少数民族汉语考试作文自动评分的特征提取研究

蔡黎，彭星源，柯登峰，赵军

中国科学院自动化研究所模式识别国家重点实验室，北京 100080

E-mail: lcai@nlpr.ia.ac.cn

摘　要：随着计算机的普遍应用以及计算机技术的飞快发展，计算机自动性测试和计算机自适应性测试都已先后成为现实，这样计算机自动评分就成为了人们所希望的下一代计算机自动工具。中文自动评分系统的研究尚处于起步阶段，据我们了解还没有一个能大规模使用的系统。本文研究了许多前人的工作，并复现部分特征，但是特征的相关度不是很理想。在本文中，我们利用统计自然语言处理和信息检索技术提取作文写作水平和作文主题特征。实验表明，利用本文提出的特征的相关度好于以前的特征。

关键词：自动评分系统，汉语，主题特征，写作水平特征

Research of the feature for Automated Essay Scoring System for Chinese Proficiency Test for Minorities

Li CAI, Xingyuan PENG, Dengfeng KE, Jun ZHAO

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, 100080

E-mail: lcai@nlpr.ia.ac.cn

Abstract: With the widespread application of computer and fast development of computer techonlogy, computer aided test and computer adopted test have turned into realization. Automated essay scoring system (AES) have become the next generation of computer aided tools in people's expectation. Chinese AES is still in its infant stage. As we have known, there is even no Chinese AES which can be widely used. We have done a lot of research on previous works. And we extracted some features described in previous work. However, the result was not promising. In this paper, we use the technology of statistical natural language processing and information retrieval to extract topic and writing level features of the essays. The experiment shows that our features are more effective than previos ones.

Keywords: automated essay score; Chinese; topic feature; writing level feature

1　引言

通过作文考试可以检测应试者综合运用语言的能力。然而，大规模作文阅卷面临两大难题：其一，阅卷需要耗费大量人力、物力等资源；其二，评判作文质量具有很强的主观性，阅卷的信度和效度不强[11]。

自动作文评分是计算机技术在语言测试方面的最新应用，也是语言技术发展的必然趋势。自动作文评分中核心的问题是，机器可用的、高信度的评分特征的选取[11]。

本文针对这个问题，利用自然语言处理和信息检索技术，从作文写作水平和主题相关两个方面，通过词的信息提取了稳定，有效的特征。

作者简介：蔡黎（1981-），男，湖北，学生，博士研究生 email: lcai@nlpr.ia.ac.cn

本文按如下方式进行组织：第二节中介绍了国外的自动评分系统，第三节中介绍了利用自然语言处理技术和信息检索技术，提取作文主题和写作水平的特征，第四节介绍和分析了实验设计以及实验的结果，第五节中对本文以及已有工作的问题进行了总结，同时展望了进一步的工作的方向。

2 相关工作

国外对自动评分系统早在1960起就开始了，比较著名的自动评分系统有 Project Essay Grade[9][10]，Intelligent Essay Assessor[7][8]，IntelliMetricTM[5][6]，E-rater[3][4]等。

PEG[9][10]是在 1966 年 EllisPage 应美国大学委员会的要求而研发的。和李亚男[12]提取的特征相似，PEG 主要依靠对文章的浅层语言学特征的分析（例如，作文长度，介词、关系代词等，词长的变化等等），然后对作文进行评分。基本上，没有涉及写作水平、句子结构，文章内容，文章措辞等高级特征。最后该系统无法给出对学生有意义的指导意见。

IEA[7][8]是 20 世纪 90 年代末由 Pearson Knowledge Analysis Technology 公司利用潜在语义分析[2](latent semantic analysis)技术开发的。潜在语义分析，是 1988 年 S.T. Dumais 等人提出了一种新的信息检索代数模型，是用于知识获取和展示的计算理论和方法，它使用统计计算的方法对大量的文本集进行分析，从而提取出词与词之间潜在的语义结构，并用这种潜在的语义结构，来表示词和文本，到达消除词之间的相关性和简化文本向量实现降维的目的。潜在语义分析的基本观点是：把高维的向量空间模型表示中的文档映射到低维的潜在语义空间中。这个映射是通过对项/文档矩阵的奇异值分解（SVD）来实现的。

