Computer Science and Application 计算机科学与应用, 2018, 8(9), 1372-1381
Published Online September 2018 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2018.89149

Hans汉斯

An Overview of the Methods on Automatic Detection of Grammatical Errors

Gong Chen[1], Jian Wang[2]

[1]School of International Studies, University of International Business and Economics, Beijing
[2]Hangzhou Lvwan Network Technology Co. Ltd., Hangzhou Zhejiang
Email: chengong813@126.com

Received: Aug. 24th, 2018; accepted: Sep. 6th, 2018; published: Sep. 13th, 2018

Abstract

Automatic detection of grammatical errors is an important research topic in the field of natural language processing. Although previous researches have accumulated a wealth of experience in this field, the automatic detection of grammatical errors for Chinese EFL learners' English is still in its early stages. By reviewing the theories and methods of the relevant studies, this paper elaborates the two main methods used in the previous researches, *i.e.* the rule-based approach and the statistic-based approach, and makes a comment on their advantages and potential problems. Finally, the paper puts forward a possible hybrid method of building a system and makes a test, aiming to help develop a grammatical error checking system for Chinese EFL learners.

Keywords

Grammatical Errors, Automatic Grammar Checking, Rule-Based, Statistic-Based

自动语法检查方法研究述评

陈　功[1]，王　健[2]

[1]对外经济贸易大学英语学院，北京
[2]杭州绿湾网络科技有限公司，浙江　杭州
Email: chengong813@126.com

收稿日期：2018年8月24日；录用日期：2018年9月6日；发布日期：2018年9月13日

摘　要

语法错误自动检查是自然语言处理领域的重要研究课题。尽管前人已经在该领域积累了丰富的经验，但

文章引用: 陈功, 王健. 自动语法检查方法研究述评[J]. 计算机科学与应用, 2018, 8(9): 1372-1381.
DOI: 10.12677/csa.2018.89149

是针对中国学生的英语语法错误自动检查研究尚处在初级阶段。本文对国外自动语法检查系统的理论和方法进行了梳理，详细描述了已有研究所采用的两大主流方法——基于规则的方法和基于统计的方法，并对这两者的优势和问题进行了评述。最后本文指出了今后自动语法检查系统可能采用的方法，并进行了试验，旨在对中国学生英语语法错误自动检查系统的开发提供启示。

关键词

语法错误，自动语法检查，基于规则，基于统计

Copyright © 2018 by authors and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

书面语中的语言错误多种多样，如拼写错误、语法错误、语义错误等。一直以来，这些错误只能由人工进行检查和纠正。20 世纪 70 年代之后，随着自然语言处理技术的不断进步，研究者开始探索利用计算机对各种语言错误进行自动检查，自动查错成为研究热点。尤其是 20 世纪 90 年代以来，随着越来越多的人开始借助计算机进行语言书写和交流，自动语法检查研究的应用范围不断拓展，并被广泛运用于文字处理系统、语言校对系统，以及计算机辅助语言学习(CALL)等领域。除了作为文字处理系统(如 Microsoft Word)必备的内嵌模块，语法检查系统开始以独立软件的形式出现，服务于更多的用户。

2. 语言错误类型

为了实现语言错误的自动检查，首先需要对语言错误进行归类，以确定自动检查的目标或范围，进而选择合适有效的检查技术。根据不同的研究条件或研究目的，语言错误大多可以被分为以下四类[1] [2]：1) 拼写错误。拼写错误的自动检查技术目前已经较为成熟，而且已经成为文字处理系统中必有的功能模块，例如，Microsoft Office 系列产品。此外还有很多独立的拼写检查系统，如 *Ispell*[1]、*Aspell*[2]、*Hunspell*[3]等。2) 语法错误。通常指的是一个句子违反语法规则所出现的问题，语法检查系统检查的就是该句子的合语法性。和拼写检查不同的是，语法检查的实现需要结合一定的语法规则或语境信息。例如，在句子“**Jill is younger then than Jack.*”中，拼写检查无法查出“*then*”是错误的，因为该词的拼写完全正确。若要对这类错误进行检查，语法信息必不可少。3) 文体错误，指使用了与文体不相符的词汇或较为复杂的结构，从而使得文本难以理解[1]。该类错误与文体高度相关，例如，朋友之间的邮件通常是比较随意或非正式的，而技术报告则需要用语严谨、正式。如果将后者的语言用于朋友间的书信交流，就会导致文体错误。4) 语义错误。语义错误本质上是词汇之间和句子之间意义的连贯问题[2]。与前三类错误的自动检查相比，语义错误自动检查的实现难度较大。要解决语义问题，除了要让计算机了解语言本身的意义，还要为系统配备广泛的世界知识。就目前的研究情况来看，语义错误的自动检查还处在起步阶段。

除了上述分类方法，也可以从词类的角度划分语言错误。例如，介词错误[3]、冠词错误[4]等等。该类错误划分标准较为简单，在此不做赘述。

[1]http://fmg-www.cs.ucla.edu/fmg-members/geoff/ispell.html.
[2]http://aspell.sourceforge.net/.
[3]http://hunspell.sourceforge.net/.

