文章编号:1000－2367(2016)06－0149－10 DOI:10.16366/j.cnki.1000－2367.2016.06.027

基于 WordNet 语义知识库的英语学习者作文局部连贯自动评价

刘国兵

(河南师范大学 外国语学院;语料库研发中心,河南 新乡 453007)

摘 要:首先从中国学生英语口笔语语料库中随机抽取 80 篇同题作文作为研究语料,利用 WordNet 语义知识库,从中提取与学习者作文局部连贯相关的预测因子;然后,将其与学习者作文局部连贯人工评分相比,通过多元线性回归分析的方法创建作文局部连贯自动评价模型.性能试验结果显示,学生作文语篇局部连贯人工评分与局部连贯评价模型预测值之间的线性关系较强;与同类模型相比,预测结果更为可靠.

关键词:WordNet;学习者作文;局部连贯;评价模型

中图分类号:TP391;H08 **文献标志码**:A

国内外有关作文语篇质量评价的研究大致可以分为两类:一类是以研究和评价作文的语言特征为主,兼顾其内容及组织结构.此类研究多数集中在对学习者作文整体质量的自动评价上,也就是传统意义上的学习者作文自动评分.第二类是通过观察和提取文本表现出的各种特征,以发现的连贯特征多少为标准,对学生的作文连贯性进行测量与评价[1].纵观以上两类研究,随着计算机技术的飞速发展,前者的发展速度相对较快,文本特征提取技术与评价指标体系也较为完善,因此评价结果基本能够满足当前大规模作文自动评分需要.但相比之下,由于过去四五十年来人们对于语篇连贯的认识存在分歧,很多学者认为它是一个相对较为主观的概念[2],专门针对学习者作文连贯性进行评价的研究为数不多.也有学者认为,连贯是自然语篇内部的总体效应[3－5],与语篇的其他外在特征相比较难以找到客观的衡量与评价标准,因此自然语言处理及计算语言学领域的很多学者对此望而却步.这是造成学习者书面语语篇连贯测量研究发展相对滞后的直接原因[6].本研究以中国英语学习者作文语篇中的局部连贯性为突破口,发现和提取与作文局部连贯相关的文本特征,试图创建学习者作文局部连贯自动评价模型,以期在书面语语篇连贯测量方面取得新的进展.

1 相关研究

作文自动评分(Automated Essay Scoring,简称 AES)是指使用计算机作为评分员,自主对作文进行综合质量评估的一种全新作文评分方式.它也可以指以统计方法为基础对作文进行评价或评分的计算机技术[7].它以人工评分结果为基础,在此基础上训练计算机学习、模拟人工对影响作文质量的参数进行自动提取,经过一定运算处理后给出相应分数.实质上讲,机器自动评分系统的评分流程就是对人工评分过程的模拟与复制[8].目前国际上影响较大的 AES 系统包括 PEG,IEA,E-rater,IntelliMetricTM,BETSY 等[9].下面分别从测量内容、评分方法、提取参数及验证方法 4 个方面对以上系统进行概述.首先是测量内容.从 PEG 到 BETSY,AES 系统的测量内容经历了从简单到复杂、从片面到全面的发展过程.从一开始只注重语言形式 ,到后来注重文本内容,再发展到对语言、内容和结构3个方面分别测量[10].但遗憾的是,直到现在,以上

收稿日期:2016-03-29;**修回日期**:2016-09-22.

基金项目:国家社会科学基金项目(14BYY084);河南省哲学社会科学项目(2013CYY025);河南省科技创新人才支持计划项目(教社科〔2014〕295 号).

作者简介(通信作者):刘国兵(1978－),男,河南滑县人,河南师范大学校聘教授,主要从事语料库语言学及计算语言学研究,E-mail:liuguobing55@163.com.

AES 系统均没有把语篇连贯包含到测量内容中去. 其次是评分方法. AES 系统评分主要有两种办法,一是从待评作文中提取特征,代入创建好的方程模型,根据待评作文中符合要求的特征数量多少计算得分. PEG, Erater 采用的是这一方法[11]. 二是每次评分之前,需要提供若干人工已评作文,作为训练语料对系统进行重新训练. 训练过程中,系统根据所能提取特征的多少与人工评分结果之间的关系对已有算法进行完善、调整,以适应新的评分任务. 其评分标准主要是考查待评作文与训练集作文在各项特征上表现出的相似性与符合度. 相似性或符合度越高,得分越高;反之越低. 采用第二种评分方法的主要有 IEA, IntelliMetricTM 和 BETSY3 个系统. 综合来看,不管是采用哪一种评分方法,AES 系统在本质上都是考查机器评分与人工评分之间的一致性. 这是 AES 系统构建的一条基本原则,也是创建学习者书面语局部连贯评价模型的主要标准. 三是提取参数. 各个系统从文本中提取的参数与其测量内容相对应. PEG 由于只关注语言,因此该系统提取的参数主要是作文的浅层语言特征,如文本长度、词长、句长、介词与关系代词使用情况等. 而 IEA 重点则放在对作文内容的考查上. 该系统利用潜在语义分析技术,针对每篇待评作文建立向量,因此该系统没有涉及语篇的表层语言特征[12]. 其他系统所考查参数也都在几十种以上,但这些参数均没有涵盖语篇连贯这一重要指标. 最后是验证方法. 目前主流的 AES 评分系统大多采用机器评分与人工评分的相关度和一致性来对系统评分的有效性进行验证. 一般来说,相关度反映机器与人工对同一批待评作文进行质量排序的相似程度. 以上 AES 系统对相关度的计算有两种方式:一种方式是计算机器与单个评分员评分的相关度,之后取相关度的平均值;另一种是计算机器与多个评分员评分的相关度,也就是说,把全部有效评分相加之后得出平均值,然后与机器评分相比进行计算. 在人工评分具有较高信度的情况下,这两种计算方式产生的结果差别不大. 但如果不能确保人工评分具有较高内部一致性,那么采取第二种方法则更为有效.

