逻辑与科学方法论

doi: 10.3969/j. issn. 1674-8425(s). 2016. 08. 002

主持人语:

中国逻辑学会会长 邹崇理 研究员

《组合范畴语法(CCG)的计算语言学价值》一文,比较深刻地刻画了 CCG 的主要性质。其观点切中了当前计算语言 学界的弊端,即大部分研究者都埋头做关于自然语言处理系统的具体工作,这些是容易收到实效的"低枝果实",而不愿意 去仰望理论的星空,做具有科学探险性质的理论探讨。CCG 就是计算语言学者理性主义思考的结果,值得计算语言学和 逻辑学界的关注。

为了深化对社会选择理论的认识,《基于理由的偏好防策略投票方案》一文,以丹尼尔·欧歇尔森和斯科特·温斯坦 定义的基于理由的偏好作为基本的偏好关系,研究基于理由的偏好聚合问题,并尝试构造一个基于理由的偏好投票方案, 进而对其是否具有防策略性进行了研究,具有一定的理论创新意义和现实价值。

《论逻辑真理的范围》一文,通过考察两大类逻辑系统——实在论逻辑系统和反实在论逻辑系统,发现 Henkin 的方法 能够给出这些系统的完全性证明。用这种方法可以构造出包含所有系统逻辑真理的集合,而这个集合就是逻辑真理的范 围。该文做出结论:完全性是一个划分逻辑真理和非逻辑真理的标准。

组合范畴语法(CCG)的计算语言学价值

鹏1,2 陈

(1. 北京语言大学 信息科学学院,北京 100083; 2. 中国社会科学院 哲学所,北京 100732)

摘要:组合范畴语法(CCG)是在 AB 演算基础上进行扩展而产生的范畴语法。从语法理论视角 看,CCG 是一种词汇形式化的方法;从计算语言学视角看,CCG 属于一类适度上下文相关文法; 从逻辑语义学视角看、CCG 在句法与语义的接口方面非常融洽。无论是 CCG 语言的、计算的、 还是逻辑的特征,都使得 CCG 非常适用于自然语言信息处理,对于计算语言学具有很好的理论 和实际价值。

关键词:组合范畴语法;语法理论:计算语言学:逻辑语义学;自然语言信息处理:适度上下文相 关文法: 句法与语义接口融洽

中图分类号:B81

文献标识码:A

文章编号:1674-8425(2016)08-0005-07

Computational Linguistic Value of Combination Category Grammar

CHEN Peng^{1,2}

- (1. School of Information Sciences, Beijing Language and Culture University, Beijing 100083, China;
 - 2. Philosophy Institute, Chinese Academy of Social Sciences, Beijing 100732, China)

收稿日期:2016-03-12

基金项目:国家社会科学基金重大项目"自然语言信息处理的逻辑语义学研究"(10&ZD073)

作者简介:陈鹏(1979—),男,江西南丰人,北京语言大学信息科学学院讲师,工学博士(计算机软件与理论专业), 中国社会科学院哲学所博士后,研究方向:逻辑学。

引用格式:陈鹏. 组合范畴语法(CCG)的计算语言学价值[J]. 重庆理工大学学报(社会科学),2016(8):5-11.

Citation format: CHEN Peng. Computational Linguistic Value of Combination Category Grammar [J]. Journal of Chongging University of Technology (Social Science), 2016(8):5-11.

Abstract: Combination Category Grammar (CCG) is a category grammar extended on the basis of AB calculus. From the perspective of grammar theory, CCG is a lexical formalization; from the perspective of computational linguistics, CCG belongs to a class of mildly context-sensitive grammar; from the perspective of logical semantics, CCG has a friendly interface between syntax and semantics. Either CCG's language characteristic or logic computational characteristic, even logic characteristic makes CCG much suitable for natural language processing, which is theoretically and practically valuable to computational linguistics.

