近五十年来自动评分研究综述*

——兼论中国学生英译汉机器评分系统的新探索

江进林

（对外经济贸易大学 商务英语与跨文化研究中心，北京 100029）

【摘要】近五十年来，国内外相继开发出多个英语作文自动评分系统，研究日臻成熟。在翻译领域，自动评分研究主要局限于机器翻译评价，人工译文自动评分研究仍处于初级阶段。近年国内建立起针对中国学生的汉译英自动评分模型，针对英译汉的自动评分研究也开始起步。由于中国学生的英译汉具有自身的特点，其评分系统在变量挖掘、模型验证等方面与已有研究不同。

【关键词】自动评分；作文；翻译；学生英译汉

【中图分类号】G40-057 **【文献标识码】**A **【论文编号】**1009—8097（2013）06—0062—05 **【DOI】**10.3969/j.issn.1009-8097.2013.06.013

引言

主观题是测量语言技能的有效方式，目前已被广泛运用于各类英语考试，其自动评分是测试领域关注的一个焦点。自二十世纪六十年代以来，国外已开发出多个作文自动评分系统，并应用于 GRE、GMAT 等大型考试中[1][2][3]。在国内，梁茂成[4]研制了适合中国英语学习者的作文自动评分系统，取得了良好的效果。在翻译领域，极少数研究也对学生汉译英的自动评分进行了尝试[5][6]。但是，针对英译汉的自动评分研究仍处于起步阶段。本文将回顾近五十年来自动评分技术的优劣，探讨英译汉机器评分系统与已有研究的异同。

一 作文自动评分系统

历史上第一个作文自动评分系统是 1966 年研制的 PEG。[7] 二十世纪九十年代以后，IEA、E-rater、IntelliMetric、MY Access 等作文自动评分系统相继出现[8][9]。近年来，自动评分扩展到医学、建筑、艺术、计算机等领域，评分对象涉及简答、绘图、口试等多种主观题[10][11]。本文仅对四个主要作文自动评分系统进行回顾，其主要特点见表 1。

表 1 主要作文自动评分系统的特点

	PEG	IEA	E-rater	梁茂成[4]的系统
测量对象	语言	内容	语言、内容、结构	语言、内容、结构
评分方法	变量提取、多元回归、计算作文分数	根据文本相似度进行机器评分	变量提取、多元回归、计算作文分数	变量提取、多元回归、计算作文分数
主要技术	统计技术、自然语言处理	潜语义分析	统计技术、自然语言处理、向量空间模型	统计技术、自然语言处理、潜语义分析
主要变量	表层形式特征（如文本长度）	语义相似度	句法结构、连接词、内容相关度	流利度、地道性、复杂度变量；语义相似度；连接词
验证方法	机器/人工评分的相关度	机器/人工评分的相关度	机器/人工评分的相关度和一致性	机器/人工评分的相关度和一致性

第一，测量对象。表 1 显示，作文自动评分系统的测量对象从语言形式发展到语义内容，再过渡到语言、内容和结构三个方面。在梁茂成[4]的研究中，内容模块主要考察作文是否紧扣主题；语言模块主要衡量作文语言形式的准确性；结构模块主要评判作文是否满足独立成篇的条件。这三个模块可以直接追溯到写作能力的构念，具有较好的效度，也更符合写作测试的评价标准[12]。

第二，评分方法。PEG、E-rater 和梁茂成[4]的系统都采用变量提取、多元回归、计算作文分数三大步骤来完成评分。首先，研究者从一批事先评分的作文中提取一系列文本特征，再以这些特征为自变量、人工评分为因变量进行多元线性回归分析，得到能够最大限度地预测分数的回归方程，最后将新作文的相关变量代入方程，获得机器给新作文评出的分数。

