

心智与计算, Vol. 1, No. 3 (2007), 333-353
文章编号: MC - 2007-032
收稿日期: 2007-08-25
出版日期: 2007-09-30
© 2007 MC - 厦门大学信息与技术学院

自然语言处理中的哲学问题

冯志伟

(香港城市大学, 香港)

zwfeng@cityu.edu.hk

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; 基于规则的方法; 基于统计的方法; 知识本体; 知识本体工程

中图分类号: H 030

文献标识码: A

Some Philosophical Problems in Natural Language Processing

FENG Zhi-wei

(City University of Hong Kong, Hong Kong, China)

zwfeng@cityu.edu.hk

Abstract: In this paper, the rationalism and empiricism and its influence to natural language processing are analyzed, the advantages and the disadvantages of rule-based rationalist approach and statistics-based empiricist approach are compared. The author proposes to combine these two approaches in order to promote the development of natural language processing. The paper also discussed the procedure from ontology concept in philosophy to ontological engineering in natural language processing.

Key words: rationalism; empiricism; rule-based approach; statistics-based approach; ontology; ontology engineering

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,从上世纪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。

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”(ontology)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(Rene Descartes, 1596-1650)、斯宾诺莎(Benict de Spinoza, 1632-1677)、莱布尼兹(Cottfried Wilhelm Leibniz, 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论理学研究,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《论理学》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(Francis Bacon, 1561-1626)、霍布斯(Thomas Hobbes, 1588-1679)、洛克(John Locke, 1632-1704)、休谟(David Hume, 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幻想和个人的机智活动”^①。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”^②。休谟运用实

^① 《十六——十八世纪西欧各国哲学》,第23页。

^② 洛克,《人类理解论》,商务印书馆,第598页。

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”^③，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”^④，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”^⑤。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”^⑥。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”^⑦。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（Shannon）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”（noisy channel）或者“解码”（decoding）。山农还借用热力学的术语“熵”（entropy）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

^③ 休谟，《人类理解研究》，商务印书馆，第27页。

^④ 《十六——十八世纪西欧各国哲学》，第634页。

^⑤ 塞耶编，《牛顿自然哲学著作选》，商务印书馆，第212页。

^⑥ 塞耶编，《牛顿自然哲学著作选》，商务印书馆，第8页。

^⑦ 《十八世纪法国哲学》，商务印书馆，第59页，1963年。

的经验主义倾向^⑧。

这种基于统计的经验主义的倾向到了乔姆斯基 (Noam Chomsky) 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”(formal language theory) 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, 形式语言理论也成为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”^⑨。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”。

生成语法创立 50 年来, 在句法理论模式方面经过几次重大的变化, 不停顿地向新的方向发展。在这样的发展过程中, 赋予生成语法以生命活力的是生成语法的语言哲学理论。其中, 最为重要的是关于人类知识的本质、来源和使用问题。

乔姆斯基把语言知识的本质问题叫做“洪堡特问题”(Humboldt's problem)。德国学者洪堡特 (W. Humboldt) 曾经提出“语言绝不是产品 (Ergon), 而是一种创造性活动 (Energeria)”, 语言实际上是心智不断重复的活动, 它使音节得以成为思想的表达。人类语言知识的本质就是语言知识如何构成的问题, 其核心是洪堡特指出的“有限手段的无限使用”。语言知识的本质在于人类成员的“心智/大脑”(mind/brain) 中, 存在着的一套语言认知系统, 这样的认知系统表现为某种数量有限原则和规则体系。高度抽象的语法规则构成了语言应用所需要的语言知识, 由于人们不能自觉地意识到这些抽象的语法规则, 乔姆斯基主张, 这些语言知识是一些不言而喻的或者无意识的知识。我们应当把语言知识和语言的使用能力区分开来。两个人拥有同一语言的知识, 他们在发音、词汇知识、对于句子结构的掌握等方面是一样的。但是, 这两个人可能在语言使用的能力方面表现得非常不同。因此, 语言知识和语言能力是两个不同的概念。语言能力可以改进, 而语言知识则保持不变。语言能力可以损伤或者消失, 而人们并不至于失去语言知识。所以, 语言知识是内在于心智的特征和表现, 语言能力是外在行为的表现。生成语法研究的是语言的心智知识, 而不是语言的行为能力。语言知识体现为存在于心智/大脑中的认知系统。

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、

^⑧ 我国学者冯志伟在 20 世纪 70 年代末期, 模仿山农的研究, 用统计方法估算出汉字的熵为 9.65 比特, 也是采用经验主义的方法。

^⑨ Chomsky, N. Quine's Empirical Assumptions, In *Words and Objections*, Dordrecht: Reidel, 1969.

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。这些认知系统叫做“心智器官”(mental organs)。决定构成人类语言知识的是心智器官中的一个系统，叫做“语言机能”(language faculty)。这个语言机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(initial state)和获得状态(attained state)。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做“普遍语法”(Universal Grammar, 简称UG)，语言机能的获得状态叫做“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做“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(science of body)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”(thinking substance)。“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”(cognitive psychology)，最终属于“人类生物学”(human biology)。它实际上应当叫做“生物语言学”(biolinguistics)。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”(methodological naturalism)。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，叫做“I 语言”。这里，字母 I 代表内在的 (Internal)、个体的 (Individual)、内涵的 (Intensional) 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 (internalist) 的，而结构主义语法则是在外在主义的 (externalist)。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主义的语言观。乔姆斯基认为，根据结构主义语法的外在主义语言观，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” (Galilean-Newtonian style)。“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。伽利略告戒人们，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是全面地、广泛地、客观地描写语言事实和现象，而是探索和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 (minimalism) 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 (methodological minimalism) 和实体性最简单主义 (substantive minimalism)。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。

生成语法在 20 世纪 60 年代末到 70 年代时期在国际语言学界风靡一时，生成语法对于自然语言的形式化描述方法，为计算机处理自然语言提供了有力的武器，大力推动了自然语言处理的研究和发展。

生成语法的研究途径在一定程度上克服了传统语言学的某些弊病，推动了语言学理论和方法论的进步，但它认为统计只能解释语言的表面现象，不能解释语言的内在规则或生成机制，远离了早期自然语言处理的经验主义的途径。这种生成语法的研究途径实际上全盘承继了理性主义的哲学思潮。