Key words: Combination Category Grammar; grammar theory; computational linguistics; logical semantics; natural language processing; mildly context-sensitive grammar; friendly interface between syntax and semantics

一、引言

肯尼斯・丘奇(Kenneth Church)于 2011 年在 《语言技术中的语言学问题》杂志上发表了一篇题 为 A Pendulum Swung Too Far 的文章[1]。文中通 过整理、分析和研究自1940年以来关于自然语言 处理的文章,发现了一个非常有趣的规律,即如果 将自然语言处理的文献划分为"理性主义"和"经 验主义"两大类的话,双方的优势局面出现连续振 荡的现象,而且存在一个每20年一个周期的规 律。两大类研究的巅峰期分别是:(1)20 世纪50 年代,以香农(Shannon)、斯金纳(Skinner)、弗斯 (Firth)和哈里斯(Harris)为代表的经验主义高峰 时期;(2)20世纪70年代由乔姆斯基(Chomsky)、 明斯基(Minsky)所主导的理性主义全盛时期;(3) 20 世纪 90 年代由 IBM 语音团队 (IBM Speech Group)、AT&T 贝尔实验室(AT&T Bell Labs)所引 发的新一轮经验主义巅峰时代。

当然,对丘奇的研究,或许我们可以说,理性 主义和经验主义并没有非常清晰的界定。此外, 许多基于算法的统计方法和深度学习方法在传统 分析数学的视角上是一种经验主义的做法,但是 若将图灵机模型本身看成是一个理性主义产物, 那么很多的所谓"经验主义"方法其实仍然不失为 一种"理性主义"成果。姑且抛开对丘奇研究的一 些争议,他的文章对当前学界的一些批判和反思 是完全值得我们深思的:

(1) 在实用主义的驱动下,计算语言学领域 的教与学的工作向统计学极度倾斜,而逻辑、代数 等基础学科却得不到应有的发展。 (2) 在经验主义的诱惑下,学者倾向于采用统计学方法把唾手可得的低枝果实采摘下来,而极少人愿意去攀登更具挑战性和科研风险的理性主义高峰,为人类的自然语言处理甚至人工智能开拓一条新的道路。

尽管已经过去近 5 年,然而丘奇所忧虑的问题却日益凸显。当今的计算技术发展日新月异,大数据技术甚至推动整个科学领域发生范式转换,形成所谓的"数据密集型科学研究"。在这种环境下,在整个计算语言学领域,"经验主义"更是以绝对压倒性的优势浩荡前进,似乎只要在摩尔定律^①、吉尔德定律^②的指导下,只要计算机速度不断加快、存储容量不断扩充、网络速度不断提升,一切问题都会迎刃而解。

然而,事实并非如此。实事求是讲,近些年,随着计算和网络速度与效率的提升,自然语言信息处理能力得到极大的改善,能够处理的语言信息容量、处理的速度和效率也取得了较大的突破。正如丘奇所说,这些更多的是简单地应用新技术工具而唾手可得的,那些计算语言学中的"硬核"问题,例如语义问题,并未得到解决。应该说,如果不解决计算语言学中的"硬核"问题,那么我们在语言信息处理方面的成果只能是量变,而不能够形成质变。

反观近些年来自然语言信息处理的"理性主义"路线,相比而言,这条研究路线冷清了许多,仅仅是少量的欧洲传统的自然语言逻辑学派和一些

⑤ 摩尔定律:集成电路的复杂度(可被间接理解为芯片上可容纳的晶体管数目)每两年增加1倍,性能也将提升1倍。

② 吉尔德定律:主干网带宽的增长速度至少是运算性能增长速度的3倍。

拥有坚定"理性主义"信念的研究者在从事这方面的研究工作。但是,在这些研究工作中,我们发现了一条非常有趣且颇具价值的研究路径:它源于理性主义,但又不局限于理性主义,其根基已经逐步蔓延到经验主义的土壤中,并从中吸取养分,在诸多的大规模自然语言处理中获得广泛应用。同时,对于一些自然语言处理的"硬核"问题也展开了探索。这项研究就是基于组合范畴语法(CCG)的自然语言信息处理应用。

二、组合范畴语法 CCG

组合范畴语法 CCG 应该算作是 20 世纪末的 理性主义产物,它在 20 世纪 80 至 90 年代开始出现,在 AB 演算基础上进行扩展而产生,其核心的扩展在于"组合",即基于范畴语法增添了函子范畴的组合运算,从而增强了表达与描述能力。另一方面,由于组合规则与柯里的组合算子非常接近,因此每个组合规则在分析过程中都具有一个语义解读,这样使得句法派生的同时,又能够构造谓词一论元结构作为语义解读。

2000 年以来,CCG 就已经广泛地应用在计算语言学的各个方面,可以说 CCG 是计算语言学中的一个全栈模型,从自然语言的分析、转换到生成等各方面都得到普遍应用^[2]。之所以如此,主要有如下两大方面的原因。

