中国学生大规模汉译英测试机助评分模型的研究与构建*

扬州大学 王金铨 北京外国语大学 文秋芳

提要:以往的翻译自动评价系统局限于机器译文评判,缺乏人工译文评价系统。本研究旨在研制信度可靠、运行稳定的汉英人工译文自动评分系统,实现大规模汉译英测试的评分自动化。本研究构建了三种比例训练集的大规模测试评分模型,模型预测分值与人工评分的相关系数均高于0.85,当训练集达到100篇时,模型预测分值与人工评分基本一致,不存在显著性差异。研究结果表明,本研究提取的变量预测能力较强,机助测试评分模型表现良好,能够比较准确地预测中国二语学习者的汉译英成绩。

关键词:汉译英、大规模英语测试、机助评分模型
[中图分类号] H319 [文献标识码] A [文章编号] 1003-6105(2009)04-0415-06

1. 引言

在大规模外语考试中,主观题是测试外语学习者语言能力的一个重要手段。为了能够真实反映学习者的外语水平,各级各类英语水平测试中都包含了主观题,主观题自动评分成为研究者的关注焦点。早在二十世纪六十年代,国外研究者(Page 1968)开始尝试研制英语写作自动评分系统。到了九十年代,BETSY,IEA,IntelliMetric 等写作自动评分系统相继推出,并取得了很好的效果(Shermis & Burstein 2003)。美国教育考试处(ETS)研制的 E-rater 目前已经完全商业化,被正式用于 GMAT 等大规模语言测试(Dodigovic 2005:104)。在国内,梁茂成(2005)对中国学生英语作文自动评分进行了有益的尝试,并取得了初步成果。

但是写作自动评分系统并不适用于对翻译文本的自动评分。其主要原因有三:(1)作文可长可短,而译文长度受到原文制约,一般存在一个阈值。作文可以在大主题的范围内对文本内容进行任意选择,且不影响文章的总体质量,而在翻译中,如果出现内容缺失则定会影响译文质量;(2)翻译比作文更注重对意义的把握,意义的比重应该超过形式(Nida 1982:11),甚至认为翻译即译意;(3)写作评分侧重于总体评分,即对整个篇章进行评分,而翻译测试评分既可能针对整个篇章,也可能针对单句,因而要求翻译自动评分系统既能够满足篇章评分的需要,也能够实现对单句译文进行评分的需求。因此,研制独立的人工译文自动评分系统非常必要。

在翻译评价领域,目前的研究主要集中于对机器译文的评价上(Papineni & Roukos 2002;Banerjee & Lavie 2005;Hovy *et al.* 2002;Turian *et al.* 2003),而对于汉译英人工译文自动评价的研究,国内外直接相关的研究在现有文献中未有报告。

本研究将综合英语写作自动评分和机器翻译自动评价两个领域的知识,尝试构建中国学生大规模汉英翻译自动评分系统。本研究将利用语料库语言学、信息检索、统计学以及自然语言处理领域中的相关知识,通过提取学生译作中与译文质量相关的多种文本特征,进行多元回归分析,构建有效的大规模测

* 感谢《现代外语》匿名审稿专家所提出的宝贵修改意见。本研究为教育部人文社会科学重点研究基地研究项目"大规模考试主观题(英汉互译)自动评分系统的研制"(项目编号 07JJD740070)的部分成果。

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

试评分模型,应用于大规模翻译测试评分。

本研究大致可以分为四个阶段:数据收集→人工评分→模型构建→模型验证。数据收集和人工评分属于建模前的准备阶段。在模型构建阶段,本研究将学生译文按照不同比例分为训练集和验证集,基于对训练集样本的深度分析,提取与译文质量相关的预测变量,以人工评分作为因变量,文本特征作为自变量进行多元回归分析,反复尝试,构建对学生汉英译作质量具有较强预测能力的统计模型。在模型验证阶段,利用所构建的统计模型,计算验证集其它学生译作的得分,然后将机器预测得分与人工评分进行统计分析,构建预测力最佳的评分模型。