在自然语言处理中的理性主义方法是一种基于规则的方法 (rule-based approach)，或者叫做符号主义的方法 (symbolic approach)。这种方法的基本根据是“物理符号系统假设”(physical symbol system hypothesis)。这种假设主张，人类的智能行为可以使用物理符号系统来模拟，物理符号系统包含一些物理符号的模式 (pattern)，这些模式可以用来构建各种符号表达式以表示符号的结构。物理符号系统使用对于符号表达式的一系列的操作过程来进行各种操作，例如，符号表达式的建造 (creation)、删除 (deletion)、复制 (reproduction) 和各种转换 (transformation) 等。自然语言处理中的很多研究工作基本上是在物理符号系统假设的基础上进行的。

这种基于规则的理性主义方法适合于处理深层次的语言现象和长距离依存关系，它继承了哲学中理性主义的传统，多使用演绎法 (deduction) 而很少使用归纳法 (induction)。

在自然语言处理中，在基于规则的理性主义方法的基础上发展起来的技术有：有限状态转移网络、有限状态转录机、递归转移网络、扩充转移网络、短语结构语法、自底向上剖析、自顶向下剖析、左角分析法、Earley 算法、CYK 算法、Tomita 算法、复杂特征分析法、合一运算、依存语法、一阶谓词演算、语义网络、框架网络等

在 20 世纪 50 年代末期到 60 年代中期，自然语言处理中的经验主义也兴盛起来，注重语言事实的传统重新抬头，学者们普遍认为：语言学的研究必须以语言事实作为根据，必须详尽地、大量地占有材料，才有可能在理论上得出比较可靠的结论。

自然语言处理中的经验主义方法是一种基于统计的方法 (statistic-based approach)，这种方法使用概率或随机的方法来研究语言，建立语言的概率模型。这种方法表现出强大的后劲，特别是在语言知识不完全的一些应用领域中，基于统计的方法表现得很出色。基于统计的方法最早在文字识别领域取得很大的成功，后来在语音合成和语音识别中大显身手，接着又扩充到自然语言处理的其他应用领域。

基于统计的方法适合于处理浅层次的语言现象和近距离的依存关系，它继承了哲学中经验主义的传统，多使用归纳法 (induction) 而很少使用演绎法 (deduction)。

这个时期自然语言处理中的经验主义派别，主要是一些来自统计学专业和电子学专业的研究人员。在 20 世纪 50 年代后期，贝叶斯方法 (Bayesian method) 开始被应用于解决最优字符识别的问题。1959 年，布莱德索 (Bledsoe) 和布罗宁 (Browning) 建立了用于文本识别的贝叶斯系统，该系统使用了一部大词典，计算词典的单词中所观察的字母系列的似然度，把单词中每一个字母的似然度相乘，就可以求出字母系列的似然度来。1964 年，墨斯特莱 (Mosteller) 和华莱士 (Wallace) 用贝叶斯方法成功地解决了在《联邦主义者》(The Federalist) 文章中的原作者的分布问题，显示出经验主义方法的优越性。

20 世纪 50 年代还建立了世界上第一个联机语料库：布朗美国英语语料库 (Brown corpus)。这个语料库包含 100 万单词的语料，样本来自不同文体的 500 多篇书面文本，涉及的文体有新闻、中篇小说、写实小说、科技文章等。这些语料是布朗大学 (Brown University) 在 1963—64 年收集的。随着语料库的出现，使用统计方法从语料库中自动地获取语言知识，成为了自然语言处理研究的一个重要方面。

20 世纪 60 年代，统计方法在语音识别算法的研制中取得成功。其中特别重要的是隐马尔可夫模型 (Hidden Markov Model) 和噪声信道与解码模型 (Noisy channel model and decoding model)。这些模型是分别独立地由两支队伍研制的。一支是杰里内克 (Fred Jelinek)，巴勒 (Bahl)，梅尔塞 (Mercer) 和 IBM 的华生研究中心的研究人员，另一支是卡内基梅隆大学 (Carnegie Mellon University) 的拜克 (Baker) 等。AT&T 的贝尔实验室 (Bell laboratories) 也是语音识别和语音合成的中心之一。

在自然语言处理中，在基于统计的方法的基础上发展起来的技术有：噪声信道理论、贝叶斯方法、最小编辑距离算法、加权自动机、Viterbi 算法、A*解码算法、隐马尔可夫模型、概率上下文无关语法、同步上下文无关语法、反向转录语法等。

不过，在 60 年代至 80 年代初期的这一个时期，在自然语言处理领域的主流方法仍然是基于规则的理性主义方法，经验主义方法并没有受到重视。

这种情况在 80 年代初期发生了变化。在 1983—1993 年的十年中，自然语言处理研究者对于过去的研究历史进行了反思，发现过去被忽视的有限状态模型和经验主义方法仍然有其合理的内核。在这十年中，自然语言处理的研究又回到了 50 年代末期到 60 年代初期几乎被否定的有限状态模型和经验主义方法上去，之所以出现这样的复苏，其部分原因在于 1959 年乔姆斯基对于斯金纳 (Skinner) 的“言语行为” (Verbal Behavior) 的很有影响的评论在 80 年代和 90 年代之交遭到了学术界在理论上的强烈反对，人们开始注意到基于规则的理性主义方法的缺陷。

这种反思的第一个倾向是重新评价有限状态模型，由于卡普兰 (R. M. Kaplan) 和凯依 (M. Kay) 在有限状态音系学和形态学方面的工作，以及丘奇 (Church) 在句法的有限状态模型方面的工作，显示了有限状态模型仍然有着强大的功能，因此，这种模型又重新得到自然语言处理学界的注意。

这种反思的第二个倾向是所谓的“重新回到经验主义”；这里值得特别注意的是语音和语言处理的概率模型的提出，这样的模型受到 IBM 公司华生研究中心的语音识别概率模型的强烈影响。这些概率模型和其他数据驱动的方法还传播到了词类标注、句法剖析、名词短语附着歧义的判定以及从语音识别到语义学的联接主义方法的研究中去。

统计方法在语音处理中取得的成就使得一些学者头脑发热起来，他们开始对于基于规则的理性主义方法进行攻击。IBM 公司华语音研究组的杰里内克 (Fred Jelinek) 于 1988 年 12 月 7 日在自然语言处

