

试论自然语言处理中的哲学问题

吴昊磊

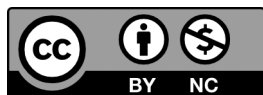
东北大学，沈阳

摘 要 | 分析了哲学上的理性主义和经验主义及其在自然语言处理的研究中的影响，对比了自然语言处理中基于规则的理性主义方法和基于统计的经验主义方法的优点和缺点，主张把这两种方法结合起来以推动自然处理研究的发展。本文还分析了哲学中知识本体的概念发展到知识本体工程的过程，说明了哲学对于自然语言处理的深刻影响。

关键词 | 哲学；自然语言处理；理性主义；经验主义

Copyright © 2022 by author (s) and SciScan Publishing Limited

This article is licensed under a [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/). <https://creativecommons.org/licenses/by-nc/4.0/>



1 引言

1992 年 6 月在加拿大蒙特利尔举行的第四届机器翻译的理论与方法国际会议（即 TMI-92）上，宣布会议的主题是“机器翻译中的经验主义和理性主义的方法”。所谓“理性主义”，就是指以转换生成语言学为基础的方法，所谓“经验主义”，就是指以大规模语料库的分析为基础的方法。可见，在国外的自然语言处理研究中，从 20 世纪 90 年代开始，就注意到了哲学中的理性主义与经验主义，试图从哲学的高度，来考察当前自然语言处理的发展趋势与动向。

作者简介：吴昊磊，东北大学本科学士。

文章引用：吴昊磊. 试论自然语言处理中的哲学问题 [J]. 哲学新探, 2022, 1 (1) : 1-11.

<https://doi.org/10.35534/nrp.0101001c>

我们自然语言处理的研究人员终日埋头于各种具体的研究工作中（这是“形而下”的工作），平时很少考虑哲学问题（这是“形而上”的问题）。在本文中，我们愿意抬起头来仰望一下哲学这一片包含了人类聪明和智慧的天空，从哲学的高度，来考察自然语言处理中的理性主义和经验主义，并进一步分析它们的利弊得失。同时，我们也从哲学的角度，研究哲学中的“本体知识”研究在自然语言处理中的作用。

2 哲学中的理性主义与经验主义

语言学中的理性主义来源于哲学中的理性主义（Rationalism）。在欧洲，这种理性主义源远流长，到了16世纪末至18世纪中期更加成熟，出现了笛卡儿（ReneDescartes, 1596–1650）、斯宾诺莎（BeneticdeSpinoza, 1632–1677）、莱布尼兹（CottfriedWilhelmLeibniz, 1646–1716）等杰出的理性主义哲学家。笛卡儿改造了传统的演绎法，制定了理性的演绎法，他认为，任何真理性的认识，都必须首先在人的认识中找到一个最确定、最可靠的支点，才能保证由此推出的知识也是确定可靠的。他提出在认识中应当避免偏见，要把每一个命题都尽可能地分解成细小的部分，直待能够圆满解决为止，要按照次序引导我们的思想，从最简单的对象开始，逐步上升到对复杂事物的认识。斯宾诺莎把几何学方法应用于伦理学研究，使用几何学的公理、定义、命题、证明等步骤来进行演绎推理，在他的《伦理学》的副标题中明确标示“依几何学方式证明”。莱布尼兹把逻辑学高度地抽象化、形式化、精确化，使逻辑学成为一种用符号进行演算的工具。笛卡儿是法国哲学家，斯宾诺莎是荷兰哲学家，莱布尼兹是德国哲学家，他们崇尚理性，提倡理性的演绎法。他们都居住在欧洲大陆，因此，理性主义也被称为“大陆理性主义”。

仰望一下哲学这片无比广阔的天空，我们发现，除了“理性主义”之外，在欧洲还存在着“经验主义”（Empiricism）哲学。经验主义以培根（FrancisBacon, 1561–1626）、霍布斯（ThomasHobbes, 1588–1679）、洛克（JohnLocke, 1632–1704）、休谟（DavidHume, 1711–1776）为代表，他们都是英国哲学家，因此，经验主义也被称为“英国经验主义”。培根批评理性派哲学家，他说，“理性

