外研社英语语料库研究系列

A Study of Dependency Rules for Automatic Grammatical Error Detection in Written English

面向英语书面语误自动检测的依存规则研究

刘 磊 著

外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS

外研社英语语料库研究系列

A Study of Dependency Rules for Automatic Grammatical Error Detection in Written English

面向英语书面语误自动检测的依存规则研究

刘 磊 著

外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS
北京 BEIJING

图书在版编目 (CIP) 数据

面向英语书面语误自动检测的依存规则研究 / 刘磊著. —— 北京：外语教学与研究出版社，2016.7
（外研社英语语料库研究系列）
ISBN 978-7-5135-7916-2

Ⅰ. ①面… Ⅱ. ①刘… Ⅲ. ①英语－计算语言学－研究 Ⅳ. ①H31 ② H087

中国版本图书馆 CIP 数据核字 (2016) 第 185359 号

出 版 人 蔡剑峰
责任编辑 付分钗
封面设计 锋尚设计
出版发行 外语教学与研究出版社
社　　址 北京市西三环北路 19 号（100089）
网　　址 http://www.fltrp.com
印　　刷 北京九州迅驰传媒文化有限公司
开　　本 650×980 1/16
印　　张 15
版　　次 2016 年 8 月第 1 版 2016 年 8 月第 1 次印刷
书　　号 ISBN 978-7-5135-7916-2
定　　价 46.90 元

购书咨询：（010）88819926 电子邮箱：club@fltrp.com
外研书店：https://waiyants.tmall.com
凡印刷、装订质量问题，请联系我社印制部
联系电话：（010）61207896 电子邮箱：zhijian@fltrp.com
凡侵权、盗版书籍线索，请联系我社法律事务部
举报电话：（010）88817519 电子邮箱：banquan@fltrp.com
法律顾问：立方律师事务所 刘旭东律师
　　中咨律师事务所 殷 斌律师
物料号：279160001

该书得到以下项目的资助：

1）项目名称：英语学习者语误自动检测系统的研制（11JJD740011）
项目类别：教育部人文社科重点研究基地重大研究项目
资助单位：教育部

2）项目名称：基于依存树库的链语法词典构建（13SKA007）
项目类别：燕山大学青年教师自主研究计划课题
资助单位：燕山大学

3）项目名称：面向自动语法检查的依存规则研究（B908）
项目类别：燕山大学博士启动基金
资助单位：燕山大学

序

自计算语言学问世以来，其研究方法就深受理论语言学主流方法的影响。

在计算语言学发展早期，由于语言使用数据匮乏，加之理论语言学研究由乔姆斯基倡导的普遍语法一统天下，计算语言学研究被基于规则的方法所主导。因为缺少数据支撑和验证，这种方法常常被一些人戏称为“玩具”。到了上世纪末、本世纪初，上世纪四五十年代曾经风靡一时的田野调查语言学研究方法又一次走到语言研究的前台，计算语言学领域的主流方法也因此发生重大变化，早期基于规则的语言分析方法逐渐让位于基于统计的分析方法。人们欣喜地发现，基于统计的方法其性能远远高于基于规则的方法。计算语言学领域的一些极端主义者甚至声称，理论语言学研究对于计算语言学毫无用途。IBM语音识别项目组负责人Frederick Jelinek甚至说，“每开除一位语言学家，我们的语音识别性能就会提升一大步”。随着大数据时代的到来，大型语料库不断问世，基于语料库大数据建立的各种语言模型对语言使用的解释能力不断提高，大数据的优势越来越得以显现，基于统计的方法得以大行其道，语言研究因而进入了一个崭新的、大数据驱动的时代。

然而，片面夸大基于统计的方法也并不合理，基于规则的语言研究方法也并非一无是处。恰恰相反，基于规则的方法或许是数据匮乏时代最为合理的产物，其中沉淀着众多语言学家的语言直觉。仅仅因为统计模型的强大而否定前人研究的价值是不可取的，这无异于把孩子和洗澡水一起倒掉。因此，在当今的语言学研究领域，越来越多的学者倡导基于数据的实证方法和基于理性的内省方法相结合的混合研究方法，而在

此为试读，需要完整PDF请访问：www.ertongbook.com

计算语言学研究领域，众多学者不断尝试将基于规则的方法和基于统计的方法结合起来。

刘磊博士的新作正是此类研究的一个代表。该研究主要目的是改进链语法，以便更好地识别英语文本中的语法错误。链语法自动识别语法错误主要依赖两个资源，即链语法核心算法和链语法规则。刘磊博士的研究目的正是基于语料库中的语言使用数据来完善链语法规则，依此改进链语法的性能。自从Sleator & Temperley（1991）提出链语法以来，一直不断地完善链语法规则，至2016年8月15日，系统版本号已经升级至5.3.7，其语法规则库不断地打补丁，经过若干次升级，规则库如今已经成了“老和尚的百衲衣”，维护起来十分困难。更为重要的是，这些规则大多源自研究者的直觉，缺乏系统性，因而规则之间常常发生冲突，使得链语法性能的提升达到了一个瓶颈阶段。刘磊博士研究的意义在于探索了一种途径，可以从大规模语料库中自动挖掘语法规则，并将语法规则转换成链语法规则。研究结果表明，这种方法十分可行。