理评测的一次讨论会上曾经不无讽刺地说：“每当一个语言学家离开我们的研究组，语音识别率就提高一步。”帕尔默（Palmer）和费宁（Finin）在1990年介绍这个讨论会时，没有写下这段引文；一些当时参加会议的人回忆，杰里内克讲的话更为尖刻，他说：“每当我解雇一个语言学家，语音识别系统的性能就会提高一步。”可见，杰里内克对于基于语言学规则的理性主义方法是嗤之以鼻的，他的意见相当偏颇。

尽管有这些偏颇的看法，使用统计方法取得的成绩还是很鼓舞人心的，因此，从20世纪90年代开始，自然语言处理进入了一个新的阶段。1993年7月在日本神户召开的第四届机器翻译高层会议（MT Summit IV）上，英国著名学者哈钦斯（J. Hutchins）在他的特约报告中指出，自1989年以来，机器翻译的发展进入了一个新纪元。这个新纪元的重要标志是，在基于规则的技术中引入了语料库方法，其中包括统计方法，基于实例的方法，通过语料加工手段使语料库转化为语言知识库的方法，等等。这种建立在大规模真实文本处理基础上的机器翻译，是机器翻译研究史上的一场革命，它将会把自然语言处理推向一个崭新的阶段。

在过去的四十多年中，从事自然语言处理系统开发的绝大多数学者，基本上都采用基于规则的理性主义方法，这种方法主张，智能的基本单位是符号，认知过程就是在符号的表征下进行符号运算，因此，思维就是符号运算。

著名语言学家弗托（J. A. Fodor）在《Representations》一书（MIT Press, 1980）中说：“只要我们认为心理过程是计算过程（因此是由表征式定义的形式操作），那么，除了将心灵看作别的之外，还自然会把它看作一种计算机。也就是说，我们会认为，假设的计算过程包含哪些符号操作，心灵也就进行哪些符号操作。因此，我们可以大致上认为，心理操作跟图灵机的操作十分类似。”^⑩ 弗托的这种说法代表了自然语言处理中的基于规则（符号操作）的理性主义观点。

这样的观点受到了学者们的批评。塞尔（J. R. Searle）在他的论文《Minds, Brains and Programmes》^⑪中，提出了所谓“中文屋子”的质疑。他提出，假设有一个懂得英文但是不懂中文的人被关在一个屋子中，在他面前是一组用英文写的指令，说明英文符号和中文符号之间的对应和操作关系。这个人要回答用中文书写的几个问题，为此，他首先要根据指令规则来操作问题中出现的中文符号，理解问题的含义，然后再使用指令规则把他的答案用中文一个一个地写出来。比如，对于中文书写的问题Q1用中文写出答案A1，对于中文书写的问题Q2用中文写出答案A2，如此等等。人并不是计算机，这样的事情对于计算机来说也许易如反掌，但是，对于一个活生生的人来说，这显然是非常困难的几乎是不能实现的事情，而且，这个人即使能够这样做，也不能证明他懂得中文，只能说明他善于根据规则做机械的操作而已。塞尔的批评使自然语言处理中基于规则的理性主义的方法受到了普遍的怀疑。

理性主义方法的另一个弱点是在实践方面的。自然语言处理的理性主义者把自己的目的局限于某个十分狭窄的专业领域之中，他们采用的主流技术是基于规则的句法分析技术和语义分析技术，尽管这些应用系统在某些受限的“子语言”（sub-language）中也曾经获得一定程度的成功，但是，要想进一步扩大这些系统的覆盖面，用它们来处理大规模的真实文本，仍然有很大的困难。因为从自然语言系统所需要装备的语言知识来看，其数量之浩大和颗粒度之精细，都是以往的任何系统所远远不及的。而

^⑩ J. A. Fodor, *Representations*, MIT Press, 1980.

^⑪ J. R. Searle, *Minds, Brains and Programmes*, In *Behavioral and Brain Sciences*, Vol.3, 1980.

且,随着系统拥有的知识在数量上和程度上发生的巨大变化,系统在如何获取、表示和管理知识等基本问题上,不得不另辟蹊径。这样,就提出了大规模真实文本的自然语言处理问题。1990年8月在芬兰赫尔辛基举行的第13届国际计算语言学会议(即COLING'90)为会前讲座确定的主题是:“处理大规模真实文本的理论、方法和工具”,这说明,实现大规模真实文本的处理将是自然语言处理在今后一个相当长的时期内的战略目标。为了实现战略目标的转移,需要在理论、方法和工具等方面实行重大的革新。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”。这里的所谓“理性主义”,就是指以转换生成语法为基础的基于规则的方法,所谓“经验主义”,就是指以大规模语料库的分析为基础的基于统计的方法。从中可以看出当前自然语言处理关注的焦点。当前语料库的建设和语料库语言学的崛起,正是自然语言处理战略目标转移的一个重要标志。随着人们对大规模真实文本处理的日益关注,越来越多的学者认识到,基于语料库的分析方法(即经验主义的方法)至少是对基于规则的分析方法(即理性主义的方法)的一个重要补充。因为从“大规模”和“真实”这两个因素来考察,语料库才是最理想的语言知识资源。

在这样的情况下,人们开始深入地思考,乔姆斯基的“普遍语法”是否是真正的语言规则?是否能够经受大量的语言事实的检验?语言规则是否应该和语言事实结合起来考虑,而不是一头钻入理性主义的牛角尖?

乔姆斯基作为一位求实求真、虚怀若谷的语言学大师,最近他也开始对于理性主义进行了反思,表现了与时俱进的勇气。在最近他提出的“最简方案”中,他认为,所有重要的语法原则直接运用于表层,不同语言之间的差异通过词汇来处理,把具体的规则减少到最低限度,开始注重对具体的词汇的研究。可以看出,乔姆斯基的转换生成语法也开始对词汇重视起来,逐渐地改变了原来的理性主义的立场,开始与经验主义妥协,或者悄悄地向经验主义复归。

在20世纪90年代的最后五年(1994-1999),自然语言处理的研究发生了很大的变化,出现了空前繁荣的局面。概率和数据驱动的方法几乎成为了自然语言处理的标准方法。句法剖析、词类标注、参照消解和话语处理的算法全都开始引入概率,并且采用从语音识别和信息检索中借过来的评测方法。理性主义君临天下的局面已经被打破了,基于统计的经验主义方法逐渐成为自然语言处理研究的主流。

