总　序

一、引言

科学研究方法大致有二：其一，归纳法。归纳法指根据一类事物的部分对象的属性推知该类事物的所有对象皆具有某种属性。比如，早期的人类在多次与狼邂逅的过程中，逐渐意识到这种体型匀称协调、四肢修长、头腭尖形、鼻端突出、耳尖直立、善于快速奔跑的野生动物具有极强的攻击性，不可为伍，需要敬而远之或群起而杀之。显然，人类是在经历了多次这样的邂逅之后才意识到了狼的危险性，每一次邂逅都为人类积累了经验、加深了印象，终于在总结若干次教训之后形成了结论：所有的狼都是危险的。诚然，人类在形成结论之前不可能邂逅了所有的狼，但照样可以得出正确的结论。其二，演绎法。演绎法指从一般性的（general）前提出发，通过推导得出具体的（specific）结论。比如，在人们把“所有的狼都是危险的”这一命题视作为一般性前提时，每次邂逅一匹狼，必然会立刻意识到眼前这匹狼是危险的。这其中包含了三个论断，即：所有的狼都是危险的；这是一匹狼；这匹狼是危险的。归纳法是由具体到一般的过程，而演绎法是由一般到具体的过程。

语言研究也不例外，其方法概括起来也不外乎有归纳法和演绎法。演绎法依据可靠的前提进行严密推导，常常可以直击结论。对这种研究方法的运作逻辑我们暂且不做讨论。对于归纳法，其中有若干要素需要考虑。首先，狼有很多特征，哪些特征才具有区别性？哪些属性才是狼的致命属性？比如说，狼嚎是否是我们应该考虑的特征？其次，人类需要与狼邂逅多少次，得出来的结论才是可靠的？返回到语言研究中，前一个问题是语言学家最为关注的问题。语言分析可以从多种语言特征入手，但哪些语言特征才是最有意义的？我们又该如何选择、提取和分析这些语言特征呢？后一个问题是实证研究中的样本问题，即，我们需要观察多大的语言样本，才可以得出可靠的结论？

自20世纪后半叶语料库语言学问世以来，研究者越发对自然发生的语言数据产生了依赖，因而产生了“经验主义语言学”、“概率语言学”、“数据驱动语言学”等说法，语料库语言学也随之兴起。就其实质而言，语料库语言学采用的是典型的归纳法。语料库是大量自然语言样本的汇集，解决了以上的第二个问题，即实证研究中的样本问题。有了大样本，充分观察成为可能，归纳而得到的结果变得更为可靠甚至可以反复验证。此外，作为方法论的语料库语言学还包含一整套分析方法和分析工具，因而解决了以上第一个问题，即如何提取和分析语言特征的问题。关于选择何种语言特征进行分析，我们将在下面讨论。

总之，有了语料库，我们可能“邂逅”的语言事实更为真实、丰富、全面，这也使得通过归纳法得出的结论更为可靠、经得起验证，不需要像Edward Sapir那样亲力亲为地走入印第安部落之中去采集各式各样的语言数据，也不需要像Charles Fries那样随身携带录音机，甚至不需要像Otto Jespersen那样不失时机地以卡片形式随时记录阅读和日常生活中接触到的各种语言事实。

基于语料库数据进行语言研究，这种方法与演绎法最重要的区别之一在于，研究者在研究中所使用的所有数据均为实际发生的语言事实，而不是靠想象编造出来的句子：

The rat the cat the dog chased killed ate the malt.

Colorless green ideas sleep furiously.

Sincerity admires John.

Golf admires John.

显然，以依据研究者的直觉编造出来的句子作为研究数据，所得结果需要以语言事实来加以验证。正因为语料库语言学研究中的全部数据皆源于事实，结果也更为可靠，因而受到了越来越多研究者的青睐。在这一理念的主导下，我们近年来进行了若干项研究，目的在于利用语料库和语言大数据，对一些语言理论问题进行深入探讨，并试图解决中国外语教育中的一些现实问题。基于这些研究，我们编辑出版了这一套丛书。

二、语料分析中的语言特征选择

正如狼的所有特征并非同等重要一样，语言特征的选择在语言的量化研究中也至关重要。在前语料库时代，虽有研究者关注语言事实，但大部分研究者常常根据自己的直觉选择一些特征进行研究。到了语料库