派哲学家只是从经验中抓到一些既没有适当审定也没有经过仔细考察和衡量的普遍例证,而把其余的事情都交给了玄想和个人的机智活动”。他提出“三表法”,制定了经验归纳法,建立了归纳逻辑体系,对于经验自然科学起了理论指导作用。霍布斯认为归纳法不仅包含分析,而且也包含综合,分析得出的普遍原因只有通过综合才能成为研究对象的特殊原因。洛克把理性演绎隶属于经验归纳之下,对演绎法作了经验主义的理解,他认为,一切知识和推论的直接对象是一些个别、特殊的事物,我们获取知识的正确途径只能是从个别、特殊进展到一般,他说,“我们的知识是由特殊方面开始,逐渐才扩展到概括方面的。只是在后来,人心就采取了另一条相反的途径,它要尽力把它的知识形成概括的命题”。休谟运用实验推理的方法来剖析人性,试图建立一个精神哲学体系,他指出,“一切关于事实的推理,似乎都建立在因果关系上面,只要依照这种关系来推理,我们便能超出我们的记忆和感觉的见证以外”,他认为,“原因和结果的发现,是不能通过理性,只能通过经验的”,经验是我们关于因果关系的一切推论和结论的基础。

现代自然科学的代表人物牛顿(Isaac Newton, 1642-1727)建立了经典力学的基本定律即牛顿三定律和万有引力定律,使经典力学的科学体系臻于完善。他的哲学思想也带有明显的经验主义倾向。他认为自然哲学只能从经验事实出发去解释世界事物,因而经验归纳法是最好的论证方法。他说:“虽然用归纳法来从实验和观察中进行论证不能算是普遍的结论,但它是事物本性所许可的最好的论证方法,并随着归纳的愈为普遍,这种论证看来也愈有力”。他把经验归纳作为科学研究的一般方法论原理,认为,“实验科学只能从现象出发,并且只能用归纳来从这些现象中推演出一般的命题”。正是由于牛顿遵循经验归纳法,才在物理学上取得了划时代的伟大成就。

法国启蒙运动的代表人物伏尔泰(Voltaire, 1694-1778)也有明显的经验主义倾向。他以洛克的经验主义为武器去反对教会至上的权威,否定神的启示和奇迹,否认灵魂不死。他赞美经验主义哲学家洛克:“也许从来没有一个人比洛克头脑更明智,更有条理,在逻辑上更为严谨”。他积极地把英国经验主义推行到法国,推动了法国的启蒙运动。

因此,当我们仰望哲学这片天空的时候,除了理性主义,还不能忽视经验主义。我们应当使用唯物辩证法的武器来分析和评价西方哲学中的理性主义和经验主义,权衡它们的利弊和得失,从而推动自然语言处理研究的发展。

3 自然语言处理中的理性主义与经验主义

早期的自然语言处理研究带有鲜明的经验主义色彩。1913 年,俄国科学家马尔可夫(A.Markov, 1856–1922)使用手工查频的方法,统计了普希金长诗《欧根·奥涅金》中的元音和辅音的出现频度,提出了马尔可夫随机过程理论,建立了马尔可夫模型,他的研究是建立在对于俄语的元音和辅音的统计数据的基础之上的,采用的方法主要是基于统计的经验主义的方法。

1948 年,山农把离散马尔可夫过程的概率模型应用于描述语言的自动机。他把通过诸如通信信道或声学语音这样的媒介传输语言的行为比喻为“噪声信道”或者“解码”。山农还借用热力学的术语“熵”作为测量信道的信息能力或者语言的信息量的一种方法,并且他采用手工方法来统计英语字母的概率,然后使用概率技术首次测定了英语字母的不等概率独立链的熵为 4.03 比特。后来,他有估算出英语字母的极限熵是 1.03 比特。山农的这些创造性研究为尔后自然语言处理的研究奠定了坚实的基础。山农的研究工作基本上是基于统计的,也带有明显的经验主义倾向。这种基于统计的经验主义的倾向到了乔姆斯基那里出现了重大的转向。

1956 年,乔姆斯基从山农的工作中吸取了有限状态马尔可夫过程的思想,首先把有限状态自动机作为一种工具来刻画语言的语法,并且把有限状态语言定义为由有限状态语法生成的语言,建立了自然语言的有限状态模型。乔姆斯基根据数学中的公理化方法来研究自然语言,采用代数和集合论把形式语言定义为符号的序列,从形式描述的高度,分别建立了有限状态语法、上下文无关语法、上下文有关语法和 0 型语法的数学模型,并且在这样的基础上来评价有限状态模型的局限性,乔姆斯基断言:有限状态模型不适合用来描述自然语言。这些早期的研究工作产生了“形式语言理论”这个新的研究领域,为自然语言和形式语言找到了一种统一的数学描述理论,形式语言理论也成了计算机科学

