作文中的连词数）
句子谓词前面部分单词的个数所占的比例
句子谓词后面部分单词的个数占比
形容词数（去重后）
形容词比率
文章中动词的总数与文章中动词的总数（去重后）的比值平方根
名词比率
文章中介词的总数与文章中介词的总数（去重后）的比值平方根
其他词性比率
句号比率
简单句个数
平均每段包括的单词数
动词短语的个数
动词短语比率
动词短语难度值的均值
动词+任意词+介词组合的个数（比如 make use of）
动词+to+动词组合的个数（比如 plan to go）
动词+任意词+任意词+介词组合的个数（比如 be in love with）
拼写错误数
由 0 到 9 个单词构成的句子个数

TABLE III. 实验结果一

特征		**实验结果**	
特征类型	***特征组合***	***准确率 P***	***召回率 R***
单特征	错误句子数	92	70
	错误单词比率	86	50
两个特征	错误单词比率、错误句子数	94	90
	错误单词比率、句子中谓词后面部分单词数占整个句子单词数的比率	92	90
三个特征	错误单词比率、错误句子数、字符个数大于 7 的单词比率	96	100
	错误单词比率、错误句子数、句子中谓词前面部分单词数占整个句子单词数的比率	96	100
	错误单词比率、错误句子数、句子中谓词后面部分单词数占整个句子单词数的比率	96	100
	错误单词比率、错误句子数、形容词比率	96	100
	错误单词比率、错误句子数、简单句个数	96	100
	错误单词比率、错误句子数、动词短语的个数占全部短语个数比率	96	100
四个特征	错误单词比率、错误句子数、字符个数大于 7 的单词比率、动词+to+动词的组合数	100	100
	错误单词比率、错误句子数、三级词汇与一级词汇的比值、动词+to+动词的组合数	100	100
	错误单词比率、错误句子数、字符个数大于 7 的单词比率、连词数	98	100
	错误单词比率、错误句子数、字符个数大于 7 的单词比率、动词短语数	98	100
	错误单词比率、错误句子数、句子中谓词后面部分单词数占整个句子单词数的比率、三级词汇与一级词汇的比值	98	100
五个特征	错误单词比率、错误句子数、字符个数大于 7 的单词比率、动词短语数、三级词汇与一级词汇的比值	100	100
	错误单词比率、错误句子数、字符个数大于 7 的单词比率、动词+to+动词的组合数、三级词汇与一级词汇的比值	100	100
	错误单词比率、错误句子数、字符个数大于 7 的单词比率、动词短语数、简单句个数	100	100
	错误单词比率、错误句子数、动词+to+动词的组合数、三级词汇与一级词汇的比值、简单句个数	100	100
	错误单词比率、错误句子数、动词+to+动词的组合数、三级词汇与一级词汇的比值、动词+任意单词+介词（形如 make use of 形式）的动词短语个数	100	100
	错误单词比率、错误句子数、动词+to+动词的组合数、三级词汇与一级词汇的比值、动词短语的个数占全部短语个数比率	100	100

单特征：在训练集中，每次使用一个特征与是否标记为低分作文(1 或 0)，用 SPSS 来进行回归分析，得到回归方程。然后在测试集中计算准确率和召回率。在使用单特征进行实验得到的结果中，我们找出两种特征对于识别英语作文是否为低分的效果较好。在前面单特征实验的基础上，接下来，我们把单特征扩展为两两组合，三三组合等多特征组合，进行上面的实验过程，找出最符合低分作文的特征组合。

双特征组合：在训练集中，每次使用两个特征与是否标记为低分作文(1 或 0)，用 SPSS 来进行回归分析，得到回归方程。然后在测试集中计算准确率和召回率。在使用双特征组合进行实验得到的结果中，我们找出两种特征组合对于识别英语作文是否为低分的效果较好，准确率和召

回率都超过了 90%，其中最好的组合为错误单词比率、错误句子数的组合，如上表 3 所示。

三个特征组合：在训练集中，每次使用三个特征与是否标记为低分作文(1 或 0)，用 SPSS 来进行回归分析，得到回归方程。然后在测试集中计算准确率和召回率。在使用三特征组合进行实验得到的结果中，我们找出六种特征组合对于识别英语作文是否为低分的效果较好，第一次出现召回率 100%，如上表 3 所示。