在本研究中,由于机器评分的所有依据都来自对人工评分结果的学习,人工评分的作用非常重要,直接决定了机器评分的有效性。为了构建汉译英大规模测试评分模型,本研究进行了两次评分尝试。第一次评分用于提供非常详细的语义和形式评分结果,作为第二次评分的数据支撑,第二次评分的有效性完全取决于与第一次评分结果之间的信度分析;第二次评分仅对译文中有区分度的语义点进行评分,其结果将用于大规模测试评分模型的构建,本研究的评分模型均在第二次评分结果基础上构建。

2. 研究方法

2.1 研究问题

问题一:可提取的变量能否预测汉译英的质量?预测力有多大?

问题二:大规模测试评分模型中,由译文质量预测因子构建的模型的预测能力如何?

问题三:哪种模型更符合汉译英大规模测试自动评分模型的需要?

2.2 语料

本研究使用的翻译语料来自 PACCEL(中国大学生英汉汉英口笔译语料库)(文秋芳、王金铨 2008),译者为国内三所不同水平层次的大学英语专业三、四年级学生,共计 300 篇限时 60 分钟的汉英笔译材料,翻译文本为 340 字左右的叙事文。在语料收集前,按照句意,把整个语篇分为 9 个汉译英句子。语料收集时,首先呈现给学生的是语篇,接着是单句。语篇便于学生整体把握翻译材料,单句则便于他们书写译文,也利于以后的语料整理。该语料共包含 68179 个形符,2542 个类符。

2.3 人工评分标准体系

本研究建立了两套评分标准体系(见王金铨、文秋芳 2009),进行了两次评分尝试。第一套评分标准体系借鉴了专业英语八级口试的评分模式,采用语义内容评判和语言形式评判相结合的方法。语义内容通过语义单位中的语义点进行评分。评判语义内容时,研究者邀请了在语言测试和翻译教学方面有着较高水平的专家对句子的语义内容进行了划分,提取了一些与句意密切相关的语义点,作为对译文语义内容的评分依据。

第二套评分标准体系是为了适应大规模考试评分的需要,按照"点面结合,突出重点"的原则由专家对译文中的语义点进行选择,只挑选能够区分译文优劣的语义点作为评分目标,选取时既注意译文中语义点的覆盖程度,也注重语义点的区分度。经过筛选,第二次评分使用的语义点共计 17 个,比第一次评分时采用的语义点减少将近三分之二。

第一套评分标准复杂细致,第二套标准简洁明了。第一套评分标准是基础,是前提,缺乏第一套评分标准的数据支撑,第二套评分标准就毫无意义。

2.4 人工评分过程

构建自动评分系统的关键是高信度的人工评分。为了取得高信度的评分,首要工作是评分员选择。本研究邀请了三位具有多年英语教学和研究经验的在读博士研究生作为评分员,他们都参加过国内不同等级的大型考试阅卷工作,本研究两次评分的评分员相同。

评分员培训是评分工作前的关键步骤。

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

通过培训，评分员可以更客观地把握评分尺度，减少评分过程中可能出现的过于宽松或过于严格的现象，做到用一把尺子去衡量不同译文的优劣。在前期语义内容量化以及语义单位切分的时候，所有评分员就加入到评分工作之中，经过充分讨论最终形成了上述语义和形式的评分标准。由于本研究是按句给分，评分时间较长，一次培训远远不够。在实际操作过程中，每次评改新句前都要进行培训讨论，以决定每句的评分标准和评改要求。

本研究的第一次评分始于 2006 年 11 月，评分过程是按句（共九句）评改，每次只评一句，每句的译作数量为 300，共计 2700 句，先评语义内容，后评语言形式。每次评改前，研究者将详细的评分标准和学生译作（以句为单位）提交给评分员。在评分过程中，评分员对句子得分做独立评判，不做相互讨论。由于分析性评分比较耗时，整个评分过程持续约 50 个小时。第一次评分耗时耗力，不符合大规模测试评分的效率要求，研究团队进行了第二次评分，仅以具有区分度的语义点为评分依据评改译文的语义内容，耗时约 10 小时，为第一次的五分之一。提交给评分员评改的 300 篇译文包括了训练集和验证集，评分工作的一次性完成保证了评分信度。评改后，所有得分被输入 SPSS 中计算分数和进行信度分析。