3. 自动语法检查系统的实现方法

早期语法检查系统主要是基于字符串匹配(string matching)或模式匹配(pattern matching)，例如，最早广泛使用的语法检查工具 *Unix Writer's Workbench* [5]。由于模式匹配的方法具有较高的信度，因此，在20世纪80年代得到较为广泛的使用。同年代，基于句法分析(parsing)的语法检查也成为自然语言处理的重要方法之一[6]，句法分析所基于的规则主要通过语言学家人工编写实现。采用了该技术的系统包括：Houghton Mifflin 公司推出的 *CorrecText*、Aspen 软件公司推出的 *Grammatik*、IBM 公司的 *Epistle* 和 *Critique*，以及 *LINGER* 等。但是由于该方法需要占用大量的计算机资源，受当时计算机技术条件的限制，未能获得进一步的发展。Pusack [6]指出，如果使用语法分析技术，“哪怕是一个最简单的系统都可能占用一台中型计算机的全部资源”。不过，随着计算机技术的飞速发展，一直以来牵制语法分析技术发展的硬件问题迎刃而解。基于句法分析技术的自动语法检查开始成为主流。可以说，现有语法检查系统几乎全部采用了基于句法分析的技术，只是程度有所不同[7]。例如，Lingsoft 公司开发的瑞典语语法检查工具 *Grammatifix* 和挪威语语法检查工具 *Grammar Checker for Norwegian*，以及 Open Source 团队开发的开源文字处理系统 *AbiWord*[4] 中的语法检查模块。

20世纪90年代之后，基于句法分析的语法检查工具在具体实现方式上逐渐多元化。随着大规模语料库的陆续建成，词性赋码工具(如 CLAWS)的不断成熟，以及宾州树库的推出及其应用的不断增多，以统计(或数据驱动)为基础的语法分析开始快速发展。由于基于统计方法省时省力，很快受到自然语言处理界的青睐。不过，在语法错误自动检查研究领域，这一发展趋势还不是很明显[7]。

对于自动语法检查实现方法的分类，Naber [1]认为自动语法检查的实现主要依靠三种基本方法：基于句法的(syntax-based)、基于统计的(statistics-based)和基于规则的(rule-based)。不过笔者认为，和自然语言处理领域的大趋势一样，这些方法仍旧可以归为两大类：基于规则的方法和基于统计的方法。

3.1. 基于规则的方法

基于规则的方法是自然语言处理领域最为常用的理性主义方法[8]，主要通过研究者内省的方式获得有关输入语言的语法知识，并人工编制形式化规则，以便实现系统对输入语言的自动处理。就自动语法检查来说，基于规则的方法主要表现为：基于正确规则的语法检查和基于错误规则的语法检查。

1) 基于正确规则的语法检查

在基于规则的方法占主导的时代，使用计算机识别并处理书面语语法错误的方法有三种：或采用模式匹配，或使用句法分析器(parser)，或将两者相结合[9]。

模式匹配的理念非常简单。在匹配模式下，所有的模式必须与相对应的搜索空间完全一致，否则匹配失败。这就是最简单的模式匹配，有的研究也称之为“字符串匹配”[7]。不过，基于正确模式匹配的自动语法检查非常少见。已有的语法检查研究多是在错误模式匹配的基础上进行的。

基于正确规则的语法检查在自动语法检查研究领域具有重要地位。该方法需要的是一部正确的语法，或者说是一部本族者语法。句法分析器[5]根据配备的语法知识对输入文本做完全的句法分析，每一个句子输出一个树形图。如果语法分析失败，则认为该句输入有误。简单来说，句法分析是将一维的线性串转换为二维结构的一种过程[10]，而实现该过程的计算机程序就是句法分析器。Holland *et al.* [11]将句法分析器称为“自然语言处理引擎”(NLP engine)，“句子输入之后，句法分析器会将其分解为不同的成分，并与计算语法进行对照，最终输出句子的结构分析结果。根据该结果，我们可以判定该句是否符合语法

[4] http://www.abisource.com/.
[5] 若无特殊说明，文中所说的“句法分析器”均指基于规则的句法分析器。

规则，或符合什么样的语法规则”[11]。Schulze [12]考察了当时基于句法分析的 CALL 系统，发现使用最为普遍的句法分析器有一个共同特点，即它们都配置了基于语言学规则的陈述性知识。也就是说，它们的语法和算法是分开的[6]。“所谓语法和算法分开，就是要把语言分析和程序设计分开，程序设计工作者提出规则描述的方法，而语言学工作者使用这种方法来描述语言的规则。……它非常有利于程序设计工作者与语言工作者的分工合作，为面向计算机的语言研究指出了方向”[8]。这一理念有利于我们将研究重心放在语言规则的描述上，进一步完备语法知识，提高语法分析的准确率。此外，这样的句法分析器允许研究者采用与之相适应的语法理论，对其知识库进行扩充。