现在我们已经进入21世纪,语料库方法已经渗透到了机器翻译研究的各个方面,一些基于语料库的机器翻译系统已经建立起来,有的系统把基于语料库的方法和基于规则的方法巧妙地结合起来,取得了可喜的成绩。

2000年,在约翰·霍普金斯大学(Johns Hopkins University)的暑假机器翻译讨论班(Workshop)上,来自南加州大学、罗切斯特大学、约翰·霍普金斯大学、施乐公司、宾西法尼亚州大学、斯坦福大学等学校的研究人员,对于基于统计的机器翻译进行了讨论,以年轻的博士研究生奥赫(Franz Josef Och)为主的13位科学家写了一个总结报告(Final Report),报告的题目是《统计机器翻译的句法》(**Syntax for Statistical Machine Translation**),这个报告提出统计机器翻译方法的有效途径。

奥赫在国际计算语言学2002年的会议(ACL2002)上发表论文,题目是:《统计机器翻译的分辨训练与最大熵模型》(**Discriminative Training and Maximum Entropy Models for Statistical Machine Translation**),进一步提出统计机器翻译的系统性方法,获ACL2002大会最佳论文奖

目前,统计机器翻译已经成为机器翻译研究的主流。

根据 Google 的调查,统计机器翻译论文发表的情况如图 1 所示:

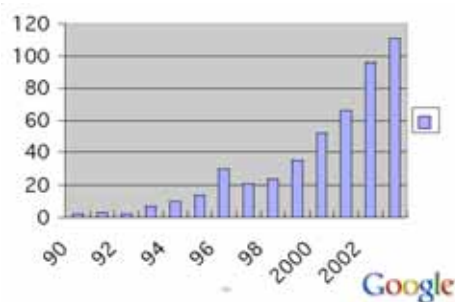


图 1 统计机器翻译论文增长情况

Fig.1 The paper growth status of statistical machine translation

从图 1 中可以看出,统计机器翻译的论文是成线性增长的,其增长速度越来越快。

根据美国 NIST (National Institute of Standardization & Technology)组织的统计机器翻译评测,汉语-英语机器翻译系统和阿拉伯语-英语机器翻译系统的 BLEU 指标如图 2 所示¹²:

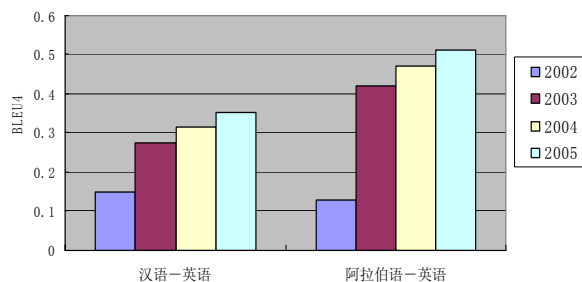


图 2 统计机器翻译系统的 BLEU 指标逐年提高

Fig.2 The BLEU index in statistical machine translation system increases annually

从图 2 中可以看出,统计机器翻译的质量正在逐年提高。

统计机器翻译的质量与语言模型的规模有密切关系。随着语言模型训练数据的增大,机器翻译的译文质量相应提高。如图 3 所示:



图 3 英语-阿拉伯语机器翻译系统的质量与训练数据规模的关系

Fig.3 The relationship between quality and training data scale in the English - Arabic machine translation system

2003 年 7 月,在美国马里兰州巴尔的摩 (Baltimore, Maryland) 由美国商业部国家标准与技术研

¹² Bleu 是机器翻译自动评测的一个指标,它综合地考虑了机器翻译译文的忠实度和流畅度。Bleu 的数值越大,机器翻译的译文越好。

究所 NIST/TIDES (National Institute of Standards and Technology) 主持的评比中, 奥赫获最好成绩, 他使用统计方法, 在很短的时间之内就构造了阿拉伯语和汉语到英语的若干个机器翻译系统。伟大的希腊科学家阿基米德(Archimedes)说过:“只要给我一个支点, 我就可以移动地球。”(“Give me a place to stand on, and I will move the world.”) 而现在奥赫也模仿着阿基米德说:“只要给我充分的并行语言数据, 那么, 对于任何的两种语言, 我可以在几小时之内给你构造出一个机器翻译系统。”(“Give me enough parallel data, and you can have translation system for any two languages in a matter of hours.”) 这反映了新一代的机器翻译研究者朝气蓬勃的探索精神和继往开来的豪情壮志。看来, 奥赫似乎已经找到了机器翻译的有效方法, 至少按照他的路子走下去, 也许有可能开创出机器翻译研究的一片新天地, 使我们在探索真理的曲折道路上看到了耀眼的曙光。过去我们研制一个机器翻译系统往往需要几年的时间, 而现在采用奥赫的方法构造机器翻译系统只要几个小时就可以了, 研制机器翻译系统的速度已经大大地提高了, 问题是目前机器翻译的质量仍然还不能令人满意, 要实现高质量的机器翻译, 我们确实还要进行艰苦的探索。

可以看出, 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(long-distance dependencies)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(subject-verb agreement)问题, wh 移位(wh-movement)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双向性的, 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” (ripple effect)，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。

不过，尽管基于规则的方法有这样的或那样的不足，这种方法终究是自然语言处理中研究得最为深入的技术，它仍然是非常有价值和非常强有力的技术，我们决不能忽视这种方法。事实证明，基于规则的方法的算法具有普适性，不会由于语种的不同而失去效应，这些算法不仅适用于英语、法语、德语等西方语言，也适用于汉语、日语、韩国语等东方语言。在一些领域针对性很强的应用中，在一些需要丰富的语言学知识支持的系统中，特别是在需要处理长距离依存关系的自然语言处理系统中，基于规则的方法是必不可少的。

基于统计的方法的优点是：

使用基于统计的方法来训练语言数据，从训练的语言数据中获取语言的统计知识，可以有效地建立语言的统计模型。这种方法在文字和语音的自动处理中效果良好。

基于统计的方法的效果在很大的程度上依赖于训练语言数据的规模，训练的语言数据越多，基于统计的方法的效果就越好。

基于统计的方法很容易与基于规则的方法结合起来，从而处理语言的约束问题，以提高系统的效能。

基于统计的方法很适合用来模拟那些有细微差别的、不精确的、模糊的概念（如“很少、很多、若干”等），而这些概念，在传统语言学中需要使用模糊逻辑 (fuzzy logic) 才能处理。