第一次评分结果显示，评分员之间在单句形式评分结果的相关系数均在.80 以上，且都具有统计学的显著意义。在形式评分方面，评分员之间的相关系数要高于语义评分，这也说明了语义评分相对于形式评分更为复杂，判断也更为困难。表 1 显示，评分员的语义、形式总体评分之间的相关系数都非常高，数值都在 .95 以上，且都具有统计学的显著意义，评分员间的内部一致性也非常好，alpha 值均在 .98 以上。

表 1　评分员间语义总分的相关系数及 alpha 系数

	评分员一/评分员二	评分员二/评分员三	评分员一/评分员三	评分员间 alpha 系数
语义总分	.959**	.966**	.951**	.986
形式总分	.988**	.977**	.978**	.993

** 相关性均在 .01 水平（双侧）上有显著意义。

评分员间的均值和标准差都比较接近，反映了评分员之间良好的一致性。

第二次评分是以有区分度的语义点为依据进行评分，评分员间的相关系数和 alpha 系数如下：

表 2 显示，评分员间的相关系数都非常

表 2　第二次评分评分员间相关系数及 alpha 系数

	评分员一/评分员二	评分员二/评分员三	评分员一/评分员三	评分员间 alpha 系数
语义总分	.917**	.944**	.928**	.975

** 相关性均在.01 水平（双侧）上有显著意义。

高，数值均在.9 以上，且都具有统计学的显著意义，评分员间的内部一致性也非常好，alpha 值在.97 以上。在本研究中，第一次评分对译文的语义、形式进行了非常细致的综合评分，第二次评分仅以具有区分度的语义点作为评分依据，因此第二次评分的有效性取决于与第一次评分结果之间的相关程度。两次评分结果比较如下：

从表 4 可见，评分员两次评分结果之间的相关系数均在 .85 以上，且都具有统计意义。两次语义评分之间的相关系数达到了.906 **，充分表明了第二次评分的有效性，也表明以具有区分度的语义点作为评分依据不仅省时省力，而且信度可靠。

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表 3 第二次评分的均值和标准差

	N	Mean	Std. Deviation
评分员一	300	42.5017	6.56477
评分员二	300	41.5923	6.72527
评分员三	300	45.0033	7.07828

表 4 评分员两次评分之间的相关系数

	评分员一	评分员二	评分员三
评分员一	.867**		
评分员二		.889**	
评分员三			.904**

**相关性均在 .01 水平(双侧)上有显著意义。

2.5 文本处理工具及统计方法

本研究使用了大量与文本处理和数据统计分析相关的工具,大致分类如下:(1)文本预处理工具,主要为自行编写的 Perl 程序,用于将译文语篇分割成句、字符清理、格式统一、矩阵生成等。(2)文本分析工具,主要用来对译文文本进行分析并提取与译文质量相关的文本特征项,包括形式特征和语义特征,提取形式特征时使用的工具包括 WordSmith Tools 4.0、Range32、Claws4 词性赋码工具以及一些字表;提取语义相关特征时使用的工具有 R 统计软件和自行编写的 Perl 程序。R 是用于统计分析的自由软件,能够为数据分析,特别是矩阵分析提供快速高效的运行效果。不过由于 R 是面向对象的统计编程语言,使用时需要编写合适的命令程序才能使 R 正常运行(关于 R 的详细介绍,请参阅 http://www.r-project.org/)。在本研究中,R 软件主要用来进行潜在语义分析的奇异值分解(Singular Value Decomposition, SVD)计算(详见王金铨 2007)。本研究在提取语义特征时还采用了最佳译文集,该集合包括了专家译文和学生优秀译文,通过这个集合来衡量其他待测译文,数据越接近最佳译文表明该译文质量越高。(3)数据分析工具,主要为 SPSS 统计软件,主要用途有三:a. 确定提取的文本特征项是否与译文质量相关以及相关程度;b. 构建自动评分模型;c. 用于验证评分模型的有效性。

