信息技术与测试学研究

大规模测试中学生英译汉机器评分模型的构建

江进林[1]，文秋芳[2]
(1. 对外经济贸易大学 英语学院 北京，100029; 2. 北京外国语大学 中国外语教育研究中心 北京，100089)

摘 要：本文旨在研制有效、可靠的英译汉学生译文机器评分系统，实现大规模测试的自动评分。本研究针对三种文体的译文，分别构建了五种比例训练集的评分模型，模型预测分值与人工评分的相关系数均高于 0.8。并且，当训练集为 130 篇时，模型对说明文和记叙文译文的预测分值与人工评分非常接近；当训练集为 100 篇时，模型对叙议混合文译文的评分与人工评分最为接近。研究结果表明，本文提取的变量预测力较强，针对不同文体构建的评分模型效果良好，能够比较准确地预测学生的英译汉成绩。

关键词：大规模测试；英译汉；机器评分
中图分类号：H319.3 文献标识码：A 文章编号：1001-5795(2012)02-0003-0006

1 研究背景

主观题的自动评分是测试领域关注的一个焦点。目前英语作文的自动评分研究已经比较成熟，国外已开发出多个评分系统，并应用于 GRE、GMAT 等大型考试中(Dikli，2006；Quellmalz & Pellegrino，2009)。在国内，梁茂成(2005)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。极少数人也对汉语作文的自动评分进行了研究，发现通过潜语义分析获得的机器评分比较接近人工评分(曹亦薇等，2007：63－71)。

在翻译领域，自动评分研究主要局限于机器翻译评价(如 Papineni et al.，2002：311－318)。少数研究者也对学生译文的自动评分进行了尝试(王金铨，2008)。该研究构建了诊断性和选拔性评分模型，前者可以对译文的语义、形式质量进行细致评分并提出反馈，后者可以满足大规模测试中的评分需要。不过，该研究的文体仅限于记叙文。在英译汉自动评分方面，王立欣(2007)挖掘了词对齐数量等文本特征，采用 10 折交叉检验法来验证模型，具有一定的优势。不过，该研究所用语料为一个广告类段落，人工评分比较粗略，变量也基本上停留在词汇层面。

在前人研究的基础上，本文拟研制稳定可靠的、适用于中国学生大规模英译汉测试的机器评分模型。其基本方法是：利用语料库语言学、自然语言处理、信息检索等领域的技术，提取与译文质量相关的多种文本特征，并通过对文本特征和人工评分进行多元回归分析，构建机器评分模型，最后采用回归方程计算同一题目的其他译文的分数，并分析机器评分与人工评分的相似程度。本研究在人工评分、特征提取、语料类别方面与已有研究不同。首先，人工语义评分以原文的“翻译单位”为单元，翻译单位是符合搭配规则、意义单一、完整的多词单位，有利于评价译文的语义正误、语法性、连贯性等特点(参见江进林、文秋芳，2010a：177－184)。人工形式评分增加了“风格切合度”标准，因为英译汉的目的语是学生的母语，译文的语言形式需要采用更高的评价标准。其次，本研究提取了翻译单位对齐数量等一批新的文本特征。再次，本文对三种文体的译文分别建模。不同文体的语篇在内容、语言、风格上都具有显著差异。本研究使用的说明文结构清晰，措辞规范、严谨，句子结构复杂；记叙文运用了比喻、排比等修辞手法，语言流畅，情态丰富，抒情色彩浓厚；叙议混合文则兼有记叙文和议论文的特点。本研

作者简介：江进林：女，博士，讲师，对外经济贸易大学商务英语与跨文化研究基地研究员。研究方向：语言测试、语料库语言学。
文秋芳：女，博士，教授，博士生导师。研究方向：应用语言学。
收稿日期：2011-09-02

究探讨了对三种文体的译文质量都具有预测力的变量,有助于提高变量的推广性。

2 研究设计

2.1 研究问题

本研究拟解决以下问题:

(1) 三种文体内,不同训练集数量构建的选拔性评分模型有多大预测力? 所预测分数的信度如何?

(2) 多少训练集译文能够满足大规模测试中英译汉机器评分的需要?