IntelliMetricTM[5][6]是第一套基于人工智能(AI)的作文评分系统。它的开发商 Vantage Learning 应用了人工智能、自然语言处理和统计技术，使得 IntelliMetricTM 能够模仿人工阅卷，对作文的内容、形式、组织和写作习惯进行分别进行评分。IntelliMetricTM 需要对已经评好分数的作文集进行训练，构建模型。对于要评阅的作文，IntelliMetricTM 提取了作文中包括语义、句法、篇章 3 个方面的 300 多项特征，代入模型评分。其效果与评卷员的一致率达到了 97%至 99%。

E-rater[3][4]是由 Educational Testing Service(ETS)的 Burstein 等人在上世纪 90 年代末开发的。据我们了解，E-rater 是目前商用效果比较好的自动评分系统，已经在 GMAT，TOEFL 考试中商用。E-rater 系统主要由 5 个模块，其中 3 个模块用来抽取特征，一共 67 个特征，这些特征包括：句法，篇章，主题。其中的自然语言处理技术采用的是微软自然语言处理的工具包来完成。第 4 个模块，是用来构建模型，对 67 个变量中进行筛选，建立回归方程。第 5 个模块是用来计算待评分文章的最后得分，即提取作文显著特征的特征值，代入回归方程计算最后得分。

3 特征抽取和建模

一个优秀的作文自动评分系统最重要是，能从作文中挖掘出反映作文质量的，机器可

[1]http://ir.hit.edu.cn/

用的特征。从 E-rater 的文献[3][4]中，我们知道 E-rater 使用了微软自然语言处理的工具包来提取比较深层次的特征如句法识别句子的复杂度等等。同样，本文利用哈工大信息检索实验室[1]提供的自然语言处理包，也做了相关实验，但是实验的效果都不理想。原因可能主要是训练语料和测试语料的领域不相关引起的。

本文利用自然语言处理和信息检索技术，从作文写作水平和作文主题两个方面，提取了稳定，有效的特征。

3.1 作文写作水平特征

作文写作水平特征，反映的是考生使用语言的能力。

衡量一个作文的写作水平，有多个方面，最重要的是遣词造句。现在的中文自然语言处理技术还没有达到，能很准确的提取句子的特征如句式，句中词语搭配的质量等。本文主要从词方面入手，提取作文写作水平特征。该特征是基于以下的常理：越常见的词，越是易用词，越不常见的词，越是难用词。图 1 给出了作文写作水平特征提取的算法流程。

算法：作文写作水平特征提取
输入：分词后的大语料L，分词后的作文E，词频阈值$limit$
输出：作文E的写作水平特征值S
方法：
1. 对大语料L进行统计词频，词W_i的词频记为f_{w_i}
2. 对于每个$W_i \in L$，如果$f_{w_i} < limit$，把$f_{w_i} = limit$，以避免语料库的稀疏性
3. 对于每个$W_i \in L$，计算W_i的使用难度系数$\lambda_{w_i} = 1/log f_{w_i}$
4. 作文E的写作水平特征为该篇文章所有词的使用难度系数之和
$$S = \sum_{i=0}^{n} \lambda_{w_i},$$ n为文章的词数
5. 返回S

图 1 作文写作水平特征提取算法描述

Fig.1 the description of algorithm of extracting writing level feature from the essay.

3.2 作文主题特征

作文主题特征，反映的是考生作文内容的扣题程度。

主题在作文评分中的重要性是不言而喻的。因为本文测试的对象是汉语作为第二语言学习者，所以作文中文不对题，背范文的现象还是很严重的。我们请两位经验丰富的评卷员对随机抽取的 500 篇作文，进行跑题作文和非跑题作文的分类。分类结果如表 1。

表 1 评卷员对作文跑题的分类结果

Tab.1 The labelers' classification result of the topic and off-topic essays

评卷员	跑题篇数	比例
评卷员 A	104	20.8%
评卷员 B	138	27.6%

从表 1 中，可以看出跑题作文的比例还是较大的。用现在的自然语言处理技术，提取

整篇文章的语义基本上很难做到。在这种情况下，主题特征就成为了衡量文章内容主题相关性很重要的特征。

作文主题特征提取，Burstein[4]利用作文内容词向量和作文题目词向量的相似度作为判别作文是否跑题的标准。Burstein 方法的问题是现代考试的作文题目是多样的，不是所有作文题目都是文字的，比如看图说话就没法用以上的算法。