时代，特征的选择方法发生了根本性变化。

在语料库时代，人们将语料库中的连续文本制作成词表或多词词表，甚至制作成词类（POS, part of speech）列表或词类序列（POS sequence）列表，然后对基于不同语料库制作而成的此类列表通过精巧的算法进行频率对比，进而有效地发现语料库中更为有意义的语言特征，特别是词语使用方面的特征。这种方法是语料库语言学研究中常用的主题词分析（keywords analysis），研究中几乎总会使用到一个观察语料库和一个参照语料库，并将由这两个语料库析出的词表进行对比，差异较大的词语（即语言特征）会自动浮现出来。这种特征选择方法虽有人工参与，但研究者的主观性和偏好得到了有效控制，因而研究结果也更为可靠，研究也可以重复验证。在有些研究中，人们还在两个语料库中查询自己感兴趣的语言现象，然后对所得频数进行对比，以发现两语料库间的差异。此外，人们还可以编写复杂的正则表达式，从语料库中提取比词表更复杂的语言特征，如名词短语、介词短语、动宾结构、定中结构、关系从句等，甚至涉及意义单位。

上文中描述的基于语料库的语言研究是当今最为常见的语言研究方法之一，其源头至少可以追溯到20世纪八九十年代，也有研究者将此种研究范式视为盛行于20世纪50年代的美国结构主义的延续和发展，甚至也有研究者将语料库之源头追溯到更为久远的时代。笔者认为，基于语料库的研究最早也只能追溯到电子语料库问世之日。正是随着电子语料库的问世，语言研究所需的研究素材在量（quantity）和质（quality）（即语言的真实性）两方面才有了真正的突破。基于语料库的语言研究是时代发展的必然，也为语言研究带来了新视野和新维度。在研究过程中，文本的质和量是研究的基础，而文本分析技术和对比算法起到了关键的作用，可以帮助我们发现最有意义的语言特征。

到了当今的大数据时代，情况又有了新的变化。计算机技术的发展推进了网络技术和互联网的普及，而网络的普及就意味着越来越多的人会花费更多的时间浏览越来越多的网页、上传越来越多的内容，发帖、回帖、发表评论，等等，这一切几乎无时无刻不在发生。智能手机的出现和普及更加推进了这一进程，登录网络、发表言论不再受时间和空间的限制。而所有这一切活动中最为常见的媒介正是我们研究的对象——语言。如此发展下去，网络上的语言资源会越来越多，沉淀也会越来越深，长尾效应也越来越明显。在这一背景之下，语言学家自然不应该满足于原来规模的语料库，他们与计算机领域的专家联手，设计出了各种

工具（常称为网络爬虫），可以从网络上获取大量的文本，彻底颠覆了传统语料库的概念。如今，语料库规模已经由原来的百万词级增大到动辄几千万词或数亿词级，甚至达到几十亿或百亿词级。如此规模的语料库，其优势自然毋庸置疑，长尾效应更扩展了研究维度，基于这样的语料库所得到的研究结果也更为可靠、更为多样化，对语言变化的预测能力也更强。然而，在这样的语料库中查询语言特征或由如此规模的语料库生成词语、词类、各类序列或结构列表变得不再那么容易，对这些海量语料库通过主题词分析法进行对比则更加困难。在大数据时代，我们所面临的问题已经不再是语言研究素材的不足。恰恰相反，数据量过于庞大为语言特征的提取带来了新的挑战，原来的文本分析技术和对比算法不再适用。研究者不得不另辟蹊径。

三、大数据时代的语言研究

大数据给语料库语言学者带来了新问题和新挑战。

数据量（volume）庞大是大数据时代最为显著的特征，但这并不是大数据的唯一特征。数据传输和变化之快，即大数据的速度（velocity）使得研究所依赖的数据几乎没有确定的形态，时刻处于变化之中，体量也不断增大，这也是我们必须面对的另一问题。除此之外，大数据的庞杂性（variety）也是一个棘手的问题。以上三个V被公认是大数据的典型特征。在大数据时代，语料库的创建、语言分析工具的开发、统计分析方法的更新和完善、统计结果的呈现等多个问题都将面临一场革命性的变化。