3. 结果与讨论

本研究构建了大规模测试评分模型,限于篇幅,文中所列模型都是经过反复优化的模型。建模时,我们遵循了三个原则:(1)决定系数 R^2 和相关系数 R 达到最高,共线性数值最低;(2)β 值与相关系数同向,且 t 检验结果显示都具有统计意义;(3)进入模型的变量间相关系数不得大于 0.8。这些建模原则保证了模型的各项指标都建立在稳定性的基础上,并以此为出发点去追求模型的高信度。

大规模汉译英测试评分模型包含八个语义预测因子,其中语义点变量由第二次人工评分时依据的有区分度的语义点组成,数量减少,但更具区分性。本研究以第二次评分结果为依据构建了三种比例训练集的大规模测试评分模型,分别为 50 篇训练集、100 篇训练集和 150 篇训练集,模型数据如下:

表 5 显示三种模型的拟合情况都比较好,相关系数 R 均在.85 以上,决定系数 R^2 均大于 .73,模型的回归效果显著。表中的 VIF 值(variance inflation factor,方差膨胀系数)是构建模型时所有变量的 VIF 均值,主要用来观测模型是否存在共线性问题。“当回归方程中一个或多个自变量与另一个自变量或自变量的线性组合相关过高则会出现共线性问题”(Poirier 1995:567),共线性问题会“损害一个自变量的统计意义,使得标准误升高,变量估计出错,进而降低模型的稳定性和预测力”(Kidwell & Brown 1982,转引自 Walker 2003: 127)。在文献中,处理共线性数据有一些可供参考的原则:(1)如果最大的方差膨胀系数(VIF)大于 10,则需关注共线性问题(Myers 1990;Bowerman & O'Connell 1990);(2)如果方差膨胀系数(VIF)均值远远大于 1,则回归方程有共线性问题(Bowerman & O'Connell 1990)。

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 5　三种比例大规模测试评分模型

训练集	R	R Square	Adjusted R Square	Std Error of the Estimate	VIF
50 篇	0.922	0.850	0.843	5.434057	2.046
100 篇	0.858	0.737	0.731	5.492171	1.857
150 篇	0.873	0.762	0.758	5.109109	1.564

表 5 中的模型都是经过优化的最佳模型，三种模型的 VIF 值均处于正常范围，不存在共线性问题，50 篇训练集模型的共线性数值略高于其他两个模型。表 6 列出了三种比例模型的机器评分与人工评分之间的相关系数：

表 6　人工评分与机器评分之间的相关系数

	机器评分（50 篇）	机器评分（100 篇）	机器评分（150 篇）
人工语义评分（第二次）	.870**	.878**	.897**
人工语义评分（第一次）	.795**	.858**	.819**
人工评分总分（六四比例[1]）	.767**	.843**	.806**

**. 相关性均在 .01 水平（双侧）上有显著意义。

表 6 中，三种比例的模型能够较好地预测验证集译文的成绩，150 篇训练集模型的预测分值与人工语义评分（第二次）的相关最高，100 篇模型的预测分值与人工语义评分（第一次）以及总分（六四比例）的相关系数最高，训练集达到 50 篇后，模型性能提高有限。结合配对样本 t 检验的结果（见表 7），以 100 篇训练集构建的评分模型能够满足对 300 篇译文评分的需要。