2) 基于错误规则的语法检查

对于语法检查系统而言，模式匹配通常指的是“错误模式”与输入语言错误的匹配。这种“错误模式”是研究者从错误语料中总结出的词或词性搭配错误的规则，并加以形式化描述的结果[13]。有不少语法检查系统是全部或部分基于模式匹配的，例如，Liou [14]和 Rypa & Feurman [15]等。从句法分析的角度来看，句法分析器除了需要做出最基本的合语法性判断，还需要在句法分析结果中提供足够的信息，来说明句子存在什么样的语法错误，错在什么地方，应该怎么修改。Matthews & Fox [16]认为，传统的句法分析器所包含的正确的语法规则属于“专家模型”；若想对学习者错误进行检查，还需要加入包含有“学习者中介语规则”的“学生模型”。Leacock *et al.* [7]也认为，语法检查系统中的语法需要有较好的“容错性”，要能够分析含有典型语法错误的句子，并给出足够的诊断信息。综合已有研究可以发现，在语法检查系统实现方法中，“错误规则”和“约束松弛”起到了非常重要的作用。Vandeventer-Faltin [2]就曾指出，“错误语法[7]”是 CALL 领域十分常用的句法检查技术，而“约束松弛”则可能是错误检查研究中使用最为广泛的技术。

其一，约束松弛。

作为不合语法句子分析中最常用到的技术，约束松弛首先预设的是，语言中存在着各种各样的约束。这些约束条件是语法的一部分，可以判断输入文本是否具有合语法性。简单来说，英语中要求主语和谓语的数要一致，“数”就可以看作是主谓一致的约束条件。输入句法分析器的文本一旦无法满足该约束，分析过程就会终止。为了使分析继续，最终输出对学生有用的信息，我们需要对该约束进行松弛，让分析器将其忽略，直至完成分析。之后，通过查找松弛的约束条件，便可确定错误所在。使用了约束松弛技术的语法检查系统包括：IBM 公司开发的 *Epistle* 和 *Critique*，Chen & Xu [17]的 *Grammar-Debugger*，Vosse [18]的荷兰语语法检查系统，以及 Catt & Hirst [19]的 *Scripsi* 等等。

其二，错误语法。

Sanders & Sanders [20]将“错误语法”定义为：能反映典型错误句子结构的语法。句法分析器首先调用正确语法对输入文本进行分析，若分析失败则会转向错误语法。错误语法或错误规则必须叠加在正确语法或正确规则的基础之上。使用错误规则的语法检查系统较多。Weischedel *et al.* [21]的德语学习系统是较早采用该技术的研究，主要利用错误规则来检查词序错误。该系统语法规模较小，只有不到两百个词汇，但却证明了用错误规则检查句法错误的可能性。Schwind [22]的错误分析、解释系统“*EAES*”采用错误规则检查三类错误：低层次的句法错误，高层次句法错误和语义错误。采用了错误规则的语法检查系统还有：Yazdani [23]的 *LINGER*，Schneider & McMoy [24]为美国手语使用者设计的英语学习系统 *ICICLE*，以及 Carlberger *et al.* [25]的瑞典语语法检查系统 *Granska* 等等。

[6]有的句法分析器的语法规则和算法是集成在一起的，比如，扩充转移网络(Augmented Transition Networks，简称 ATN)，这种句法分析器叫做程序性句法分析器。

[7]“错误语法”即“错误规则”，英文表达可以是，“error grammar”，“error rules”，“mal-rules”，“bug/buggy rules”或“error production rules”。

3) 基于规则的方法：优势与问题

基于正确规则的句法分析方法之所以被语法检查领域的研究者持续关注，是因为该方法体现了两大优势：第一，可以检查出句子中长距离依存关系错误，而不仅仅局限于局部错误；第二，如果语法足够完备，则能够分析各种类型的句子，而不需要预置其他信息。不过，构建一部完备的语法绝非易事。这也就意味着，一旦输入的句子结构是分析器没有涵盖的语言知识，句法分析器就无法输出句子结构的分析结果。因此，构建一部覆盖所有语言知识的语法，应该是不断追求的目标。

基于错误模式匹配的语法检查之所以还在发挥作用[26]，是因为该方法有三大优势：第一，部分错误模式的总结和提取较为容易，且模式匹配的实现较为容易；第二，对于部分错误来说，该方法查错准确性高，可信度好。错误模式一旦与输入文字相匹配，则可说明输入的文本有误；第三，系统运行速度快。

