doi:10.13582/j.cnki.1672-7835.2023.02.007

面向 AI 新时代的多模态组合范畴语法

罗丹
(湖南工学院 马克思主义学院,湖南 衡阳 421002)

摘　要:在大数据与大知识双轮驱动的 AI 新时代,自然语言处理遭遇到前所未有的冲击和挑战。多模态组合范畴语法是标准组合范畴语法的升级与优化。它继承了标准组合范畴语法的内部构造和运行方式,保留了原有的计算特性,同时增加了利用新策略的可能性。它最重要的创新在于,添加了模态算子并且融入了广义斯科仑项技术,使之更加符合普遍语法特征,具有跨语言通用性,具有更强的描写能力和解题功能,能够更好地实现自然语言处理,契合新一代 AI 的发展趋势。

关键词:AI 新时代;自然语言处理;多模态组合范畴语法;广义斯科仑项;量化辖域消歧
中图分类号:H030　　**文献标志码**:A　　**文章编号**:1672-7835(2023)02-0050-07

一　双轮驱动的 AI 新时代及其挑战

进入 21 世纪以来,AI 发展日新月异,取得了一系列重大突破。特别是基于深度神经网络的基础模型技术,引导 AI 走向了大数据与大知识的双轮驱动,自主智能化成为主流,AI 越来越多地走进日常生活,AI 主流技术的第四次创新到来了①。但在一路高歌猛进的热潮背后,我们应该清醒地认识到 AI 的技术瓶颈,特别是在自然语言理解、语言决策分析等基础层面举步维艰,亟待新的突破。究其原因,部分在于当前的 AI 缺少对信息的深度加工、理解和思考,所做的只是相对简单的比对与识别,仍停留在对虚拟符号特定关系的“感知”,尚未达到对物理世界的“认知”,更谈不上真实生活场景的“具身认知”。因此,AI“自主智能化”才刚刚起步,依然任重道远。

AI“自主智能化”的前提条件是,AI 必须能够像人类一样理解语言,通过对语言的“认知”来实现对世界的“认知”。人类借助语言来表征世界、交流思想。借用维特根斯坦的话来说,语言是世界的图像,语言的界限是世界的界限,也是思想的界限。人类语言不仅是一套符号系统,具有相对稳定的句法结构,而且还承载着丰富的语义信息,与世界保持着密切联系。理解语言的关键在于理解语言所携带的语义信息。如何理解或者捕获语言所携带的语义信息呢?一种常见且简单的办法是诉诸句法分析。此种进路预设了句法优先并且句法与语义之间存在一一对应关系。它的优点在于能够建立高度形式化的句法系统,揭示出自然语言的无穷生成机制,有利于对自然语言进行大规模处理。但问题也恰恰在于,自然语言从来不是一种清晰明确的交流工具②,它具有复杂多样性,存在许多句法与语义不对称现象。所以,单纯的句法分析是远远不够的。那么,能否完全抛弃句法分析而直接进行语义分析呢?答案是否定的。因为句法是构成语言的基础,语义依附于(但不依赖于)句法。所以,自然语言理解的关键在于如何协调二者之间的关系,建立一套高效的句法—语义匹配机制。

另外,人类对自然语言的理解通常受到语句所处的上下文和认知者(个体或群体)所拥有的背景知识、认识视野甚至身体构造等因素影响,具

收稿日期:2022-09-25
基金项目:湖南省哲学社会科学基金项目(19YBA128);湖南省教育厅科学研究优秀青年项目(21B0799)
作者简介:罗丹(1983—),女,湖南常德人,博士,讲师,主要从事语言逻辑、语言哲学研究。
①《潘云鹤:AI 走向数据与知识的双轮驱动》,http://news.sohu.com/a/582059088_120632774.
②《Yann Lecun:语言的有限性决定了 AI 永远无法比肩人类智能》,http://k.sina.com.cn/article_2118746300_7e4980bc02001b2ps.html.