第三，主要技术。四个评分系统都采用多种技术来提取变量。其中，IEA 和梁茂成的系统都采用了潜语义分析方法（Latent Semantic Analysis）。其基本假设是，文本中隐藏着一个潜在的语义空间，是所有词汇的语义之和。由于语言中存在大量多词同义和一词多义现象，语义空间往往带有许多噪

音，需要通过特征过滤、选择、抽取来进行压缩。具体做法是：首先，研究者使用停词表过滤信息量很少的词汇；其次，选择一批与主题相关的文本（如专家作文、主题知识材料）构建词频矩阵，并根据词频对词汇赋予不同权重。词汇出现的次数越多，表示信息量越小，权重越低；最后，使用奇异值分解技术（Singular Value Decomposition）对矩阵进行降维。这种技术类似于主成分分析法，压缩后的矩阵既保留了原矩阵的重要信息，又排除了干扰信息，代表作文主题的典型潜在语义空间[13]。潜语义分析具有提取语义内容的优势，甚至能够处理创造性的记叙文。不过，它忽略了词汇顺序、句法、逻辑等信息，不能反映学生的全部知识[8]，因而需要与反映语言形式的变量结合使用。

与潜语义分析不同，E-rater 使用向量空间模型（Vector Space Model）来判定文本内容的相关度。[9]不过，这是一种基于主题词分析的技术，难以达到潜语义分析的降维、消除噪音等效果。[14]

第四，主要变量。各个系统使用的主要变量与其测量对象对应。例如，梁茂成采用流利度、地道性、复杂度方面的变量来考察语言形式质量，采用语义相似度来衡量语义质量，采用连接词等特征来评判作文结构质量。

第五，验证方法。上述系统主要采用相关度和一致性来检验机器评分与人工评分的接近程度。相关度反映机器与人工排序的相似性，既包括机器与单个评分员评分的相关，也包括机器与多名评分员平均分数的相关。第一种相关度不一定可靠，因为单个评分员的评分可能具有偏差（bias），内部一致性难以保证[15]；第二种相关度更有价值，因为多名评分员对同一名学生的平均评分接近其真分数（true score）[7]。

一致性反映具体评分等级的一致程度，包括绝对一致（exact agreement）和相邻一致（adjacent agreement）百分比[12]。前者指机器与人工所评等级相同的文本数量占所评文本总数的比例，后者指机器与人工所评等级相差 1 级的文本数量占所评文本总数的比例，两者各有所长。当评分结果为离散数据且等级较少时，往往使用绝对一致百分比；当评分等级较多时，相邻一致百分比更适合[15]。E-rater 和梁茂成的研究对两种百分比都进行了统计。

除了上述系统外，极少数人还对汉语作文的自动评分进行了初步研究[16]。不过，该研究仅探讨了潜语义分析技术在自动评分中的应用，不够全面。

总之，现有作文自动评分系统在评分步骤、主要技术和变量挖掘方面对英译汉的机器评分研究具有重要启示。研究结果表明，不管考生处于哪个年龄段、作文话题如何变化，上述系统的评分与人工评分的相关度都在 0.7-0.9 之间，一般为 0.8-0.85，可以代替一名评分员使用。

二 翻译自动评分系统

翻译自动评分系统有两种，分别对机器翻译和人工译文进行评价，下面分两部分进行述评。

1 机器翻译评价系统

机器翻译评价主要采用两种方法：

第一，基于 N 元组（Ngram）的评价。其主要思想是：高质量的机器译文应与人工译文具有较多相同的语言片段。BLEU 和 NIST 是该方法的主要代表。BLEU 通过计算机器翻译与一组参考译文内 N 元组的相似度来考察机器译文的质量，即 N 元组的匹配数量所占机器译文 N 元组的比例。如果机器译文比它最接近的参考译文短，相似度的结果还需要乘以长度罚分比（Brevity Penalty），以接受一定的惩罚[17]。在 BLEU 的基础上，NIST 根据 N 元组在参考译文中出现的频率，对它们赋予不同的信息权重。频率越低，则信息量越大，权重越大[18]。BLEU 和 NIST 不仅方法简单，所评分数与人工评分也高度相关，可供英译汉机器评分系统借鉴。

