doi:10.6043/j.issn.0438-0479.201805007

现代汉语句子的扩展模式语法模型的构建

王笑盈[1],冯志伟[2],张 丹[1],瞿云华[1*]

(1.浙江大学外国语言文化与国际交流学院,浙江 杭州 310058;2.杭州师范大学外语学院,浙江 杭州 311121)

摘要:由于现有适用于自然语言处理的语法分析体系在分析汉语句式时无法准确体现汉语句子的特点,导致对汉语句子本体研究不够深入,限制了汉语自然语言处理各种应用的精度和速度.鉴于此,为了服务自然语言处理和汉语句式研究,提出构建一种新的分析汉语句式的语法体系——现代汉语句子的扩展模式语法模型.本模型对 Susan Hunston 提出的模式语法进行扩展,植入广义话题理论,对二者进行整合,凸显汉语句子特征,并设置了标点句分析模块和话题非自足句分析模块.现代汉语句子的扩展模式语法模型能够准确而全面地描述和归纳汉语句式规则,体现汉语句式中虚词与实词之间的限定关系,反映句式的线性序列,并提高汉语流水句的分析质量.

关键词:现代汉语句子;扩展模式;语法模型;汉语句子特征;自然语言处理

中图分类号:H 043 **文献标志码**:A **文章编号**:0438-0479(2018)06-0859-08

句子是语言的基本运用单位,以特定句式构造而成[1].句式是句子的语法结构格式,它是从大量句例中抽象概括出的、具有共性和本质性的语法结构格式.句式研究是汉语语法研究中至关重要的一部分,是汉语学界的重要议题.对汉语句式进行研究,进而构建汉语句式系统,一方面可满足深层理解汉语的自身需求;另一方面可直接服务于汉语自然语言处理的各种应用,如汉字输入、语音识别、文本检索、汉语自动分词、信息抽取、机器翻译与问答系统等.但是,现有适用于自然语言处理的语法分析体系在分析汉语句式时都无法准确体现出汉语句子的 3 个重要特点:1) 汉语作为非形态语言,语序是其意义的载体,汉语语序既相对固定,又便于灵活变换构成新的结构关系[1-3];2) 虚词是汉语的重要语法手段,虚词对实词的使用具有选择和限定的关系[2,4];3) 汉语注重话题,结构松散,多流水句[5-7].现有语法分析体系的缺陷不仅导致对汉语句子本体研究不够深入,而且也限制了汉语自然语言处理各种应用的精度和速度.鉴于此,急需开发一套能够准确全面描述和分析汉语句式的语法体系、归纳汉语句式规则、服务汉语句式研究和自然语言处理的系统.

目前国内外自然语言处理领域中应用最广的语法分析体系是短语结构语法和依存语法.短语结构语法[8](phrase structure grammar)是乔姆斯基用数学方法研究自然语言和人工语言的语法理论,其基本思想是句子由短语结构组成.短语结构分为两大类型:名词性短语结构(NP)和谓词性短语结构(VP),S 代表句子,S=NP+VP.短语结构语法能够识别出句子的语序、层次和词类信息.方立等[9-11]介绍了短语结构语法在汉语中的应用;也有一些学者利用短语结构语法对汉语进行分析,姚小烈[12]探索了汉语“的”字结构,郑友阶[13]考察了汉语同位句,张莹[14]研究了汉语人称代词相关句法问题,刘亮[15]分析了汉语轻动词体系,上述研究考察了某些特定的汉语句子结构.同时,短语结构语法在汉语句法自动分析中有广泛的应用,如宾州中文树库(Chinese Penn treebank)、斯坦福中文句法分析器(Stanford parser)、Readworld 语言信息处理研究院研制的短语结构语法分析器.短语结构语法能够体现句子线性顺序和层级结构,但也有不尽如人意之处:它的树形图是单标记的(如词类标记 N,V;词组类型标记 NP、VP 等),这使得它难以表达纷繁复杂的自然语言现象,分析能力过弱[16];短语结构树中

收稿日期:2018-05-04 **录用日期**:2018-10-21

基金项目:国家社会科学基金(17BYY002)

* **通信作者**:qu163hua@163.com

引文格式:王笑盈,冯志伟,张丹,等.现代汉语句子的扩展模式语法模型的构建[J].厦门大学学报(自然科学版),2018,57(6):859-866.

Citation:WANG X Y,FENG Z W,ZHANG D,et al. Modern Chinese Sentence Extended Pattern Grammar Model for Natural Language Processing[J]. J Xiamen Univ Nat Sci,2018,57(6):859-866.(in Chinese)

标有大量不在句中出现的非终极结点(NP、VP 等),层次过多;短语结构树无法体现句中各终极结点之间的支配关系[17].