2 学习者作文局部连贯自动评价模型的设计

2.1 流程

本研究使用的语料选自中国学生英语口笔语语料库 1.0 版. 本研究采取随机抽取的方法,从 80 篇研究语料中抽出 50 篇作为训练集,用于构建中国英语学习者书面语语篇连贯性评价模型;然后把剩余的 30 篇作文作为验证集,用来验证语篇连贯性评价模型的有效性. 此外,为了确保评价模型的稳定性与准确性,本研究采用了交叉验证的方法. 之后对语料进行预处理,包括对文本进行清洁消噪、统一格式以及规范文件名等. 同时挑选具有大量作文评分经验的教师对作文的局部连贯进行评分,保存人工评分结果. 之后用英文自动分句工具 BFSU Sentence Segmenter 1.1[13]对语料进行批量分句,建立句对库. 接下来的工作就是对语料进行处理、提取参数并分析数据,根据数据结果及人工评分结果进行建模. 为了查看评价模型的准确性与稳定性,最后还需要对模型进行验证,根据验证结果进行必要的调整. 具体研究流程见图 1.

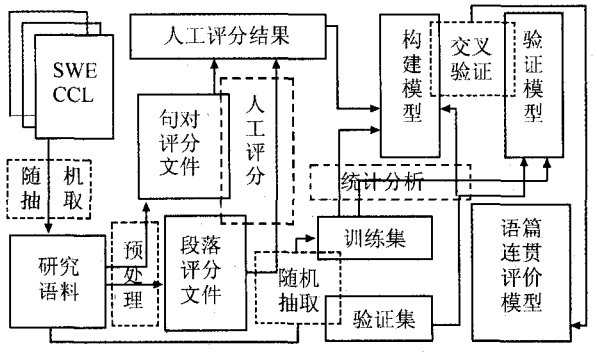

图1 研究流程示意图

高信度的人工评分是构建中国英语学习者书面语语篇连贯性评价模型的关键环节,也是保证自动评分结果可靠、有效的基本前提. 句子作为构成语篇的基本单位与基本元素,它们之间的连贯性是语篇实现局部连贯的重要条件. 句子的连贯性除了具有以上 4 个方面的特征之外,还具有一些其他独立特征. 例如,句子与句子之所以能够实现连贯,主要是因为两个句子中语义成分之间、各个命题之间在意义上以某种关系联系起来,逻辑和推理过程符合人们的正常思维习惯. 由此可以看出,不管是语篇的哪个层次,其连贯性很大程度上

都体现在意义的内部一致性上.基于这些认识,制定并完善了语篇局部连贯性人工评分标准.

之后开始挑选可靠的评分员.评分员的教育背景、工作经历以及参与评分的经验等都会对评分结果产生影响.为了把以上因素造成的负面影响降到最低,研究中尽量挑选教育背景与工作经历相似、且具有多次参与大型考试作文评分经验的评分员.基于以上考虑,最终选定的评分员共有 12 名,评分之后选取信度较高的 8 名评分员,把其评分结果作为本研究的有效控制变量.评分员选取的标准主要参考 alpha 系数与 Pearson 相关系数.如果上述两个系数较低,达不到统计学要求,其评分结果就会被排除.所有评分员均为获得博士学位的大学英语教师,均有 5 年以上大学英语教学经验,且多次参加全国大学英语四、六级或英语专业四、八级考试作文评分,有着非常丰富的评分经验.表 1 是 8 名评分员所评结果之间 Pearson 系数及 alpha 系数情况.从表 1 中数据可以看出,评分员给出的评分结果均具有较高的信度,可以用作本研究构建模型的数据.

表 1　评分员间 Pearson 系数及 alpha 系数

评分员	1	2	3	4	5	6	7	8
1	1	0.941	00.929	0.939	0.939	0.935	0.940	0.928
2	0.892**	1	0.931	0.938	0.944	0.942	0.940	0.935
3	0.868**	0.872**	1	0.932	0.934	0.928	0.929	0.921
4	0.888**	0.882**	0.874**	1	0.941	0.939	0.938	0.929
5	0.888**	0.893**	0.877**	0.888**	1	0.938	0.934	0.932
6	0.886**	0.891**	0.871**	0.887**	0.884**	1	0.946	0.937
7	0.890**	0.887**	0.869**	0.884**	0.877**	0.898**	1	0.937
8	0.881**	0.883**	0.864**	0.872**	0.876**	0.882**	0.886**	1

注:1) ** 指相关性在 0.01 水平(双侧)上有显著意义,* 指相关性在 0.05 水平(双侧)上有显著意义;2)表中左下方数据为 Pearson 系数,右上方数据为 alpha 系数.

