本文载《暨南大学华文学院学报》2006 年，第 1 期（总 21 期）

当前自然语言处理发展的四个特点

冯志伟
（教育部语言文字应用研究所）

摘要：本文分析了当前自然语言处理发展的四个特点：基于句法-语义规则的理性主义方法受到质疑，随着语料库建设和语料库语言学的崛起，大规模真实文本的处理成为自然语言处理的主要战略目标；自然语言处理中越来越多地使用机器自动学习的方法来获取语言知识；统计数学方法越来越受到重视；自然语言处理中越来越重视词汇的作用，出现了强烈的“词汇主义”的倾向。

关键词：自然语言处理，语料库，机器自动学习，统计数学，词汇主义。

二十一世纪以来，由于国际互联网的普及，自然语言的计算机处理成为了从互联网上获取知识的重要手段，生活在信息网络时代的现代人，几乎都要与互联网打交道，都要或多或少地使用自然语言处理的研究成果来帮助他们获取或挖掘在广阔无边的互联网上的各种知识和信息，因此，世界各国都非常重视自然语言处理的研究，投入了大量的人力、物力和财力。

我认为，当前国外自然语言处理研究有四个显著的特点：

第一，　基于句法-语义规则的理性主义方法受到质疑，随着语料库建设和语料库语言学的崛起，大规模真实文本的处理成为自然语言处理的主要战略目标。

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，这种方法的哲学基础是逻辑实证主义，他们认为，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J. A. Fodor在《Representations》[1]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。”Fodor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。J. R. Searle 在他的论文《Minds, Brains and Programmes》（1980，载《Behavioral and Brain Sciences》, Vol.3）[2]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。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。Searle 的批评使基于规则的理性主义的观点受到了普遍的怀疑。

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-语

义分析，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”。所谓“理性主义”，就是指以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。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。

这种大规模真实的语料库还为语言研究的现代化提供了强有力手段。我在 20 多年前曾经测试过汉字的熵（即汉字中所包含的信息量），这是中文信息处理的一项基础性研究工作。为了计算汉字的熵，首先需要统计汉字在文本中的出现频度，由于 70 年代我们还没有机器可读的汉语语料库，哪怕小规模的汉语语料库也没有，我只得根据书面文本进行手工查频，用了将近 10 年的时间，对数百万字的现代汉语文本（占 70%）和古代汉语文本（占 30%）进行手工查频，从小到大地逐步扩大统计的规模，建立了 6 个不同容量的汉字频度表，最后根据这些不同的汉字频度表，逐步地扩大汉字的容量，终于计算出了汉字的熵。这是一件极为艰辛而烦琐的工作。如今我们有了机器可读的汉语语料库，完全用不着进行手工查频，频度的统计可以在计算机上进行，只要非常简单的程序就可以轻而易举地从语料库中统计出汉字的频度并进一步计算出汉字的熵。语言研究工作的效率成百倍、成千倍地提高了！尽管学问是从苦根上长出来的甜果，但是，现代化的手段不仅可以帮助我们少吃很多的苦，而且也还能把学问做得更好。手工查频犹如赶着老牛破车在崎岖的山路上跋涉，使用语料库犹如乘宇宙飞船在广阔的太空中翱翔。这是我从前根本不敢想象的。大规模机器可读语料库的出现和使用，把语言学家从艰苦繁重的手工劳动中解放出来，使语言学家可以集中精力来研究那些更加重要的问题，这对于促进语言学研究的现代化具有不可估量的作用。

第二，　　自然语言处理中越来越多地使用机器自动学习的方法来获取语言知识。

传统语言学基本上是通过语言学家归纳总结语言现象的手工方法来获取语言知识的，由于人的记忆能力有限，任何语言学家，哪怕是语言学界的权威泰斗，都不可能记忆和处理浩如烟海的全部的语言数据，因此，使用传统的手工方法来获取语言知识，犹如以管窥豹，以蠡测海，这种获取语言知识的方法不仅效率极低，而且带有很大的主观性。传统语言学中啧啧地称道的所谓“例不过十不立，反例不过十不破”的朴学精神，貌似严格，实际上，在浩如烟海的语言数据中，以十个正例或十个反例就轻而易举地来决定语言规则的取舍，难道就能够万无一失地保证这些规则是可靠的吗？这是大大地值得怀疑的。当前的自然语言处理研究提倡建立语料库，使用机器学习的方法，让计算机

自动地从浩如烟海的语料库中获取准确的语言知识。机器词典和大规模语料库的建设，成为了当前自然语言处理的热点。这是语言学获取语言知识方式的巨大变化，作为二十一世纪的语言学工作者，都应该注意到这样的变化，逐渐改变获取语言知识的手段。

2000 年，在美国约翰.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、宾西法尼亚州立大学、斯丹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，以德国亚琛大学（Aachen university）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”），这个报告提出了把基于规则的方法和基于统计方法结合起来的有效途径。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。