依存语法[18](dependency grammar)是泰尼埃提出的基于词间关系的语法,强调动词为句子中心,其他词汇依存于动词.依存语法的优势体现在依存树的层次少、结点少,能够清晰地表示句中各词之间的依存关系.汉语学者曾对依存语法进行了综述,如冯志伟[19-20]、刘海涛[21];不少学者采用依存语法研究汉语中的名词短语[22]、句法歧义结构[23]、并列结构[24],或构建长句分析多视图汉语树库[25]等.但是,依存语法忽视了自然语言的语序特点:结点之间的支配关系不能直接推导出它们之间的前于关系.

总之,短语结构语法和依存语法在分析汉语句子时均存在以下缺陷:未能凸显虚词对实词的限定关系,对汉语流水句的处理也差强人意.

模式语法①(pattern grammar)是 Hunston 等[26]提出的语言描述模型.模式是以特定方式共现的、两个或更多虚词与词类以及词类之间的组合,能够呈现语言真实的线性序列.如:标点句"曾是个运动员"的模式为"曾…vshi…个…n".模式语法采取新颖的编码方式,使用具体词形(曾、个)、词类标签(vshi、n)而非传统的语法功能范畴标签(NP、VP),清晰地描述了词类之间的关系及具体虚词与实词词类之间的关系,呈现了句子结构信息.在一定程度上,增补了上述语法缺少具体虚词对于实词词类限定关系描述的功能.利用这一限定关系,可在汉语自然语言处理中提高对实词的预测准确度,降低计算复杂度与计算量,从而提升汉语句式分析质量.此外,模式的深度浅、长度短、出现频率高,更适合作为语言特征构建自然语言处理中的语言模型.

采用模式理论对英语进行的语法分析已经发挥了重要作用.英国陆续出版了基于模式的词典与参考语法书,如《The Collins cobuild English language dictionary》[27]、《The Collins cobuild English grammar》[28]、《The Collins cobuild English dictionary》[29]等,这些成果侧重描述在真实语言中常见的、典型的英语模式,增强学习者识别和使用英语词汇及结构的能力.目前模式语法在国内的应用仍处于起步阶段:王勇[30]对模式语法进行了介绍;陈功等[31]综述了模式语法的产生、特点和应用价值.个别研究者将模式语法应用在英语研究中,如:Huang 等[32]将模式语法运用到学习者语法检查系统中;陈功[33]构建了一个面向中国学生的英语书面语动词形式错误检查系统;熊思尘[34]在不同语域中研究了英语"it 模式";于涛[35]开发了一套自动识别和抽取英语动词模式的程序等.但迄今为止,还没有研究者将模式语法应用于汉语句式研究.但模式语法也有其局限性:短语结构语法和依存语法的研究对象是整个句子,而 Hunston 等[26]描述的模式不包括主语和表示时间、地点及方式的状语,因此模式语法的研究范围仅限于句子片段.若将模式语法应用于汉语句式研究,必须对其理论进行扩展和补充以适应汉语句子层面的研究.

综上所述,为了对汉语句式进行深入研究,同时将研究成果服务于自然语言处理,本研究提出构建能够凸显汉语句子特征的"现代汉语句子的扩展模式语法模型"(后文简称为扩展模式语法模型),全面描述汉语句式规则,构建汉语句式体系.

1 扩展模式语法模型构建

1.1 理论基础

针对汉语句式研究,本研究构建扩展模式语法模型.该模型以模式语法[26]和广义话题理论[36]为基础,并在汉语句式研究背景下对二者加以扩充、改进和整合.

模式语法诞生于语料库研究,基于大量真实语料对语言进行可靠性描述.模式语法是对 Firth[37]的"搭配(collocation)"研究和 Sinclair[38]的"成语原则(idiom principle)"思想的继承和发展.该语法继承了 Sinclair 的思想,认为语言具有短语倾向,即词语不是孤立存在,而是通过它们的共选关系而获得意义.与此同时,模式语法将词汇看成语言的核心,句法结构和词项之间具有共选关系,不能将其分开考察,具体表现为:一方面,特定的句法结构通常与特定的词项共现;另一方面,词项通常只出现在有限的结构中[26].模式语法最初在描述时着眼于短语层面,因此模式缺失了表示主语和表示时间、地点及方式的状语等元素.但这些元素在汉语句式研究中是不可或缺的.鉴于此,本研究以模式语法的核心思想为理论指导,扩充模式元素,进而提出扩展模式语法,将扩展模式的研究范围提升至句式研究层面.