2.2　局部连贯预测因子的提取

构建英语学习者作文局部连贯评价模型的关键环节就是提取与确定语篇局部连贯性预测因子.对于语篇连贯参数,在进行提取时均以段落为单位,利用界面化程序语言 Delphi 设计程序,实现对所有参数的自动提取.本研究所涉及的原始参数,主要从 WordNet 语义知识库中提取.具体方法是,首先从普林斯顿大学官方网站下载 WordNet 数据库;之后编写程序,与数据库建立接口;利用赋码软件对学生作文进行自动赋码,得到每篇语料中的实义词,即名词、动词、形容词和副词 4 类;利用程序从数据库中调用词与词之间的语义关系,基于语义关系对学生作文中的 4 类词汇逐一检索,并记录不同单位内出现的频数信息.

以上程序并不意味着抛弃语篇信息而完全把学生作文当作是段落组合.提取过程中,所有参数数据均以段落为单位进行呈现,但同时对段落的语篇信息进行标记,即该段来自于研究语料中的哪一篇作文,属于该篇作文的第几段等.本研究从学生作文中提取出与语篇局部连贯有关的参数共计 31 个.按照参数性质不同,把其分为 5 类,分别是:基础类、语义类、连系类、连结类及综合类.具体参数名称、参数数目以及从中提取的语篇连贯预测因子情况见表 2.

对提取到的 31 个原始参数进行标准化处理后,分析其与语篇局部连贯人工评分之间的相关关系,进而确定局部连贯预测因子.从表 1 可以看出,在提取的所有语篇局部连贯相关参数中,共有 18 个参数由于与语篇局部连贯人工评分显著相关而成为语篇局部连贯预测因子.综合参数中包含预测因子最多,共有 7 个,分别是连结密度、句子数、中心句百分比、非中心句子数、非中心句百分比和两个连系构成的连结数.连系强度系数与 3 个连系构成的连结两个参数,因与语篇局部连贯人工评分之间相关性太低而没有成为预测因子,因此不会作为自变量进入下一步的回归分析.其次是连系类参数.此类参数中的全部 4 个参数均与语篇局部连贯人工评分显著相关,因此都将作为语篇局部连贯预测因子.第 3 类是语义关系参数.该类参数中,派生词复现、二级上义词和二级下义词 3 个参数为语篇局部连贯预测因子,其余 7 个参数与语篇局部连贯性之间均无显著相关关系.第 4 类是连结关系参数,该类参数中的连贯预测因子是连结总数与非相邻连结数两个参数,相邻连结数与语篇主题相关连结数两个参数被排除在语篇局部连贯性预测因子范围之外.最后一类是基础参数,除了把一元单位复现这一参数手动剔除以外,其余两个参数即二元单位复现和词元复现两个参数均与语篇局部连贯显著相关,因此成为有效预测因子.

表 2 提取的语篇整体连贯参数及预测因子

序号	参数类型	包含参数	参数数目	预测因子数
1	基础参数	一元单位(TR)，二元单位(BR)*，词元(LR)**	3	2
2	语义参数	派生词(DR)**，同义词(SYN)，反义词(ANT)，上义词(HYE)，一级上义词(HYE1)，二级上义词(HYE2)*，下义词(HYO)一级下义词(HYO1)，二级下义词(HYO2)*，部分与整体(MR)，整体与部分(HL)	11	3
3	连系参数	连系总数(NL)**，相邻连系数(NAL)**，非相邻连系数(NNL)*，主题相关连系数(KL)**	4	4
4	连结参数	连结总数(NB)** 相邻连结数(NAB)，非相邻连结数(NNB)** 主题相关连结数(KB)	4	2
5	综合参数	连结密度(DenB)**，句子数(NS)*，中心句子数(NCS)*，中心句百分比(NCS%)**，非中心句子数(NMS)**，非中心句百分比(NMS%)**，连结强度(SC)，两个连系构成的连结(BiL-Bond)**，三个连系构成的连结(TriL-Bond)	9	7

注:表中带星号的参数为语篇局部连贯预测因子. ** 和 * 分别表示相关性在 0.01 和 0.05 水平上意义显著. 无星号的参数与语篇整体连贯性之间不存在显著相关关系.

2.3 建模过程及结果

本研究的最终目的是构建信度较高性能稳定且能够应用于大规模英语考试的学习者作文局部连贯自动评价模型. 建模的每一个环节都直接影响到评价系统的准确性与稳定性. 因此在构建模型时，会反复尝试，争取构建出对学习者作文语篇局部连贯性具有较强预测力的评价模型. 构建学生作文局部连贯评价模型时采用的统计分析方法是多元逐步回归. 语篇局部连贯方程模型共经历了 3 次构建过程. 在前两次建模中，在对方程模型的 t 值、容忍度、方差膨胀系数以及条件指数等数据进行分析时，发现模型中存在负抑制参数及共线性问题. 因此，剔除这些负抑制参数及共线性较高的参数之后，需要对自变量与因变量重新做回归分析以构建新的评价模型. 以下是具体的建模过程及模型的性能分析结果.

2.3.1 建模过程

以学生作文语篇局部连贯人工评分为因变量，以表 1 中的语篇连贯预测因子为自变量，利用 SPSS 19.0 对其进行逐步回归分析. 分析后得到一个初步模型 M_1. 在此次建模过程中，共有 5 个参数作为有效预测因子进入了回归方程，且进入模型的参数与学生作文局部连贯人工评分结果之间存在显著相关关系.

$$M_1 = 6.963 - 0.03N'_{M,S} + 0.052K_L - 0.452N_{C,S} + 0.042L_R + 0.102B_R,$$

其中，6.693 为模型方程的常数项，$N'_{M,S}$，K_L，$N_{C,S}$，L_R 和 B_R 分别代表非中心句比例、语篇主题相关连系数、中心句子数、词元复现和二元单位复现等 5 个参数得分.