1. 在基于 CCG 的自然语言信息处理系统中, 很好地协调了计算、规则和算法几方面的因素

现代的自然语言信息处理系统都可以抽象为一个三元组〈R,C,O〉中,其中,R代表规则、C代表计算、O表示Oracle,规则是整个语言信息中的内核,例如上下文无关文法(CFG)或者组合范畴文法(CCG),计算是在计算系统中实现的算法,O表示一些经验性的语料或者人为的干涉等。

要想系统发挥最佳作用,〈R,C,O〉3个部件需要相互协同。例如,如果选择了一种上下文无关文法来描述自然语言,那么在对该语言进行分析时,就需要设计一种多项式复杂性的计算算法,同时能够有效地进行消歧。甚至,如果可能的话,需要对上下文无关文法的模型进行概率化,引入一些优选机制来对生成规则进行排序,从而提高处理效率和扩展处理的规模。

〈R,C,O〉三元组3个部件相互协同,体现了计算语言学中理性主义与经验主义的调和。正如我国计算语言学家冯志伟所主张的,自然语言处理应该将理性主义与经验主义结合起来^[3]。在自然语言处理中,理性主义与经验主义各有优缺点。理性主义更贴近自然语言本身,更注重自然语言本身的规则与规律,能有效处理远距离的主语和谓语动词之间的一致关系(subject-verb agreement)、wh 移位(wh-movement)等远距离依存关系(long-distance dependencies)问题;经验主义在大规模和工程化方面具有显著优势,结合强大的计算和信息处理能力,可以进行语言的自动学习和统计分析^[3]。

自然语言信息处理系统通常需要这 3 个部件之间的协同,同时也受到这 3 个部件之间的彼此约束。例如,如果系统采用了上下文相关文法(1型文法),那么基于该文法,对自然语言进行分析过程的计算复杂性通常都是 NP。反过来,如果你选择一个计算复杂性较低的文法,例如正则文法(0型文法),那么又存在该文法在描述和表达能力上不够强的问题。详细的关联情况如图 1 所示。

计算复杂度文法	Р	NP	不可计算
3型文法	$\sqrt{}$		
2型文法			
1型文法		V	
0型文法		V	

图1 文法形式化与计算复杂性的关联

说明:在文法描述上,存在 0 型文法到 3 型文法,其中 涉及一个表达能力和计算能力的折衷;在计算复杂度方面,存在多项式可计算(P)、指数可计算(NP)、不可计算 (图灵机意义下的不可计算)3 个不同的层次。

CCG 在规则与计算上做了一个很好的折衷。 CCG 在文法的描述和表达能力上是介于上下文无 关文法(2 型文法)和上下文相关文法(1 型文法) 之间,属于一类适度上下文相关(mildly context sensitive)文法^[4]。所谓的适度上下文相关语言具 有如下特点:

- (1) 有限的交叉依赖。
- (2)连续增长,即如果存在一个界值 k,只要有两个语句之间的长度差异超过 k,那么必然存在一个语句,其长度介于这两个语句之间。
 - (3) 分析的时间复杂度是多项式复杂度。

除了很好地调和计算、算法和规则几个因素 之外,CCG 本身具有一些非常有益于自然语言信 息处理的特性。

2. CCG 在文法形式化、语言与计算和逻辑语义等方面都具有非常有益于自然语言的计算机信息处理的特性

从自然语言的信息处理视角来看,CCG的优势主要可以从以下几个方面来阐述:

- (1) 从语法理论方面来看, CCG 是词汇形式 化的思路,是以词作为单位的形式化方法。在进 行大规模的自然语言信息处理过程中, CCG 的词 汇形式化在处理的信息规模、计算效率和复杂性 方面都有比较明显的优势。
- (2) 从计算语言学方面来看,CCG 属于一类适度上下文相关文法。适度上下文相关文法在描述和表达能力上要明显优于上下文无关文法,能描述一些在自然语言中经常出现的交叉依赖(Crossing dependencies)现象。分析适度上下文相关语言的时间复杂度通常是在多项式时间复杂度上,这对于计算而言是非常融洽的。
- (3) 从逻辑语义学方面来看, CCG 是一种组合性的文法。此外, 句法与语义之间融洽的接口使得 CCG 在对自然语言的语义进行分析和计算时非常便捷。