有动态性、情境性和具身性等特征。新一代 AI 技术融合了大数据和大知识,充分模拟人类对自然语言的动态理解。大数据意味着自然语言处理的经验主义方法,它贴近真实文本,“感知”人类实际需要;大知识意味着自然语言处理的理性主义方法,它旨在构建规则,便于机器推理和“认知”。在大数据的基础上提取新知识,并用已有知识来规范海量数据,这就相当于在“感知”的基础上形成“认知”,在“认知”的视域下统摄“感知”。总之,自然语言处理是“人工智能皇冠上的明珠”,“下一步人工智能要害的地方就是想办法让机器理解人类的语言”①。这要求计算机对自然语言的处理必须融合逻辑与经验、大数据与大知识、经验主义与理性主义,助力实现新一代 AI“自主智能化”的新目标。

为了全面提升计算机的“认知”能力,首先需要自然语言具备高度解释力的形式化,以获得计算机对语义深刻透彻的理解。计算机理解语言(识别句法、掌握语义)、获得“认知”的过程,实质是一种“符号操作”模式,以形式化的句法系统为框架。所以,无论语言学、逻辑学还是人工智能领域,都致力于寻求最适宜于 AI 新时代需求的语法形式化理论。其中,组合范畴语法(combinatory categorial grammar,简称 CCG)表现突出,句法衍生和语义组合之间的联系紧密,并且设计出功能强大的句法分析器,是最具影响力的形式语法之一②。但是,CCG 在语义分析方面不够理想,存在生成过多之类的问题。它的升级版本——多模态组合范畴语法(Multi-modal combinatory categorial grammar,简称 MMCCG)试图克服 CCG 的不足,是语言、逻辑与计算“高能”交叉与深度融合的产物。MMCCG 的句法—语义匹配性更好,解析功能更加强大。目前,已有学者设计出基于 MMCCG 的分析软件并将其推广运用于 AI 对话系统,甚至开发出高质量的语料库,使之能够应用于大规模大范围的语句实现③。

二　CCG 的应对策略及其局限性

CCG 是在范畴语法(categorial grammar,简称 CG)基础上演变而来,其创始人为 2018 年国际计算语言学协会 ACL 终身成就奖获得者、英国著名计算语言学家马克·斯蒂德曼(Mark Steedman)。CCG 最初是作为一个心理学模型提出来的,主要面向心理学家感兴趣的句式,后来受到乔姆斯基的观点启发,注意到语义从一开始就在人类理解语言过程中扮演着重要角色,因而将其改造成一种新型的语法理论,以应用于计算机自然语言处理④。相较于以往的范畴语法,它跳出对语言片段式研究的藩篱,在描述能力上可以涵括更为广泛的语言现象,形成一种“覆盖式”的语言生成能力。

CCG 普遍应用于自然语言分析、转换和生成等各个方面,堪称计算语言学中的一个全栈模型⑤。CCG 之所以具有这些良好特性,主要得益于它在经典范畴语法基础上增加了对应组合逻辑中三个“组合子”(一种高阶函项运算)B、T、S 的范畴运算规则,用十分简洁的方式扩展了经典范畴语法的生成能力。从逻辑语义学角度看,CCG 是一种组合性语法,它贯彻句法与语义的透明性,实现了句法与语义的并行推演。从计算应用层面看,CCG 属于柔和的上下文相关语法,在描述和表达力上要明显优于上下文无关语法,能够处理一些在自然语言中经常出现的节点提升、宾语提取、交叉依存等现象。因此,基于 CCG 的自然语言处理系统能够较好地协调计算、规则与算法三要素。

尽管 CCG 力图融合经验主义和理性主义,但以“自主智能化”为目标的新一代 AI 对自然语言处理提出了更高要求。一方面,基于神经网络的大数据与深度学习技术的突飞猛进和广泛应用,促成了经验主义的新高峰,给 CCG 带来了前所未有的挑战。其一,随着深度学习技术的发展,在知识图的语义解析与归纳领域,CCG 的语法解析器已被基于深度神经网络的训练模型所超越。深度神经网络方法无需获取普遍的语义表征甚至普遍

①《清华自然语言处理科学家孙茂松:深度学习碰壁之后,我们还能做什么?》,https://www.sohu.com/a/352736821_651893.