(3) 三种文体的评分模型内,相同特征和不同特征有哪些? 为什么?

2.2 研究工具

本研究使用了大量文本分析和数据分析工具:

(1) 文本预处理工具,主要为自编的 perl 程序,用于对文本中的不规则输入进行整理,并对文本进行随机编号、句子整合。

(2) 文本分析工具,用于提取与译文语义质量相关的文本特征,包括 R 软件和 perl 程序。R 是一款统计分析软件,本研究使用自编的 R 程序进行潜语义分析,它通过奇异值分解来压缩词语 - 文本矩阵,构建潜藏的文本语义空间(桂诗春, 2003: 76 - 84)。研究者还使用 perl 程序提取一到四元组匹配数量、评分点对齐数量等文本特征。这些特征的参照对象是最佳译文集,包括 30 篇专家译文和优秀学生译文,待测译文与该集合越接近,译文质量越高。

(3) 数据分析工具,主要是 SPSS,用于计算文本特征与分数的相关度,利用回归分析构建评分模型,以及验证模型的有效性。

2.3 研究步骤

本研究可分为五个阶段: 语料收集、人工评分、特征提取、模型构建、模型验证,前三个是建模前的准备阶段,下面分别进行介绍。

2.3.1 语料收集

本研究使用了三组语料,包括一篇说明文、记叙文、叙议混合文的汉语译文各 300 多篇,是国内三所不同水平高校英语专业三、四年级学生的限时翻译测试译文(60 分钟)。三篇原文各包含约 300 个词,按照句意可分为 15、15、13 个句子。在收集语料时,首先呈现语篇,便于学生从整体上把握原文; 接着呈现单句,要求学生在各单句下面写出译文,便于整理。

2.3.2 人工评分

在自动评分研究中,高信度的人工评分是保证机器评分有效、可靠的前提。本研究组织三名有经验的评分员先后进行细致型和简化型评分。细致型评分以"信、达、切"为标准,从语义和形式两个方面分别对译文进行评判。语义评分主要考察"信",评分员以"翻译单位"为单元,判断每个翻译单位译文的忠实度; 形式评分主要衡量"达"和"切",评分员以句为单位,评价每句译文的语法性、地道性和风格切合度。评分共持续约 240 小时。

由于第一次评分费时费力,不适应大规模考试的效率要求。评分结束一年后,本研究进一步采用简化型评分,仅对有区分度的评分点进行语义评价。评分点由国内两位翻译研究专家确定,三篇原文中各有 33、35、28 个评分点,分别占形符数的 1/7、1/8、1/9 左右。这次评分约耗时 32 个小时。

表 1 显示,在三组语料的细致型评分过程中,三名评分员对篇章译文语义评分的相关系数均值都在0.89 以上,alpha 系数在 0.95 以上; 形式评分的相关系数均值在 0.85 以上,alpha 系数在 0.94 以上,可见三名评分员具有良好的一致性。在简化型评分过程中,评分员的相关系数和 alpha 系数也令人满意。

由于第一次评分对译文的语义进行了穷尽性评价,第二次评分大大简化,其有效性取决于它与第一次语义评分的相似程度。统计显示,三组语料中两次语义平均评分之间的相关度分别达到 0.924、0.932 和 0.963,可见基于评分点的评分方法效果良好,也说明以往对评分法的二维划分,即整体评分法(holistic scoring)和分析评分法(analytic scoring)过于简单。整体评分法只需要评出一个整体印象分数,而分析评分法需要对目标技能的不同组成部分单独评分。已有研究指出,分析评分法的信度高于整体评分法,但是费时、花费高(Weigle, 2002: 121)。不过,分析评分法可能有不同的"度",比如本研究中的细致型和简化型评分; 其中,简化型分析评分法不仅信度可靠,还具有较

表 1 篇章译文评分信度

	第一次评分						第二次评分		
	语料 1 语义	语料 1 形式	语料 2 语义	语料 2 形式	语料 3 语义	语料 3 形式	语料 1	语料 2	语料 3
相关均值	0.891**	0.857**	0.956**	0.870**	0.973**	0.909**	0.944**	0.939**	0.956**
alpha	0.957	0.946	0.985	0.948	0.986	0.951	0.980	0.979	0.978