要想完全检验统计模型是否有效，需要对进入方程的参数进行进一步检验，以确定方程中不包含负抑制参数，且各参数之间不存在明显线性关系. 负抑制参数的出现是多元线性回归分析经常遇到的问题，它会使模型的参数估计出错，进而直接影响模型的稳定性与预测力. 检验方程模型是否存在负抑制参数，主要通过观察自变量与因变量之间的相关关系与回归关系方向是否一致的方法来实现. 通过对各参数进行的 t 检验结果显示，由 5 个自变量与 1 个因变量构建的局部连贯评价模型中，$N_{C,S}$这一参数的非标准化回归系数β(−0.452)及标准化回归系数 β 值(通过计算为−0.300)是负数. 但前期数据统计结果显示，$N_{C,S}$与语篇局部连贯性之间的相关系数为 0.137，呈显著的正相关关系. 因此，$N_{C,S}$这一参数与语篇局部连贯之间在相关关系与回归关系方向上不一致，这说明该参数在语篇局部连贯评价初步模型中属于负抑制参数. 负抑制参数的存在通常与方程模型的共线性问题联系在一起[14]. 因此在接下来的研究中，需要把 $N_{C,S}$从自变量中剔除，然后对方程模型进行重新构建.

剔除了中心句子数这一参数之后，与学习者作文局部连贯显著相关的参数剩余 17 项. 以这 17 项语言特

征为自变量，以局部连贯人工评分为因变量，重新对其进行多元线性回归分析，分析后得到一个二次模型 M_2.

$$M_2 = 6.966 - 0.03N'_{M,S} + 0.052K_L - 0.453N_S + 0.042L_R + 0.102B_R + 0.454N_{M,S},$$

其中，6.966 为方程常数. $N'_{M,S}$，K_L，N_S，L_R，B_R，$N_{M,S}$ 分别为非中心句比例、语篇主题相关连系数、段落的句子数、词元复现、非中心句子数和二元单位复现等参数的得分.

从方程 M_2 可知，在此次建模中，共有 6 个参数作为有效预测因子进入方程模型. 对第二次构建的方程模型进行有效性验证. 有效性验证结果表明，二次构建的模型中自变量仍存在的问题：$N'_{M,S}$ 与 $N_{M,S}$ 两个参数，从理论上来讲对方程的贡献作用应该一致；根据对文本进行分析之后的判断，这两个参数与语篇局部连贯性之间应该是负相关关系. 为了验证判断的正确性，对参数 $N_{M,S}$ 与局部连贯性进行了相关分析，结果显示二者之间确实存在显著负相关关系. 根据以上情况可以断定，$N_{M,S}$ 在该方程模型中属于负抑制参数. 它的存在从一定程度上说明，去掉 $N_{C,S}$ 这一负抑制参数之后，在重新构建模型的过程中，又产生了新的负抑制参数. 负抑制参数是引起共线性的主要原因，二者通常会联系在一起[14]. 通过对共线性数据进行分析，发现该模型确实存在潜在的共线性可能. 在进入模型的 6 个参数中，$N_{M,S}$ 这一参数的容忍度值为 0.107，方差膨胀系数是 9.334. 不管是容忍度还是方差膨胀系数，均在临界值以外. 基于以上判断，决定对进入模型的参数再次进行优化，基于新的参数组合与语篇局部连贯人工评分，对语篇局部连贯评价模型进行第 3 次构建.

把学习者作文局部连贯人工评分和与提取出的连贯预测因子分别作为因变量与自变量，采取逐步回归的方法进行第 3 次回归分析. 表 3 总结了本次回归分析产生的最后一个方程模型的拟合情况. 数据显示，去除 NMS 这一负抑制参数后，进入模型的其他参数没有发生变化，二次建模中的 5 个参数全部进入方程模型. 这 5 个参数分别是 $N_{M,S}$，K_L，N_S，L_R 和 B_R. 评价模型的相关系数为 0.604，说明 $N'_{M,S}$，K_L，N_S，L_R 和 B_R 这 5 个参数整体上与语篇局部连贯性之间存在明显的相关关系. 该模型的决定系数是 0.365，即相关系数 R 的平方，表明以非中心句子比例为代表的 5 个参数能够解释语篇局部连贯人工评分 36.5%的方差.

表 3　局部连贯自动评价模型拟合情况表

R	R^2	校正 R^2 系数	标准误
0.604	0.365	0.351	1.199 34

a. 预测因子：(常数)，$N'_{M,S}$，K_L，N_S，L_R，B_R

b. 因变量：语篇局部连贯人工评分

为了进一步检验模型的有效性，避免方程模型中出现负抑制参数，对局部连贯评价模型中各参数再次进行 t 检验. 表 4 对第 3 次建模过程中依次构建的 5 个方程中的参数检验结果进行了总结.

