中国学生英译汉机器评分模型的研究和构建

江进林 著

高等教育出版社

ZHONGGUO XUESHENG YINGYIHAN JIQI PINGFEN MOXING DE YANJIU HE GOUJIAN

中国学生英译汉机器评分模型的研究和构建

江进林 著

高等教育出版社·北京

图书在版编目（CIP）数据

中国学生英译汉机器评分模型的研究和构建／江进林著．--北京：高等教育出版社，2016.1
ISBN 978-7-04-044316-5

Ⅰ.①中… Ⅱ.①江… Ⅲ.①英语－翻译－评分－机器识别－识别模型－研究 Ⅳ.①H315.9-39

中国版本图书馆CIP数据核字（2015）第297114号

策划编辑 贾巍巍　项目编辑 王代军　责任编辑 刘 瑾　封面设计 王 洋
版式设计 魏 亮　责任校对 王春玲　责任印制 赵义民

出版发行 高等教育出版社
社　　址 北京市西城区德外大街4号
邮政编码 100120
印　　刷 北京市密东印刷有限公司

开　　本 787mm×1092mm 1/16
印　　张 12.25
字　　数 224千字
购书热线 010-58581118

咨询电话 400-810-0598
网　　址 http://www.hep.edu.cn
　　http://www.hep.com.cn
网上订购 http://www.hepmall.com.cn
　　http://www.hepmall.com
　　http://www.hepmall.cn
版　　次 2016年1月第1版
印　　次 2016年1月第1次印刷
定　　价 28.00元

物 料 号 44316-00

本研究受到北京市哲学社会科学基金项目（批准号：15WYC064）、对外经济贸易大学学科建设专项经费（批准号：324-811005120501）和教学实验研究课题（批准号：X14520）的资助，以及对外经济贸易大学英语学院学术出版基金的资助。

序 言

已有的人工译文自动评分系统主要评价汉译英，极少数针对人工英译汉的自动评分研究还处于初等阶段。本研究旨在构建中国学生英译汉的机器评分模型，在人工评分标准、特征提取、语料和人机评分差异的分析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、完整并具有一定区分度的多词单位，有利于评价译文的语义正误、语法性、连贯性、地道性等。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些新特征。再次，以往研究采用单一文体，本研究对三种文体的译文分别建模，并确定了对每种译文评分最有效的质量预测因子。此外，本研究对人机评分差异较大的译文进行了质性分析，发现主要原因是机器依据的词表欠完美、语料预处理欠仔细。此发现为完善初始模型指明了方向。

创建评分模型需要五个步骤：语料收集、人工评分、特征提取、模型构建、模型验证。本研究收集了说明文、记叙文、叙议混合文译文各300多篇，译文评分按句进行，包括细致型和简化型两种。细致评分从语义和形式两个方面评价译文质量，简化评分仅对区分度较高的评分点进行语义评价。两种评分结果分别用于构建诊断目的和选拔目的的评分模型。评分结束后，提取多个语义和形式方面的文本特征，并在一半译文中进行多元线性回归分析，确定文本特征对人工评分预测力最大、共线性最小的方程。之后，利用方程计算另一半译文的得分，并考查机器与人工评分的相关性和一致性。最后，分析人机评分差异较大的译文，究其原因并提出改进措施。

本研究的主要结果如下：

1. 在三组语料的诊断性语义、形式评分模型及选拔性模型中，既有相同特征也有不同特征。其中，翻译单位对齐数量的贡献最突出。

2. 在诊断性模型中，篇章译文语义、形式评分模块的信度良好。交叉检验结果显示，三组语料篇章译文机器语义评分与人工语义评分的相关系数均值分别为0.846**、0.881**、0.925**，机器形式评分与人工形式评分的相关系数均值分别为0.625**、0.690**、0.774**。机器与人工评分比较一致。此外，与单句机器评分相加的篇章译文分数相比，机器对整体篇章的评分更稳定，更接近人工评分。

3. 以50、100、130、150、180篇训练集译文构建的选拔性评分模型都能较好地预测译文成绩。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机器评分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，还能满足大型评分的需要。

本书的主体部分是笔者的博士论文。衷心感谢我的导师文秋芳教授的精心指导。她踏实严谨的治学态度和批判性的思维能力一直鞭策着我，使我受益匪浅。感谢王立非教授、梁茂成教授对我论文撰写提出的宝贵意见。感谢参加我博士论文开题报告和预答辩的冯志伟教授、宋柔教授、陈国华教授、王克非教授、罗选民教授、王小捷教授，他们对本研究提出了很多精辟见解。感谢秦颖博士提供技术上的帮助。晏小琴、马晓雷博士、王金铨教授帮助评判学生译文，胡德香教授、熊兵教授、范娜、王东志博士提供专家译文，特此感谢。张莎、马玉学、陈功、蒋岳春、杨志红、刘长珍博士帮忙校对论文，一并致谢。

江进林

2015年12月7日

目 录

表　目

第六章

第七章

第八章

图 目

绪 论

0.1 引言

自20世纪60年代以来，国内外已相继开发出多个英语作文自动评分系统，相关研究日臻成熟（Burstein 2003; Landauer et al. 1998, 2003; Page 1968, 2003; 梁茂成 2005）。近年来，少数研究者对中国学生译文的自动评分进行了尝试。其中，汉译英自动评分系统采用两种人工评分分别构建诊断性（diagnostic）和选拔性（selective）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，后者满足大规模测试中汉译英自动评分的需要。两类模型的评分都比较接近人工评分，具有良好的效果（王金铨 2008）。不过，该研究的文体仅限于记叙文，且采用保留样本法（hold-out method）[1]验证模型，结果受到验证集译文的影响。在英译汉自动评分方面，已有研究挖掘了词对齐数量等特征，采用10折交叉检验（10-fold cross validation）[2]来验证模型，具有一定的优势。不过，该系统还处于初级阶段，比如语料仅涉及一个广告类段落；人工评分比较粗略；变量基本上属于词汇层面，且参考译文很少（王立欣 2007）。

基于以上不足，本研究将构建稳定、可靠的中国学生英译汉自动评分模型。研究采用细致型和简化型两种人工评分分别构建诊断性和选拔性评分模型。诊断性模型包括篇章、单句译文的语义、形式评分模块，构建单句模型的目的是对单句译文机器评分相加的篇章评分模型和整体篇章译文评分模型进行比较，选择篇章译文的最佳评分方法；选拔性模型包括篇章译文的语义评分模块。本研究在四个方面与已有研究不同。第一，人工评分在“忠实、通顺”的基础上，进一步考查译文的风格是否符合原文，并以“翻译单位”为单元对译文的语义质量进行穷尽性评价。第二，改进以往研究中的一些变量，并提取一些新的变量，比如翻译单位对齐数量。第三，为三种文体的译文分别构建评分模型。第四，采用2折交叉检验法验证模型，并深入分析人机评分差异较大的译文，究其原因并提出一系列减少机器评分偏差的措施。

总之，本研究将改进已有英译汉自动评分研究的诸多不足，并对汉译英自动评分系统的完善提供方法论指导，具有重要的理论意义。此外，研究构建的评分模型可以提高评分效率，帮助学生自学，具有很高的实践价值。

0.2 本研究的理论和实践意义

本节包括两个部分，分别概述本研究的理论和实践意义。

0.2.1 理论意义

本研究的理论意义主要有三点：

1 该方法将译文随机分为两半，一半进行训练，一半进行验证。

2 K折交叉检验是衡量机器学习结果推广性的常用方法，5折或10折交叉检验的平均结果通常最接近机器的真实性能（Breinman & Spector 1992）。