为了解决这个问题，本文利用信息检索里成熟的词频技术。

词频技术是一种用于信息搜索和信息挖掘的常用加权技术。词频技术的主要思想是，如果某个词或短语在一篇文章中出现的频率较高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。词频指的是某一个给定的词语在该文件中出现的次数。本文利用这项技术背后的思想，即在样本作文中出现次数越多的词，越是主题相关词。图 2 给出了作文主题特征提取的算法流程。

算法：作文主题特征提取
输入：分词后的大语料L，分词和词性标注后的N篇作文样本语料，分词后的预测作文E，词频阈值$limit$
输出：作文E的主题特征值S
方法：
1.对大语料L进行统计词频，词W_i的词频记为F_{w_i}
2.对于每个$W_i \in L$，如果$F_{w_i} < limit$，把$F_{w_i} = limit$，以避免语料库的稀疏性，总词数为H
3.对分词和词性标注后的N篇作文样本语料，去掉辅助功能的词，只保留名词，动词，形容词，副词。
4. 对N篇作文样本语料统计词频，词W_i的词频记为f_{w_i}，总词数为h
5. 计算词W_i的主题相关系数$\lambda_{w_i} = (\mathrm{f}_{w_i}/h)/(F_{w_i}/H)$
6. 作文E的主题特征值为该篇文章所有词的主题特征值之和$S = \sum_{i=0}^{n} \lambda_{w_i}$，$n$为文章的词数
7. 返回S

图 2 作文主题特征提取算法描述

Fig.2 the description of algorithm of extracting topic feature from the essay.

4 实验结果及分析

4.1 实验数据集简介

实验中，本文使用的语料来自 2008 年 5 月在内蒙古举行的少数民族汉考。我们从 18000 篇真实考生作文中，随机挑选了 772 篇录入。作文总分为 6 分。考试作文的题目是读一段短文，然后写感想。这种半开放的考试题目在汉语考试很常见。本文用作文的最终评分作

为试卷的分数，即有两个评卷员评分，如果两个评卷员的评分相差大于 1 分，就请一个资深评卷员评分，最终分数是以资深评卷员评分为主，综合前面两个评卷员的评分。如果两个评卷员的评分相差小于或等于 1 分，就取两个评阅员评分的中间值作为最终分数。

4.2 评测指标

实验中，评测采用了相关度。相关度代表的是现象之间是否相关、相关的方向和密切程度，一般不区别自变量或因变量。相关度的计算方法和数学的协方差一样。本文采用相关度作为主要的评测指标，是因为在提取特征后，建模手段是回归分析。回归分析是通过规定因变量和自变量来确定变量之间的因果关系，建立回归模型，所以特征的相关度越高，回归模型预测会越好。

4.3 实验及分析

实验中，本文首先以李亚男[12]的一些特征为基线系统，然后实现了作文写作水平特征和作文主题特征做对比。实验中，我们使用哈工大信息检索实验室[1]提供的自然语言处理包对作文进行分词和词性标注。

李亚男[12]主要通过提取浅层的特征，如文章的总词数，包含大纲中甲级词数(考试大纲中根据词的难易分级)，包含大纲中乙级词数等等，共 44 个，后进行回归。浅层特征的相关度不是很高，基本都在 0.3 以下。本文用 772 篇测试语料，实验了李亚男的几个特征，特征相关度见表 2。

表 2 李亚男特征测试结果

Tab.2 The test result of Li's features

特征	词数	甲级词数	乙级词数	丙级词数
相关度	0.2865	0.1770	0.2731	0.1301

从表 2 中是李亚男论文中特征度相关比较高的特征，可以看出这些特征的相关度仍然小于 0.3

4.3.1 作文写作水平特征实验

在进行作文写作水平特征抽取实验时，本文采用人民日报（1998 年 1-6 月）的语料，作为大语库，测试语料是 772 篇作文语料。772 篇这个数量，足够能保证，特征相关度的稳定性和有效性。对于防止语料稀疏性的词频阈值，我们通过实验来选取。表 3 反映的是，特征相关度随着词频阈值变化的情况。

表 3 作文写作水平特征测试结果

Tab.3 The test result of writing level feature from the essay

阈值	10	20	30	40	50	60	70	80	90
相关度	0.4398	0.4437	0.4455	0.4464	0.4477	0.4486	0.4488	0.4471	0.4453