表 4　局部连贯评价模型各系数的 t 检验结果

模型	非标准化系数		标准化系数	t	Sig.
	β	Std. Error	β		
常数	6.286	0.213	—	29.537	0.000
$N_{M.S}$	−0.015	0.003	−0.343	−5.943	0.000
K_L	0.047	0.012	0.218	3.831	0.000
N_S	−0.220	0.042	−0.369	−5.225	0.000
L_R	0.037	0.010	0.249	3.608	0.000
B_R	0.099	0.033	0.185	2.983	0.003

a. 预测因子：常数　　$N'_{M,S}$＝非中心句比例

K_L＝语篇主题相关连系数　　N_S＝句子数

L_R＝词元复现数　　B_R＝二元单位复现数

b. 因变量：语篇局部连贯人工评分

从表 4 数据可以看出，此次进入方程模型的参数为：$N'_{M,S}$，K_L，N_S，L_R 和 B_R. 在进入方程模型的 5 个参数中，有 3 个参数的非标准化回归系数为正数. 这 3 个参数是 K_L，L_R 与 B_R，对应的 β 值分别为0.047、0.037 和 0.099. $N'_{M,S}$ 与 N_S 两个参数的非标准化回归系数为−0.015 和−0.220. 与之前的相关分析结果对比，发现进入方程模型的 5 个自变量与因变量之间在回归关系与相关关系上方向一致. 由此可以得出结论，在此次构建的方程模型中没有出现负抑制参数. 另外，从各参数的 t 值与其对应的显著性(p 值均小于 0.01)来看，

自变量与因变量之间的线性关系显著. 因此,从整个 t 检验结果来看,语篇局部连贯评价模型具有统计学意义.

2.3.2 建模结果

进行 3 次回归分析之后,得到了一定程度上能够有效预测学生作文局部连贯性的回归方程,常数项及其相关参数系数见回归方程 M_3.

$$M_3 = 6.286 - 0.015N'_{M,S} + 0.047K_L - 022N_S + 0.037L_R + 0.099B_R,$$

方程 M_3 中,$N'_{M,S}$, K_L, N_S, L_R, B_R 分别为非中心句百分比、语篇主题相关连系数、段落的句子数、词元复现、非中心句子数等参数的得分. 从方程 M_3 来看,$N'_{M,S}$,K_L,N_S,L_R 以及 B_R 复现 5 个参数对学生作文语篇局部连贯性有着重要的预测作用. 其中,K_L,L_R 和 B_R 这 3 个参数的回归系数为正数,说明学生作文中以上 3 种参数出现次数越多,其局部连贯性就越强. 与之相反,$N'_{M,S}$与段落的 N_S 两个参数的回归系数为负数(−0.015 和−0.220),表明这两类参数在学生作文中出现次数越多,其局部连贯性就越差. 从本研究构建的学习者作文局部连贯评价回归方程来看,相对来说,段落的 N_S 这一参数对学生作文的语篇局部连贯性影响最大,回归系数达到了−0.220;其次为 B_R,K_L 以及 L_R 这 3 个参数,其对应的回归系数分别为 0.099、0.047 和 0.037. 最后是 $N'_{M,S}$,回归系数为−0.015.

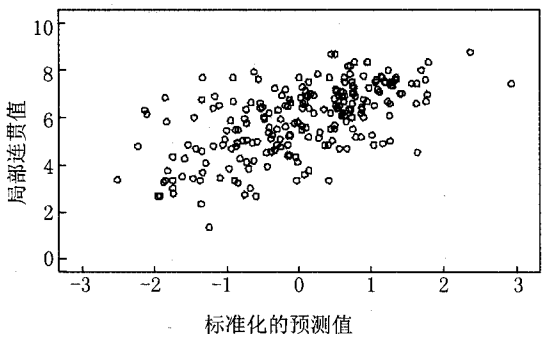

图2 模型标准化预测值与作文局部连贯人工评分关系图

图 2 直观地反映了学习者作文局部连贯评价模型标准化之后的预测值与局部连贯人工评分之间的关系. 从图 2 中观测量的分布情况来看,局部连贯标准化之后的预测值与语篇局部连贯人工评分之间呈现出明显的线性关系,因为图 2 中圆点在整体分布上呈现出从左下方到右上方的线性趋势. 同时,从横坐标轴(回归模型标准化之后的预测值数据轴)来看,散点图中绝大多数观测量分布在以分值为中心、以左右±2 为边界的有效区域内,说明语篇局部连贯评价模型标准化之后的预测值整体上服从正态分布. 以上情况表明,学生作文语篇局部连贯人工评分与局部连贯评价模型预测值之间的线性关系较强,一致性较好,模型预测结果可靠.

3 性能试验

本研究利用双重交叉验证的方法对构建的模型进行性能试验,即先用构建的方程模型给验证集作文进行评分,之后进一步利用验证集构建模型再给训练集进行评分. 最后把机器评分结果与人工评分结果进行相关分析,以验证这两个方程模型的客观性与准确性.

3.1 机器评分与人工评分相关性分析

学生作文的语篇连贯性自动评分结果由两部分构成:用训练集构建的模型给验证集作文的评分以及用验证集构建的模型给训练集作文的评分. 计算相关系数与 alpha 系数时,机器评分结果需与人工评分(详细情况见 2.1)配对进行计算. 计算结果如表 5 所示.

表 5　机器评分与人工评分的相关系数及 alpha 系数

比较参数	相关系数	alpha 系数
机器评分人工评分	0.708**	0.801

根据文献[11]的研究，机器模拟人工评分对信度的要求通常不低于 0.70. 表 5 显示，语篇局部连贯机器评分与人工评分的相关系数为 0.708，说明语篇局部连贯评价模型能够很好地预测学生作文的连贯性，且评分模型具有良好的评分信度. 另外，从机器评分与人工评分的 alpha 系数来看，局部连贯评价模型为 0.801，这说明该评价模型的预测分值与人工评分结果之间具有良好的一致性，能够满足学习者作文语篇局部连贯自动评价的要求.