在语料库创建方面，巨量语料库的提纯是一个至关重要的问题。由于网络文本的多样性，粗暴而盲目地堆砌文本、追求语料库的大容量，会使得语料库变得十分地异质、庞杂，因而是不可取的。为此，人们汲取了网络爬虫技术，并加以改造，推出了Web as Corpus技术并开发了专用软件，依据网络页面中的关键词快速创建各种专题语料库。这种技术必将成为大数据时代语言研究中的重要技术。另外，专题语料库固然重要，但对于语言研究者而言，语体差异性、文本的时代性等问题也是语言研究中必须考虑的因素。与语体差异性、文本时代性等密切相关的问题之一是，我们应该如何通过各种途径有效获取文本的外部属性（即元信息），这也是大数据时代的语言研究中面临的又一重大挑战。只有挖掘网络文本的元信息特征，研究者才可以利用文本的各种社会属性（如语种、产生年代、作者身份、作者性别、语体特征、领域特征等），使语言

研究特别是文本差异（text variation）研究得以深入。

在语言分析工具方面，由于大量文本都存储于网络或云端，加之语料库规模不断扩大，原先广泛使用的WordSmith Tools、AntConc等单机版的文本分析工具逐渐会变得不再适用，基于网络或云端的工具或许将会成为技术开发的重点之一。此外，在语料库加工方面，基于大数据和深度学习（Deep Learning）技术设计的系统（如谷歌公司开发的句法标注工具SyntaxNet）将代表主流的研究方向，标注的准确率也会有明显提高。

从标注语料库中提取和统计语言特征时，原先广泛使用的统计方法不再适用，主题词分析方法随着语料库规模的增大也必将变得越来越困难，逐渐取而代之的是更为复杂的数据科学（Data Science），聚类、因子分析、复杂回归分析等成为语言分析的常用方法，分析工具也由原来常用的SPSS等工具变成R等更为复杂的系统。R软件的优势不仅在于可以分析大数据，还将编程和统计融合起来，使研究者可以定制各式各样的分析手段。

在统计结果呈现方面，语料库研究常见的图表呈现方式仍然会被广泛使用，但与此同时，随着数据量的增大，数据的可视化将成为呈现研究结果的重要方式，这种呈现方式将更为直观、便于理解。相信在不远的未来，语料库研究的结果将会使越来越多的人受益。

四、结语

随着大数据时代的到来，语料库语言学必将得到更多研究者的重视和青睐，大数据时代的特点将在语言研究中逐渐显现。我们希望通过本系列丛书的出版推进语言研究的不断科学化，推动我国外语与外语教育研究的发展。

本套丛书是教育部人文社会科学重点研究基地北京外国语大学中国外语与教育研究中心“十三五”规划重大项目“大数据视野下的外语与外语学习研究”（编号：17JJD740003）的研究成果，特此鸣谢。

梁茂成

二〇一七年三月

前　言[1]

自动语法纠错旨在综合运用语言学知识与自然语言处理技术检查和纠正文本中的语法错误，以提高语言表达的规范性、可读性和地道性。常见的错误类型主要有标点符号、拼写、词法、句法、词汇选择与搭配等。该项任务可广泛应用于二语学习、写作与编辑、语言翻译、搜索引擎以及语音识别等多种场景。

迄今为止，自动语法检查与纠错研究大致经历了语法规则、数据驱动的句法剖析、机器学习分类器、统计机器翻译（Statistical Machine Translation）、神经机器翻译（Neural Machine Translation）、神经序列标注（Neural Sequence Tagging）等主要发展阶段。自2011年以来，该研究领域迎来了快速发展期，主要表现在：计算语言学学会连续组织了HOO、CoNLL-2013、CoNLL-2014、BEA-2019等多项共享任务。从研究对象来看，多数研究以英语学习者语法纠错为主，一方面是因为英语学习者的群体庞大，研究成果的应用范围最广；另一方面是因为公开可用的大规模数据集以英语学习者语料库为主。

随着神经网络与深度学习技术的应用，英语学习者语法纠错研究取得了重要进展，模型的综合纠错性能已接近或达到人工修改语法错误的平均水平。尽管如此，现有研究仍存在明显的不足之处。研究者试图构建高性能的通用型模型以解决英语学习者的所有语法错误问题，忽视了学习者母语背景对语法错误的影响。譬如，就中国学习者而言，其错

1　本书得到烟台大学哲学社会科学学术著作出版基金的资助，特此鸣谢。

误类型的数量及其分布有着不同于其他国家英语学习者的显著特点。此外，现有研究面临的另一个重大问题是，语法纠错系统性能评估主要以CoNLL-2014测试集为基准，过度使用该测试集作为评估标准导致模型难以很好地泛化至其他领域，现有研究中纠错模型的性能存在被高估的可能。