基于统计的方法的缺点是：

使用基于统计的方法研制的自然语言处理系统，其运行时间是与统计模式中所包含的符号类别的多少成比例线性地增长的，不论在训练模型的分类中或者是在测试模型的分类中，情况都是如此。因此，如果统计模式中的符号类别数量增加，系统的运行效率会明显地降低。

在当前语料库技术的条件下，要使用基于统计的方法为某个特殊的应用领域获取训练数据，还是一件费时费力的工作，而且很难避免出错。基于统计的方法的效果与语料库的规模、代表性、正确性以及加工深度都有密切的关系，可以说，用来训练数据的语料库的质量决定了基于统计方法的效果。

基于统计的方法很容易出现数据稀疏的问题，随着训练语料库规模的增大，数据稀疏的问题会越来越

越严重，这个问题需要使用各种“平滑”（smoothing）的技术来解决。

自然语言中既有深层次的现象，也有浅层次的现象，既有远距离的依存关系，也有近距离的依存关系，自然语言处理中既要使用演绎法，也要使用归纳法。因此，我们主张把理性主义和经验主义结合起来，把基于规则的方法和基于统计的方法结合起来。我们认为，强调一种方法，反对另一种方法，都是片面的，都无助于自然语言处理的发展。近年来，在统计机器翻译（Statistical Machine Translation，简称 SMT）中，一些学者开始把短语知识、句法知识逐渐地导入到系统中，致力于研制“基于句法的统计机器翻译”（Syntax-based Statistical Machine Translation，简称 SSMT），有效地改善了统计机器翻译的译文质量，这是非常可喜的。

美国南加州大学信息系统研究所的纳依特（Kevin Knight）教授对于机器翻译从 1954 年到 2004 年的发展过程做了如下的总结（图 4）：

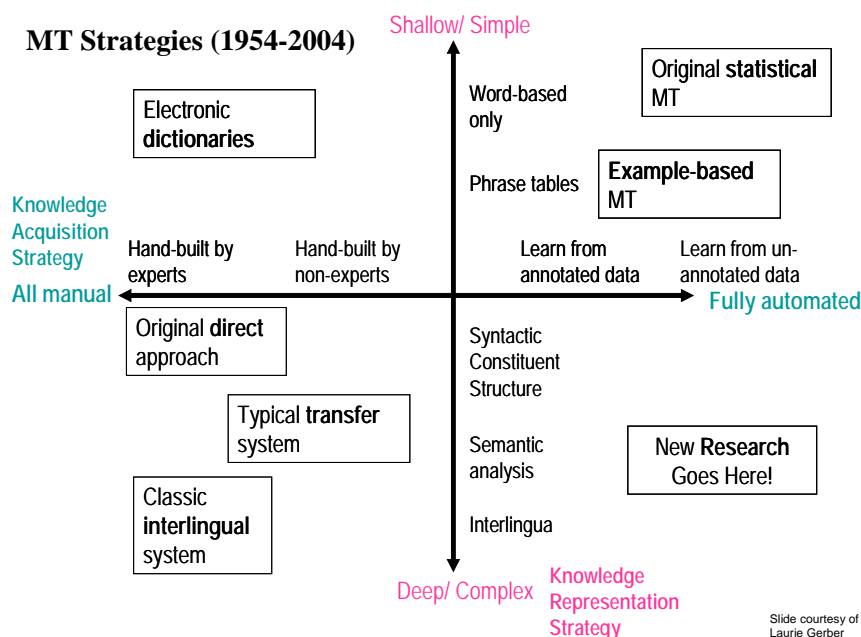


图 4 机器翻译发展 50 年的图式

Fig.4 The map of the 50 years' machine translation development

横轴表示自动化的程度，从 1954 年到 2004 年，由 full manual 发展到 fully automated；纵轴表示加工的深度，从 1954 年 2004 年，由 shallow/simple 发展到 deep/complex。当前的任务是，自动化程度要高，加工的程度要深。这就应当把语言知识融入到统计机器翻译中，把理性主义的方法和经验主义的方法结合起来。

英国经验主义哲学家培根既反对理性主义，也反对狭隘的经验主义，他指出，由于经验能力和理性能力这两方面的“离异”和“不和”，给科学知识的发展造成了严重的障碍，为了克服这样的弊病，他提出了经验能力和理性能力联姻的重要原则。他说，“我以为我已经在经验能力和理性能力之间永远建立了一个真正合法的婚姻，二者的不和睦与不幸的离异，曾经使人类家庭的一切事务陷于混乱”¹³。他生动而深刻地说道：“历来处理科学的人，不是实验家，就是教条者。实验家像蚂蚁，只会采集和使用；推论家像蜘蛛，只凭自己的材料来织成丝网。而蜜蜂却是采取中道的，它在庭园里和田野里从花朵中采

¹³ 《十六——十八世纪西欧各国哲学》，第 8 页。

集材料，而用自己的能力加以变化和消化。哲学的真正任务就正是这样，它既非完全或主要依靠心的能力，也非只把从自然历史和机械实验收来的材料原封不动，囫圇吞枣地累置于记忆当中，而是把它们变化过和消化过放置在理解力之中。这样看来，要把这两种机能、即实验的和理性的这两种机能，更紧密地和更精纯地结合起来（这是迄今还未收到的），我们就可以有很多的希望”¹⁴。

培根是著名的哲学家，他的主张是值得我们深思的。我们不能采取像蜘蛛那样的理性主义方法，单纯依靠规则，也不能采取像蚂蚁那样的经验主义方法，单纯依靠统计，我们应当像蜜蜂那样，把理性主义和经验主义两种机能更紧密地、更精纯地结合起来，推动自然语言处理的发展。

5 知识本体在自然语言处理中的指导作用

哲学对于自然语言处理的影响还表现在哲学中“知识本体”（ontology）的研究逐渐发展成为为了具有工程性特征的“知识本体工程”（ontology engineering）的研究。

知识本体本来是一个哲学概念。如果我们对于一个领域中的客体进行分析，找出这些客体之间的关系，获得了这个领域中不同客体的集合，这一个集合可以明确地、形式化地、可共享地描述这个领域中各个客体所代表的概念的体系，它实际上就是概念体系的规范，这样的概念体系规范就可以看成这个领域的“知识本体”（ontology）。