从表 3 中可以看出，阈值对作文写作水平特征相关度的影响不是特别明显，作文写作水平特征相关度在不同阈值下还是比较稳定的。从表 3 中可以看出，作文写作水平特征在阈值 50-70 时，达到相对比较高的阶段，达到 0.45 左右。

4.3.2 作文主题特征实验

表 4 作文主题特征测试结果

Tab.4 The test result of topic feature from the essay

阈值	1	3	5	10	15	20	25
相关度	0.5015	0.4934	0.4901	0.4859	0.4793	0.4751	0.4759

在进行作文主题特征的抽取实验时，本文也采用人民日报（1998 年 1-6 月）的语料，作为大语库，随机选 100 作文作为作文样本语料，测试语料是 772 篇作文语料。对于防止语料稀疏性的词频阈值，我们通过实验来选取。表 4 反映的是，特征相关度随着词频阈值变化的情况。

从表 4 中可以看出，阈值对作文写作水平特征相关度的影响不是特别明显，作文写作水平特征相关度在不同阈值下还是比较稳定的。从表 4 中可以看出，作文写作水平特征在阈值 1-5 时，相对比较高，达到 0.5 左右。通过上面实验，对比表 2，表 3 和表 4，可以看出本文提出的作文写作水平和作文主题特征的相关度都在 0.4 以上，最好的还达到 0.5 以上，比表 2 中的特征相关度要高，且提高幅度很大。

5 结语

本文利用统计自然语言处理和信息检索技术，提出了作文自动评分系统中作文写作水平和作文主题特征，并进行了特征提取与评估的相关实验。实验证明，本文所提出的特征稳定且有效。

相对英文作文自动评分系统，中文作文自动评分系统还处于起步阶段，未来还有很多工作需要完善。可能存在以下几个方面：

(1) 利用更高级的自然语言处理和信息检索技术，从作文中挖掘出更多跟作文质量好坏相关的特征。

(2) 防作弊技术，中文作文自动评分如果要想在商业应用上取得成功，很完善的防作弊技术是必不可少的。

(3) 特征建模技术，如何利用现有特征建模，提高作文自动评分系统预测的精度。

计算机作文作文评分是一个复杂的过程，需要总结前人的经验并不断汲取新的理念、利用最新的技术。这样，才能不断的提高计算机作文作文评分的精度。

参 考 文 献

[1] Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2002). Comparing the validity of automated and human scoring of essays. Journal of Educational Computing Research, 26, 407–425.

[2] Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. Discourse Processes, 25, 259–284.

[3] Burstein, J. (2003). The e-rater scoring engine: Automated Essay Scoring with natural language processing. In M. D. Shermis and J. C. Burstein (Eds.), Automated Essay Scoring: A cross disciplinary approach (pp. 113–121). Mahwah, NJ: Lawrence Erlbaum Associates.

[1] http://ir.hit.edu.cn/

[4] J. Burstein, K. Kukich, S. Wolff, C. Lu, M. Chodorow,L. Braden-Harder, and M. D. Harris. Automated scoring using a hybrid feature identification technique.In Proceedings of the 17th international conference on Computational linguistics, pages 206 - 210, Morristown,NJ, USA, 1998. Association for Computational Linguistics.

[5] Elliot, S., & Mikulas, C. (2004). How does IntelliMetricTM score essay responses? A mind based approach. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

[6] Elliot, S. M. (2001, April). IntelliMetric: From here to validity. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

[7] Landauer, T. K., Laham, D., & Foltz, P. W. (2001, February). The intelligent essay assessor: Putting knowledge to the test. Paper presented at the Association of Test Publishers Computer-Based Testing: Emerging Technologies and Opportunities for Diverse Applications conference, Tucson, AZ.

[8] Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), Automated essay scoring: A cross-disciplinary perspective (pp. 87–112). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

[9] Page, E. B. (2003). Project Essay Grade: PEG. In M. D. Shermis & J. Burstein (Eds.), Automated essay scoring: A cross-disciplinary perspective (pp. 43–54). Mahwah, NJ: Lawrence Erlbaum Associates.

[10] Page, E. B. (1994). Computer Grading of Student Prose ,Using Modern Concepts and Software, Journal of Experimental Education, 62, 127–14

[11] 梁茂成，文秋芳. 国外作文自动评分系统评述及启示.外语电话教学，2007, No.117.

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

[13] 李莉，张太红. LSA 在中文短文自动判分系统中的应用研究，计算机工程与应用，2007，43(20):177-180.