第二，基于测试点的评价。其主要思想是：模拟标准化考试的方法，不评价整句，而是通过设置测试点简化测试目标。测试点分 6 组：词汇量测试、固定词组测试、词法测试，以及初、中、高级句法测试。研究者采用描述语言对各句的测试点进行句法描述，使评测可以全自动完成。程序评估机器译文中各个测试点的翻译质量，加权平均后获得最终的机器翻译评价结果[19]。由于翻译中有些语言点的区分度比较高，基于测试点的评价方法能够有效缩短评价时间，值得借鉴。

2 人工译文评价系统

除了机器翻译评价，人工译文自动评价研究也已起步，国内已建立起针对中国英语学习者的汉译英自动评分模型[5]。下面从六个方面进行介绍。

第一，语料来源。该研究使用国内英语专业三、四年级学生的 300 篇英译汉译文。原文为记叙文，包括 9 个句子，约 300 字。为满足研究需要，测试时既呈现整个篇章供学生整体理解，也提供单个句子让学生逐句翻译，单句译文合并即可获得篇章译文。

第二，模型设计。该系统按用途分为诊断性和选拔性评分模型，采用分模块设计。诊断性模型包括篇章和单句译文的语义内容、语言形式评分模块，通过提取各个模块对应的文本特征，可以分别构建它们的评分模型，并提供有针对性的诊断性信息。选拔性模型仅包括篇章译文的语义评分模块，可以对大规模测试中的汉译英进行评分。

第三，人工评分。该研究采用两次人工评分。第一次评分比较细致，以“忠实、通顺”为标准，分别对译文的语义内容和语言形式进行评价，结果分别用于构建诊断性语义、形式评分模型。语义评分时，先将原文各句划分为 2~3 个语义单位，逐个单位进行评价。形式评分以句为单位，衡量语

言的准确性和恰当性[5]。第二次评分比较简化，仅对具有较大区分度的语义点进行评价，结果用于构建选拔性评分模型。

第四，变量挖掘。该研究采用语料库工具、自然语言处理、信息检索技术和统计方法，挖掘了多个文本特征。其中，反映译文语义质量的变量有三类：N 元组匹配数量及其百分比、语义相似度和语义点对齐数量。语义点对齐技术考察译文对区分度较高语言点的翻译能否与正确译文表匹配，和俞士汶等[19]使用的针对测试点的评价方法有相似之处。该研究还提取了字词、句子、篇章三个层面的形式变量。表 2 对主要变量进行了总结。

表 2 汉译英自动评分研究中的主要变量

类别		变量	提取方法
语义		N 元组数量及其百分比	对照 25 篇最佳译文中的 N 元组提取
		SVD 值	学生译文与最佳译文集的语义相似度，通过潜语义分析获得
		语义点对齐数量	以 100 篇学生译文中的正确语义点译文为参照
形式	字词	形符、类符、类形符比	学生译文变量与最佳译文集相应变量均值的差值绝对值
		分级词汇形符、类符等	参照 Range 中的二、三级词表提取
		平均词长、词长标准差	以字母数量计
		各词性百分比	各词性数量/译文形符
	句子	句子数、平均句长等	以单词数量计，采用学生译文变量与最佳译文集相应变量均值的差值绝对值
	篇章	过渡词语	根据英语过渡词表提取

第五，模型构建。该研究首先在事先评分的一半译文（训练集）中计算所提取的变量与相应人工评分之间的相关度，采用与人工评分相关的变量作为译文质量预测因子；然后进行多元线性回归分析，选择性能最佳的模型，作为预测因子与人工评分之间的关系方程。数据表明，诊断性篇章译文语义、形式评分模型的决定系数 R^2 分别为 0.794、0.547。该研究进一步使用三种数量的训练集译文（30、50、100 篇）构建了选拔性评分模型，模型的相关系数 R 都在 0.8 以上。