话题现象是汉语语法的重要特点.赵元任[6]指出在汉语中,把主语和谓语当作话题和说明来看待比较合适.考虑到话题是汉语句子的突出特征,本模型在扩展模式语法的基础上引入"广义话题理论"[39]以深入挖掘汉语句式的话题信息.该理论根据汉语

① 由于本文中旨在研究汉语句式,因此仅关注模式语法与句式相关的特征.

篇章特点,将实体、时间/处所、状性/谓性/推理前提等纳入话题范围[36],以边界明确的标点句为基本单位,阐述了汉语话题结构和话题句特征[40]. 话题自足句对于汉语篇章信息处理有重要意义,话题自足句中话题与说明成对出现,结构相对完整,许多上下文信息在话题自足句中已经聚集在一起,处理话题自足句可以提高汉语信息处理应用系统的性能[41]. 若仅把标点句作为完整的句子来处理显然会严重影响汉语信息处理系统的性能,是汉语句子自动分析和机器翻译质量较差的主要原因[36]. 本模型的话题非自足分析模块将标点句转为语法通顺、语义清楚的话题自足句,从而分析和归纳汉语句式类型和特征.

1.2　构建思路

在模式语法与广义话题理论的基础上构建扩展模式语法模型,以弥补短语结构语法、依存语法与模式语法在描述汉语句式时的局限性. 本研究在构建扩展模式语法时,利用基于短语结构语法和依存语法的句法分析器分析出句法树,并将句法树转换为扩展模式语法的表现形式. 在转换过程中,保留了短语结构语法能够体现句子线性顺序和层级结构的优势,同时借鉴了依存语法非终极结点少的优点,并在此基础上融入了扩展模式语法的编码方式,一方面凸显虚词与实词的限定关系,另一方面直观反映句子各元素的线性序列而非非终极结点的序列. 从本质上讲,扩展模式语法与短语结构语法、依存语法是不矛盾的,扩展模式语法在继承此两种传统语法精髓的基础上,进行了表现形式的变化,为汉语句式研究提供了新的观察视角.

本研究利用现代汉语书面语与口语平衡语料库,通过对把字句扩展模式进行实例分析,探究在利用扩展模式语法分析汉语句式时较传统语法的优势,从而提出现代汉语句子的扩展模式语法类型. 如果在本模型增加句子必有成分与汉语特色结构,对语料进行模式标注,将实词标为词类,虚词标为具体词形,就有望在今后最终建立汉语句子模式树库,总结出一套凸显汉语特征的句子模式规则.

1.3　语法模型

为了同时服务于汉语句式理论研究和自然语言处理的实际应用,本研究提出建立新的句式分析模型——“现代汉语句子的扩展模式语法模型”. 本模型是一种基于模式语法的考察汉语句式全貌的语言描述模型,关注汉语的语序特点和虚词对实词的限定关系,同时提高汉语流水句的分析质量,以期最终服务于自然语言处理. 本模型包含两大模块:标点句分析模块与话题非自足句分析模块,能够对标点句和话题非自足句的句式进行统一分析. 标点句分析模块对模式语法进行扩充,将表示主语、时间、地点及方式的介词短语和副词短语纳入考察范围,本模块能够描述汉语句式构成规则,构建汉语句式体系;在此基础上,话题非自足句分析模块引入广义话题理论,补全标点句话题,考察汉语流水句的话题-句式特征,进而提高流水句的自动处理质量. 模型具体内容见图 1.

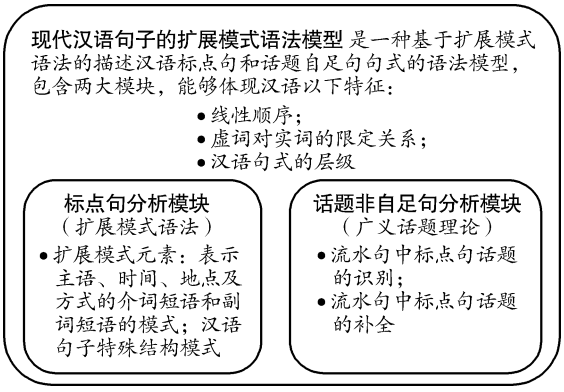

图 1　扩展模式语法模型

Fig. 1　Sentence extended pattern grammar model

1.4　语法模型与句式研究

句式研究是现代汉语语法研究中的核心课题之一. 早在 20 世纪 70 年代,以朱德熙、吕叔湘、陆俭明、范晓、张斌为首的语法学家都非常关注汉语句式研究. 朱德熙[42]曾将句式定义为“代表这个句子的有一定的层次构造和内部结构关系的抽象的词类序列”;张斌等[43]在《现代汉语句子》一书中指出句式反映了句子内部层次、关系、成分和特殊标志;陆俭明[44]认为同一句式通常具有相同的词类序列、词语、构造层次和内部语法结构,并指出范晓对汉语句式及其特征做了较为全面的综述;范晓[45-46]基于前人对汉语句式的研究成果,总结出汉语句式的四大特征,即线条性、有序性、层次性和抽象性;句式的形式主要以词类、词类排序、特定词语、层次分合、固定格式等体现.