三、CCG 的性质:词汇形式化以及 适度上下文相关特性

首先,CCG 是一种基于词汇的形式化理论,即 CCG 将自然语言生成过程凝缩在词条的范畴构造 上^[5]。例如:

(1) $S \rightarrow NP VP$

 $VP \rightarrow TV NP$

TV → {喜欢,爱,…}

(2) 喜欢:= (S\NP)/NP

式(1) 是一个上下文无关文法所表达的产生 式规则,式(2)是对单个词指派范畴。可以说通过 式(2)中所指派的词法范畴,捕获了式(1)中的句法规则。

通过式(2)中的范畴指派,将及物动词"喜欢"定义为一个函数,并说明了其论元的类型和方向以及结果的类型。例如:"喜欢"作为一个函数,其从右边接受一个类型为 NP 的论元,同时计算结果的类型为 S\NP。

CCG 体现的是一种词本位的思想,形式化聚 焦在词条上,而规则是相对简洁和紧致的。这种 词汇形式化特性在自然语言信息处理上具有如下 优势:

1. 可以为每一种自然语言构建一个 CCG 范畴 语料库

CCG 范畴语料库中的内容包括覆盖每一个词的范畴库(通常一个词汇对应 1 个或者多个范畴)、一些典型语句的加标记 CCG 范畴派生树库。

这样的 CCG 范畴语料库可以为自然语言处理提供如下用途:

- (1)作为词法-范畴字典,可以在语料库中检索和查找任何一个词所对应的范畴(当然,有可能出现多个范畴)。
- (2)作为自然语言分析过程中的训练语料和 测试模型。在开发学习器和分析器的过程中,可 以使用语料库进行学习器的训练语料,同时也作 为测试分析器的精度和准确性的测试样本。

目前,已经有许多语种都开发出相应的 CCG 范畴语料库,有些是重新构建,有些是基于以往的一些语料库进行自动转换而来。例如:从宾州树库转换而得到的英语 CCGBank、汉语 CCGBank、德语 CCGBank,清华大学汉语树库转换而来的清华 CCGBank 等。

2. 极大地促进大规模自然语言分析工程化的 可行性

在基于文法的大规模自然语言分析应用中, 普遍存在着歧义性问题,通常每一个句子成分都 会对应大量的分析,从而使得解析空间爆炸式增 长,极大地提升了各类复杂性,使得大规模应用难 以实施与开展。

基于 CCG 的自然语言分析过程大致可以为 分成两阶段:第一阶段是将句子中的词指派词法 范畴,第二个阶段便是使用 CCG 组合规则组合这 些范畴。在第一阶段中,由于有些词对应的可能 范畴多达上百个^①,如果采取完全指派,那么大规模应用显然是行不通的。因此,在 CCG 范畴语料库基础上,采用了一种称为"超级标记器"(supertagging)的技术来减少范畴规模。

超级标记器是在进行自然语言分析之前,使用统计序列标记技术(statistical sequence tagging techniques)为语句中的每一个词都择优指派少量的词法范畴,其择优标准采取的是一种概率模型:

$$p(y \mid x) = \frac{1}{Z(x)} e^{\sum_i \lambda f_i(y,x)}$$

其中, f_i 代表一种特征, λ_i 是其对应的权重,Z(x)是一个规范化常量。语境是围绕目标词的 5-词的 窗口,特征通过窗口内的每个词和每个词的词性来定义。

这种超级标记技术,极大地提高了分析的速度和效率。斯蒂芬·克拉克(Stephen Clark)和詹姆斯·柯伦(James R. Curran)的研究表明,采用超级标记技术的 CCG 分析器比同类型的分析器的速度提升了一个量级(根据他们的实验,在他们给定的实验环境下,采用超级标记技术的分析器比未采用超级标记技术的分析器的速度提高了77倍)^[6]。