不过，错误模式匹配也具有一些无法克服的问题：第一，所关注的语言错误太过具体化，而且无法涵盖学习者语言中的其他问题，因此，检查中可能会遗漏很大一部分语法错误；第二，模式匹配仅考虑输入语言在线性顺序上的问题，而无法关注语言的层次性或结构方面的错误；第三，即便是线性顺序上的语法错误，仅凭模式匹配也无法解决，如主谓一致问题。显然，对于这类错误，采用一例一错的模式匹配反而是最不经济的办法。

和模式匹配相类似，基于错误规则的句法分析查错准确率高。只要预先将错误规则写入语法，句法分析器就能检查到相关错误。同样，约束松弛的查错准确性也比较高。不仅如此，约束松弛技术还能够确保句法分析器顺利完成整个句子的句法分析[2]，语法检查系统可以根据分析过程中的约束松弛记录，来抓取句子错误信息。此外，约束松弛还免去了为语法添加新规则的麻烦，也避免了规则添加所导致的其他问题。

但是，在考虑采用错误规则和约束松弛技术时，需要注意三方面问题：第一，学习者错误很难穷尽。为了尽可能覆盖所有错误，我们必须编写足够多的规则，而有些错误规则的编写较为困难；第二，错误规则的条件太过于具体，一条规则只能对应一类错误，研究者以有限的时间和精力难以完成。Murphy，Krüger & Grieszl [27]指出，“颗粒度细化的唯一方法就是制定非常具体的规则，而过度具体化的错误规则只能覆盖非常少量的语言错误。所以，设计错误规则的核心任务就是找一个合适的度。这个度既要足够宽泛，能让同一条错误规则尽可能覆盖多个错误；又要足够具体，以便给用户详细的反馈”。Sanders [28]更是直言，“一部错误语法至少应该和分析器配备的主体语法规模相当”。这个任务在短时间内是不可能完成的。第三，如果采用约束松弛技术，那也就意味着系统搜索空间的增大，可能造成过度生成的问题。因此，在设计约束松弛的时候，还要考虑如何限制结构过度生成的问题。

3.2. 基于统计的方法

在自然语言处理领域，基于统计的方法由来已久。不过，具体到自动语法检查的实际应用领域，直到上个世纪 90 年代，基于统计的方法才开始逐渐应用于研究当中。该方法主要依赖于大型语料库来获取语法知识，系统通过对输入文本的关键特征进行统计或计算，然后通过句子成分的权重来判断句法关系，实现语法检查的目的。

基于统计的语法检查系统的实现可以从两个视角来考察：一是分类方法(classification approach)；二是语言模型(language modeling)的方法[7]。

1) 基于分类方法的语法检查

基于分类的方法本质上属于机器自动学习的范畴，而大多数基于分类的语法检查系统采用的是有指导的学习方式。其主要思想是：首先使用事先定义好的类别或范畴标记对文本中的实例进行人工标注；标注好的文本作为训练数据对分类器进行训练；之后，机器根据标注好的训练数据进行自动学习，获得

某一类词，如介词的用法模型，包括一些上下文特征。在实际使用中，计算机会根据学习得到的知识对新输入的文本进行分类，计算新文本中介词的用法是否与训练数据提供的特征相近，如果差距过大，则说明可能有错。

不同研究所使用的分类器和训练数据有所不同。对于分类器而言，主要有：最大熵分类器，支持向量机，以及决策树等。早期的研究主要使用的决策树(如 Knight & Chander [29])，而现在使用较为广泛的则是最大熵分类器(如 Chodorow *et al*. [30]等)。从训练数据的角度来看，有的研究采用的是不含任何语法错误的正确文本(well-formed text) (如 Gamon *et al*. [31]；Bergsma *et al*. [32])，有的使用的则是含有语法错误的文本(ill-formed text)或是经过错误标注的文本(如 Han *et al*. [3])。

2) 基于语言模型的语法检查

统计语言模型就是表示语言基本单位的分布函数[33]。基于统计的语言模型将自然语言近似地看成随机过程，以概率形式对语言进行描写，并做进一步的定量分析，发现蕴藏其中的规律和特征。这种方法是建立在大量语料的统计基础之上的，能够较为细致地从概率特征方面对语言进行描写[34]。

基于统计语言模型的语法检查理念较为简单。对于统计语言模型来说，语言中的任何一个句子(符号串)都是可以接受的，只是接受的可能性(概率)不同。因此，在实际操作中，研究者需要为合语法的句子设定一个最低阈限，输入文本一旦低于这个值就会被判定是错误的。

将语言模型运用到语法检查研究始于 Atwell [35]。他在研究中提出了将词性赋码工具 CLAWS 用作语法检查工具的可能性。其主要思路是：首先，使用 CLAWS 为文本标注词性码；其次，提取核心词及其左右邻词的词性标注码串，并计算这些标注码共现的概率；如果低于系统设定的阈值，则认为该处有语法错误。有的研究采用了语言概率模型。Nagata *et al*. [36]的研究以日本的英语学习者的冠词错误为目标，为了实现错误检查的目的，他们以不同上下文特征中冠词的条件概率为基础，通过概率模型呈现了冠词使用的各种情况。还有的研究用到了 N 元组模型。例如 Lee & Seneff [37]的研究就采用 N 元组模型来生成最佳 N 元组集，以帮助检查并纠正冠词、动词的时态体等错误。