3.2　机器评分与人工评分一致性分析

检验机器评分结果是否可靠，除了 3.1 部分进行的相关性分析外，还有与人工评价结果的一致性等指标. 一致性也叫吻合百分比(Percent Agreement)，是一种较为常见的信度测量方法[11]. 一致性能够较好地反映二者评分等级的一致程度，包括绝对一致百分比与相邻一致百分比两个参数[15]. 本研究把绝对一致性与相邻一致性两种数据合并成一种，统称为相对一致性. 下面从绝对一致百分比与相对一致(Relative Agreement)百分比两个方面来分析语篇整体连贯与局部连贯机器评分与人工所评等级的一致情况.

表 6　机器与人工评分等级绝对一致的文本段落数量及比例

	等级	1 分	2 分	3 分	4 分	5 分	绝对一致	总数
数量	1	1	16	26	38			
	比例	0.28%	0.28%	4.51%	7.32%	10.70%	196	
语篇连贯性评分	等级	6 分	7 分	8 分	9 分	10 分		355
	数量	63	42	8	1	0	55.21%	
	比例	17.75%	11.83%	2.25%	0.28%	0		

表 6 显示，在语篇局部连贯性的评价上，机器评分与人工评分等级完全一致的段落共有 196 段，占全部语料的 55.21%. 机器评分与人工评分等级绝对一致情况的分布相对来说较为分散，主要分布在 3 分、4 分、5 分、6 分和 7 分 5 个等级，在总数 55.21%中占到 52.11%. 在分析与评价评分一致性时，与单独的绝对一致性相比，很多情况下相邻一致性与绝对一致性作为一个整体，其应用性更强[15]. 在大型考试的作文人工评分标准中，除了规定详细的评分档次与具体要求之外，通常在解释部分加上“根据具体情况，在确定基本档次的基础上可上下浮动 1 分”. 这也间接体现了实际评分中相邻一致的重要性. 此外，国内外自动评分方面的一些研究如文献[16－18]等在验证机器评分信度时，通常参考等级绝对一致百分比和相邻一致百分比两个参数. 因此，本研究除了报告机器与人工评分等级绝对一致情况外，同时报告二者的相对一致情况. 在分析机器与人工评分等级相对一致性时，对于所有的机器评分与人工评分，将其合并为 5 个等级. 每个等级为一个分数段，其涵盖的分值大约为 2 分，如 1 分档包括至 1.9 之间的所有分数，3 分档包括 2.0 至 3.9 之间的所有分数. 表 7 总结了局部连贯评价方面机器与人工评分等级相对一致的情况.

表 7　机器与人工评分等级相对一致的文本(段落)数量及比例

等级	档次	1 分	3 分	5 分	7 分	9 分	相对一致	总数
	分数	0～1.9	2.0～3.9	4.0～5.9	6.0～7.9	8.0～10		
语篇连贯性评分	数量	1	45	125	110	5	28680.56%	355
	比例	0.28%	12.68%	35.21%	30.99%	1.41%		

在语篇局部连贯性评价方面，机器与人工评分出现相对一致的段落数为 286，占到研究语料总段数的 80.56%. 研究声称，E-rater 与人工评分之间的绝对一致及相邻一致吻合率一般稳定在 75%～80%之间[11]. 由此可以看出，语篇局部连贯评价模型的自动评分与人工评分在相对一致性方面已经超过 E-rater，一致性指标的结果令人满意. 此外，具有相对一致性的语篇及段落，从其分布情况来看，与绝对一致性有相似之处. 表 7 显示，语篇局部连贯性的机器评分与人工评分等级相对一致的情况主要分布在 3 分、5 分和 7 分这 3 个等级上，这 3 个等级中相对一致性段落占到总数的 78.88%.

3.2 机器评分与向心法则测试结果对比

3.2.1 向心法则及其运作模式

向心理论是广泛应用于自然语言处理领域的语篇局部连贯理论.该理论的核心是向心法则,主要用来探讨语篇的指称方式选择、注意焦点和语篇局部连贯之间的关系.它通过关注语篇注意焦点的变化,为语篇局部连贯提供一个较为理想的量化评价模式.文献[19－20]表明,与语篇连贯研究领域的其他理论相比,利用向心法则对书面语语篇连贯进行评价的突出优势在于,它可以在瞬间完成对整个语篇局部连贯性的量化评价,且可信度较高.向心法则的主要内容可归结为 3 项制约条件与 2 条规则.它的具体实施条件及判断方式见表 8.

1) 3 个制约条件,即:在由语句 U_1, U_2,…,U_i 组成的语篇 D 中,每个语句 U_i:1)只能有一个回指中心 $C_b(U_i, D)$;2)前瞻中心 $C_f(U_i, D)$中的每个元素都必须在 U_i 中实现;3)回指中心 $C_b(U_i, D)$是在 U_i 中实现的 $C_f(U_i-1, D)$中显著程度最高的那个成分.

2) 2 条规则,即:1)如果 $C_f(U_i-1, D)$中有一个元素在 U_i 中实现为代词,那么 $C_b(U_i, D)$也应实现为代词;2)过渡状态是有序排列的.延续过渡优于保持过渡,保持过渡优于流畅转换过渡,流畅转换过渡优于非流畅转换过渡.