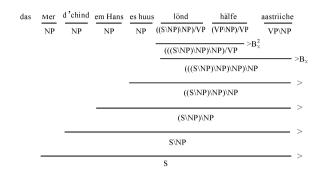
其次,CCG 是一种适度上下文相关文法。关于自然语言究竟位于乔姆斯基形式文法的哪一个层级尚存在争议。首先,乔姆斯基本人否定自然语言是正则语言,但是他也不确认自然语言是否是上下文无关语言。有许多学者根据自然语言出现的一些复杂交叉依赖的现象,认为自然语言必定是超越上下文无关语言,而接近于上下文相关语言。乔希(Joshi)在1985年对自然语言的形式化层次做了一个假设:人类的自然语言是适度上下文相关的^[7]。总之,虽然对自然语言究竟属于哪一形式语言层次尚未有定论,但大多数学者还是倾向于自然语言应该介于上下文无关语言和上下文相关语言之间,类似于适度上下文相关语言。

CCG 就是一种适度上下文相关文法,其优势是处理一些内在于语言构造的远距离依赖现象。在使用基于 CCG 的分析器进行语句分析的时候,能够直接融入远距离依赖分析,而不需要像其他一些分析器那样,在后处理阶段中再去处理远距离依赖。例如,在 CCG 中可以处理一些非常复杂的交叉依赖现象。

例 1 das mer d'chind em Hans es huus lönd hälfe aastriiche.

> that we let the children help Hans paint the house

我们让孩子们帮助汉斯粉刷房间



类似例 1 这种交叉依赖是不能够由上下文无 关文法来描述的。

CCG 的适度上下文相关文法特性使得它在描述与表达能力和计算复杂度之间取得一个较好的折衷。我们可以基于上下文无关文法的移位一归约(Shift-Reduce)算法,通过修改和增加操作符:{SHIFT, COMBINE, UNARY, FINISH},便可以获得基于 CCG 的移位一归约分析算法[8]。

例2 我爱真理



分析见图 2。

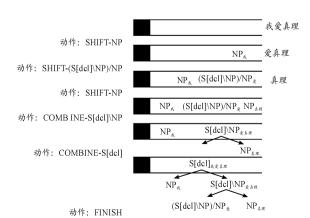


图 2 基于 CCG 的移位-归约算法示意:"我爱真理"

① 例如,在一个我们从宾州汉语树库转换而来的汉语 CCGBank中,"的"一词对应的范畴就有181个之多。

四、CCG 性质:组合性以及句法与 语义接口的融洽性

CCG 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。CCG 只需要在词条项中增加语义标记,并解释少量的组合规则,便能提供组合语义。基于 CCG 透明的句法一语义接口,分析器能够直接或者间接地访问谓词一论元结构。CCG 的组合性以及句法与语义接口的融洽性使得分析、处理大规模自然语言的语义成为可能。

组合原则是逻辑语义学的基本原则,是其基础和出发点。CCG 直观地体现了意义组合原则, 其规则中语法与语义严格对应,如表1所示。

表1 CCG 中范畴与语义规则严格对应

规则 类型	范畴规则	对应的语义 规则	缩写
应用	X/Y Y →X	$f a \rightarrow f a$	(>)
	$Y X \setminus Y \rightarrow X$	$a f \rightarrow f a$	(<)
组合	$X/Y Y/Z \rightarrow X/Z$	$f g \to \lambda x. f(gx)$	(> B)
	$Y \setminus Z X \setminus Y \to X \setminus Z$	$g f \rightarrow \lambda x. f(gx)$	(< B)
类型 提升	$X \to T/(T\backslash X)$	$a \rightarrow \lambda f. fa$	(>T)
	$X \to T \setminus Z$	$a \rightarrow \lambda f. fa$	(< T)
替换。	$\begin{array}{c} (X/Y)/Z \\ Y/Z \to X/Z \end{array}$	$f g \to \lambda x. fx(gx)$	(>S)
	$\frac{Y \setminus Z}{(X \setminus Y) \setminus Z \to X \setminus Z}$	$g f \to \lambda x. fx(gx)$	(< S)

CCG 通常可以进行句法与语义并行推演,例如:

例 3 Utah borders Iaho

Utah: = NP:utah

Idaho: = NP:idaho

borders: = $(S\NP)/NP : \lambda x. \lambda y.$

borders(y,x)

UtahbodersIdahoNP:utah $(S\NP)\NP:\lambda x.\lambda y.borders(y,x)$ NP:idaho $(S\NP):\lambda y.borders(y,idaho)$

S:borders(utah,idaho)