高的评分效率，这与前人的结论不同，也为大规模翻译测试中基于评分点的分析评分法提供了有力的效度证据。

笔者进一步运用多面 Rasch 模型对人工评分进行了分析（江进林、文秋芳，2010b：14－18）。结果显示，各评分员的评分没有出现趋中性；不过，三名评分员的严厉度具有显著差异，这是考试中不应出现的情况。本研究权且采用三名评分员的平均分，降低了评分员差异对评分结果的影响。

2.3.3 特征提取

本研究提取了 N 元组匹配数量、词对齐数量等语义特征。①N 元组匹配数量以最佳译文集为参照，分别检索最佳一到四元组在学生译文中出现的频率。N 元组是以词为单位的线性序列，对译文内容进行了最大限度的利用。不过，它不一定是完整的语义单位，没有充分考虑语境因素。②词对齐数量以英汉词典为基准，利用同义词词林扩展版对词典译文进行补充，并考虑了一对多、多对一、多对多等匹配情况，对学生译文中译对的词语进行统计。该变量可以衡量译文的漏译、误译等情况（文秋芳等，2009：3－8）。③评分点对齐数量模拟大型考试阅卷中按采分点给分的方法，将评分点的专家译文和其他正确译文制成词典，在学生译文中进行匹配，对译文质量的区分性较强。④语义相似度的计算采用潜语义分析法，衡量学生译文与最佳译文集的近似程度。这些变量各有所长，与译文分数显著相关的变量将作为质量预测因子，进入模型构建环节。

3 结果与讨论

本研究的建模方法是，以简化型人工评分为因变量、与该分数显著相关的文本特征为自变量，进行多元线性回归分析。模型的确立需要反复尝试、不断修正，评价模型优劣的标准有三个：第一，进入模型的自变量间相关系数不超过 0.8，以免出现共线性（collinearity）。共线性指回归方程中两个或多个自变量高度相关，或者一个自变量解释的方差基本上可以由其他多个自变量解释（Ryan，2009）。第二，模型的决定系数 R2 达到最大、共线性数据最合理。考察共线性的统计标准主要有容忍度、方差膨胀因子和条件指数（秦晓晴，2003）。第三，自变量的系数正负性与它和因变量的相关情况同向。如果不同向，该变量为“负抑制变量”（negative suppressor），往往与共线性问题联系在一起（Ryan，2009）。

表 2 所列是经过反复优化的模型，共线性数据都在可接受的范围内，自变量的系数也与它和因变量的相关性一致。限于篇幅，这些数据暂不呈现。

表 2 显示，在三组语料中，五种训练集所构建模型的相关系数都在 0.8 以上，表明模型中的变量能够较好地解释译文的成绩。在说明文语料中，训练集为 50 篇译文时，模型的相关系数最高；随着训练集文本逐渐增加，模型的相关系数整体上呈下降趋势。不过，训练集译文越少，模型受具体译文的影响越大，越不稳定，因而不能断定 50 篇译文能够满足大规模评分的需要。在记叙文语料中，训练集为 100 篇时，模型的相关系数最高；训练集减少或增加时，模型的相关系数变化很小。在叙议混合文语料中，训练集为 50 篇时，模型的相关系数达到 0.965；训练集增加至 100 篇时，相关系数降至 0.935；训练集进一步增加时，模型的相关系数差异很小。由此可见，仅根据拟合数据难以确定最佳的训练集译文数量，需要比较模型的评分效果。

本研究将验证集中的语义变量代入相应训练集所构建的回归方程，获得验证集译文的机器评分。然后，计算机器评分与人工评分的相关度和 alpha 系数，结果见表 3。

表 3 显示，在三组语料中，不同模型的预测分数与人工评分的相关系数和 alpha 值都在 0.8 以上，表明模型都能有效预测验证集译文的成绩。在说明文语料中，训练集为 50～150 篇译文时，验证集人机评分的相关度逐渐上升；训练集为 150 篇时，相关度达到0.862。结合表 2 可以发现，模型解释的训练集分数方差与模型在验证集中的表现并不同步。训练集越少，模型对