表 8 向心法则的实施条件及连贯判断方式

语句条件	$C_b(U_i)=C_b(U_i-1)$	$C_b(U_i)\neq C_b(U_i-1)$
$C_b(U_i)=C_p(U_i)$	CONTINUE 延续过渡	SMOOTH-SHIFT 流畅转换
$C_b(U_i)\neq C_p(U_i)$	RETAIN 保持过渡	ROUGH-SHIFT 非流畅转换

从表 8 可以看出,4 种连贯过渡方式的关键区别在于两个方面:一是从 U_{i-1} 到 U_i 的 C_b 是否发生改变,二是回指中心 $C_b(U_i)$是否与优选中心 $C_p(U_i)$相同.可以把以上规则转化为公式 F_1 和 F_2:

F_1: $C_b(U_i)=C_b(U_{i-1})$, $C_b(U_{i-1})=[?]$,

F_2: $C_b(U_i)=C_p(U_i)$.

以上公式说明:F_1 和 F_2 同时成立,即表示相邻两句的回指中心相同,过渡方式为延续过渡,作者在延续同一个实体的话题;如 F_1 成立,F_2 不成立,表示语句 U_i 的回指中心与优选中心不相同,过渡方式为保持过渡,预示着作者将选择一个新话题;如 F_1 不成立,说明作者的话题已经发生转换,此时如果 F_2 成立,相邻两句的过渡方式为流畅转换;如果 F_1 和 F_2 都不成立,则为非流畅转换.延续过渡、保持过渡、流畅转换与非流畅转换 4 种方式,对于某一特定语篇来说,其连贯性呈降序排列.且对于特定情况,计算机对多项参数进行计算,最后对语篇连贯性进行赋值.

截至目前,向心法则已经被广泛应用于自然语言处理、语篇自动生成及连贯性评价等多个方面,评价结果较为准确、客观,目前已经被人们广泛接受.为了进一步验证本研究所构建语篇连贯评价模型的有效性,下面分别应用该模型以及向心模型对相同语料进行处理,并对其语篇连贯性进行评分.

3.2.2 测试结果对比

模型建好以后,从中国学生英语口笔语语料库中,利用随机的方法,再次抽取学生作文 10 篇.分别利用该研究构建的语篇连贯自动评价模型(因构建模型参数时主要使用了 WordNet 语义知识库,因此简称为语义知识模型)和向心规则模型分别对这 100 篇作文的局部连贯性进行评价,之后把评价结果进行比较.为了增加二者的可比性,在进行比较之前,把这两个评价模型的输出结果参数进行了调整,均采用百分制的计分方式.比较结果如图 3.

从图 3 可以看出,图中代表评分结果的两条曲线趋于吻合.这说明在对随机抽取的 100 篇学生作文的连贯性评价上,本研究构建的语篇连贯评价模型与向心规则评价模型所取得的评价结果具有较高一致性.因为利用向心法则对自然语篇连贯性进行评价,其结果已被人们广泛了解和接受,并被应用到自然语言处理和计算语言学等各个领域.因此,与向心规则模型评分结果取得较高一致性,充分说明本研究所构建的书面语语篇连贯自动评价模型在英语语篇连贯自动评价方面又前进了一步,从另外一个方面验证了语篇连贯自动评价的可能性与可行性.

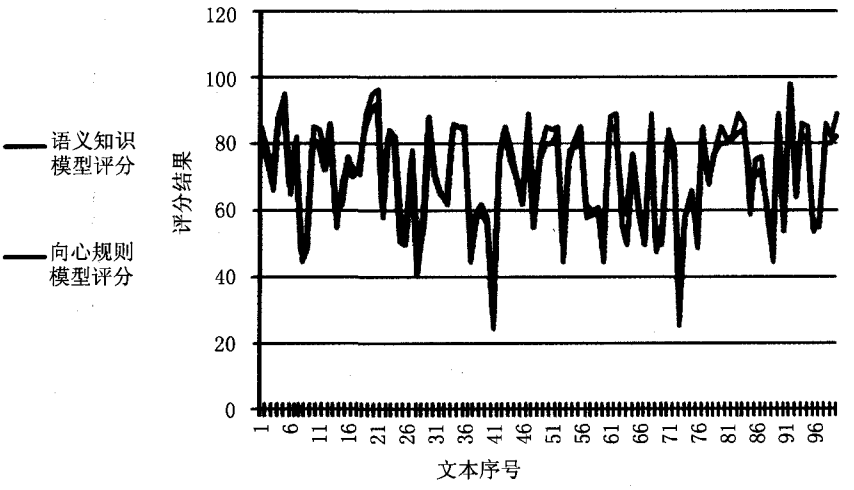

图3 语义知识模型评分与向心规则模型评分对比图

4 结　论

学习者作文局部连贯自动评价模型构建之后，把机器评分结果与人工评分结果分别从相关性、绝对一致性及相对一致性三个方面进行了对比. 相关分析结果发现，局部连贯机器评分与人工评分的相关系数为 0.708，alpha 系数为 0.801；学习者作文局部连贯机器评分与人工评分等级绝对一致的段落共有 196 段，占全部作文语料的 55.21%；从机器评分与人工评分的相对一致性表现来看，局部连贯机器评分与人工评分相对一致的段落数为 286，占到研究语料总段数的 80.56%. 之后，利用随机抽取的 100 篇语料，把本研究构建的语篇连贯评价模型与向心原则评价模型的评分结果做了对比，发现二者在存在较高一致性. 以上数据表明，英语学习者作文局部连贯评价模型完全能够应用于学习者书面语语篇连贯性的自动评价，自动评分与人工评分在相对一致性方面已经达到令人满意的效果.