本研究针对中国英语学习者语法错误，在语言迁移理论视域下利用深度学习技术探索语料库数据优化组合对模型构建的影响，最终建成高性能的中国学习者专用语法纠错模型。研究过程包括以下步骤：通过数据优化与过滤算法提升语料库数据的质量，整合并建立大规模的中国学习者、国际学习者、本族语者英语语法错误平行语料库；经两位高校英语专业教师标注来自中国高校英语学习者的错误语句，涵盖低、中、高三种语言水平，建成中国英语学习者语法错误测试集；基于建成的学习者语料库与长短时记忆循环神经网络构建四个语法纠错模型，通过评估和对比分析找出最佳模型及其不足之处；最后，针对深度学习模型的缺陷，人工编写语法规则进一步提升模型的纠错能力。

研究结果表明，词性特征可以显著提升基于神经机器翻译的语法纠错模型的性能，人工编写的语法规则可以弥补模型的不足。基于中国学习者与占比30%的本族语者语料库相结合的词–码模型纠错性能最佳，其精确率、召回率与F0.5分别达到75.63%、46.47%与67.19%；针对流水句、无主句、形容词形式以及主谓一致等错误类型编写的语法规则可以在一定程度上进一步提升深度学习模型的纠错能力，精确率、召回率与F0.5分别提高至76.47%、48.56%与68.65%。同时，本研究还发现，面向中国学习者构建的模型在低、中、高三种水平上的纠错效果呈逐渐递增的趋势，这在一定程度上“折射”出学习者逐渐克服母语负迁移的学习过程。因此，在母语迁移理论视域下构建的语法纠错模型可以较好地解决中国学习者语法纠错问题，数据驱动与人工规则相结合的方式可以达到最佳的纠错效果，为单母语背景学习者语法纠错研究提供了可行路径，同时为基于大数据和深度学习的二语习得研究提供了有益启示。

以下简要介绍本书的结构。第一章为引言，主要介绍本研究的主要背景、理论与实践意义、研究目的、主要研究问题以及研究步骤。第二章为文献综述，系统综述了英语学习者语法纠错研究现状并对相关问题进行了评析。第三章论述了构建中国英语学习者语法纠错模型所必需的平行语料库建设过程，主要内容包括：学习者平行语料库的建设背景、语料库设计与组成结构以及文本清洁与处理的主要原则；用于模型性能

评估的测试语料库标注与建设方法及过程；建成后的语料库规模、形式与组成。第四章论述了构建语法纠错模型的方法，主要包括构建基于深度学习的语法纠错模型所采用的方法、工具、模型结构以及实验参数的配置与优化方法；语法规则编写与纠错方法。第五章分析了模型在具体错误类型上的性能表现并讨论了本研究所构建的语法纠错模型的创新点、优势、缺陷以及启示。第六章汇报了人工编写的语法规则对深度学习模型的提升效果，并简要分析了原因。第七章为本书的结论。

在书稿即将付梓之际，首先要感谢导师梁茂成教授。从研究选题到收集数据再到构建深度学习模型，全过程环节无不受益于导师的悉心指导和帮助。他儒雅、幽默、睿智、学识渊博，在为人、处事、治学等多方面言传身教，使本人终身受益。还要感谢计算语言学家冯志伟教授、语料库语言学家卫乃兴教授、易绵竹教授、毕玉德教授和常宝宝教授，他们对本研究提出了重要的建设性意见。感谢李文中教授、熊文新教授和许家金教授对本研究的认可和指导。感谢外语教学与研究出版社的领导、审稿专家和编辑老师们，他们为本书的顺利出版提供了大量帮助，在此对他们的辛勤努力和敬业精神，谨致以深切的谢忱。由于时间仓促，加之本人学识水平有限，虽已成书但难免存在纰漏之处。敬请同行、专家学者及读者朋友们不吝赐教，并批评指正。

目　录

表　目

图　目

缩略语表

ACL	Association for Computational Linguistics
ANN	Artificial Neural Network
API	Application Programming Interface
BiLSTM	Bidirectional Long Short-Term Memory
BP	Back Propagation
BPTT	Back Propagation Through Time
CA	Contrastive Analysis
CLC	Cambridge Learner Corpus
CLEC	Chinese Learner English Corpus
CNN	Convolutional Neural Network
CoNLL	Conference on Computational Natural Language Learning
EA	Error Analysis
EFCAMDAT	EF-Cambridge Open Language Database
ERRANT	Error Annotation Toolkit
FCE	First Certificate in English
FP	False Positive
FN	False Negative
GEC	Grammatical Error Correction
GLEU	Generalized Language Evaluation Understanding
JFLEG	JHU FLuency-Extended GUG corpus
LM	Language Model