3) 基于统计的方法：优势与问题

其实，不论是基于分类的方法，还是基于语言模型的方法，它们在研究中的广泛使用都说明，最近几年基于统计的方法开始备受关注。究其原因，主要是该方法克服了基于规则方法的几点不足：第一，避免了手工编写规则的耗时耗力；第二，避免了基于规则的系统所面临的容错性问题。因为对于基于统计的系统来说，输入文本不论对错只有概率上的不同。

不过，基于统计的方法也有自身较难克服的问题：第一，最突出的就是数据稀疏的问题，即无法获得足够多的人工标注语料来训练系统。Leacock *et al*. [7]指出，基于统计的方法在自动语法检查研究中的使用还不是很广泛，这可能和数据稀疏问题有很大关系，因为语言错误的人工标注执行起来困难较大，而通过少量包含错误标注的语料对系统进行训练，又不足以建立一个可推广的模型。Yarowsky [38]将该问题称为“知识获取瓶颈”。不少研究者对这一问题的解决办法进行了探索。有的研究者采用人工引入错误的语料(如 Foster & Andersen [39]等)；有的研究者则采用机器翻译的方法来解决该问题，例如，Lee *et al*. [40]就探讨了机器翻译译文是否可以有效替代学习者文本作为训练语料的问题。此外，还有的研究者转而寻求网络资源的帮助(如 Yi *et al*. [41]；Bergsma *et al*. [32]等)。另外一些研究者则试图通过技术手段来解决这个问题，例如，采用半指导的学习和无指导的学习(如 Chodorow & Leacock [42])等。这些新的尝试在不同程度有助于问题的解决，但也势必带来其他一些问题。第二，基于统计的方法在处理长距离语言约束时可能会存在问题[7]，这或许是“基于统计的方法在语法检查研究中使用较少”[7]的原因之一。第三，基于统计的方法对语法错误的判断取决于研究者对阈值的设定。这会导致两方面的问题。首先，

阈值设定的合理性问题还需要进一步论证。其次，不论阈值是否合理，在实际语法检查中，都会有“一刀切”的问题，系统可能无法区分不常用但正确的句子和不常用也不正确的句子。

4. 对构建中国学生英语语法错误自动检查系统的启示

学习者语法错误形形色色，与母语者差异较大[43]，其中既有深层问题，也有浅层问题；既有远距离错误，又有近距离错误。因此，处理问题的方法也要多样化，既要使用演绎法，也要使用归纳法，基于规则的方法和基于统计的方法要适时地结合起来，共同实现语法检查的目的。总之，不管是规则的方法也好，统计的方法也罢，其本质都是要更多地、同时又是更系统、更有效地给计算机灌输有关人类自然语言的知识。

近几年来，语法检查研究者们将两种方法相结合的呼声越来越高。Gamon *et al.* [44]就指出，对于不同类型的语法错误，我们可以使用不同类型的模型进行检查。Leacock *et al.* [7]也强调，任何一个鲁棒的语法错误检查系统都应该是一个多种方法混合构建的系统。比如，对主谓一致错误的检查，还是基于规则的句法分析器更占优势，而冠词和部分介词错误的检查则是基于统计的方法更胜一筹。

综合考虑上述方法的优劣，要想构建一个更好地适用于中国学生的英语语法错误自动检查系统，采用基于正确规则的方法构建语法检查系统较为理想。同时考虑到基于统计的方法的优势，下一步研究可采用通过大规模语料库统计得到的语法规则，即“采用规则和语料库统计相结合的方法”[45]。该方法与以往研究中基于正确规则的方法有所不同，主要表现在以下两个方面：

第一，规则来源不同。下一步研究所基于的正确规则来自语料库驱动的型式语法(Pattern Grammar) [46] [47]。型式语法来自大规模真实语料，也就是说，所有型式(patterns)的背后都有大规模语料的支撑。可以说，作为语料库研究的理论成果，型式语法对语言细节描写的充分性和客观性是其他语言学理论所不及的。以往研究中的规则通常是基于直觉的规则，大多来自语言学家的直觉和内省，尤其是乔姆斯基学派的研究者们，更是“置内省证据于首要地位，注重研究者自己的直觉”[48]。这样的规则是建立在非自然语言基础之上的，因此，很难对自然语言做出充分客观的描写。