本研究提出的扩展模式语法模型中,句式的形式以具体虚词、实词词类、具体词和词类排序、层次分合、汉语特殊结构等体现,凸显汉语虚词对实词的限定作用,能够体现出汉语学界认可的句式特征:

1) 线条性. 句式一般由一个或多个汉语扩展模式组成,表现为扩展模式按照语流的先后顺序排列成线. 如扩展模式"他…挺…adj"("他挺高"),扩展模式元素他、挺、adj 按照先后顺序成串出现.

2) 有序性. 句式内部的扩展模式不是孤立的,而是互相联系互相制约的,有一定的排列规则. 改变了排列次序句式也就改变了,表达的意义也随之改变. 如扩展模式"她…v…n"("她教学生"),如果改变为"n…v…她"("学生教她"),该句式的句式意义随之改变.

3) 层次性. 句式中扩展模式的内部结构关系具有层次性,在本模型中体现为模式流、模式线和模式环:模式间没有重叠称为模式线(pattern string),有重叠称为模式流(pattern flow),大模式嵌套小模式称为模式环(pattern loop). 如标点句"你应该掌握提问的技巧",其句式的扩展模式层次体现为:

你…v

v…v…n

v…的…n

扩展模式"r…v"和"v…v…n"具有上下层级的关系,且这两个模式有重叠部分"v",因此它们共同组成了模式流;v…的…n"体现了"n"的内部构成,因此它们共同组成模式环.

4) 抽象性. 句式是不同内容的具体句的相同语法组合格式的集合. 同一句式下的具体句中同一位置上的词或词类具有替换关系. 例如标点句"我把苹果吃了"和标点句"你把作业写了",其句式的扩展模式类型均为"你…把…n…v…了". 在该扩展模式中,同一位置上的"你""我"可统一抽象为元素"v",同理"苹果"和"作业"抽象为"n","吃"和"写"抽象为"v".

2 实例分析

在扩展模式语法模型分析句式的框架下,本文中对汉语"把"字句的句式进行了分析. "把"字句是现代汉语中极其常用且比较复杂的句式,一直是汉语语法学界一个重要研究课题. 张伯江[47]将"把"字句句式概括为"A 把 B v C",其意义为由 A 作为起因,针对选定对象 B 以 v 的方式进行的使 B 变化为 C 的行为. 本文中基于前人对"把"字句的研究,在语料库中提取"把"字句,在扩展模式语法模型下探索"把"字句更加细化和准确的句式类型.

2.1 语料库

本研究采用总规模为 220 万词次的浙江大学现代汉语书面语与口语平衡语料库中的普通小说和新闻评论子库. 普通小说子库约 5 万词次,新闻评论子库约 5.3 万词次. 语料均经过中国科学院计算技术研究所 ICTCLAS 系统分词、标注,经过人工检查和修正后其准确率达 98%以上.

2.2 提取方法

在提取"把"字句扩展模式时,开放词类标注为词性,封闭词类标注为具体词形,其中开放词类包括名词、动词、形容词、描摹类副词、区别词、处所词、状态词、拟声词、时间词,封闭词类包括介词、助词、连词、代词、方位词、描摹类之外的副词、数词、量词、感叹词、语气词. 在自建的 220 万词次现代汉语书面语与口语平衡语料库中提取句子扩展模式,得到汉语句子扩展模式库. 基于扩展模式树库,在普通小说和新闻评论子库中抽取"把"字句模式. 提取时排除"把"作为量词的结构,如"二把手""推了一把";排除"把"作为动词的结构,如"严把质量关". 共得到"把"字句 265 句,其中 150 句来自普通小说子库,115 句来自新闻评论子库.

2.3 归纳分析

通过对普通小说和新闻评论子库中抽取的 265 个"把"字句的总结,共归纳出"把"字句扩展模式类型 11 类,模式实例和具体"把"字句见表 1,包含了"把"字句中时间、地点及方式的介词短语和副词短语的模式分析. 对于这些句式实例化后的话题非自足句补足话题分析,详见扩展模式语法优势第 3 点.