诚然，刘磊博士的研究结果尚有提高的空间。由于研究需要依赖可靠的句法剖析语料库（即树库），而句法剖析的准确率至今仍然未达到令人满意的水平。为了保证规则的准确性，研究中只能满足于规模相对不大、经过人工校对的句法剖析语料库。笔者认为，该研究最大的意义在于探索了一种自动挖掘链语法规则的方法。我们相信，随着自然语言处理技术的提高，语法剖析的准确性必将取得突破性进展，可靠的大规模语法剖析语料库将成为可能，链语法的性能也必将随之获得更大的提高。

阅读全书后，深感刘磊博士具有文理兼通的研究者素质。虽然研究方法复杂，但全书叙述得井井有条，特别是其中转换规则的编写渗透着作者的睿智和丰富的语言学知识。我能先睹为快，十分欣喜。

梁茂成

2016年8月于北京

前　言

在英语教学和测试领域，作文是检测英语学习者语言能力的重要指标。评测学习者英语作文通常依靠教师或评分员人工评阅。这一过程需要耗费大量的人力和物力，同时很难保证作文评测的信度和效度（梁茂成、文秋芳 2007）。为了克服上述弊端，国内外学者近年来开始借助自然语言处理技术，利用计算机自动评测学习者的作文质量。其中，语法错误的自动检测和修改是作文质量自动评测的重要环节。目前，学习者作文的语法检测主要采用基于语法规则和基于统计两种方法，前者有可靠的语言学理论基础，充分考虑语言的线性和层级结构，但依靠语言学家直觉编写的语法规则覆盖面有限，无法检测涉及搭配和冠词、介词等类别的语法错误；后者以大规模的真实语料为知识来源，避免了人工编写规则的繁琐，且覆盖面大，但这种方法对语言的层级结构考虑不够。本研究拟结合上述两种方法的优点，使用混合法进行自动语法检查，从大规模经过词性和句法标注的语料库中提取词汇-语法信息构建模型，提高现有语法检测系统的准确率。

链语法是一套用计算机分析自然语言句法结构的形式化模型，由词典和算法两部分构成：词典记录每个词条的句法链接方式，算法利用词典分析句子中各单词的链接组合，符合语法的句子会构成一个完整的链条。链语法分析器可以用来自动检测英语书面语的语法错误，但由于其词典依靠人工编写，存在以下两点不足：（1）词典编写依靠编纂者的主观判断，缺乏系统性；（2）词典中单词链接规则的描写不全面，无法检测出某些主谓不一致类、动词+介词类和动词+动词补语类语法错误。本研究旨在通过数据驱动的方法从依存树库中自动提取链语法词典，改善上述链语法词典的缺陷，提高英语学习者书面语语法错误自动检测的准确率。

本研究以库容为100万词的宾州英语树库作为训练语料，通过以下三个步骤重建链语法词典：（1）将宾州英语短语结构树库转换为依存树库；（2）完善现有依存树库标注体系，并按该体系调整和细化宾州英语依存树库中的依存关系；（3）从依存树库中提取链语法词典，使用中国英语学习者语料库中的504例错误例句作为测试集，检验新建词典在自动语法检查时的准确率、召回率和F值。

研究结果表明，利用依存树库构建的链语法词典避免了人工编写词典缺乏系统性的弊端；从真实语料中获取的词-语法规则更加全面，能够检测出原有链语法词典无法检测出的学习者书面语语法错误。与原链语法词典相比，本研究新建链语法词典检查测试语料中语法错误的准确率、召回率和F值分别提高了5.9%、19.9%和13.4%。

英语书面语误的自动检测是一个历久弥新的研究问题。早期研究通过编写规则模板匹配语法错误，但是这种方法所能识别的错误种类有限，无法识别搭配和冠词、介词等与语境密切相关的语法错误。因此，近期的研究引入了基于统计和大数据的方法，通过构建基于概率的语言模型弥补规则模板的不足，涌现了一批规则和统计相结合的研究（Ng et al. 2014）。笔者认为，为了进一步提高语误自动检测的准确率，应该融合语言规则和统计模型两种方法的优点，本研究正是在这一思路指导下做出的初步探索。

在本书的撰写过程中，我的导师梁茂成教授在理论选取、数据分析和结果讨论等诸多层面对我的教导使我获益匪浅。没有导师的指引，是不可能完成该书写作的。我还要感谢计算语言学家冯志伟教授、浙江大学的刘海涛教授、解放军外国语学院的易绵竹教授、北京外国语大学中国外语教育研究中心的李文中教授、许家金教授和熊文新教授，他们从繁忙的工作中抽出时间阅读了本文的初稿，提出了许多宝贵的意见和建议。同时，感谢教育部和笔者所在单位燕山大学为本研究提供经费支持，以及在场地、设备上予以协助，保证了研究的顺利进行和书稿的最终出版。

由于笔者水平所限，书中难免有纰漏之处，恳请各位读者不吝赐教，有不妥之处，敬请批评指正！

缩略语表

APSG	Augmented Phrase Structure Grammar
CFG	Context-Free Grammar
CKY	Cocke-Kasami-Younger algorithm
CLEC	Chinese Learner English Corpus
DAG	Directed Acyclic Graph
HPSG	Head-driven Phrase Structure Grammar
PCFG	Probabilistic Context-Free Grammar
PML	Prague Mark-up Language
PSG	Phrase Structure Grammar
PTB	Penn Tree Bank
XML	eXtensible Mark-up Language

目 录

此为试读，需要完整PDF请访问：www.ertongbook.com

图 目

此为试读，需要完整PDF请访问：www.ertongbook.com