第二，规则获取方式不同。型式规则的获取涉及两方面的资源：一是大规模真实语料；二是语言学家自身的语言学修养。具体来说，在整个语法系列的编撰过程中，除了语料库驱动的方法，语言学家的人工处理也占据了很大的比重。Francis & Sinclair [49]也明确表示，型式语法创建的工作离不开语言学家的直觉。他们提倡使用真实语料，忠实语言事实，但这并不等于对直觉的压制或抛弃。语料库在辅助人们的语言直觉和内省的同时，离不开研究者本人的语言分析和判断能力，关键是要找到一个合理的平衡点[50]。与以往研究单纯依靠内省和直觉获取规则的方式不同，型式语法规则的获取是经验主义和理性主义的结合；不仅具有统计学意义，还具有很强的人文性，可以更好地服务于英语教学。

为了对上述假设进行验证，本研究选取了型式语法中的两个动词型式(V *so/not*, V n *for* n)，及其型式中的少量动词进行了试验。这样的动词型式错误常出现在中国学习者的作文中，学生很容易将动词和介词(或副词)的搭配用错，但是很多语法检查软件却无法查出。实验大致步骤如下：

首先是选取试验动词。从Francis *et al.* (1996)中随机选取V *so/not*、V n *for* n这两个型式下的动词若干。

其次，将链语法分析器(Link Grammar Parser)作为句法分析工具。根据链语法分析器的要求对动词下标进行相应的设置和添加。链语法词典中的单词具有唯一性，因此，有些用法较多的单词需要以下标作为区分，如 *run*. n 和 *run*. v 的下标就是用来区分词性的。由于型式语法对单词的划分主要是通过型式，因此笔者将动词所在的型式设计成了它们的下标，以传达型式语法“词汇和语法不可分”的理论主张。那么，动词型式 V *so/not* 和 V n *for* n 所属动词的分别是：“.v-sonot”和“.vn-for”，例如，*think*. v-sonot 和 *bring*. vn-for。

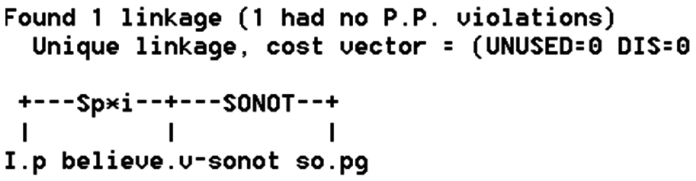

Figure 1. The analysis of verb pattern "V *so/not*"
图 1. 动词型式"V *so/not*"的分析结果

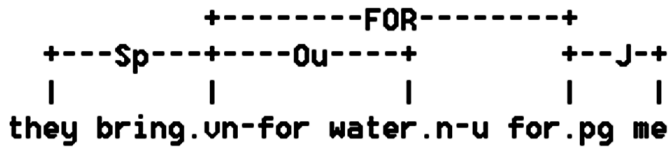

Figure 2. The analysis of verb pattern "V n *for* n"
图 2. 动词型式"V n *for* n"的分析结果

然后，为动词型式中具体的词项设计链名。由于所选取的两个动词型式中包含具体词项，如 *so/not*、*for*，而链语法形式化体系并没有为它们设计链接子，因此，笔者将这些具体词项本身设计成了它们的链接子，以表示它们在动词型式中的特殊作用。

将相关动词内容存入链语法词典之后，运行链语法分析器；分别输入两个包含有目标动词型式的句子："*I* ***believe*** *so.*"和"*They* ***bring*** *water for me.*"，获得结果如图 1、图 2。

由上述分析结果可以看出，将动词型式用于句法分析以及错误检查的方法是可行的。增加了动词型式规则的链语法分析器在分析包含目标动词的句子时，准确调用了本研究新编写的链语法规则，生成了非常具体的"SONOT"链和"FOR"链，为下一步的查错以及纠错提供了明确的指向。

基金项目

国家社科基金青年项目"西方媒体和中国外宣媒体笔下的"中国故事"叙事语篇结构关系研究"(项目编号：17CYY016)。

参考文献

[1] Naber, D. (2003) A Rule-Based Style and Grammar Checker. MS Thesis. Universität Bielefeld, Bielefeld, 4, 7.

[2] Vandeventer-Faltin, A. (2003) Syntactic Error Diagnosis in the Context of Computer-Assisted Language Learning. PhD Dissertation, University of Geneva, Geneva, 9, 43, 71.

[3] Han, N., Tetreault, J., Lee, S., *et al*. (2010) Using Error-Annotated ESL Data to Develop an ESL Error Correction System. *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (*LREC*), 763-770.

[4] Han, N., Chodorow, M. and Leacock, C. (2006) Detecting Errors in English Article Usage by Non-Native Speakers. *Natural Language Engineering*, **12**, 115-129. https://doi.org/10.1017/S1351324906004190

[5] MacDonald, N.H., Lawrence, T.F., Patricia, S.G., *et al*. (1982) The Writer's Workbench: Computer Aids for Text Analysis. *IEEE Transactions on Communications*, **30**, 172-179. https://doi.org/10.1109/TCOM.1982.1095380

[6] Pusack, J.P. (1984) Answer Processing and Error Correction in Foreign Language CAL. In: Wyatt, D.H., Ed., *Computer-Assisted Language Instruction*, Pergamon Press, Oxford, 63.