表 2 五种训练集的评分模型

训练集	语料 1			语料 2			语料 3		
	R	R2	Adjusted R2	R	R2	Adjusted R2	R	R2	Adjusted R2
50 篇	0.908	0.824	0.816	0.908	0.824	0.812	0.965	0.931	0.926
100 篇	0.852	0.726	0.717	0.912	0.831	0.826	0.935	0.875	0.871
130 篇	0.835	0.698	0.690	0.909	0.827	0.823	0.940	0.884	0.881
150 篇	0.834	0.696	0.690	0.903	0.815	0.811	0.936	0.877	0.873
180 篇	0.848	0.720	0.715	0.894	0.799	0.796	0.940	0.884	0.882

表3　人机评分的相关性和 alpha 系数

训练集	语料1		语料2		语料3	
	相关性	alpha	相关性	alpha	相关性	alpha
50 篇	0.832**	0.909	0.893**	0.943	0.910**	0.952
100 篇	0.837**	0.908	0.885**	0.935	0.923**	0.959
130 篇	0.860**	0.917	0.883**	0.933	0.916**	0.955
150 篇	0.862**	0.919	0.888**	0.935	0.923**	0.959
180 篇	0.851**	0.910	0.895**	0.941	0.942**	0.967

训练集分数的预测作用越强，而稳定性也越差，对验证集分数的预测有效性越低。可见，训练集需要达到一定的数量，才能保证模型的有效性。在记叙文语料中，训练集为 180 篇时，验证集人机评分的相关度最高，达到 0.895。不过，五种模型中人机评分的相关度差异很小。同样，在叙议混合文语料中，训练集为 180 篇时，验证集人机评分的相关度高达 0.942。

上述结果优于已有的口语自动评分研究（人机评分的相关度在 0.5～0.7 之间）（Xi et al.，2008；Chen & Zechner，2011），但略低于汉译英自动评分模型（王金铨，2008）。在王金铨的研究中，训练集为 50、100、150 篇时，验证集人机评分的相关系数分别为 0.870、0.878 和 0.897，比表 3 的结果约高出 0.03。由于本研究的对象为汉语译文，而汉语为意合语言，自动评分取得如此结果已属不易。与已有的英译汉自动评分研究相比（人机评分的相关度为 0.75）（王立欣，2007），本评分模型的效果更好。

本研究进一步采用配对样本 t 检验考察了人机评分的差异性，结果如表 4：在说明文语料中，训练集为 50 和 100 篇译文时，验证集人机评分的差异均值分别为 1.085 和 1.288，且具有显著意义。当训练集增至 130 篇时，人机评分的差异均值降至 0.471，没有统计意义。训练集继续增加时，人机评分的差异均值进一步降低，统计意义更不显著。因此，130 篇训练集译文基本满足机器评分的需要。

在记叙文语料中，各种模型的人机评分差异均值都在 0.1 左右，都没有显著意义。其中，训练集为 130 篇时，人机评分的差异均值最小（－0.051）。

在叙议混合文语料中，训练集为 50、130 和 150 篇译文时，验证集人机评分的差异都具有显著意义。训练集为 100 篇时，差异均值降至 0.481，没有显著意义。训练集增至 180 篇时，差异均值最小（0.165），此时人机评分最为接近。不过，考虑成本因素，100 篇训练集译文已能满足评分需要。

表4　人机评分的配对样本 t 检验

	Model	Paired Differences		t	Sig.（2-tailed）
		Mean	Std. Deviation		
语料1	机器-人工评分（50 篇）	1.085	4.293	4.076	0.000
	机器-人工评分（100 篇）	1.288	4.045	4.614	0.000
	机器-人工评分（130 篇）	0.471	3.873	1.631	0.105
	机器-人工评分（150 篇）	0.434	3.886	1.412	0.160
	机器-人工评分（180 篇）	0.184	4.077	0.514	0.608
语料2	机器-人工评分（50 篇）	0.168	3.953	0.699	0.485
	机器-人工评分（100 篇）	0.150	3.984	0.560	0.576
	机器-人工评分（130 篇）	－0.051	4.132	－0.170	0.865
	机器-人工评分（150 篇）	－0.093	4.106	－0.295	0.769
	机器-人工评分（180 篇）	0.053	4.090	0.153	0.879
语料3	机器-人工评分（50 篇）	0.615	4.057	2.395	0.017
	机器-人工评分（100 篇）	0.481	3.896	1.745	0.082
	机器-人工评分（130 篇）	0.815	3.876	2.740	0.007
	机器-人工评分（150 篇）	－0.747	3.661	－2.499	0.014
	机器-人工评分（180 篇）	0.165	3.098	0.583	0.561