第六，模型验证。该研究使用从训练集中获得的多元回归方程，计算另一半学生译文（验证集）的机器评分，然后分析机器与三名评分员平均评分的相关度和一致性。研究结果表明，诊断性篇章译文机器语义、形式评分与人工评分的相关度分别为 0.842**、0.741**。在选拔性模型中，机器与人工评分的相关度都在 0.8 以上。若需提高评分效率，以 100 篇译文构建的评分模型就能满足大规模测试评分的需要。

总之，该研究探索了诊断性与选拔性评分模型的区别，构建的模型能够准确、有效地评价中国学生的汉译英译文。不过，研究也存在一定的不足：（1）不同文体的原文及其译文在内容、语言、风格上都具有显著差异[20]，该研究使用记叙文译文构建模型，难以判定译文质量预测因子在其他文体中有效。（2）人工语义评分主要针对信息量较大的语义点进行，当学生未译或误译某处次级信息时，自动评分模型难以进行诊断性反馈。（3）采用保留样本法，训练集一直用于建模，验证集一直用于检验模型，结果在一定程度上受到译文分集的影响。

除了汉译英自动评分研究外，王立欣[21]对英译汉的自动评分进行了初步探讨。该研究的原文是一个广告段落，译文有 230 份，模型构建也经过变量提取、多元回归、计算新译文分数等步骤。研究采用 10 折交叉检验法，使用 9 成语料构建模型，1 成语料进行验证，经过 10 次循环计算的人机评分相关度均值为 0.75**。研究中使用的主要变量见表 3。

表 3 已有英译汉自动评分研究中的主要变量

变量	提取方法
句长	分别以词语数量、字符数量计
句长对齐概率	基于“译文句长应与原文句长对应”的假设
词对齐数量	基于词典的词对齐，运用模糊匹配
词性比例	原文各词性的词汇译为相同词性的比例
句子概率	基于词汇与上下文词汇的联系紧密度
句法分数	使用句法分析器，与 4 篇参考译文比较得出学生译文的句法分数
N 元组	BLEU 方法，采用 4 篇参考译文

该研究采用的一些变量值得借鉴，模型的验证方法比较科学。不过，研究也存在一些不足：（1）原文为广告文体，难以判断模型中的变量对其他文体的译文起作用。（2）未采用分模块设计，机器仅对译文质量进行整体评分，难以对语

义、形式质量及其分项特征进行有针对性的反馈。（3）采用机器翻译评价的惯例，仅使用 4 篇参考译文，而人工译文的多样性和复杂性都远远超过机器译文，这种做法对变量的有效性造成了一定影响。

除了王立欣外，Tian 等人[22]也考察了关键词匹配和语义相似度对英译汉译文语义质量的预测力。不过，该研究的语料为句子译文，容易忽略篇章层面的质量预测因子，并且研究挖掘的变量比较有限。

三 英译汉机器评分系统的新探索

笔者将借鉴上述研究的经验，构建适用于中国学生英译汉的机器评分系统。该系统与已有研究的区别在于以下几个方面：

首先，人工评分。（1）汉译英的目的语是学生的外语，译文达到“忠实、通顺”已属不易，因而汉译英自动评分研究的人工评分以“忠实、通顺”为标准[5]。而英译汉的目的语是学生的母语，译文在语言形式上往往比较通顺，需要采用更高的标准来衡量。本研究将在“忠实、通顺”的基础上加入“风格切合度”，对语言形式进行更高层次的评价。（2）本研究将原文划分为符合搭配规则、意义单一、完整的多词单元，即翻译单位[23][24]。评分员对每个翻译单位的译文逐个进行评分，能够更全面地衡量译文的语义质量，也便于机器对译文的语义优劣进行更细致的反馈。