②满海霞:《组合范畴语法:通向人工智能语义理解的一种逻辑经验主义路径——访马克·斯蒂德曼教授》,《哲学动态》2022 年第 1 期。

③陈鹏:《汉语组合范畴语法研究——基于交叉学科的视角》,中国社会科学出版社 2022 年版,第 212 页。

④满海霞:《组合范畴语法:通向人工智能语义理解的一种逻辑经验主义路径——访马克·斯蒂德曼教授》,《哲学动态》2022 年第 1 期。

⑤Steedman M.“The lost combinatory”. *Computational Linguistics*, 2018, 44(4): 9.

的思维语言,只需使用小的数据集来归纳语义解析,再通过端到端的深度神经网络的强大算力来解析归纳。其二,CCG 的突出优势在于它强大而自然的生成能力,但由此导致了生成过多、效果不理想、效率不高之类的问题。因为 CCG 不仅能生成许多合乎语法的句子,同时也能推出许多不合语法的句子,严重制约了它的实际应用价值。其三,CCG 局限于很小的语言处理范围,规则的设定主要以英语、荷兰语为参照,涵括的语言类型不丰富,普遍语法特征表现不明显,并未真正具备跨语言通用特性。其四,量词辖域歧义问题是衡量语法理论可行性的试金石,是自然语言处理过程中绕不过的难题,影响着机器对自然语言的分析和理解。CCG 在量化语义的表征方面存在明显短板,特别是处理难题的能力较弱。

另一方面,尽管深度学习算法在 NLP 领域异军突起,尤其在强大算力方面的效果十分明显,但是这种算法在分析句法结构、识别句法成分以及针对自然语言中的大量长尾现象(如非成分并列、主语抽取、交叉依存)解析方面的表现差强人意,有时甚至无能为力。也就是说,所谓的"深度学习"并未真正理解人类语言,不是真正学习句法,而是学习一个巨大的有限状态转换器或者一个由软增强的过渡网络①。所以,无论"经验主义"的钟摆摆得多远,"理性主义"的摆幅仍不可忽视。在自然语言处理过程中,CCG 所呈现出的结构化表征方法仍有不可或缺的理论地位。关键是,在新一代 AI 浪潮冲击下 CCG 如何克服自身不足?我们不能把对人类语言的探索交给机器的无穷算力,也不能闭门造车,固守现有的语法理论。因此,为了适用新一代 AI 自主智能化的要求,一方面,应该继续秉承经验主义与理性主义融合之路,但另一方面需要寻求比 CCG 更加强大的形式语法理论。MMCCG 正是在这样的背景下应运而生,成为应对时代挑战最理想的逻辑语义工具。

三　MMCCG 对 CCG 的继承与优化

鲍德里奇(Baldridge)将组合范畴语法与范畴类型逻辑相融合,搭建了一个混合范畴框架——MMCCG。该框架一方面保持了 CCG 良好的计算优势,同时结合范畴类型逻辑所采用的资源敏感性方法与精细度控制手段。所谓资源敏感性方法,即用最简洁的方式取代 CCG 中针对不同语言所采取的规则上的特定设置。精细度控制手段,即在 CCG 语法体系中增添四种模态算子(★、•、◇、×),以满足不同结合律和置换律的函项运算要求。虽然 MMCCG 在 CCG 的基础上获得了改进和优化,但它仍然是柔和的上下文相关语法,并没有添加新的组合规则。也就是说,MMCCG 是以提升 CCG 的通用语法功能与解题功能为目的,并非背道而驰。