LSTM	Long Short-Term Memory
MLC	Machine Learning Classifier
M2	MaxMatch
NMT	Neural Machine Translation
NNLM	Neural Network Language Model
NUCLE	the NUS Corpus of Learner English
OOV	Out of Vocabulary
PDP	Parallel Distributed Processing
RNN	Recurrent Neural Network
seq2seq	Sequence to Sequence
SMT	Statistical Machine Translation
SOTA	State of the Art
TP	True Positive
TN	True Negative
WaC	Web as Corpus
WfC	Web for Corpus

第一章　引　言

1.1　研究背景

学习者语法纠错（Grammatical Error Correction，GEC）研究是利用计算机算法对学习者语言产出中的错误进行自动识别与纠正的跨学科研究，内容涉及自然语言处理、计算语言学、应用语言学、语料库语言学等多个学科领域。GEC研究成果可应用于计算机辅助写作、作文自动评分、学习者自主学习等外语教学任务，尤其对提升学习者的写作能力具有直接促进作用。因此，GEC研究成果因其受益面颇广而备受学界关注（Dale & Kilgarriff 2011：242；Dale et al. 2012：54；Ng et al. 2013：1，2014：1）。近年来，随着机器学习与深度学习技术的发展，GEC研究已经成为自然语言处理研究领域的热点话题之一（Rozovskaya & Roth 2016），也成为人工智能技术赋能语言学研究的重要探索方向之一。

自2011年第一次自动语法纠错共享任务（Helping Our Own Shared Task）（Dale & Kilgarriff 2011：242）以来，随着研究方法的不断更新，纠错系统的性能得到迅速提升。特别是神经网络和深度学习技术的应用

使得GEC系统在CoNLL-2014公开测试集[1]上的纠错效果不断超越前人研究，纠错性能已达到较高水平（Ge et al. 2018）：精确率为74%，召回率为36.30%，F0.5为61.34%。综观近十年国内外英语学习者自动语法纠错研究发现，该领域呈现出以下几个重要特征：

第一，通用性。GEC研究以两个“所有”为研究目标，即识别与纠正所有英语学习者及其书面语中出现的所有语法错误，研究者致力于构建通用的GEC系统。具体来讲，现有的GEC研究一般不区分英语学习者的母语背景和语言水平，而是基于现有可用的英语学习者错误标注语料库训练模型并在公开的测试集上检验模型的纠错效果。此外，神经机器翻译技术在GEC研究中取得突破以后，研究者不再以某种（如介词、冠词等）或某类（如实词、语法形式等）错误类型为研究对象，而是面向所有语法错误类型开展建模研究。

第二，技术性。近十年来，GEC研究方法随着自然语言处理技术的发展而快速更新与迭代，具有显著的技术性特征。迄今为止，构建GEC系统的主流方法大致经历了以下三个主要阶段，即机器学习分类器（Machine Learning Classifier，MLC）、基于短语的统计机器翻译（Statistical Machine Translation，SMT）以及基于深度学习的神经机器翻译（Neural Machine Translation，NMT）。GEC研究方法几乎与自然语言处理技术同步发展，研究者致力于将最新的技术迁移到GEC研究中，进而探索建模算法的改进与完善，以实现提升纠错精确率、召回率以及F值的目标。

第三，竞争性。大多数研究以提升GEC系统的纠错性能、刷新最佳纠错效果（State of the Art，SOTA）的百分点为目标，具有很强的竞争性。为了推动自动语法纠错技术的发展，计算语言学学会先后三次（CoNLL-2013、CoNLL-2014和BEA-2019）设置GEC专项共享任务或工作坊，来自全球的研究者或团队以竞赛的形式参与任务或工作坊，最终按照系统评估指标由高到低排序。这一特性大大促进了研究方法的更新与创新，使得GEC系统的纠错效果逐渐接近人工水平。

1　CoNLL-2014公开测试集是CoNLL-2014（the Eighteenth Conference on Computational Natural Language Learning）会议上，学习者自动语法纠错共享任务中的标准测试集。该测试集由两位英语本族语者标注，用于评估参赛的语法错误纠错系统性能。另外，Bryant & Ng（2015）曾组织10人对CoNLL-2014测试集进行重新标注，有研究（Ge et al. 2018）在10人标注版测试集上达到甚至超过了人工纠错水平。