其次，变量挖掘。（1）为拟合人工评分过程，研究者将根据翻译单位的最佳译文和正确译文列表，提取学生译文中的翻译单位对齐数量。由于翻译单位符合搭配规则，并具有单一和完整的意义，能够较好地评价译文的语法性、连贯性和地道性[24]。（2）由于英汉语言表达的差异和汉语分词的影响，一个英语词汇可能对应一个或多个汉语词语，也可能出现多对一、多对多的情况。同时，少数英语词汇的汉语翻译呈分离状态，如 as quickly as 的译文“像……一样快”，中间间隔一个或多个词语。此外，中国学生的英译汉译文中大量使用同义词和近义词，如 in radiant bloom 的翻译包括“盛开”、“开花”、“开放”、“绽放”、“怒放”等。针对以上特点，本研究的词对齐不仅考察英、汉语的一一对应，还将进行一对多、多对一、多对多的对齐，同时考虑英语词汇与汉语分离结构对应的情况，还将嵌入同义词词林，考察原文词汇与词典译文的同义词、近义词对应的情况。初步研究发现，这种词对齐的效果优于仅仅基于词典的词对齐技术[25]。

再次，文体类别。本研究将采用说明文、记叙文、议论文三种文体，分别使用 300 多篇学生译文构建机器评分模型。通过比较，进一步挖掘对三种文体的译文质量都具有预测力的文本特征，以便提高系统的迁移性。

最后，验证方法。本研究将对机器与人工评分差异较大的译文进行质性分析并究其原因，在此基础上提出改进变量、提高系统性能的方法，以便减少大规模测试中机器评分的偏差。

四 结语

本文对近五十年来自动评分系统的优缺点进行了回顾和评价，探讨了现有技术对英译汉机器评分系统的启示，以及其中可资借鉴的方面。在总结以往经验的基础上，研究者将针对中国学生英译汉译文的特点，在人工评分、变量挖掘、文体类别，以及对机器评分结果的分析方面进行新的探索，以便构建适用于中国学生英译汉的机器评分模型。

参考文献

[1]Dikli, S. An overview of automated scoring of essays[J]. Journal of Technology, Learning, and Assessment, 2006, (1): 3-35.

[2]Quellmalz, E. S. & Pellegrino, J. W. Technology and testing[J]. Science, 2009, (2): 75-79.

[3]Williamson, D. M. A framework for implementing automated scoring[R]. Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education. San Diego, CA, U.S.A., 2009,4:13-17.

[4]梁茂成.中国学生英语作文自动评分模型的构建[D].南京:南京大学,2005.

[5]王金铨.中国学习者汉译英机助评分模型的构建[D].北京:北京外国语大学,2008.

[6]王金铨,文秋芳.中国学生大规模汉译英测试机助评分模型的研究与构建[J].现代外语,2009,(4):415-420.

[7]Page, E. B. Project Essay Grade: PEG[A]. In Shermis, M. D. & Burstein,J.C. (eds.).Automated Essay Scoring: A Cross-Disciplinary Perspective[C]. NJ: Lawrence Erlbaum Associates, 2003: 43-54.

[8]Landauer, T. K., Laham, D. & Foltz, P. W. Automated essay scoring and annotation of essays with the Intelligent Essay Assessor[A]. In Shermis, M. D. & Burstein, J. C. (eds.). Automated Essay Scoring: A Cross-Disciplinary Perspective[C]. NJ: Lawrence Erlbaum Associates, 2003: 87-112.

[9]Burstein, J. The E-rater Scoring Engine: Automated essay scoring with natural language processing[A]. In Shermis, M. D. & Burstein, J. C. (eds.). Automated Essay Scoring: A Cross-Disciplinary Perspective[C]. NJ: Lawrence Erlbaum Associates, 2003: 113-121.

[10]Mislevy R. J. et al. Making sense of data from complex assessment[J]. Applied Measurement in Education, 2002, 15(4): 363-389.