(一)MMCCG 对 CCG 的继承

1.谓词—论元结构②

对自然语言进行形式化解析,就是从逻辑视角对个体及其关系进行描述,这种描述形成了 MMCCG 表征语义的基本特征:"谓词—论元"结构。每个"谓词—论元"结构的形成完全是从谓词的涵义出发(基于词义的分析),即使是同一个谓词,如果语义发生变化,那么相应的"谓词—论元"结构也会发生变化。这种变化的影响直接映射为自然语句中主语和宾语成分之间关系的变化。一个典型的例子是汉语动词"死",当表达常规义"死亡"时,具有不及物动词属性,不能携带宾语,为"主+谓"式结构,范畴表述为"S\NP:λx.死亡′x"。当"死"的涵义为"失去"义③时,"死"具有及物动词属性,其"谓词—论元"结构由"主+谓"式变为"主+谓+宾"式。此时,"死"的范畴表述应为"(S\NP)/◇NP:λx λy.失去′xy"。④MMCCG 承继 CCG,主张从"谓词—论元"结构来表征语义关系,符合心理现实性,便于机器操作语言。

①Steedman M."The lost combinatory",*Computational Linguistics*,2018,44(4): 15.

②其他范畴语法一般认为,语义完全可以从表层结构中的句法关系推导出来,表层结构是唯一可以依赖的表达层。斯蒂德曼则对此提出质疑,他认为基于表层结构的范畴语法缺乏必要的心理现实性,在现实中无法兑现。因而它在 CCG 中提出"谓词—论元"结构才是唯一的表达层。参见:Ades A E , Steedman M ."On the Order of Words", *Linguistics and Philosophy*,1982,4(4): 517-558.

③如"王冕死了父亲"中"死"之涵义。参见韩玉国:《词汇主义视阈下的汉语非连续结构研究——以范畴语法为纲》,北京语言大学出版社 2017 年版,第 30 页。

④范畴"SWP:λx.死亡′x"显示它只需与一个左方向的论元结合便可成句,而"(S\NP)/◇NP:λx λy.失去′xy"显示它要先与一个右方向的论元结合,再与一个左方向的论元结合,才能组合成合法句。逻辑表达式"λx λy.失去′xy"揭示,它有两个虚空位置待填充。参见:Steedman M.*Combinatory Categorial Grammar:An Introduction*.UK:The SOMESUCH Press,2017, pp.39-40.

2.句法—语义并行推演

MMCCG 沿袭 CCG 句法与语义接口融洽性特点。这一特点保证 MMCCG 能够为句法提供非常直观的组合语义,使得句法与语义的接口是透明的,实现了句法与语义的并行推演。只需要在词条上增加语义标记,以极简的带语义解释的组合规则予以运算,便能产生组合语义。句法—语义并行推演的过程,还能对自然语言表达式是否是合式进行判定。透明的句法—语义接口,分析器能够便捷地访问谓词—论元结构,不仅包括局部依存关系,而且还包括远距离依存关系。这使得处理大规模的自然语言的语义成为可能。句法运算距离语义越近,我们距离理解自然语言的目标也就越近①。

3.更加彻底的词汇主义

MMCCG 与 CCG 一样,是基于词汇主义的形式化理论,即将自然语言生成过程凝缩在词条的范畴构造上②。形式化聚焦于词条,而规则是相对简洁和紧致的。在词库中为自然语言词汇赋予相应的范畴,词条蕴涵着丰富的句法、语义、类型化模态等信息,所有的句子成分都一一对应带有模态算子的句法范畴,以及映射相应的语义表达式。在传统语法中交由规则来处理的信息,在此都交付词库来完成。词本位思想是范畴语法的基本特征,在 MMCCG 中体现地更为明显。为了限制语句过度生成,MMCCG 所采取的方式不是对已有规则进行添加或修改,而是通过在词条的斜线算子上增添类型模态词。这种多模态扩张实质是更加彻底的词汇主义,更好地迎合当下计算机处理自然语言的要求。MMCCG 所提供的由词条出发,基于少量规则进行推演分析的方式,有利于构建简洁高效的计算语言③,极大地提升大规模自然语言分析工程化的可能性。