表 7 显示，100 篇模型和 150 篇模型的机器评分与人工语义评分之间的区别没有统计意义，而 50 篇模型的机器评分与人工语义评分存在显著性差别。在表 6 中，当译文训练集达到 50 篇时，模型预测分值与人工评分的相关系数就达到了.870**，但是机器预测分值与人工评分之间存在显著性差异。当训练集达到 100 篇时，虽然相关系数提高甚微，该分值与人工评分基本一致，不存在显著性差异。按照逻辑推理，可接受译文的数量不是无限的。如果把一篇文章切分为若干个语义点，一个语义点的可接受译文一般为 5、6 个，加上本研究在进行 SVD 和语义点计算时对译文进行了词形还原，少量的语义点可以覆盖形式各异的译文可接受形式，100 篇译文所包含的形态各异的语言单位基本能够满足测试评分的需要。当然这个观点的成立还需要大规模样本做进一步测试。

总体看来，大规模测试评分模型的表现令人满意。首先，从评分过程来看，第一次评分标准体系中共包含 20 个语义单位，49 个基本语义点，整个评分过程耗时约 50 小时，而大规模测试评分模型所依据的第

表 7　机器评分与人工评分之间的配对样本 t 检验

Model	Paired Differences			*t*	*df*	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean			
机器评分-人工评分（50 篇）	2.402880	5.114862	.323492	7.428	249	.000
机器评分-人工评分（100 篇）	.060062	5.403636	.382095	.157	199	.875
机器评分-人工评分（150 篇）	−.3999613	5.29393	.432248	-.925	149	.356

[1] 在翻译中，意义所占比重较大，因此，本研究采用的总分合成方式为：语义占 60%，形式占 40%。

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

二套评分标准只使用了17个有区分度的语义点,耗时约10小时,评分过程仅为前者的五分之一,大大节约了时间和人力。其次,从评分效果来看,第二次语义评分与第一次语义评分虽然时隔1年多,但是两者之间的相关系数达到了.906**,充分表明了以有区分度的语义点作为评分依据的可行性。最后,从模型统计数据来看,大规模测试评分模型的各项数据与人工评分一致性较高,且模型的共线性数值完全符合稳定模型的要求,能够胜任大规模测试汉译英评分工作的要求。

4. 结语

本研究创建了中国学生汉译英测试自动评分模型,解决了模型创建过程中遇到的理论问题和实践问题,具有一定的初创性。研究结果显示,三种比例训练集的大规模测试评分模型的预测分值与人工评分的相关系数均高于.85,当训练集达到100篇时,模型预测分值与人工评分基本一致,不存在显著性差异,数据表明本研究提取的变量预测能力较强,机助测试评分模型表现良好,能够比较准确地预测中国二语学习者的汉译英成绩。随着系统研发的不断深入,我们还将构建诊断性测试评分模型,用于日常翻译训练和考试模拟。

参考文献

Allen, M. P. 1997. *Understanding Regression Analysis* [M]. New York: Plenum Press.

Banerjee, S. & A. Lavie. 2005. *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments* [P]. Presented at the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization.

Bowerman, B. L. & R. T. O'connell. 1990. *Linear Statistical Models: An Applied Approach (Second Edition)* [M]. Boston: Pws-Kent Publishing Company.

Dodigovic, M. 2005. *Artificial Intelligence in Second Language Learning: Raising Error Awareness* [M]. Buffalo, NY: Multilingual Matters.

Hovy, E., M. King & A. Popescu-belis. 2002. Principles of context-based machine translation evaluation [J]. *Machine Translation* 16:43-75.

Kidwell, J. & L. Brown. 1982. Ridge regression as a technique for analyzing models with multicollinearity [J]. *Journal of Marriage and the Family* 44: 287-299.

Myers, R. 1990. *Classical and Modern Regression with Applications (2nd edn.)* [M]. Boston: Duxbury Press.

Nida, E. A. 1982. *Translating Meaning* [M]. San Dimas, California: English Language Institute.

Page, E. B. 1968. The use of computer in analyzing student essays [J]. *Int'l Rev. Education* 14: 210-225.

Papineni, K. & S. Roukos. 2002. *Bleu: A Method for Automatic Evaluation of Machine Translation* [C]. Philadelphia:311-318.

Poirier, D. J. 1995. *Intermediate Statistics and Econometrics: A Comparative Approach* [M]. Cambridge, Mass.: MIT Press.