在计算语言学中,存在着一些"硬核"任务,或

者说是最为困难的任务,其中之一就是对自然语言的语义分析,即将自然语言语句映射为表征其意义的形式化(通常是某种逻辑式)。通过语义分析,将一个自然语言句子,按照特定的句法,解析成逻辑表达式,基于这些逻辑表达式可以实现逻辑和知识操作,并构建相应的顶层应用,例如自动问答系统和知识推理系统等。

CCG 在语义分析方面具有较好的优势,除了能够结合一些统计、学习的方法之外,CCG 还能够进行一个规则映射。一方面,CCG 可以通过类似概率 CCG 的模型来解决歧义解析问题。另外,CCG 的句法与语义接口的融洽性非常有助于语义学习,例如,卢克·泽特尔莫耶(Luke S. Zettlemoyer)等人基于 CCG 开发了一个语义分析框架^[8],在该框架中使用了一些规则,将逻辑式反向映射为范畴与语义,其规则为:

a. 对于一个常元 c,那么其输出的范畴是:

NP:c

b. 对于一个一元谓词 p, 那么其输出的范畴是:

 $N : \lambda x. p(x)$

或者

S\NP: $\lambda x. p(x)$

 $N/N : \lambda g. \lambda x. p(x) \wedge g(x)$

c. 对于一个二元谓词 p, 那么其输出的范畴是:

 $(S\NP)/NP : \lambda x. \lambda y. p(y,x)$

 $(S\NP)/NP:\lambda x. \lambda y. p(x,y)$

 $N/N : \lambda g. \lambda x. p(x,c) \wedge g(x)$

d. 对于一个二元谓词 p 和一个常元 c ,其输出的范畴是:

 $N/N: \lambda g. \lambda x. p(x,c) \wedge g(x)$

e. 对于一元函数 f, 其输出的范畴是:

 $NP/N: \lambda g. argmax/min(g(x), \lambda x. f(x))$

 $S/NP:\lambda x. f(x)$

例如,对于例3中对应的逻辑式:

borders (utah, idaho)

由于 utah 和 idaho 都对应两个常元,因此有:

NP:utah

NP:idaho

此外, borders 对应一个二元谓词,那么有:

 $(S\NP)/NP : \lambda x. \lambda y. borders(y,x)$

CCG 的反向映射不仅能够辅助大规模的语义 学习,而且使得 CCG 能够作为机器翻译中的中介 语言,实现由源语言到逻辑表达式,再由逻辑表达 式到"范畴和语义词项",再到目标语言的一个三 阶段机器翻译的方法,这样的机器翻译将使得"意 义保真"。

五、结论

以词汇为核心,能够描述一类适度上下文相 关语言,同时句法与语义可以透明地相互转换是 组合范畴语法 CCG 的几个非常显著的特征,这些 特征结合一些统计和学习方法,能够对自然语言 信息处理中一些非常复杂的"硬核"问题的解决有 所帮助。

参考文献:

- CHURCH K. A Pendulum Swung Too Far[J]. Linguistic Issues in Language Technology, 2011, 6(5):1-27.
- [2] The tombinatory categorial grammar site [EB/OL]. [2016-03-27]. http://groups. inf. ed. ac. uk/ccg/publications. html.
- [3] 冯志伟. 自然语言处理中的理性主义和经验主义

- [C]. 全国民族语言文字信息学术研讨会,2007.
- [4] EDWARD P S. Varieties of crossing dependencies; structure dependence and mild context sensitivity [J]. Cognitive Science, 2004, 28:699-720.
- [5] 邹崇理. 关于组合范畴语法 CCG[J]. 重庆理工大学学报(社会科学),2011,25(8):1-5.
- [6] CLARK S, CURRAN J R. The importance of supertagging for wide-coverage CCG parsing [C]. International conference on computational linguistics. Association for computational linguistics, 2004;282-288.
- [7] JOSHI A K. How much context sensitivity is necessary for characterizing structural descriptions-tree adjoining grammars [C]. Natural Language Processing, 1983.
- [8] ZHANG Y, CLARK S. Shift-Reduce CCG Parsing [C]. Meeting of the association for computational linguistics: human language technologies, proceedings of the conference. Portland: [S. l.], 2011:683-692.
- [9] ZETTLEMOYER L S, COLLINS M. Learning to map sentences to logical form; structured classification with probabilistic categorial grammars [J/OL]. [2016-01-16]. Eprint Arxiv, http://arxiv. org/abs/1207. 1420, 2012; 658-666.

(责任编辑 张佑法)