最重要的理论基石。乔姆斯基在他的著作中明确地采用理性主义的方法，他高举理性主义的大旗，把自己的语言学称之为“笛卡儿语言学”，充分地显示出乔姆斯基的语言学与理性主义之间不可分割的血缘关系。乔姆斯基完全排斥经验主义的统计方法。在1969年的Quine's Empirical Assumptions一文中，他说：“然而应当认识到，‘句子的概率’这个概念，在任何已知的对于这个术语的解释中，都是一个完全无用的概念”。他主张采用公理化、形式化的方法，严格地按照一定的规则来描述自然语言的特征，试图使用有限的规则描述无限的语言现象，发现人类普遍的语言机制，建立所谓的“普遍语法”。

语言知识的来源问题，是西方哲学中的“柏拉图问题”（Plato's Problem）的一个特例。所谓“柏拉图问题”是：我们可以得到的经验明证是如此贫乏，而我们是怎样获得如此丰富和具体明确的知识、如此复杂的信念和理智系统呢？人与世界的接触是那么短暂、狭隘、有限，为什么能知道那么多的事情呢？刺激的贫乏和所获得的知识之间为什么会存在如此巨大的差异呢？与“柏拉图问题”相应，人类语言知识的来源问题是：为什么人类儿童在较少直接语言经验的情况下，能够快速一致地学会语言？乔姆斯基认为，在人类成员的心智/大脑中，存在着由生物遗传而天赋决定的认知机制系统。在适当的经验引发或一定的经验环境下，这些认知系统得以正常地生长和成熟。这些认知系统叫作“心智器官”。决定构成人类语言知识的是心智器官中的一个系统，叫作“语言机能”。这个语言机能能在经验环境引发下的生长和成熟，决定着人类语言知识的获得。语言机能有初始状态和获得状态。初始状态是人类共同的、普遍一致的；获得状态是具体的、个别的。语言机能的初始状态叫作“普遍语法”（Universal Grammar，简称UG），语言机能的获得状态叫作“具体语法”（Particular Grammar，简称PG）。对普遍语法UG的本质特征及其与具体语法PG的关系的研究和确定，是解决关于语言知识的“柏拉图问题”的关键。

乔姆斯基把语言知识的使用问题叫作“笛卡儿问题”（Cartesian problem）。基于机械论哲学的物质概念，法国哲学家和数学家笛卡儿（Descartes）认为，所有非生命物质世界的现象、动物的生理与行为、大部分的人类器官活动，都能

够纳入物质科学的范畴。但是,笛卡儿又指出,某些现象不能处于物质科学的范畴之内,其中最为显著的就是人类语言,特别是“语言使用的创造性方面”,更是超出了机械论的物质概念所能够解释的范围。所以,对于语言的正常使用,是人类与其他动物或机器的真正区别。为了寻求对于语言这一类现象的解释,笛卡儿设定了一种“第二实体”的存在,这种第二实体就是“思维实体”。“思维实体”明显地不同于物质实体,它与物质实体相分离,并通过某种方式与物质实体相互作用。这一种“思维实体”就是心灵或者心智。语言知识的使用是内在于心智/大脑的,因此,对于这样的问题是很难解决和回答的。语言使用问题对于当年的笛卡儿来说是神秘的,目前对于我们而言也同样是神秘的。乔姆斯基认为,我们应当首先解决语言知识的本质问题和语言知识的来源问题,在这样的基础上,才有可能对于语言的使用问题进行有意义的探索。乔姆斯基坚持认为,语言机能内在于心智/大脑,对语言的研究是对心智的研究,最终是在抽象的水平上对大脑结构的研究。因此,生成语法研究在学科归属上属于“认知心理学”,最终属于“人类生物学”。它实际上应当叫作“生物语言学”。这是生成语法与其他任何传统的语言研究的根本区别。生成语法追求的目的,就是在理想化和抽象化的条件下,构建关于语言和心智的理论,它期待着与主体自然科学的统一,生成语法通过抽象的关于普遍语法、语言获得机制、所获得的状态以及语言与其他认知系统的关系的研究,不管好与坏、正确与错误,都是自然科学的组成部分。这就是生成语法“方法论的自然主义”。生成语法的这种自然主义的研究与自然科学的研究,在本质上是完全一致的。乔姆斯基力图把对于语言、心智的研究合成对于大脑的研究统一在一个共同的理论原则之下,最后把它纳入自然科学的总体研究之中。

乔姆斯基主张,语言是语言机能或者语言器官所呈现的状态,说某个人具有语言L,就是说他的语言技能处于状态L。语言机能所获得的状态能够生成无限数目的语言表达式,每一个表达式都是语音、结构和语义特征的某种排列组合。这个语言机能所获得的状态是一个生成系统或者运算系统。为了与一般人理解的外在语言相区别。乔姆斯基把这样的运算系统,叫作“I语言”。这里,字母I代表内在的、个体的、内涵的等概念。这意味着,I语言是心智的组成部分,