Shermis, M. D. & J. *Burstein. 2003. Automated Essay Scoring: A Cross-disciplinary Perspective* [M]. NJ: Lawrence Erlbaum Associates.

Turian, J. P., L. Shen & I. D. Melamed. 2003. *Evaluation of Machine Translation and Its Evaluation* [M]. New Orleans, U.S.A.

Walker, D. A. 2003. Suppressor variable(s) importance within a regression model: An example of salary compression from career services [J]. *Journal of College Student Development* 44,1: 127-133.

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

王金铨、梁茂成、俞洪亮,2007,基于N-gram和向量空间模型的语句相似度研究[J]。现代外语(4):405-13。

文秋芳、王金铨,2008,中国大学生英汉汉英口笔译语料库[M]。外语教学与研究出版社。

王金铨、文秋芳,2009,学习者汉英翻译分析性评分细则的制定[J]。外语教学(4):96-99。

收稿日期:2009-01-12
作者修改稿,2009-03-27
本刊修订,2009-09-17

通讯地址:225009 江苏省扬州市 扬州大学外国语学院 <bfsuwjq@yahoo.com.cn> (王)
100089 北京外国语大学中国外语教育研究中心 <wenqiufang@bfsu.edu.cn> (文)

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

design investigated the effects of instruction on the development of EFL learners' sociopragmatic competence in English requests. Treatments included explicit instruction, visual enhancement, and input-output activities. Results from a retrospective interview show that explicit instruction was most effective in making learners' pragmatic decision explicit. But written DCT results show that both explicit instruction and input-output activities facilitated these learners' development of sociopragmatic competence as revealed in their request strategy distribution in a variety of situations. Visual enhancement seemed to be less effective in this respect. These results confirmed the effectiveness of instruction in EFL learners' sociopragmatic development, and pointed to the fact that not all procedures are equally effective in teaching the same type of pragmatic features among adolescent beginners.

The influence of pointing distance and pointing mode on the choice of spatial demonstratives, by Xu Xueping and Zhou Rong, p.408

Through the experiment of table-top elicitation tasks, this study has found that the proximal-distal distinction represented by Chinese spatial demonstratives is not directly related to any concrete spatial distance. The real factor that determines the choice of proximal and distal demonstratives is not physical distance but the subjective mental construction of the speaker. The idealized cognitive model (ICM) of deixis is the psychological basis on which the speaker makes the choice of demonstratives when performing the deictic speech act. The proximal-distal distinction encoded by spatial demonstratives is related to human perceptual experience of spatial distance. It is based upon the innate properties of human mind and is therefore fundamentally embodied. The meaningful world in which human beings live is composed of language, reality and human beings who own cognitive ability. This meaningful world cannot exist without any of these three components.

A model for the computer-assisted scoring of Chinese EFL learners' Chinese-English translation, by Wang Jinquan and Wen Qiufang, p.415

Although there are multiple algorithms applied in the evaluation of machine translation, no system is available for human translation. This study aims to construct a reliable and stable statistical model for the computer-assisted scoring of Chinese EFL learners' Chinese-English (C-E) translation. It is hoped that the model proposed in this study, once implemented in the form of a computer program, can be used to score C-E translation papers in large-scale examinations. This study constructed 3 tentative selection models with different sizes of training sets. The result of the Paired-Samples T-Test showed that the computer scores produced by the 100 training set and the 150 training set bore no statistically significant difference from the human scores. Statistical results indicate that the variables extracted in this research are very effective with high predictive power, and the selection models can produce reliable scores for Chinese EFL learners' C-E translation.

The Processor-based Emergentism: Interviewing William O'Grady, by Yang Mei, p.421

This article reports an interview with Professor William O'Grady, who puts forward Processor-based Emergentism to work on language and language acquisition. In the interview, Professor O'Grady provides enlightening answers to a few questions about Processor-based Emergentism, which helps us to better understand the emergentist nature of language and language acquisition, as well as the application of Emergentism in the domain of applied linguistics.

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net