参 考 文 献

[1] Shermis M D, Burstein J. Automated Essay Scoring: A Cross-Disciplinary Perspective[M]. New Jersey: Lawrence Erlbaum Associates, 2013:88-91.

[2] Harabagiu S, Record M. From Lexical Cohesion to Textual Coherence: A Data Driven Perspective[J]. International Journal of Pattern Recognition & Artificial Intelligence, 2015(13):247.

[3] Polio C G. Research Methodology in Second Language Writing Research: The Case of Text-Based Studies [C]. London: Lawrence Erlbaum Associates, 2015:91-115.

[4] Chan S W K. Extraction of Salient Textual Patterns: Synergy Between Lexical Cohesion and Contextual Coherence[J]. IEEE Transactions on Systems, Man & Cybernetics, 2014(34):205-218.

[5] Trabasso T, Suh S, Payton P. Explanatory Coherence in Understanding and Talking about Events[C]//Coherence in Spontaneous Texts. Amsterdam: John Benjamins Publishing Company, 2015:189-214.

[6] Long L, Wilson J, Hurley R, et al. Assessing Text Representations With Recognition: The Interaction of Domain Knowledge and Text Coherence[J]. Journal of Experimental Psychology: Learning, Memory & Cognition, 2014(32):816-827.

[7] Shermis M D, Barrera F D. Exit Assessments: Evaluating Writing Ability through Automated Essay Scoring[M]. Amsterdam: John Benjamins Publishing Company, 2012:289-383.

[8] Ducheneaut N, Leon W, In Search of Coherence: A Review of E-Mail Research. Human-Computer Interaction[M]. Oxford: Oxford University Press, 2015:11-48.

[9] Cornish F. Coherence and Anaphora[M]. Amsterdam: Benjamins Publishing Company, 2015:38-46.

[10] Shermis M D, Burstein J. Automated Essay Scoring: A Cross-Disciplinary Perspective[M]. New Jersey: Lawrence Erlbaum Associates, 2013:122-129.

[11] Stemler S E. A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability[J]. Practical Assessment, Research & Evaluation, 2014(4):328-345.

[12] Valenti S, Neri F, Cucchiarelli A. An Overview of Current Research on Automated Essay Grading[J]. Journal of Information Technology Education, 2012(2): 319-330.

[13] Xu J, Jia Y. BFSU Sentence Segmenter[M]. Beijing: Foreign Language Teaching and Research Press, 2012.

[14] Wackerly D, Mendenhall W, Scheaffer R. Mathematical Statistics with Applications [M]. 10th ed. Honolulu: University of Hawaii Press, 2014: 10-28.

[15] Cohen J. Cohen P. Applied Multiple Regression/Correlation Analysis for Behavioral Sciences[M]. Hillsdale, NJ: Lawrence Erlbaum Associates, 1983: 58.

[16] Valenti S, Neri F, Cucchiarelli A. An Overview of Current Research on Automated Essay Grading[J]. Journal of Information Technology Education, 2003(2): 319-330.

[17] 江进林. 中国学生英译汉机器评分模型的研究与构建[M]. 北京:外语教学与研究出版社,2010:88-92.

[18] 梁茂成. 中国学生英语作文自动评分模型的构建[M]北京:外语教学与研究出版社,2010:28-32.

[19] Suri L, McCoy K, De Cristofaro J. A methodology for extending focusing frameworks[J]. Computational Linguistics, 1999, 25(2): 173-194.

[20] 洪 明. 向心理论在英语写作质量评价中的应用[M]. 上海:复旦大学出版社,2011:40-48.

Automatic Evaluation of Local Coherence in Chinese EFL Learners' Essays with WordNet

LIU Guobing

(School of International Studies; The Corpus Research Center, Henan Normal University, Xinxiang 453007, China)

Abstract: We selected 80 essays with the same title and extracted the coherence predicting factors with WordNet; then we built an automatic local coherence evaluating model of learners' essays. The result of confirmatory test shows there is significant linear relation between the predicting scores by the evaluating model we built and the human-made scores.

Keywords: WordNet; learners' essays; local coherence; evaluating model

(上接第 129 页)

[29] Rejon C R. Cytogenetics and molecular analysis of themultiple sex chromosome system of *Rumexacetosa*[J]. Heredity, 1994, 72: 209-215.

Establishment of DOP-PCR Amplification System for Microdissecting Single Chromosome in *Humulus japonicas*

QIN Ruiyun, LI Shuangshuang, LI Shufen, YUAN Jinhong, GAO Wujun, DENG Chuanliang

(College of Life Science, Henan Normal University, Xinxiang 453007, China.)

Abstract: *Humulus japonicas* is dioecious plants with XX/XY_1Y_2 sex chromosome system, which is one of model materials for studying sex chromosome evolution. In this study, the single chromosome of *Humulus japonicaswas* microdissected and amplified from roots tip on mitosis metaphases. Then, the DOP-PCR products amplified from the single chromosome were labeled with Alexa Flour-488 by nick translation method and hybridized to the chromosomes of female *Humulus japonicasplants*. The fluorescence signals were distributed on all chromosomes. The results showed that the single chromosome of *Humulus japonicaswas* successfully microdissected and amplified, which found a basis for microdissect the X or Y chromosome of *H. japonicas*.

Keywords: *Humulus japonicas*; single chromosome; microdissection; degenerate oligonucleotide primer-PCR