最终表现于大脑的神经机制之中，因此，I语言是“内在的”；I语言直接与个体有关，与语言社团存在见解的联系，语言社团的存在取决于该社团的成员具有相似的I语言，因此，I语言是“个体的”；I语言是一个函数或者生成程序，它生成一系列内在地表现与心智/大脑中的结构描写，因此，I语言是“内涵的”。

根据这种对于I语言的认识，乔姆斯基指出，基于社会政治和规范目的论因素之上的关于语言的通常概念，与科学的语言学研究没有任何关系，这些概念都不适合于用来进行科学的语言研究。生成语法对于语言的科学认识是内在主义的，而结构主义语法则外在于人类的心智/大脑。结构主义语法研究的方法，是在广泛搜集语言材料的基础上，通过切分、归类、替换等程序，概括出有关语言的语法规则。这些结构规则存在与外部世界，外在于人类的心智/大脑。结构主义语法研究的方法是经验主义的方法，这种方法的基础是外在于人类的语言观。乔姆斯基认为，根据结构主义语法的外在于人类的语言观，人们不能正确地认识和揭示人类语言的本质特征，不能解释人类语言知识获得的过程。只有内在主义的语言观才有可能正确地、全面地认识和解释人类语言知识的本质、来源和使用等问题。

乔姆斯基认为，生成语法的研究应当遵循自然科学研究中的“伽利略-牛顿风格”。“伽利略风格”的核心内容是：人们正在构建的理论体系是确实的真理，由于存在过多的因素和各种各样的事物，现象序列往往是对于真理的某种歪曲。所以，在科学研究中，最有意义的不是去考虑现象，而应当去寻求那些看起来确实能够给予人们深刻见解的原则。伽利略告诫人们，如果事实驳斥理论的话，那么，事实可能是错误的。伽利略忽视或无视那些有悖于理论的事实。“牛顿风格”的核心内容是：在目前的科学水平下，世界本身还是不可理解的，科学研究所要做的最好的事情就是努力构建可以被理解的理论，牛顿关注的是理论的可理解性，而不是世界本身的可理解性，科学理论不是为了满足常识理解而构建的，常识和直觉不足以理解科学的理论。牛顿摒弃那些无助于理论构建的常识和直觉。因此，“伽利略-牛顿风格”的核心内容是：人们应当努力构建最好的理论，不要为干扰理论解释力的现象而分散精力，同时应当认识到，世界与常识直觉是不相一致的。

生成语法的发展过程，处处体现着这种“伽利略－牛顿风格”。生成语法的目的是构建关于人类语言的理论，而不是描写语言的各种事实和现象。语言学理论的构建需要语言事实作为其经验的明证，但是，采用经验明证的目的是为了能够更好地服务于理论的构建，生成语法所采用的经验明证一般是与理论的构建有关的那些经验明证。因此，生成语法研究的目的不是全面地、广泛地、客观地描写语言事实和现象，而是探索 and 发现那些在语言事实和现象后面掩藏着本质和原则，从而构建解释性的语言学理论。所以，在生成语法看来，收集和获得的语言客观事实材料越多，越不利于人们对于语言本质特征的抽象性的把握和洞察。这是生成语法与当今广为流行的语料库语言学的根本区别。

最简单主义是生成语法的一个重要原则。最简单主义可以分为方法论最简单主义和实体性最简单主义。方法论最简单主义是从一般性科学方法论的思想和概念出发的，实体性最简单主义是就研究对象本身而言的。

方法论最简单主义要求人们在科学研究中创建最好的理论，而好的理论的主要标准就是最简单性。这种最简单性的表现是：在科学研究中使用最小数量的理论原则和理论构件；最大限度地减少复杂性，消除冗余性，增加理论原则的抽象性和概括性；构建最简单的理论模式和最具有解释性的理论；寻求理论的对称性和完美性。实体性最简单主义要求科学研究对象本身在设计 and 结构方面具有简单性、优化性和完美性。在最简单主义的思想原则下，生成语法的理论构建过程是一个逐步抽象化、概括化和最简单化的过程。在生成语法构建的早期，乔姆斯基就指出，虽然洪堡特在很早就认识到语言的本质是“有限规则的无限使用”，但是，由于当时缺少相应的技术手段，使得洪堡特的这种见解难以得到很好的发展。现代数学和逻辑学的发展，为生成语法提供了有力的形式化描述手段，使得生成语法在表达形式上与其他自然科学研究取得了一致。