(二)MMCCG 对 CCG 的优化

虽然 MMCCG 与 CCG 具有相同的规则集,但它的资源敏感性方法使其具有跨语言通用的规则。它将所有的跨语言变化存储在词库中,以语法的类型学视角(非转换环境下更彻底的词汇化)来描述与解释语言,从而形成更精简的语言分析④。MMCCG 保留了 CCG 诱人的计算特性,同时还增加了利用新策略的可能性,使其具有更强的兼容性和解释力。概括起来,MMCCG 对 CCG 的改进主要体现在两个方面:

1.添加模态算子

虽然 CCG 具有极强的计算优势和生成能力,但它面临着生成过度和适用范围狭窄等问题⑤。为了避免语句的过度生成,CCG 所采取的方案是在规则的基础上增加额外的限制条件,并根据实际情况进行不断调适⑥。这种在规则上针对不同语言所采取的措施,又会加重词库负担,失去计算优势,既违背普遍语法所坚持的极简原则,又背离自然语言处理所采取的“大词库、小规则”策略,仍然不具有跨语言通用性。MMCCG 另辟蹊径,避免从规则上做文章而在词库上下功夫,通过在范畴斜线上添加模态算子,对 CCG 做了进一步优化。鲍德里奇为 MMCCG 提出了四个基本的模态假设,将★、●、◇、×作为斜线算子的下标。不同模态标记的算子适用于不同的推演规则,而词条被指派的函子范畴可能含有不同模态标记的斜线算子,这样适用于函子范畴的推演规则直接通过对词条的范畴指派表现出来⑦。经过此番改进,大大增加了斜线的表达力与适用性。一方面,不同的斜线类型可以满足规则在句法毗连上的不同要求,不因语言不同而变化,相当于不变的“普遍语法”;另一方面,对于特定语言中的语词的具体范畴指派,所适用的斜线类型由该语言的句法特点

①满海霞:《组合范畴语法:通向人工智能语义理解的一种逻辑经验主义路径——访马克·斯蒂德曼教授》,《哲学动态》2022 年第 1 期。

②邹崇理:《关于组合范畴语法 CCG》,《重庆理工大学学报(社会科学版)》2011 年第 8 期。

③张璐:《汉语形名结构的范畴语法研究》,中国社会科学院博士论文,2013 年。

④Baldridge J, Kruijff G J M. “Multi-Modal Combinatory Categorial Grammar”, *10th Conference of the European Chapter of the Association for Computational Linguistics*, 2003: 211-218.

⑤例如,适用于荷兰语、土耳其语的向前交叉规则>B ×并不适应于像英语这样具有较强语法特征的语言;而向后交叉复合规则<B ×在英语“重型 NP 移位”现象中适用,在其他情况下却是禁止使用的,因为允许<B ×规则的同时也会导致某些不合法语句合法化,如英语 *a nice in Edinburgh pub* 这样的混序语句、以及汉语“每个男人被所一个女人吸引”这样的病句。参见:邹崇理:《自然语言信息处理的逻辑语义学研究》,中国社会科学出版社 2018 年版,第 284 页。

⑥Steedman M, Baldridge J. “Combinatory categorial grammar”, *Non-Transformational Syntax: Formal and Explicit Models of Grammar. Wiley-Blackwell*, 2011: 181-224.