为了便于归类,表 1 第 1 列为"把"字句的扩展模式,第 2 列为具体"把"字句. 通过观察可发现依据扩展模式语法模型描述"把"字句,能够完全反映出汉语"把"字句句式的特征.

首先,"把"字句的线条性体现为各类"把"字句扩展模式均按照汉语语流线性排列,例如"把"字句"把瓦罐收起来"的扩展模式为"把…n…v…vf",模式中的各元素之间的顺序均按照语流排列;其次,"把"字句有序性体现在模式内部元素按照规则有序排列,且相互制约,这种规则体现为若改变排列次序,则句子的意义也随之改变或不符合汉语语法,例如"把"字句"把问题搞清楚",其扩展模式为"把…n…v…adj",若改变其中任意元素的位置,则会导致该句意义不明且违反了语法规则;再次,"把"字句的扩展模式也具有层次性,其中的名词性元素"n"在部分情况下可能代表名词性短语,如果深入研究"n",就能发现它与"把"字句的扩展模式形成上下层级的关系,如"把深圳建成

表 1　普通小说和新闻评论语域中“把”字句的扩展模式

Tab. 1　Ba extended patterns in general novel and news editorial subcorpora

“把”字句扩展模式	“把”字句实例
把…n…v…vf	把 瓦罐 收 起来
把…n…vf	把 门 关上
把…n…往…n…v	把 人 往 房间 拉
把…n…v…n…里	把 苦根儿 拉到 棉花地 里
把…n…v…向…里	把 帽子 抛 向 天空
把…n…v…mq	把 赵一 揍 一顿
把…n…v…v	把 男孩 弄 哭
把…n…v…adj	把 问题 搞 清楚
把…n…v…v…n	把 规章 上升 为 法律
	把 他 接到 美国
把…n…v…n	把 深圳 建成 国际城市
	把 统一进程 纳入 轨道
	把 自由 视为 生命
把…他…v	把 他 抱住
	把 他 吓死

国际城市”的扩展模式为“把…n…v…n”,其中最后一个元素 n 实质为名词短语“国际城市”,若进行具体分析可进一步将 n 扩展为“n…n”;最后,把字句的扩展模式具有抽象性,具有开放性的实词标记为词类,封闭性的虚词标记为具体词形,内容不同但语法格式相同的“把”字句可抽象为同一“把”字句模式,例如句子“把他抱住”、“把他吓死”的扩展模式均可归纳为“把…他…v”. 因此,扩展模式语法适用于汉语句式研究.

3　相对于传统语法的优势

利用扩展模式语法模型研究汉语句式相对于短语结构语法和依存语法等传统语法具有以下优势:

1) 扩展模式语法着重体现了句式中虚词与实词之间的限定关系. 短语结构语法和依存语法没有强调汉语虚词与实词的限定关系. Hunston 和 Francis[26] 指出语言中虚词对其后的实词类型有限定作用. 文本构建的扩展模式语法使用词类或具体词形的标注方式,这样的编码方式凸显了汉语虚词对实词的限定关系,利用这一限定关系可在汉语自然语言处理中提高对实词的预测准确度,消解部分歧义,降低计算复杂度与计算量. 以“把”字句“把人往房间拉”为例,图 2 为短语结构语法和扩展模式语法的分析结果.

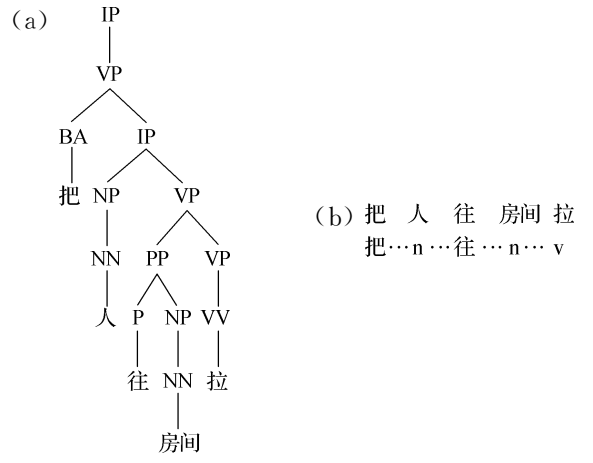

图 2　短语结构语法(a)和扩展模式语法(b)分析结果

Fig. 2　Analysis of phrase structure grammar (a) and extended pattern grammar (b)