[11]Xi, X. M, et al. Automated scoring of spontaneous speech using SpeechRater v1.0[OL].<http://www1.ets.org/Media/Research/pdf/RR-08-62.pdf.>

[12] Chung, G. K.W. K. & Baker, E. L. Issues in the reliability and validity of automated scoring of constructed responses[A]. In Shermis, M. D. & Burstein, J. C. (eds.). Automated Essay Scoring: A Cross-Disciplinary Perspective[C]. NJ: Lawrence Erlbaum Associates, 2003: 23-40.

[13]桂诗春.潜伏语义分析的理论及其应用[J].现代外语,2003, (1): 76-84.

[14]梁茂成,文秋芳.国外作文自动评分系统评述及启示[J]. 外语电化教学, 2007, (5): 18-24.

[15]Yang, Y. W. et al. A review of strategies for validating computer-automated scoring[J]. Applied Measurement in Education, 2002, (4): 391-412.

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

[17] Papineni, K. et al. Bleu: A method for automatic evaluation of machine translation[A]. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)[C], 2002: 311-318.

[18]NIST Report. Automated evaluation of MT quality using N-gram co-occurrence statistics[OL]. <www.itl.nist.gov/iad/894.01/tests/mt/doc/ngram-study.pdf.>

[19]俞士汶等.基于测试集与测试点的机译系统评估[A].载陈肇雄主编.机器翻译研究进展[C].北京:电子工业出版社,1992: 524-537.

[20]刘宓庆.文体与翻译（第二版）[M].北京:中国对外翻译出版公司, 2007: 91-95, 181-183.

[21]王立欣.翻译标准自动量化方法研究[D].博士论文,上海外国语大学,2007.

[22]Tian, Y., Lu, R. Z. & Wu, B. S. Towards on-line automated semantic scoring of English-Chinese translation[J]. Journal of Shanghai Jiaotong University (Science),2007, 12(6): 725-730.

[23]Teubert, W. The role of parallel corpora in translation and multilingual lexicography[A]. In Altenberg, B. & Granger, S. (eds.). Lexis in Contrast: Corpus-Based Approaches[C]. Amsterdam and Philadelphia: Benjamins, 2002: 189-214.

[24]江进林,文秋芳.N 元组和翻译单位对齐在学生英译汉自动评价中的比较研究[J].现代外语, 2010,(2): 177-184.

[25]文秋芳,秦颖,江进林.英语考试翻译自动评分中双语对齐技术的应用[J].外语电化教学, 2009, (1): 3-8.

Rethinking 50 Years of Studies on Automated Scoring

—Explorations of Computer Scoring System for English-Chinese Translations of Chinese Learners

JIANG Jin-lin

(Research Centre for Business English and Cross-Cultural Studies, University of International Business and Economics, Beijing 100029, China)

Abstract: In the past 50 years a number of automated English essay scoring systems have been developed at home and abroad, so research along this line has matured. However, previous attempts to automate the scoring process of translation focused on machine translation and research dealing with human translation is very rudimentary. In recent years scoring models for Chinese students' Chinese-English (C-E) translations have been constructed and a system handling English-Chinese (E-C) translations is also taking the initial step. Since the E-C translations of Chinese learners have unique features, the scoring system is different from previous studies in many aspects, such as variable mining and model validation.

Keywords: automated scoring; essay; translation; English-Chinese translations of students

---

*基金项目：本研究为国家社会科学基金项目“专用英汉互译机助评分系统的研制”（批准号：11CYY007）的部分成果，同时受到对外经济贸易大学优秀青年学者培育计划资助（批准号：2012YQ12）。

作者简介：江进林，博士，对外经济贸易大学讲师、商务英语与跨文化研究中心研究员。研究方向：机辅测试、语料库语言学。

收稿日期：2013年1月9日

编辑：李婷