⑦邹崇理:《多模态范畴类型逻辑》,《安徽师范大学学报(人文社会科学版)》2012 年第 6 期。

决定,相当于可变的"参数"①。

2.引入广义斯科仑项

虽然鲍德里奇确立了 MMCCG 的逻辑框架与运行方式,但主要是从句法角度来考量,将语义搁置一旁。斯蒂德曼全面吸纳鲍德里奇的多模态思想,试图从语义角度进一步完善 MMCCG。他对 MMCCG 的最大贡献是通过引入广义斯科仑项来优化量化语义表征工具。广义斯科仑项作为语义标签可以刻画自然语言中所有的非全称量词,以此来简化多重量词之间复杂的依存关系,用于解决量化辖域疑难问题。广义斯科仑项具有形如 $sk^{E}_{n;p;c}$ 的逻辑表达式,其中 E 代表环境(或称"参数");p 代表由λ-公式所刻画的物名化属性;索引 n 只用于区分具有相同性质的不同个体,通常情况可省略。例如"两个女孩喝奶茶,两个女孩喝咖啡"中两次出现的"女孩"就需要添加索引来加以区分;基数条件 c 表示复数量词②。上述语句中前一个出现的"两个女孩"可表述为 $sk^{\{\}}_{1;\lambda x.女孩'x;\lambda s.\ |s|=2}$,其中环境 E 为空,表示这个广义斯科仑项不受任何全称量词约束③。形如 $sk^{\{\}}_{1;\lambda x.女孩'x;\lambda s.\ |s|=2}$ 这样的逻辑表达式是经过分析推演之后才形成的,并非一开始就具有这样的广义形式,最初只是"未确指"状态。"未确指"状态还需通过一种称为斯科仑确指化的过程形成最终的广义形式④。这种确指化可以在运算推演的任意点自由出现。任意点的自由选择便于机器从多种可能的语义解读中提取最佳选择。

经过上述改进,MMCCG 具有比 CCG 更强的兼容性和解释力。MMCCG 的进一步工作是深入挖掘和拓展语义解析方面的优势,致力于与机器学习的深度融合。尽管 CCG 已经开发了运算效率极高的句法分析器,但是存在语义分析的短板,制约了它的现实可行性。所以,问题的关键在于如何克服所谓的"天花板效应"⑤,开发出句法—语义并行的 MMCCG 分析器,深化语义解析机制。

四　MMCCG 的解题功能——以汉语量化辖域消歧为例

量化辖域歧义现象历来是自然语言处理的难题。量化辖域歧义句在自然语言中并不占多数,为何显得如此重要?根据著名的奇夫定律(Zlip'law)可知,一个单词在自然语言的语料库中出现的频率与它在频率表里的排名成指数幂次反比。关注自然语料库中最频发的少量事件,可以获得常规大规模机器学习即可掌握的 80%—90%的数据,捕捉数量庞大的稀有事件,有助于突破机器学习的"天花板"⑥。所以,量化辖域歧义句是自然语言处理的试金石,MMCCG 能够攻克这一难题,即验证了它强大的解题功能。

汉语歧义现象大致可分为词汇歧义、结构歧义和辖域歧义三种。其中,辖域歧义主要是由某些特定的词(量词、限定词、否定词、情态动词、内涵动词等)在同一句子中两两共用时引起的。例如,如果语句中含有两个或多个量词,那么该句由于多个量词之间的辖域作用,会产生多种可能的解读⑦。如语句:

两位老师批改了六份作业(以下简称为例句(1))。

此句可以有三种解读:①"两位老师"统一批改了"六份作业"(复数主语作为一个集合体完成某项行为);②"两位老师"分别批改了"六份作业"(复数主语作为个体独自完成某项行为);③"两位老师"批改了相同的"六份作业",意谓甲老师批改的这"六份作业",乙老师重复批改了。第一种解读是将复数短语"两位老师"作为一个集合,取统指解(collecting reading),处于宾语位置的复数短语"六份作业"被看成是一个常项,不受复数主语约束。第二种解读是复数主语取宽辖域,作逐指解(distributive reading),复数宾语被当作一个函项,受到复数主语约束。在第三种解读中,虽然复数主语仍取宽辖域,作逐指解,但是复

①Hockenmaier J, Steedman M,"CCGbank: a corpus of CCG derivations and dependency structures extracted from the Penn Treeban",*Computational Linguistics*, 2007, 33(3): 355-396.

②Steedman M.*Taking Scope*: *The Natural Semantics of Quantifier*,Cambridge: The MIT Press, 2012,p.166.