人们很早就开始研究知识本体，因此，知识本体有很多不同的定义，这些定义有的是从哲学思辨出发的，有的是从知识的分类出发的，最近的一些定义则是从实用的计算机推理出发的。

牛津英语词典对于知识本体（ontology）的定义是：“对于存在的研究或科学”（the science or study of being），这个定义显然是非常广泛的，因为它试图研究存在的一切事物，为存在的一切事物建立科学。不过，这个定义确实是关于知识本体的经典定义，它来自哲学研究。

什么是事物（things）？什么是本质（essence）？当事物发生改变时，本质是否仍然存在于事物之中？概念（concept）是否存在于我们的心智（mind）之外？怎样对世界上的实体（entities）进行分类？这些都是知识本体要回答的问题，所以，知识本体是“对于存在（being）的研究或科学”。

远古希腊时代，哲学家就试图研究当事物发生变化的时候，如何去发现事物的本质。例如，当植物的种子发育变成树的时候，种子不再是种子了，而树开始成为了树。那么，树还包含着种子的本质吗？巴门尼德（Parmenides）认为，事物的本质是独立于我们的感官的，种子在表面上虽然变成了树，但是，它的本质是没有改变的，所以，在实质上种子并没有转化为树，只不过是我们的感官原来感到它是种子，后来感到它是树。亚里士多德（Aristotle）认为，种子只不过是还没有完全长成的树，在发育过程中，树的本质并没有改变，只是改变了它存在的形式，从没有完成长成的树（潜在的树）变成了完全长成的树（实在的树）。种子和树的本质都是一样的。知识本体就要研究关于事物的本质的问题。亚里士多德还把存在区分为不同的模式，建立了一个范畴系统（system of categories），包含的范畴有10个：substance（实体），quality（质量），quantity（数量），relation（关系），action（行动），passion（感情），place（空间），time（时间）、主动（active）、被动（passive）。这个范畴系统是最

¹⁴ 培根，《新工具》，商务印书馆，第75页。

早的概念体系。

在中世纪，学者们研究事物本身和事物的名称之间的关系，分为唯实论（realism）和唯名论（nominalism）两派。唯实论主张，事物的名称就是事物本身，而唯名论主张，事物的名称只不过是引用事物的词而已。在中世纪晚期，大多数学者都倾向于认为，事物的名称只是表示事物的符号（symbol），例如，book 这个名称只不过是用来引用一切作为实体的“书”的一个符号。这是现代物理学的一个起点，在现代物理学中，采用不同符号来表示物理世界的各种特征（如，速度的符号为 V ，长度的符号为 L ，能量的符号为 E ，等）。这些用符号表示的特征，实际上都是物理学中的概念或范畴。

1613 年，德国哲学家郭克兰纽（R. Goclenius）在他用拉丁文编写的《哲学辞典》中，把希腊语的 on （也就是 being）的复数 onta （也就是 beings）与 logos （含义为“学问”）结合在一起，创造出 ontologia 这个术语。 ontologia 也就是英文的 ontology ，这是西方文献中最早出现的 ontology 这个术语。1636 年，德国哲学家卡洛维（A. Calovius）在《神的形而上学》中，把 ontologia 看成“形而上学”（ metaphysica ；英文为 metaphysics ）的同义词，这样，他便把“ ontologia ”与亚里士多德的“形而上学”紧密地联系起来。法国哲学家笛卡尔（R. Descartes）更是明确地把研究本体的第一哲学叫做“形而上学的 ontologia ”，这样， ontologia 便成为形而上学的一个部分了。德国哲学家莱布尼兹（G. von Leibniz）和他的继承者沃尔夫（C. Wolff）更是从学科分类的角度，把 ontologia 归属为形而上学的一个分支，使 ontologia 成为了哲学中一个相对独立的分支学科。 ontologia 这个术语，在哲学中翻译为“本体论”，在自然语言处理中，从应用的角度出发，我们认为翻译为“知识本体”更为恰当。因此，在本文中，我们统一地使用“知识本体”这个术语。

德国哲学家康德（Emmanuel Kant）也研究知识本体，他认为，事物的本质不仅仅由事物本身决定，也受到人们对于事物的感知或理解的影响。康德提出这样的问题：“我们的心智究竟是采用什么样的结构来捕捉外在世界的呢？”为了回答这个问题，康德对范畴进行了分类，建立了康德的范畴框架，这个范畴框架包括 4 个大范畴：quantity（数量），quality（质量），relation（关系），modality（模态）。每一个大范畴又分为 3 个小范畴。Quantity 又分为 unity（单量），plurality（多量），totality（总量）3 个范畴；quality 又分为 reality（实在质），negation（否定质），limitation（限度质）3 个范畴；relation 又分为 inherence（继承关系），causation（因果关系），community（交互关系）3 个范畴；modality 又分为 possibility（可能性），existence（现实性），necessity（必要性）。根据这个范畴框架，我们的心智就可以给事物进行分类。从而获得对于外界世界的认识。例如，本文作者冯志伟属于的范畴是：unity, reality 和 existence，这样，我们就认识到：冯志伟是一个“单一的、实在的、现实的”人。在数据库中，我们可以根据康德的方法给事物建立一些范畴，从而根据这些范畴来管理数据。例如，我们给人事管理数据库建立“姓名，性别，籍贯，职业”等范畴，使用这些范畴进行人事管理。可以看出，康德对于范畴框架的研究，为知识本体的研究奠定了坚实的基础。不过，他的这个范畴框架不同于亚里士多德的范畴系统，康德在他的《纯粹理性批判》的著作中明确地反对亚里士多德的 10 个范畴。

1991 年，美国计算机专家尼彻斯（R. Niches）等在完成美国国防部高级研究计划局（Defense Advanced Research Projects Agency，简称 DARPA）的一个关于知识共享的科研项目中，提出了一种构建智能系统方法的新思想，他们认为，构建的智能系统由两个部分组成，一个部分是“知识本体”

(Ontology), 一个部分是“问题求解方法”(Problem Solving Methods, 简称 PSMs)。知识本体涉及特定知识领域共有的知识和知识结构, 它是静态的知识, 而 PSMs 涉及在相应知识领域进行推理的知识, 它是动态的知识, PSMs 使用知识本体中的静态知识进行动态的推理, 就可以构建一个智能系统。这样的智能系统就是一个知识库, 而知识本体是知识库的核心, 这样, 知识本体在计算机科学中就引起了学者们的极大关注。