[7] Leacock, C., Chodorow, M., Gamon, M. and Tetreault, J. (2010) Automated Grammatical Error Detection for Language Learners. Morgan & Claypool Publishers, San Rafael, 5-6, 12-13.

[8] 冯志伟. 自然语言处理的形式模型[M]. 合肥: 中国科学技术大学出版社, 2010: 21, 33, 174.

[9] Dodigovic, M. (2005) Artificial Intelligence in Second Language Learning: Raising Error Awareness. Multilingual Matters Ltd., Clevedon, 108.

[10] 刘海涛. 依存语法的理论与实践[M]. 北京: 科学出版社, 2009: 156-157.

[11] Holland, M., Maisano, R., Alderks, C. and Martin, J. (1993) Parsers in Tutors: What Are They Good for? *CALICO Journal*, **11**, 28-46.

[12] Schulze, M. (1999) From the Developer to the Learner: Describing Grammar-Learning Grammar. *ReCALL Journal*, **11**, 117-124. https://doi.org/10.1017/S0958344000002159

[13] 龚小谨, 罗振声, 骆卫华. 中文文本自动校对中的语法错误检查[J]. 计算机工程与应用, 2003(8): 98-100, 127.

[14] Liou, H. (1991) Development of an English Grammar Checker: A Progress Report. *CALICO Journal*, **9**, 57-70.

[15] Rypa, M. and Feurman, K. (1995) CALLE: An Exploratory Environment for Foreign Language Learning. In: Holland, *et al*., Eds., *Intelligent Language Tutors*. Lawrence Erlbaum, Mahwah, 55-76.

[16] Matthews, C. and Fox, J. (1991) Foundations of ICALL—An Overview of Student Modeling. *Proceedings of Eurocall*, 163-170.

[17] Chen, S. and Xu, L. (1990) Grammar-Debugger: A Parser for Chinese EFL Learners. *CALICO Journal*, **8**, 63-75.

[18] Vosse, T. (1992) Detecting and Correcting Morpho-Syntactic Errors in Real Texts. *The Proceedings of the* 3*rd Conference on Applied Natural Language Processing*, Trento, 31 March-3 April 1992, 111-118. https://doi.org/10.3115/974499.974519

[19] Catt, M. and Graeme, H. (1990) An Intelligent CALL System for Grammatical Error Diagnosis. *Computer Aided Language Learning*, **3**, 3-26. https://doi.org/10.1080/0958822900030102

[20] Sanders, A.F. and Sanders, R.H. (1989) Syntactic Parsing: A Survey. *Computer and the Humanities*, **23**, 13-30. https://doi.org/10.1007/BF00058766

[21] Weischede, R.M., Voge, W.M. and James, M. (1978) An Artificial Intelligence Approach to Language Instruction. *Artificial Intelligence*, **10**, 225-240. https://doi.org/10.1016/S0004-3702(78)80015-0

[22] Schwind, C.B. (1995) Error Analysis and Explanation in Knowledge Based Language Tutoring. *Computer Assisted Language Learning*, **8**, 295-324. https://doi.org/10.1080/0958822950080402

[23] Yazdani, M. (1991) The LINGER Project: An Artificial Intelligence Approach to Second-Language Tutoring. *Computer Assisted Language Learning*, **4**, 107-116. https://doi.org/10.1080/0958822910040205

[24] Schneider, D. and McCoy, K. (1998) Recognizing Syntactic Errors in the Writing of Second Language Learners. *Proceedings of the* 36*th Annual Meeting of the Association for Computational Linguistics* (*ACL*) *and the* 17*th International Conference on Computational Linguistics*, 1198-1204.

[25] Carlberger, J., Domeij, R., Kann, V. and Knutsson, O. (2000) A Swedish Grammar Checker. http://www.csc.kth.se/tcs/projects/granska/rapporter/compling20000419.ps

[26] Rider, Z. (2005) Grammar Checking Using POS Tagging and Rules Matching. *Proceedings of the Class of* 2005 *Senior Conference*, 14-19.

[27] Murphy, M., Kruger, A. and Grieszl, A. (1998) RECALL—Providing an Individualized CALL Environment. In: Jager, S., Nerboone, J. and Essen, A.V., Eds., *Language Teaching and Language Technology*, Routledge, London & New York, 62-73.

[28] Sanders, R.H. (1991) Error Analysis in Purely Syntactic Parsing of Free Input: The Example of German. *CALICO Journal*, **5**, 77-86.

[29] Knight, K. and Chander, I. (1994) Automated Post Editing of Documents. *Proceedings of the* 12*th National Conference on Artificial Intelligence*, Seattle, 31 July-4 August 1994, 779-784.