从图 2(a)可以看出,短语结构语法关注语法范畴 NP 与 VP 内部的层级关系,但不强调各词之间的限定关系;在图 2(b)的扩展模式“把…n…往…n…v”中,虚词“把”“往”以具体词形体现,实词“人”“房间”以词类“n”体现. 图 2(b)不仅能反映出句中 5 个组成元素的层级关系,并且能凸显其中虚词与实词之间的限定关系,尤其能体现出介词“往”对其后元素的限制:介词“往”表示动作的方向,通常将其后出现的名词限制为方位词、方位短语、处所名词、方所指示代词,且一般搭配的动词是位移性方向动词. 在本例中,介词“往”之后出现的名词是处所名词“房间”,搭配的动词是位移方向动词“拉”. 由此可见,扩展模式语法在捕捉虚词和实词的互选关系时更具优势.

2) 扩展模式语法能够直观反映句式的线性序列. 线性是语言的基本属性[48]. 尽管依存语法以层次少、结点少、体现词与词之间的支配关系的特点弥补了短语结构语法的不足,但依存树中结点之间的支配关系和前于关系是互相排斥的,只有把表示结构关系的依存树转变成表示线性关系的句子才能推导出句子结点之间的前于关系[49]. 以“把”字句“把纸飞机抛向窗户外面”为例,采用两种分析结果如下:

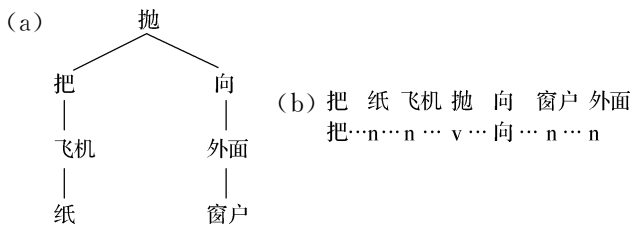

图 3　依存语法(a)和扩展模式语法(b)分析结果

Fig. 3　Analysis of dependency grammar (a) and extended pattern grammar (b)

从图 3(a)看出,依存树明确体现出词与词之间的支配关系:核心动词“抛”支配“把”和“向”,“把”和“向”分别支配“飞机”和“外面”,“飞机”和“外面”分别支配“纸”和“窗户”.但是依存树无法反映出“纸”前于“飞机”、“窗户”前于“外面”的句子真实语序.图 3(b)中的扩展模式体现了把字句的真实语序.

3) 依据扩展模式语法补全标点句话题,继而研究汉语句式,有助于提高汉语流水句的自动分析质量.补全话题后标点句在句法和语义上都是完整的单句,话题与说明成对出现,上下文信息集中,这可以解决涉及汉语局部篇章的问题[40].下文通过对比补全话题前后标点句的译文,证明补全话题有利于提高汉语流水句的自动分析质量.表 2 对比了流水句补全话题前后谷歌翻译的质量,下划线部分是补全的话题.

表 2 流水句补全话题前后译文对照表

Tab. 2 Comparison of translation performance of topic sufficient sentences and topic deficient sentences

话题	流水句	谷歌译文
补全话题	俺哥痛得厉害, 俺哥有天把俺们叫到床前, 俺哥嘱咐俺们快去找大夫.	My brother had a terrible pain, and my brother called me to the bed one day, and my brother asked me to go find the doctor.
未补全话题	俺哥痛得厉害, 有天把俺们叫到床前, 嘱咐俺们快去找大夫.	My brother suffered painful, one day we called bed, asked to find the doctor.

将表 2 中标点句补全话题后得到话题自足句.“俺哥”是流水句中各标点句的话题.观察未补全话题流水句的谷歌翻译结果,可发现译文存在句子成分缺失严重、可读性差的问题;相比之下,补全话题后的标点句译文质量明显提升,各标点句均翻译准确,语法正确,可读性高.翻译质量的提高表明机器对汉语流水句原文的分析质量显著提升,进而说明补全话题是提高汉语流水句自动分析质量的有效途径.

4 结 论

本研究构建了扩展模式语法模型.该模型内置了标点句分析模块与话题非自足句分析模块:标点句分析模块含有表示主语、时间、地点及方式的介词短语和副词短语的模式,以及汉语句子特殊结构模式的功能;话题非自足句分析模块能够识别和补全流水句的话题,并对标点句和话题非自足句的句式进行统一分析,考察汉语流水句的话题-句式特征.