综上所述，130 篇训练集译文基本满足对 180 篇说明文译文和 190 篇记叙文译文进行机器评分的需要；100 篇训练集译文即可满足 200 篇叙议混合文译文的机器评分需要。后者需要的训练集更少，可能因为原文的区分度更合理，且人工评分的信度最高（见表 1）。不过这一结论需要在大规模语料中进一步检验。最终确定的评分模型如表 5。

表5　最佳评分模型

语料1	译文成绩 = －10.988 + 0.983 × 评分点对齐数量 + 0.098 × 一元组匹配数量 + 24.163 × 语义相似度
语料2	译文成绩 = 3.325 + 1.097 × 评分点对齐数量 + 0.098 × 一元组匹配数量 + 0.053 × 词对齐数量
语料3	译文成绩 = －2.43 + 1.094 × 评分点对齐数量 + 0.065 × 一元组匹配数量 + 0.049 × 二元组匹配数量

表 5 显示，在三组语料中，评分点对齐数量和一元组匹配数量始终是译文成绩的有效预测因子。并且，评分点对齐数量在三个方程中的标准化系数都最大（分别为 0.549、0.679、0.564，限于篇幅，具体数据暂未呈现），对译文语义质量的预测力最强。评分点的内核是翻译单位的简化。由于译者通常从意义出发，将多个词组成的片段作为整体来考虑（Teubert，2002：189－214），翻译单位作为符合句法规则、具有完整意义的多词单位，能够拟合该操作过程，并且较好地考虑了语序、上下文等因素，因而能够比较全面地衡量译文质量。不过，对齐翻译单位所依据的译文词典需要专

业人员的大量工作才能制定，而评分点是具有区分度的词或短语，语言单位较小，数量也远远少于翻译单位，人工介入大大减少。表 5 表明，评分点对齐数量对英译汉具有明显的预测效果。

数据进一步显示，一元组匹配数量在三个方程中的标准化系数分别为 0.279、0.271 和 0.265（限于篇幅，具体数据暂未呈现），贡献仅次于评分点对齐数量。二元组匹配数量也在叙议混合文译文的评分方程内出现，标准化系数为 0.215。该结果证实了 N 元组匹配法对译文质量的预测作用。Papineni 等人（2002：313）指出，与参考译文一元组相同的待测译文往往符合忠实度标准；二元和多元组则能够衡量译文的流利度。N 元组匹配法简单易行，且具有区分译文的明显效果。由于一、二元组匹配数量考察学生译文中一两个字或词与参考译文相同的情况，与具有较大区分度的评分点对齐数量互为补充，在本研究中取得了良好的效果。

语义相似度在说明文译文的评分方程内出现，标准化系数为 0.146，对译文质量也产生了一定的预测力。该指标是通过潜语义分析获得的、学生译文与最佳译文集的相似度。潜语义分析能够有效消除文本噪音，通过降维揭示文本内潜藏的语义空间，在已有的作文自动评分系统（梁茂成，2005）和汉译英评分系统（王金铨，2008）中都起着重要作用。本研究的评价对象为汉语文本，语义相似度的作用仍然比较明显。

词对齐数量在记叙文译文的评分方程内出现。已有的英译汉自动评分研究采用了基于词典的词对齐，词对齐数量与译文成绩的相关系数约为 0.6**（王立欣，2007）。由于缺乏大量高质量的双语对齐语料，本研究也采用了基于词典的词对齐。不过，本文的词对齐更全面地考虑了多种英、汉语对齐情况，最终取得了较好的预测效果。

