# 自然语言处理中理性主义与经验主义

姓名：李勐　学号：071202　班级 07409

## 一、自己的观点

六七十年代自然语言处理是由经验主义主导，一个原因是由于机器计算和存储能力的限制。众所周知，统计学方法需要手机大量的语料，并在语料库上进行大量的计算比如聚类、分类，但当时计算机的存储空间有限，限制了这方面的研究。随着计算机计算能力和存储空间的提升，经验主义方法开始得到越来越多的应用，而理性主义方法则在准确虑上遇到瓶颈，因为理性主义方法基于大量先验规则和推导，要提高精确的就必须细化规则，使得时空复杂度呈指数级的上升，超过了机器的极限。而就在最近几年，由于分布式计算技术的成形以及各种云计算产品的推出，计算机的计算能力大规模提升，理性主义方法的研究又活跃起来。
总结一下就是，自然语言处理属于计算机语言学，采用的方法受到计算机本身能力的限制，会随着计算机能力的变化而变化。

## 二、自然语言处理中理性主义和经验主义方法的优缺点（摘自 www.52nlp.com）

总结自然语言处理发展的曲折历史可以看出，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，因此，我们应当用科学的态度来分析它们的优点和缺点。
我们认为，

### *1* 基于规则的理性主义方法的优点

1. 基于规则的理性主义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。
2. 基于规则的理性主义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。
3. 基于规则的理性主义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。
4. 基于规则的理性主义方法在本质上是没有方向性的，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。
5. 基于规则的理性主义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。
6. 基于规则的理性主义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理性主义方法在自然语言处理中得到有效的使用。

### *2* 基于规则的理性主义方法的缺点

1. 基于规则的理性主义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。

2. 使用基于规则的理性主义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。
3. 使用基于规则的理性主义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。
4. 基于规则的理性主义方法在实际的使用场合其表现往往不如基于统计的经验主义方法那样好。因为基于统计的经验主义方法可以根据实际训练数据的情况不断地优化，而基于规则的理性主义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，但是基于规则的理性主义方法很难模拟这种优先关系。

不过，尽管基于规则的理性主义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，基于规则的理性主义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，基于规则的理性主义方法是必不可少的。

## *3* 基于统计的经验主义方法的优点

1. 使用基于统计的经验主义方法来训练语言数据，从训练的语言数据中自动地或半自动地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，在句法自动分析和词义排歧中也初露锋芒。
2. 基于统计的经验主义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，基于统计的经验主义方法的效果就越好。在统计机器翻译中，语料库的规模，特别是用来训练语言模型的目标语言语料库的规模，对于系统性能的提高，起着举足轻重的作用。因此，可以通过扩大语料库规模的办法来不断提高自然语言处理系统的性能。
3. 基于统计的经验主义方法很容易与基于规则的理性主义方法结合起来，从而处理语言中形形色色的约束条件问题，使自然语言处理系统的效果不断地得到改善。
4. 基于统计的经验主义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。

## *4* 基于统计的经验主义方法的缺点

1. 使用基于统计的经验主义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还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。
2. 在当前语料库技术的条件下，要使用基于统计的经验主义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。基于统计的经验主义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在很大的程度上决定了基于统计的经验主义方法的效果。
3. 基于统计的经验主义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严重，这个问题需要使用各种平滑（smoothing）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。

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”。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，囫囵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”。

培根的主张是值得我们深思的。在自然语言处理的研究中，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。

# 三、哲学中的理性主义和经验主义

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（Benetict 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。

# 四、参考资料

[1] 自然语言处理中理性主义与经验主义的优缺点 from http://www.52nlp.cn/the-advantages-and-disadvantages-of-the-rationalism-and-empiricism-in-nlp
[2] 自然语言处理中的哲学问题，冯志伟，心智与计算 Vol.1,No3 (2007), 333-353