在机器学习技术、深度学习算法和竞赛任务的驱动下，尽管GEC研究取得了显著的进步，但是仍然存在以下明显的不足之处。

首先，现有研究致力于通用的英语学习者语法纠错，忽视了学习者母语背景和语言水平之间的差异。应用语言学研究表明，学习者语法错误受母语负迁移影响，即使是母语相同的学习者，其错误分布也会因语言水平不同而存在较大差异。因此，理论上讲，研究者可能无法构建一个适用于所有学习者、所有语法错误的通用GEC模型。例如，对中国英语学习者而言，其书面语中由汉语负迁移影响导致的语法错误高达70%至80%（王盈盈 2012；俞理明 2004）。通用纠错模型尽管在公开数据集上可以达到最佳性能，但是面向中国学习者时则有可能因缺乏具有针对性的训练数据而导致性能发生急剧下降。

其次，技术与竞赛驱动的特性决定了大多数GEC研究以算法为核心，仅仅依赖公开数据和先进的算法即可取得最佳纠错性能。学界对学习者语料库数据的来源、组成以及优化组合，对语言学理论和知识的作用等方面关注不足，同时缺乏对GEC模型的可解释性分析。因此，一旦在技术和算法上遇到“天花板效应”，该领域研究的发展空间有可能会受到较大限制。

最后，语法纠错系统性能评估主要以CoNLL-2014测试集为基准，学界过度使用该测试集作为评估标准，导致模型难以很好地泛化至其他领域。CoNLL-2014测试集距今已有10年时间，仅包含由新加坡国立大学25名东南亚本科生撰写的、关于两个不同主题的50篇作文，无论在时效性和代表性上均存在较大的局限性，导致GEC系统的纠错能力存在被高估的可能。

为了弥补现有研究的不足，本研究提出面向中国英语学习者基于深度学习技术构建语法纠错模型这一选题，在二语习得理论视域下，拟从以下几个方面开展研究：1）聚焦单一母语背景的中国学习者及其书面语中的语法错误，采用最新的深度学习技术构建专门面向中国学习者的适用型语法纠错模型；2）在母语迁移理论视域下，探索语料库训练数据优化、语料库数据组合与深度学习模型纠错效果之间的关系；3）分析深度学习模型的局限性并结合人工编写的语法规则加以补充，最终确立深度学习与语法规则相结合的中国学习者英语语法纠错模型。

1.2　研究意义

1.2.1　理论意义

本研究的理论意义主要有以下三个方面：

第一，探索语料库大数据优化与组合、语言学特征、人工编写的语法规则等非技术因素对基于深度学习技术的GEC模型的影响。除了复杂的计算机算法之外，训练语料库的质量、词性与句法特征、语法规则等显性因素[1]对语法纠错模型构建同样发挥重要作用。本研究利用语言学研究的经验和已有的语言学知识，尝试通过优化数据、组合语料库以及增加显性词性特征的方式建模，并辅以人工编写的语法规则探索构建面向中国英语学习者的高性能纠错模型。尤其在人工智能技术因受制于有限的人工标注语料库而接近或达到技术瓶颈的背景下，该理论意义愈发明显。

第二，推动二语习得理论视域下的学习者语法纠错研究。二语习得研究发现，母语负迁移是导致学习者语法错误现象的重要原因之一，学习者语法错误的分布因其母语不同而存在较大差异。然而，大多数GEC研究者却忽略了语言习得领域中关于母语迁移理论的重要发现，仅有个别研究者（Leacock et al. 2014：96；Rozovskaya et al. 2017）提出或尝试将母语背景因素纳入GEC模型以提升其纠错性能。随着深度学习技术的突破，GEC模型仅凭借大数据和深度神经网络就可以胜过传统的纠错模型，导致与语法纠错相关的语言学理论愈发被忽视，甚至被束之高阁。本研究中的实验结果显示，以母语迁移理论为指导构建的中国学习者GEC模型在中国学习者测试集上可以取得最佳的纠错效果。换言之，GEC研究应与二语习得理论相结合，研究者应依据学习者的不同母语背景和不同二语水平，选择合适的方法建立与之相适应的纠错模型。从这个意义上说，本研究拓宽了GEC研究的视角和路径，一定程度上推动了二语习得理论指导下的GEC研究。