通过比较可以发现，三种文体译文评分方程内的变量并不完全相同，说明文译文的评分模型内出现了语义相似度，记叙文译文的评分方程内有词对齐数量，而叙议混合文译文的评分模型内有二元组匹配数量。研究者对三种文体的原文和译文进行了细致深入的比较，暂未发现系统性的原因。笔者的初步结论是，本研究采用的变量对译文质量都具有较强的预测力，但由于不同文体、题目的译文在内容和语言上存在差异，这些变量可能会产生不同的交互效果。该结论需要进一步探讨。

4 结论

本研究利用多个领域的知识，创建了适用于大规模测试的中国学生英译汉机器评分模型。研究结果显示，说明文、记叙文、叙议混合文译文所构建的评分模型都表现良好。从人工评分的效率上看，简化型人工评分节省了约五分之四的评分时间，且与细致型人工评分的相关度和一致性很高，表明以评分点为评判依据的方法有效、可行。从评分模型的效果上看，以 50、100、130、150、180 篇训练集译文构建的评分模型都能较好地预测译文成绩，其中，说明文和记叙文译文中 130 篇训练集、叙议混合文译文中 100 篇训练集所构建模型的预测分数与人工评分非常接近，选择此类数量的训练集不仅能够节约成本，还能满足大规模测试的自动评分需要。

不过，本研究也具有一定的不足之处。首先，需要使用大规模语料，检验 130 和 100 篇训练集数量能否在其他文体、题目、数量的译文中产生同样效果。其次，某些特征并不完美。例如，对齐评分点所依据的词典并未穷尽正确译文，区分度较高和较低的评分点也没有进行区别。再次，自动评分模型难以评判少数创造性译文。本研究在这方面做出了一定努力，提取变量时基本上以 30 篇最佳译文为参照，不过这些译文无法涵盖所有创造性翻译，因此，在人机评分差异较大时，不可避免地需要人工进行干预。另外，文体与模型之间的关系需要进一步探讨。 □

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。感谢梁茂成教授、秦颖博士的技术指导。

参 考 文 献

[1] Chen，M，& Zechner，K. Computing and evaluating syntactic complexity features for automated scoring of spontaneous non－native speech [A]. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics [C]，2011.

[2] Dikli，S. An overview of automated scoring of essays [J]. *Journal of Technology，Learning，and Assessment*，2006 (1).

[3] 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.

[4] Quellmalz，E. S. & Pellegrino，J. W. Technology and tes-

ting [J]. *Science*, 2009(2).

[5] Ryan, T. P. *Modern Regression Methods* [M]. New York: John Wiley and Sons, 2009.

[6] 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.

[7] Weigle, S. C. *Assessing Writing* [M]. Cambridge: Cambridge University Press, 2002.

[8] Xi, X. M, et al. Automated scoring of spontaneous speech using SpeechRater v1.0 [R]. Retrieved August 18, 2011 from http://www1.ets.org/Media/Research/pdf/RR-08-62.pdf. 2008.

[9] 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).

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

[11] 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，2010a(2).

[12] 江进林、文秋芳. 基于Rasch模型的翻译测试效度研究 [J]. 外语电化教学，2010b(1).

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

[14] 秦晓晴. 外语教学研究中的定量数据分析 [M]. 武汉：华中科技大学出版社，2003.

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

[16] 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.

[17] 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).

Computer Scoring Models for EFL Learners' English-Chinese Translation in Large-Scale Tests

JIANG Jin-lin[1], *WEN Qiu-fang*[2]

(1. School of International Studies, University of International Business and Economics, Beijing 100029;
2. National Research Center for Foreign Language Education, Beijing Foreign Studies University, Beijing 100089)

Abstract: This study constructed computer scoring models for Chinese EFL learners' E-C translation. The models proposed in this paper, once implemented in the form of a computer program, can score E-C translation in large-scale tests. This study built five tentative scoring models with different sizes of training sets for three text types respectively, and the correlation coefficients between their computed scoring and human scoring are above 0.8. The results further indicated that computed scoring with 130 training texts in expository and narrative subsets, and 100 training texts in narration-argumentation-mixed subset were very close to human scoring. It is therefore concluded that the variables extracted in this research have high predicting power, and the constructed models can produce reliable scores for Chinese EFL learners' E-C translation of three text types.

Key words: Large-scale Test; English-Chinese Translation; Computer Scoring