1990 年, 我国学者冯志伟提出“双态理论”(di-states theory), 他主张, 在机器翻译系统中, 要把静态标记(static label)和动态标记(dynamic label)结合起来, 静态标记要表示存储在机器词典中的单词的词类特征和单词固有的语义特征, 它们是与单词所在的上下文语境无关的, 动态标记是使用静态标记经过计算机运算求出来的句法功能标记、语义关系标记、逻辑关系标记, 它们是要根据单词的上下文语境来确定的。静态信息的制定要根据词类和语义系统的规范, 动态标记的求解要根据产生式规则, 产生式规则的基本形式是“条件-动作”偶对, 因此, 面向机器翻译的语言学研究要着重阐明规则的条件。冯志伟所说的词类规范, 实际上就是语法信息的规范, 冯志伟所说的语义系统的规范, 实际上就是概念系统的规范, 也就是“知识本体”。

冯志伟的构想可以表示为:

基于语法信息和知识本体的静态标记标注的机器词典 + 基于产生式规则的动态标记求解规则 = 机器翻译系统

尼彻斯的构想可以表示为:

静态的“知识本体” + 动态的“问题求解方法” = 知识库

通过比较可以看出: 冯志伟关于静态标记与动态标记相结合的“双态理论”构想, 与尼彻斯关于静态的“知识本体”与动态的“问题求解方法”相结合的构想是非常相似的。

在 20 世纪末和 21 世纪初, 知识本体的研究开始成为计算机科学的一个重要领域。它主要任务是研究世界上的各种事物(例如, 物理客体、事件等)以及代表这些事物的范畴(例如, 概念、特征等)的形式特性和分类。计算机科学对于知识本体的研究当然是建立在上述的经典的知本体的研究的基础之上的, 不过, 有了很大的发展。因此, 我们有必要重新给知识本体下定义。下面, 我们介绍在计算机科学中对于知识本体的定义。

在人工智能研究中, 格鲁伯(Gruber)在 1993 年给知识本体下的定义是:

“知识本体是概念体系的明确规范”(An ontology is an explicit specification of conceptualization)。

这个定义比较具体, 也比较便于操作, 在知识本体的研究中广为传布。

1997 年, 波尔斯特(Borst)对格鲁伯的定义做了很小修改; 提出了如下的定义:

“知识本体是可以共享的概念体系的形式规范”(Ontologies are defined as a formal specification of a shared conceptualization)。

1998 年, 施图德(Studer)等在格鲁伯和波尔斯特的定义的基础上, 对于知识本体给出了一个更加明确的解释:

“知识本体是对概念体系的明确的、形式化的、可共享的规范”(An ontology is a formal explicit specification of a shared conceptualization)。

在这个定义中,所谓“概念体系”是指所描述的客观世界的现象中有关概念的抽象模型;所谓“明确”是指对于所使用的概念的类型以及概念用法的约束都明确地加以定义;所谓“形式化”是指这个知识本体应该是机器可读的;所谓“共享”是指知识本体中所描述的知识不是个人专有的而是集体共有的。

具体地说,如果我们把每一个知识领域抽象成一个概念体系,再采用一个词表来表示这个概念体系,在这个词表中,要明确地描述词的涵义、词与词之间的关系、并在该领域的专家之间达成共识,使得大家能够共享这个词表,那么,这个词表就构成了该领域的一个知识本体。知识本体已经成为了提取、理解和处理领域知识的工具,它可以被应用于任何具体的学科和专业领域,知识本体经过严格的形式化之后,借助与计算机强大的处理能力,可以对于人类的全部知识进行整理和组织,使之成为一个有序的知识网络。

人们对于知识本体的认识可能存在差别,因此,有不同类型的知识本体。

- 通用知识本体 (common ontology) 常常从哲学的认识论出发,概念的根结点往往是很抽象的,例如,时间、空间、事件、状态、对象等。
- 领域知识本体 (domain ontology) 对领域的知识进行抽象,概念比较具体,容易进行形式化和共享。例如,我国学者最近研制的生物学领域知识本体 (domain-specific ontology of botany)、考古学领域知识本体 (domain-specific ontology of archeology) 都是领域知识本体。
- 语言知识本体 (language ontology) 常常表现为一个词表,其中要描述单词和术语之间的概念关系,词网 (WordNet) 就是一个语言知识本体。如果语言知识本体中的概念结点是专业术语,那么,这样的语言知识本体就叫做术语知识本体 (terminology ontology)。术语是科学技术知识在自然语言中的结晶,哪里有科学技术,哪里就有术语,所以,术语知识本体对于领域知识的处理是非常重要的。
- 形式知识本体 (formal ontology) 对于概念和术语的分类很严格,要按照一定的原则和标准,明确地定义概念之间的显性和隐性关系,明确概念的约束和逻辑联系。领域知识本体或术语知识本体经过进一步的抽象和提炼,就可能发展成形式知识本体。

知识本体可以帮助我们对于领域知识进行系统的分析,把领域知识形式化,使之便于计算机处理。知识本体还可以实现人和人之间以及人和计算机之间知识的共享,实现在一定领域中知识的重复使用。在自然语言处理的语义分析中,知识本体可以给我们提供单词的各种信息,帮助我们揭示单词之间的各种语义关系,是语义分析的知识来源。

目前,支持知识本体的开发工具已经有数十种,功能各不相同,对于知识本体语言的支持能力、表达能力各有差别,可扩展性、灵活性、易用性也不一样。其中比较著名的有 Protégé-2000、OntoEdit、OilEd、Ontolingua 等。Protégé-2000 是使用比较广泛的知识本体工具,是可以免费获得的开放软件,它用 Java 语言开发,通过各种插件支持多种知识本体格式。

冯志伟在日汉机器翻译的研究中,设计了一个知识本体系统 ONTOL-MT。这个知识本体的初始概念有事物 (entity)、时间 (time)、空间 (space)、数量 (quantity)、行为状态 (action-state) 和属性 (attribute) 6 个。在这 6 个初始概念之下,还有不同层次的下位概念。