第三，启发大数据与深度学习技术应用于语言学研究。由于采用多

1　本文中使用的“显性因素”“显性语言学特征”“显性语法信息”等表述是与神经网络中的自动化特征相对而言的。深度学习与神经网络通过对词形的分布式表征，“隐性”抽取语言特征并以词向量的形式呈现语言特征；而“显性”语言学特征则是指通过模型“外部”自然语言处理工具，自动标注后得到的词性、句法、语义（语义角色、命名实体等）等特征。

层神经网络架构，本研究所构建的GEC模型具备一定的“类人”特性。因此，不同模型在语法错误类型上的纠错性能表现，可以作为中国学习者母语迁移研究和错误分析研究的有力证据，进而为研究者从大数据和人工智能的视角来全面描写中国学习者的母语负迁移现状、验证相关母语负迁移理论提供新视角。

1.2.2　实践意义

本研究是一项结合英语教学、语料库技术以及自然语言处理技术的经典研究，是从实际应用和社会需求出发开展的学术研究，研究成果可“落地”为实际产品，助力实现人工智能技术赋能英语教学实践的目标。其实践意义主要有以下四个方面：

第一，提升中国英语学习者的写作能力。据统计，中国英语学习者的数量庞大，包括接受学历教育、继续教育及其他教育形式的学习者在内，总人数多达4亿人[1]。自动语法纠错模型可部署在个人电脑、智能手机、平板电脑等多种学习设备，为学习者书面语写作提供实时反馈和修改建议，有助于提高学习者的英语写作水平。

第二，为中国英语学习者写作自动评分提供参数。语法错误的数量是衡量语言表达质量的标准之一，也是自动作文评分系统的评分参数之一。基于深度学习的语法纠错模型具有接近于人工语法检查的错误识别与纠错能力，可以为作文自动评分提供更加可靠的参数，提高自动评分系统与人工评分的一致性。

第三，为建设英语智慧课堂提供支持，助力教学模式创新。本研究构建的深度学习模型可在中国高校英语课堂“落地”，作为基于人工智能技术的教学产品或软件为英语写作教学、英语语法教学等课程“赋能”。

第四，为其他与英语写作密切相关的场景提供语法纠错支持，如学术写作、商务合同与信函写作、中译英任务的译后编辑等。

1.3　研究概述

本小节主要阐述本研究的目的、研究对象与问题以及研究步骤。

1　信息来自《中国日报》报道，网址：http://global.chinadaily.com.cn/a/201912/28/WS5e06e53ba310cf3e355813c4.html（2020年1月10日提取）。

1.3.1　研究目的

本研究以母语迁移理论为指导，将深度学习技术与语法规则相结合，构建适用于中国学习者的英语语法纠错模型。首先，收集、整理、优化并建立用于训练语法纠错模型的学习者语法错误平行语料库，人工标注并建立用于评估纠错模型的中国英语学习者语法错误测试集；其次，基于大规模语料库及其组合，利用基于循环神经网络的机器翻译技术构建四种不同的深度学习模型，进而对比分析模型之间的差异以及模型的优势和不足；最后，基于人工编写的语法规则对深度学习模型加以补充，最终确立面向中国英语学习者语法纠错的最优模型。

1.3.2　研究对象与问题

本研究聚焦中国学习者英语语法错误，探索如何利用学习者语料库大数据和深度学习技术，特别是基于循环神经网络的序列到序列（Sequence to Sequence，seq2seq）方法构建语法纠错模型，拟回答以下几个问题：

1）面向中国学习者，如何确定语法纠错模型的数据、参数、架构以及评估指标?

2）基于深度学习技术的语法纠错模型的性能如何?

3）基于深度学习技术的语法纠错模型存在哪些不足，应如何弥补?

本研究聚焦中国学习者语法错误，把研究对象明确限定在中国学习者英语书面语中，除机械错误（拼写、标点等）和词汇错误以外的语法错误。主要原因如下：

第一，目前专门针对中国英语学习者的语法错误识别的研究已取得一定进展（陈功、梁茂成 2017），但是相关的语法纠错研究仍处于起始阶段。随着深度学习技术的发展，自动语法纠错在方法上已取得重要进展，有必要探索如何利用最新的自然语言处理方法开展中国学习者语法纠错研究。