[30] Chodorow, M., Tetreault, J. and Han, N. (2007) Detection of Grammatical Errors Involving Prepositions. *Proceedings of the* 4*th ACL-SIGSEM Workshop on Prepositions*, Prague, 28 June 2007, 25-30. https://doi.org/10.3115/1654629.1654635

[31] Gamon, M., Gao, J., Brockett, C., *et al*. (2008) Using Contextual Speller Techniques and Language Modeling for ESL Error Correction. *Proceedings of the International Joint Conference on Natural Language Processing*, Hyderabad, 7-12 January 2008, 449-456.

[32] Bergsma, S., Lin, D. and Goebel, R. (2009) Web-Scale n-Gram Models for Lexical Disambiguation. *Proceedings of the* 21*st International Joint Conference on Artificial Intelligence*, Pasadena, 11-17 July 2009, 1507-1512.

[33] 邢永康, 马少平. 统计语言模型综述[J]. 计算机科学, 2003, 30(9): 22-26.

[34] 邢富坤, 程东元. 基于统计语言模型的英语易读性研究[J]. 解放军外国语学院学报, 2010, 33(6): 19-24.

[35] Atwell, E.S. (1987) How to Detect Grammatical Errors in a Text without Parsing It. *Proceedings of the* 3*rd Conference of the European Association for Computational Linguistics*, Copenhagen, 1-3 April 1987, 38-45. https://doi.org/10.3115/976858.976865

[36] Nagata, R., Iguchi, T., Masui, F., *et al*. (2005) A Statistical Model Based on the Three Head Words for Detecting Ar-

ticle Errors. *Transactions on Information and Systems*, **7**, 1700-1706. https://doi.org/10.1093/ietisy/e88-d.7.1700

[37] Lee, J. and Seneff, S. (2006) Automatic Grammar Correction for Second-Language Learners. *Proceedings of the* 9*th International Conference on Spoken Language Processing*, Pittsburgh, 17-21 September 2006, 1978-1981.

[38] Yarowsky, D. (1994) Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French. *Proceedings of the* 32*nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, 27-30 June 1994, 88-95. https://doi.org/10.3115/981732.981745

[39] Foster, J. and Andersen, Ø. (2009) GenERRate: Generating Errors for Use in Grammatical Error Detection. *Proceedings of the* 4*th Workshop on Building Educational Applications Using NLP*, Boulder, 5 June 2009, 82-90. https://doi.org/10.3115/1609843.1609855

[40] Lee, J., Zhou, M. and Lin, X. (2007) Detection of Non-Native Sentences Using Machine-Translated Training Data. *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, 22-27 April 2007, 93-97. https://doi.org/10.3115/1614108.1614132

[41] Yi, X., Gao, J. and Dolan, W.B. (2008) A Web-Based English Proofing System for English as a Second Language Users. *Proceedings of the International Joint Conference on Natural Language Processing*, Hyderabad, 7-12 January 2008, 619-624.

[42] Chodorow, M. and Leacock, C. (2000) An Unsupervised Method for Detecting Grammatical Errors. *Proceedings of the Conference of the North American Chapter of the Association of Computational Linguistics*, 140-147.

[43] 李霞, 刘建达. 适用于中国外语学习者的英文作文全自动集成评分算法[J]. 中文信息学报, 2013, 27(5): 100-106.

[44] Gamon, M., Leacock, C., Brockett, C., *et al*. (2009) Using Statistical Techniques and Web Search to Correct ESL Errors. *CALICO Journal*, **26**, 491-511. https://doi.org/10.1558/cj.v26i3.491-511

[45] 张仰森, 丁冰青. 中文文本自动校对技术现状及展望[J]. 中文信息学报, 1998, 3(12): 50-56.

[46] Francis, G., Hunston, S. and Manning, E. (1996) Collins COBUILD Grammar Patterns 1: Verbs. HarperCollins, London.

[47] Hunston, S. and Francis, G. (2000) Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English. John Benjamins, Amsterdam/Philadelphia. https://doi.org/10.1075/scl.4

[48] 梁茂成. 理性主义, 经验主义与语料库语言学[J]. 中国外语, 2010(4): 90-97.

[49] Francis, G. and Sinclair, J.M. (1994) "I Bet He Drinks Carling Black Label": A Riposte to Owen on Corpus Grammar. *Applied Linguistic*, **15**, 190-200. https://doi.org/10.1093/applin/15.2.190

[50] Xiao, R. (2009) Theory-Driven Corpus Research: Using Corpora to Inform Aspect Theory. In: Lüdeling, A. and Kytö, M., Eds., *Corpus Linguistics*: *An International Handbook*, Volume 2, Mouton de Gruyter, Berlin, 987-1007.

Hans汉斯

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2161-8801，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：csa@hanspub.org