本模型从全新角度分析汉语句式,用扩展模式的线性组合和层级构造描述句式,并凸显汉语虚词对实词的限定关系,同时可以对汉语句式进行系统的归纳和总结.通过详细梳理该模型的背景、概念、特征及优势,认为扩展模式语法模型将补充和发展现有的汉语语法体系,对机器和人学习汉语句子都具有可观的发展前景和应用价值:

1) 扩展模式语法模型的构建将为描述汉语语法提供新视角.在理论创新方面,本模型结合模式语法与广义话题理论,补充句子必有成分,关注汉语虚词对实词的限定关系以及语序特点和流水句的结构特征;在方法论方面,本模型采用语料库方法,观察和归纳汉语句子模式的结构类型,以期弥补传统汉语句法分析多基于经验、无大量数据支持的不足;与此同时,本模型运用自然语言处理中的前沿算法开发识别和补全话题、提取模式、检索统计模式的程序,能够高效分析汉语话题自足句的模式类型和结构特征.

2) 本模型研究成果将服务于自然语言处理中语言模型的构建.传统基于规则及基于统计的语言模型均未重视汉语中虚词对实词的限制作用,导致运算量过大或精度较低.扩展模式语法模型因其虚实结合、线性顺序的特点能够提高虚词对实词的预测准确度,降低计算复杂度与计算量.与此同时,本模型还能够提高汉语长句自动分析的质量和汉语句子处理系统的性能,降低汉语篇章处理的难度.

3) 本模型研究成果将为汉语作为第二语言的教学提供新视角.相较传统句法分析,扩展模式语法模型产出的句子结构规则更加简洁,可以帮助汉语学习者把握句子的典型用法、提高语言产出的准确性和流利性.

参考文献:

[1] 文炼,胡附.汉语语序中的几个问题[J].中国语文,1984,3:1-5.

[2] 连淑能.英汉对比研究[M].北京:高等教育出版社,1993:21.

[3] 王力.中国语法理论[M].北京:中华书局出版社,2015:48.

[4] 陆俭明,马真.现代汉语虚词散论(修订版)[M].北京:语文出版社,2003:1.

[5] 吕叔湘.汉语语法分析问题[M].北京:商务印书馆,1979:97-99.

[6] 赵元任.汉语口语语法[M].北京:商务印书馆,1979:

41-44.
[7] LI C N, THOMPSON S A. Subject and topic: a new typology of language[M]. New York: Academic Press, 1976:457-489.
[8] 林杏光.短语结构语法:"信息处理用语言理论讲话"第一讲[J].语言文字应用,1994(2):58-64.
[9] 方立,吴平.中心语驱动短语结构语法评介[J].语言教学与研究,2003(5):31-43.
[10] 范子衿,王惠临,张均胜.中心语驱动短语结构语法研究综述[J].现代图书情报技术,2013(4):40-47.
[11] 满海霞,梁雅梦.乔姆斯基层级与自然语言语法:从短语结构语法到非转换语法[J].外国语文,2015,31(3):84-89.
[12] 姚小烈.生成语法视角下的现代汉语"的"字结构研究[D].北京:中央民族大学,2010:60-113.
[13] 郑友阶.同位语句法研究[D].武汉:华中师范大学,2013:18-77.
[14] 张莹.现代汉语人称代词及相关句法问题[D].武汉:华中师范大学,2013:103-116.
[15] 刘亮.现代汉语轻动词体系研究[D].上海:华东师范大学,2015:43-67.
[16] 冯志伟.计算语言学基础[J].方言,2002(4):384-384.
[17] 冯志伟.应用语言学综论[M].广州:广东教育出版社,1999:254-262.
[18] TESNIÈRE L. Éléments de syntaxe structurale[M]. Paris:Klincksieck,1959:97-186.
[19] 冯志伟.特思尼耶尔的从属关系语法[J].当代语言学,1983(1):63-65.
[20] 冯志伟.泰尼埃与依存语法:纪念泰尼埃逝世 60 周年[J].现代语文(语言研究版),2014(11):4-9.
[21] 刘海涛.依存语法的理论与实践[M].北京:科学出版社,2009:76-166.
[22] 赵军,黄昌宁.汉语基本名词短语结构分析模型[J].计算机学报,1999,22(2):141-146.
[23] 苑春法,黄锦辉,李文捷.基于语义知识的汉语句法结构排歧[J].中文信息学报,1999,13(1):1-8.
[24] 赵怿怡,高松,刘海涛.基于依存语法的汉语并列结构自动分析研究[C]//中国计算机语言学研究前沿进展.北京:清华大学出版社,2009:148-153.
[25] 邱立坤,金澎,王厚峰.基于依存语法构建多视图汉语树库[J].中文信息学报,2015,29(3):9-15.
[26] HUNSTON S,FRANCIS G. Pattern grammar:a corpus-driven approach to the lexical grammar of English[M]. Amsterdam:Benjamins,2000:37-271.
[27] COLLINS J. Collins cobuild English language dictionary [M]. Glasgow:Collins,1987.
[28] COLLINS J. Collins cobuild English grammar[M]. London:HarperCollins,1990.
[29] COLLINS J. The Collins cobuild English dictionary[M]. London:HarperCollins,1995.
[30] 王勇.行走在语法和词汇之间:型式语法述评[J].当代语言学,2008(3):257-266.
[31] 陈功,梁茂成.型式语法的产生、特点及其应用价值[J].外语学刊,2017(1):17-24.
[32] HUANG C C,CHEN M H,HUANG S T,et al. Edit:a broad-coverage grammar checker using pattern grammar [C]//Meeting of the Association for Computational Linguistics:Human Language Technologies(Systems Demonstrations). Portland: Association for Computational Linguistics,2011:26-31.
[33] 陈功.中国学习者英语书面语动词形式错误自动检查:一项基于链语法的研究[D].北京:北京外国语大学,2012:1-268.
[34] 熊思尘.基于语料库的带 it 动词型式与语域的相互作用研究[D].杭州:浙江大学,2014:32-58.
[35] 于涛.基于索引行聚类的英语动词型式自动识别与提取研究[D].北京:北京外国语大学,2015:41-55.
[36] 宋柔.汉语篇章广义话题结构的流水模型[J].中国语文,2013(6):483-494.
[37] FIRTH J R. Modes of meaning[C]//Papers in Linguistics 1934—1951. London: Oxford University Press, 1957: 190-215.
[38] SINCLAIR J M. Corpus, concordance, collocation[M]. Oxford:Oxford University Press,1991:67-98.
[39] SONG R, JIANG Y, WANG J. On generalized-topic-based Chinese discourse structure[C]//CIPS-SIGHAN Joint Conference on Chinese Language Processing. Beijing: Chinese Information Processing Society of China,2010:23-33.
[40] 蒋玉茹,宋柔.基于广义话题理论的话题句识别[J].中文信息学报,2012,26(5):114-119.
[41] 卢达威,宋柔,尚英.从广义话题结构考察汉语篇章话题认知复杂度[J].中文信息学报,2014,28(5):112-124.
[42] 朱德熙.句子和主语:印欧语影响现代书面汉语和汉语句法分析的一个实例[J].世界汉语教学,1987(3):31-34.
[43] 张斌,陈昌来.现代汉语句子[M].上海:华东师范大学出版社,2000:388-438.
[44] 陆俭明."句式语法"理论与汉语研究[J].中国语文,2004(5):412-416.
[45] 范晓.关于句式问题:庆祝《语文研究》创刊 30 周年[J].语文研究,2010(4):1-11.
[46] 范晓.关于句式的几点思考[J].汉语学习,2013(4):3-12.
[47] 张伯江.论"把"字句的句式语义[J].语言研究,2000(1):28-40.

[48] SINCLAIR J M, MAURANEN A. Linear unit grammar: integrating speech and writing[M]. Amsterdam: John Benjamins, 2006: 49-91.

[49] 冯志伟. 现代语言学流派(增订本)前言[J]. 现代语文(语言研究版), 2014(3): 161.

Modern Chinese Sentence Extended Pattern Grammar Model for Natural Language Processing

WANG Xiaoying[1], FENG Zhiwei[2], ZHANG Dan[1], QU Yunhua[1*]

(1. School of International Studies, Zhejiang University, Hangzhou 310058, China;
2. School of Foreign Languages, Hangzhou Normal University, Hangzhou 311121, China)

Abstract: The existing grammar analysis system for natural language processing cannot accurately reflect characteristics of Chinese sentences when we analyze Chinese sentence patterns, resulting in inadequate theoretical studies on Chinese grammar and limiting the precision and speed of Chinese natural language processing applications. In view of this, to serve natural language processing and Chinese sentence pattern research, we propose to construct a new grammar system for analyzing Chinese sentence patterns, which is called the modern Chinese sentence extended pattern grammar model. This model extends the pattern grammar proposed by Susan Hunston and implants the generalized topic theory. This model, with a punctuation sentence analysis module and a topic insufficient sentence analysis module, integrates these two theories and highlights Chinese sentence features. This model can accurately and comprehensively describe and generalize Chinese sentence patterns, reflecting the interaction between functional words and content words, the linear sequence of sentence patterns, and improving the analytical quality of Chinese flowing sentences.

Key words: modern Chinese sentence; extended pattern; grammar model; Chinese sentence features; natural language processing