③Steedman M.*Combinatory Categorial Grammar*:*An Introduction*.UK:The SOMESUCH Press,2017, p.111.

④姚从军,朱乐亚,邹崇理:《广义斯科伦项理论:一种新的量词理论》,《学术研究》2021 年第 5 期。

⑤基于句法的语言生成达到 90%之后,便很难提高,称"天花板效应"。斯蒂德曼指出,通往其余 10%的数据的钥匙,就在被忽略的语义之中。

⑥满海霞,崔佳悦:《组合范畴语法的量词辖域歧义研究新思路》,《哲学动态》2015 年第 8 期。

⑦方立:《逻辑语义学》,北京语言文化大学出版社 2000 年版,第 11 页。

数宾语被处理为一个常项,不受复数主语约束。

从逻辑语义研究的角度来看,如何从例句(1)的表层结构推演出这个语句的几种语义解读,历来倍受学界关注。应该说,量词辖域歧义的形式化生成问题一直是语义研究中的重要课题,也是 NLP 领域衡量一门形式语法是否先进的试金石。量词辖域歧义的生成方案最早可以追溯至蒙太格语法,以生成语法为框架也有诸多研究。但限于篇幅,我们并不打算铺陈所有方案,仅以 CCG 与 MMCCG 为例来进行比较。

在 CCG 框架下,以经典量词理论为量化语义的表征手段,针对例句(1)只能生成“统指解”的逻辑语义表达式,即:$\exists y\ \exists x[$老师′$y \wedge Ty \wedge [$作业′$x \wedge Sx \wedge$批改′$xy]]$①,推演过程如图 1 所示。

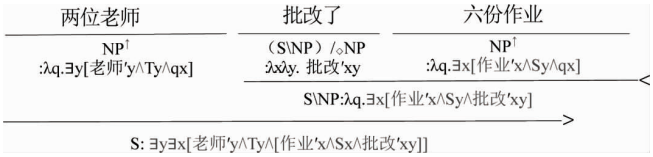

图 1　CCG 方案下的多重复数量词辖域歧义句的统指解读推演

针对例句(1),CCG 方案无法生成清晰完整的解读,语义表征比较生硬。如多重复数短语量化式,要形成向下的“逐指”行为,关键在于语义指派:要么为复数主语指派“逐指”语义范畴,要么将“逐指”义指派给谓语动词。显然,CCG 语义标签不具有这样强的表征效果。此外,CCG 无法简洁清晰地刻画诸如“两位”“六份”“多数”“少数”等这样的复数量词。

在 MMCCG 框架下,以广义斯科仑项为量化语义表征手段,复数主语、复数宾语都可以表述为广义斯科仑项(2a、2b),动词“批改”指派标准范畴(2c),可生成“统指解”,如图 2 所示:

(2)a.两位:= $NP^{\uparrow}_{3pl}/_{\diamond}N_{3pl}$: λnλp(skolem′n;λs.|s|=2)

b.六份:= $NP^{\uparrow}_{3pl}/_{\diamond}N_{3pl}$: λnλp(skolem′n;λs.|s|=6)

c.批改:= $(S\backslash NP_{3pl})/NP$:λxλy.批改′xy

d.批改:= $(S\backslash NP_{3pl})/NP$:$\lambda x\lambda y.\ \forall z[z \in y \rightarrow$批改′$xz]$

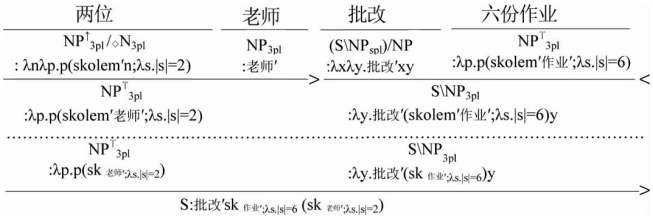

图 2　MMCCG 方案下的复数量词辖域歧义的统指解读推演

如果例句 1 更替为“每位老师批改了六份作业”,由于全称量词对存在量词的分配性,谓语动词即使取标准范畴,也能产生“逐指解”(即:每位老师分别批改了六份相同或不同的作业)。而例句 1 中谓语动词“批改”前面的量词是未确指的、可数的存在量词,若要生成逐指解读,须采用(2d)这个非标准的范畴。推演如图 3 所示。