第二，中国学习者书面语中的语法错误主要源于母语负迁移的影响，现有的通用型纠错模型缺乏针对性，实际效果难以保证。因此，需要探索适合中国学习者语法错误的建模方法。

第三，词汇错误涉及社会、文化、政治等多种因素，错误与语言创新使用之间的界限模糊，缺乏判定的具体标准。一方面，自动纠错模型对词汇错误（如词语搭配、组合错误）的纠错效果较差；另一方面，以英语本族语为标准判定词汇错误的做法缺乏理论依据，学界尚存在一定争议。据此，本研究未把该类错误作为研究对象。

1.3.3　研究步骤

本研究主要通过以下步骤构建面向中国英语学习者的语法纠错模型。

1）建立中国学习者英语语法错误平行语料库

为了探索最优的模型构建方案，本研究利用现有可用的英语学习者标注语料库，自编Python程序收集了Lang-8社交平台尚未公开发布的大规模标注数据，经过数据预处理和优化之后，分别建成国际学习者英语语法错误平行语料库、中国学习者英语语法错误平行语料库以及英语本族语者语法错误平行语料库等三个大规模语料库。

2）建立中国学习者英语语法错误测试语料库

为了评估本研究所构建的语法纠错模型的性能，构建由中国学习者英语作文组成的测试语料库。语料来自iWriteBaby中国学习者英语语料库（许家金 2019）与全国大学生作文比赛中的获奖作文，涵盖国内高职高专、本科和重点院校的英语学习者，分为低、中、高三种水平，共900个错误样本，约12,000词。语法错误的标注和纠正工作由作者及另外一位来自国内某高校的英语专业资深教师分别完成，两种纠错方案之间具有较高的一致性，共同作为评估深度学习模型性能的“黄金”标准。

3）探索并确立适合中国学习者语法错误纠正的最佳模型

为了找到在中国英语学习者测试集上表现最佳的模型，本研究采用基于循环神经网络和注意力机制的序列到序列算法作为模型基本架构，分别基于单一语料库及组合语料库训练模型，最终确定基于中国英语学习者与本族语者组合语料库的纠错模型在精确率、召回率与F0.5等三个指标上均优于其他模型。因此，该模型被认为是最适合用于中国学习者语法纠错的深度学习模型。

4）分析深度学习模型在具体错误类型上的表现及不足之处

为了深入分析深度学习模型的优势和劣势，作者对比了四种模型在具体的语法错误类型上的表现，并详细分析了其优势和不足。

5）编写语法规则进一步提升深度学习模型的纠错性能

通过分析深度学习模型的局限性，利用spaCy自然语言处理工具包，有针对性地编写少量语法规则并构建基于语法规则的纠错引擎。语法规则作为深度学习模型的补充，可进一步提升语法纠错性能。

1.4　本书结构

本书共分为七章，各章主要内容如下：

第一章为引言，主要介绍研究背景、理论与实践意义、研究目的、研究问题以及研究步骤。

第二章为文献综述，系统综述了英语学习者语法纠错研究现状，并对相关问题进行了评析。主要内容包括英语学习者自动语法检查研究的兴起过程，纠错方法与相关技术的历时演变，学习者自动语法纠错研究所必需的训练语料库以及测试语料库资源，语法错误类型界定和纠错模型评估指标等关键问题。

第三章论述了构建中国英语学习者语法纠错模型所必需的平行语料库建设过程，主要内容包括：学习者平行语料库的建设背景、语料库设计与组成结构以及文本清洁与处理的主要原则；用于模型性能评估的测试语料库标注、建设方法与过程；建成后的语料库规模、形式与组成。

第四章论述了构建语法纠错模型的方法，主要包括连接主义与神经网络的基本原理、连接主义与语言学研究之间的关系；构建基于深度学习的语法纠错模型所采用的方法、工具、模型结构以及实验参数的配置与优化方法；补充深度学习模型所使用的语法规则编写与纠错方法。

第五章汇报了深度学习模型的训练结果数据、模型整体评估结果数据以及模型在具体错误类型上的性能表现，并分析和讨论了本研究所构建的语法纠错模型的创新点、优势、缺陷以及对二语习得理论的启示。

第六章汇报了人工编写的语法规则对深度学习模型的提升效果，并简要分析了原因。

第七章为结论，主要介绍了本研究的主要发现，并总结了本研究在理论、方法以及实践方面的价值及创新之处，最后提出了研究启示、不足之处和未来的研究方向。

1.5　小结

本章为全书的引言部分，首先简要陈述了研究背景、理论与实践意义，然后对研究目的、问题和步骤进行了介绍，最后对本书结构及各章主要内容进行了简要概括。