ONTOL-MT 的基本结构如图 5 所示。

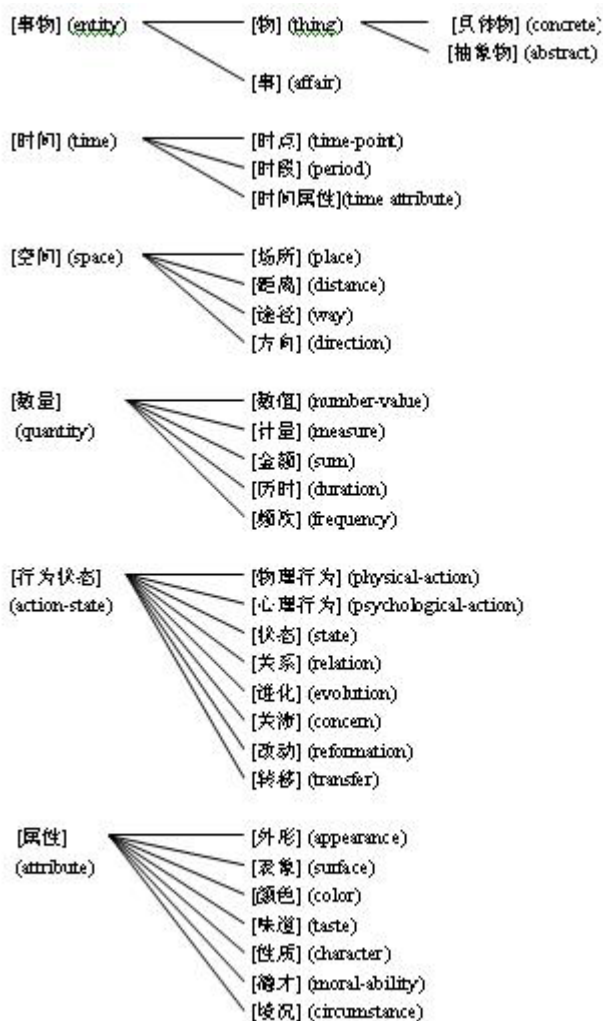


图 5 知识本体 ONTOL-MT

Fig.5 Ontology ONTOL-MT

ONTOL-MT 中的上述主要概念的涵义定义如下：

[事物] (entity) 在空间（包括思维空间）上和时间上延展的事物本体。

[物] (thing) 主要在空间（包括思维空间）上延展的事物本体。

[具体物] (concrete) 有形、有色、有质量的物。

[抽象物] (abstract) 无形、无色、无质量的物。

[事] (affair) 主要在时间上延展的事物本体。包括人类生活中的一切活动和所遇到的一切社会现象（政治、军事、法律、经济、文化、教育）或与人有关联的自然现象。

[时间] (time) 由过去、现在和将来构成的连绵不断的系统，它是物质运动和变化的持续性的表现，是物质存在的一种客观形式。

[时点] (time-point) 指时间里的某一点。

[时段] (period) 指有起点和终点的一段时间。

[时间属性] (time-attribute) 时间所具有的属性（年、月、日、小时、分、秒、毫秒等）。

[空间] (space) 事物及其运动存在的另一种客观形式，它在不同的维度上延伸。

- [场所] (place) 由长度、宽度和高度表现出来的物质存在的一种客观形式, 也就是活动的处所。
- [距离] (distance) 在空间或者时间上相隔。
- [途径] (way) 两地之间的通道。
- [方向] (direction) 指东、南、西、北、上、下等。
- [数量] (quantity) 事物的多少与计量。
- [数值] (number-value) 一个量用数目表示出来的多少。
- [计量] (measure) 温度, 长度, 重量, 使用量等。
- [金额] (sum) 钱数的多少。
- [历时] (duration) 时间在数量上的长短。
- [频次] (frequency) 事情发生的频繁程度的大小。
- [行为状态] (action-state) 人或事物表现出来的活动和形态。
- [物理行为] (physical-action) 人或事物在物理上表现出来的活动。
- [心理行为] (psychological-action) 人或动物在心理上表现出来的活动。
- [状态] (state) 人或事物表现出来的形态。
- [关系] (relation) 事物之间相互作用相互影响的状态。
- [进化] (evolution) 事物由简单到复杂、由低级到高级的变化。
- [关涉] (concern) 一事物关联或牵涉到另一事物。
- [改动] (reformation) 使事物发生变化或者差别。
- [转移] (transfer) 使事物从一方改动到另一方。
- [属性] (attribute) 事物所具有的特性和关系。
- [外形] (appearance) 人或事物的外部形体属性。
- [表象] (surface) 从外表可以观察到的现象的属性。
- [颜色] (color) 由物体发射、反射或者透过的光波通过视觉所产生的印象。
- [味道] (taste) 能使舌头得到某种味觉的特性。
- [性质] (character) 一个事物区别于另一个事物的属性。
- [德才] (moral-ability) 人的道德和才能表现出来的属性。
- [境况] (circumstance) 外界环境所具有的属性。

这里只是列出了 ONTOL-MT 中主要的上层概念, 在这些概念的下层, 还有很多其他的概念, 限于篇幅, 这里就不列出了。可以看出, ONTOL-MT 中的初始概念与亚里士多德的范畴系统中的范畴很接近, 明显地受到了亚里士多德的范畴系统的影响。这个知识本体也反映了我们的世界观: 万事万物都是在时间和空间中运动和存在的, 它们都具有一定的属性和数量。所以, ONTOL-MT 虽然是为机器翻译的技术而设计的, 但是, 它继承了亚里士多德的范畴系统, 反映了我们的世界观, 具有鲜明的哲学色彩。

ONTOL-MT 知识本体系统中的概念, 实际上也就是单词本身所固有的语义特征, 它们是独立于单词的上下文而存在的, 因此, 可以用这些概念来表示机器翻译词典中单词的固有语义特征。在日汉机器翻译系统中, 我们利用单词固有的这些语义特征进行日语同形词的判别, 效果良好。

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.
- [5] Jurafsky D, Martin J. Speech and language processing [M].冯志伟,等译. 电子工业出版社, 2005.

作者简介:

冯志伟(1939-),男,云南昆明人,教育部语言文字应用研究所研究员,语言学和 Information Science 双硕士学位,博士生导师,研究方向:计算语言学、自然语言处理、机器翻译及应用语言学