四个特征组合：在训练集中，每次使用四个特征与是否标记为低分作文(1 或 0)，用 SPSS 来进行回归分析，得到回归方程。然后在测试集中计算准确率和召回率。在使用四特征组合进行实验得到的结果中，我们找出五种特征组合对于识别英语作文是否为低分的效果较好，第一次出现准确率和召回率双 100%，其中最好的组合为错误单词比率、错误句子数、字符个数大于 7 的单词比率、动词+to+动词的组合数和错误单词比率、错误句子数、三级词汇与一级词汇的比值、动词+to+动词的组合数，这两种组合，如上表 3 所示。

五个特征组合：在训练集中，每次使用五个特征与是否标记为低分作文(1 或 0)，用 SPSS 来进行回归分析，得到回归方程。然后在测试集中计算准确率和召回率。在使用五特征组合进行实验得到的结果中，我们找六种特征组合对于识别英语作文是否为低分的效果较好，准确率和召回率均为 100%的组合增加到了 6 组，如上表 3 所示。

2) 方案二

特征选择方案二以六级作文作为训练集，四级作文作为测试集来进行低分作文特征选择的研究，根据方案一中选择出来的较好的特征组合来进行实验，从而找出实验结果较为稳定的特征组合。

在训练集中，每次使用方案一中得到较好组合的其中一组与是否标记为低分作文(1 或 0)，用 SPSS 来进行回归分析，得到回归方程。然后在测试集中计算准确率和召回率。

TABLE IV. 实验结果二

特征		实验结果	
特征数	***特征组合***	***准确率 P***	***召回率 R***
四个特征	错误单词比率、错误句子数、三级词汇与一级词汇的比值、动词+to+动词的组合数	90	75
五个特征	错误单词比率、错误句子数、动词+to+动词的组合数、三级词汇与一级词汇的比值、动词+任意单词+介词（形如 make use of 形式）的动词短语个数	90	75
	错误单词比率、错误句子数、字符个数大于 7 的单词比率、动词短语数、简单句个数	88.3333	75

在使用方案一中得到的较好组合进行实验得到的结果中，我们找到三种特征组合对于识别英语作文是否为低分的效果较好，其中准确率和召回率表现最好的组合是四特征组合和五特征组合，如表 4 所示。

也就是说，在目前实验结果的基础上，可以认为前两种是在识别英语作文是否为低分时，效果和稳定性是较好的组合。

IV. 结论

通过观察和分析实验结果，我们找到一些较好的特征组合可以用来判别英语作文是否为低分作文。首先，我们发现在这些较好的特征组合中全部出现了错误单词比率和错误句子数这两种特征，并且在双特征组合实验中，使用这两种特征的组合也有较高的准确率和召回率。其次，出现较为频繁特征是动词短语相关的特征，这说明英语作文中，动词短语的使用情况对低分作文有着较明显的影响。最后，我们发现在方案二中，表现较好的特征组合都是特征个数较多的组合，这说明随着使用特征个数的增加，判别英语作文是否为低分作文的稳定性则越强。

随着英语的学习越来越受到人们重视，作文自动评分系统也越来越受到关注。作文自动评分系统不但可以帮助老师减轻负担，而且可以及时给学生们作文批改的反馈信息。尤其可以帮助老师修改一些简单，重复，机械性的错误（例如：拼写等），用这些节省下来的时间和精力，老师可以更多地关注作文的内容、表达方式等方面，指导学生，帮助学生提高写作水平。

致谢

感谢王智超同学在实验过程中的帮助和建议。

REFERENCES

[1] Mark Warschauer and Douglas Grimes , Automated Writing Assessment in the Classroom, Pedagogies: An International Journal, January 2008.

[2] Xianchun Xie, Discussion with the validity and reliability and operability of the English essay automatic scoring, Jiangxi Normal University Journal, April 2010 (In chinese).

[3] Shili Ge, Chinese learners to explore essay automatic scoring, Foreign sector, vol 5, 2007 (In chinese).

[4] Xinhua Newspaper[EB/OL] , http://news.xhby.net/system /2009/04/14/010481858.shtml, April 11th 2009 (In chinese).

[5] Runlong Huang, Statistical analysis——Principle and Application of SPSS, Higher Education Press, July 2010 (In chinese).

[6] Shili Ge, Comparision and study of College English essay automatic scoring method, Guangdong University of Foreign Studies Journal, May 2010 (In chinese)..

[7] Puwei Huang, Analysis of English examination results via SPSS correlation analysis and linear regression analysis, China Electricity Education, October 2007 (In chinese).