在自然语言处理发展的过程中，始终充满了基于规则的理性主义方法和基于统计的经验主义方法之间的矛盾，这种矛盾时起时伏，此起彼伏。自然语言处理也就在这样的矛盾中逐渐成熟起来。

4 自然语言处理中理性主义方法和经验主义方法的利弊得失

仰望天空，总结历史，我们认为，基于规则的理性主义方法和基于统计的经验主义方法各有千秋，我们应当用科学的态度来权衡它们的利弊得失，来分析它们的优点和缺点。

基于规则方法的优点是：基于规则的方法中的规则主要是语言学规则，这些规则的形式描述能力和形式生成能力都很强，在自然语言处理中有很好的应用价值；基于规则的方法可以有效地处理句法分析中的长距离依存关系等困难问题，如句子中长距离的主语和谓语动词之间的一致关系问题、移位问题；基于规则的方法通常都是明白易懂的，表达得很清晰，描述得很明确，很多语言事实都可以使用语言模型的结构和组成成分直接地、明显地表示出来；基于规则的方法在本质上是没方向性的，使用这样的方法研制出来的语言模型，既可以应用于分析，也可以应用于生成，这样，同样的一个语言模型就可以双向使用；基于规则的方法可以在语言知识的各个平面上使用，可以在语言的不同维度上得到多维的应用。这种方法不仅可以在语音和形态的研究中使用，而且，在句法、语义、语用、篇章的分析中也大显身手。基于规则的方法与计算机科学中提出的一些高效算法是兼容的，例如，计算机算法分析中使用 Earley 算法（1970 年提出）和 Marcus 算法（1978 年提出）都可以作为基于规则的方法在自然语言处理中得到有效的使用。

基于规则的方法的缺点是：基于规则的方法研制的语言模型一般都比较脆弱，鲁棒性很差，一些与语言模型稍微偏离的非本质性的错误，往往会使得整个的语言模型无法正常地工作，甚至导致严重的后果。不过，近来已经研制出一些鲁棒的、灵活的剖析技术，这些技术能够使基于规则的剖析系统在剖析失败中得到恢复；使用基于规则的方法来研制自然语言处理系统的时候，往往需要语言学家、语音学家和各种专家的配合工作，进行知识密集的研究，研究工作的强度很大，基于规则的语言模型不能通过机器学习的方法自动地获得，也无法使用计算机自动地进行泛化；使用基于规则的方法设计的自然语言处理

系统的针对性都比较强,很难进行进一步的升级。例如,斯罗肯(Slocum)在1981年曾经指出,LIFER自然语言知识处理系统在经过两年的研发之后,已经变得非常之复杂和庞大,以至于这个系统原来的设计人很难再对它进行一点点的改动。对于这个系统的稍微改动将会引起整个连续的“水波效应”,以至于“牵一发而动全身”,而这样的副作用是无法避免和消除的;基于规则的方法在实际的使用场合其表现往往不如基于统计的方法那样好。因为基于统计的方法可以根据实际训练数据的情况不断地优化,而基于规则的方法很难根据实际的数据进行调整。基于规则的方法很难模拟语言中局部的约束关系,例如,单词的优先关系对于词类标注是非常有用的,但是基于规则的方法很难模拟这种优先关系。

知识本体本来是一个古老的哲学问题,近年来,知识本体的研究和设计成为了自然语言处理的基础性工程之一。本文列举的这些情况说明,哲学已经渗透到了自然语言处理的研究中,认真地学习哲学,已经成为自然语言处理研究必须面对的一个重要课题了。

参考文献

- [1] 冯志伟. 从知识本体看自然语言处理的人文性[J]. 语言文字应用, 2005(4).
- [2] 吴刚. 生成语法研究[M]. 上海: 上海外语教育出版社, 2006.
- [3] Chomsky N. The minimalist program[M]. Cambridge: MIT Press, 1995.
- [4] Feng Zhiwei. Evolution and present situation of corpus research in China[J]. International Journal of Corpus Linguistics, 2006, 11(2): 174-211.

On Philosophical Problems in Natural Language Processing

Wu Haolei

Northeastern University, Shenyang

Abstract: The author analyzes the philosophical rationalism and empiricism, and its effect in natural language processing research compared the rule-based rationalism in natural language processing methods and the advantages and disadvantages of the empirical method based on statistics, advocated to combine the two methods to promote the development of natural treatment research. This paper also analyzes the development of the concept of knowledge ontology from philosophy to knowledge ontology engineering and illustrates the profound influence of philosophy on natural language processing.

Key words: Philosophy; Natural language processing; Rationalism; The empirical