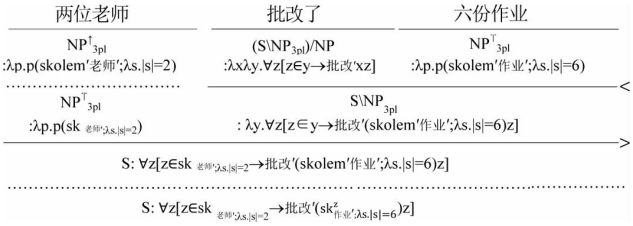

图 3　MMCCG 方案下的复数量词辖域歧义的逐指解读推演

①复数量词“两位”占宽域,且 T 表示“两”,复数量词“六份”占窄域,S 表示“六”。

因为斯科仑确指化是一个自由的运算过程,如果图 3 的复数宾语的斯科仑确指化发生在与主语毗连的运算之前,将生成一个不受复数主语约束的 $sk_{作业';\lambda s.|s|=6}$ 常项,那么可获得逐指解的第二种解读 $\forall z[z \in sk_{老师';\lambda s.|s|=2} \rightarrow 批改' sk_{作业';\lambda s.|s|=6} Z]$,意谓"六份作业,两位老师重复批改了"。

MMCCG 不仅在语义上增添了更丰富的高阶内涵逻辑手段,使解题功能变得更加强大,而且在句法方面通过添加模态算子,能够有效阻止生成不合法的量化歧义句。譬如,在 CCG 中可能会生成形如"每个男人被所一个女人吸引"这样的"病句",而 MMCCG 可以阻止生成这样的"病句"。如图 4 所示。

图 4　MMCCG 阻止不合法的被动语态的量化歧义句的生成

结语

面向新一代 AI 带来的挑战,传统的 CCG 愈加显得力不从心,从根本上无法满足大数据与大知识双轮驱动背景下 AI 实现自主智能化的目标。MMCCG 不但继承了 CCG 的衣钵,还对其做了必要的改造、升级和优化,特别是添加了模态算子,使用广义斯科仑项技术,使句法—语义联系更为紧密,更加贴近真实文本,并且能够有效阻止过度生成。这有助于计算机真正理解自然语言,使 AI 获得某种"认知"能力。然而,MMCCG 的语料库建设迄今仍未完成,并且以英语和荷兰语为样本,未能涉及汉语等其他语言类型。相比之下,国内还停留在社科汉语 CCG 语料库建设阶段,没有扩展为更具普遍意义的 MMCCG 语料库,存在语义分析不足等问题。所以,构建一个高质量、大规模的汉语 MMCCG 语料库是亟待完成的新任务,也是助力新一代 AI 发展的内在要求。

Multi-Modal Combinatory Categorial Grammar in the New Era of AI

LUO Dan

(School of Marxism Studies, Hunan Institute of Technology, Hengyang 421002, China)

Abstract: In the new era of AI driven by big data and big knowledge, natural language processing has encountered unprecedented impacts and challenges. Multi-modal combinatory category grammar is the upgrade and optimization of standard combinatory categorial grammar. It inherits the internal structure and operation of the standard combinatory categorial grammar, retains the original computational properties, and increases the possibility of using new strategies. The most important innovation lies in the addition of modal operators and the integration of generalized Skolem term technology, which makes it more consistent with universal grammatical features, cross-language generality, stronger description ability and problem solving function, better realization of natural language processing, and in line with the development trend of the new generation of AI.

Key words: the new era of AI; natural language processing; Multi-modal combinatory categorial grammar; generalized Skolem term; quantifier-scope

(责任校对　葛丽萍)