2002 年 1 月，在美国成立了 Language Weaver 公司，专门研制统计机器翻译软件（Statistical Machine Translation Software，简称 SMTS），奥赫加盟 Language Weaver 公司，作为这个公司的顾问。 Language Weaver 公司是世界上第一个把统计机器翻译软件商品化的公司。他们使用机器自动学习的技术，从翻译存储资料（translation memories）、翻译文档（translated archives）、词典（dictionaries & glossaries）、因特网（Internet）以及翻译人员（human translators）那里获取大量的语言数据，在这个过程中，他们对这些语言数据进行各种预处理（pre-processing），包括文本格式过滤（format filtering）、光学自动阅读和扫描（Scan + OCR）、文字转写（transcription）、文本对齐（document alignment）、文本片段对齐（segment alignment）等。接着，把经过预处理的语言数据，在句子一级进行源语言和目标语言的对齐，形成双语并行语料库（parallel corpus）。然后使用该公司自己开发的“LW 学习软件”（Language Weaver Learner，简称 LW Learner），对双语并行语料库进行处理，从语料库中抽取概率翻译词典、概率翻译模板以及概率翻译规则等语言信息，这些抽取出来的语言信息，统称为翻译参数（translation parameters），这样的翻译参数实际上就是概率化的语言知识，经过上述的处理，语言数据就变成了概率化的语言知识。翻译参数是该公司翻译软件的重要组成部分。为了处理这些翻译参数，该公司还开发了一个统计翻译器，叫做解码器（Decoder），这个解码器是该公司翻译软件的另一个重要组成部分，解码器和翻译参数成为了 Language Weaver 公司翻译软件的核心（core components）。解码器使用上述通过统计学习获得的翻译参数对新的文本进行机器翻译，把新的源语言文本（new source language documents）自动地翻译成新的目标语言译文（new target language translation），提供给用户使用。

Language Weaver 公司的翻译系统的工作流程如下图所示：

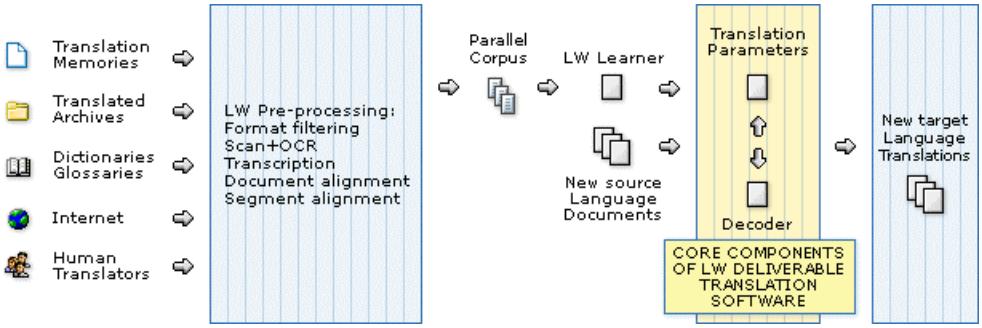

图 1　Language Weaver 统计机器翻译软件工作流程

目前，该公司开发的汉英机器翻译系统和英语—西班牙语双向机器翻译系统即将问世。他们还要使用同样的方法，开发英语—法语的双向机器翻译系统、印地语—英语以及索马里语—英语的单向机器翻译系统。

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究所 NIST/TIDES (National Institute of Standards and Technology) 主持的机器翻译评比中，奥赫获得了最好的成绩，他使用统计方法从双语语料库中自动地获取语言知识，建立统计机器翻译的规则，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，我就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."）[3]。这反映了新一代的自然语言处理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，使用机器自动学习的方法，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。过去我们使用人工编制语言规则的方法来研制一个机器翻译系统，往往需要几年的时间，而现在采用奥赫的机器学习方法，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，这是令我们感到振奋的。

第三，　统计数学方法越来越受到重视。

自然语言处理中越来越多地使用统计数学方法来分析语言数据，使用人工观察和内省的方法，显然不可能从浩如烟海的语料库中获取精确可靠的语言知识，必须使用统计数学的方法。

语言模型是描述自然语言内在规律的数学模型，构造语言模型是自然语言处理的核心。语言模型可以分为传统的规则型语言模型和基于统计的语言模型。规则型语言模型是人工编制的语言规则，这些语言规则来自语言学家掌握的语言学知识，具有一定的主观性和片面性，难以处理大规模的真实文本。基于统计的语言模型通常是概率模型，计算机借助于语言统计模型的概率参数，可以估计出自然语言中语言成分出现的可能性，而不是单纯地判断这样的语言成分是否符合语言学规则。

目前，自然语言处理中的语言统计模型已经相当成熟，例如，隐马尔可夫模型（Hidden Markov Model，简称 HMM）、概率上下文无关语法（Probabilistic Context-Free Grammar，简称 PCFG）、基于决策树的语言模型（Decision-Tree Based Model）、最大熵语言模型（Maximum Entropy Model）等[4]。研究这样的语言统计模型需要具备统计数学的知识，因此，我们应当努力进行知识更新，学习统计数学。如果我们认真地学会了统计数学，熟练地掌握了统计数学，就会使我们在获取语言知识的过程中如虎添翼。

第四，　**自然语言处理中越来越重视词汇的作用，出现了强烈的"词汇主义"的倾向**。

句法歧义问题的解决不仅与概率和结构有关，还往往与词汇的特性有关。这里讨论两个问题。

⑴　PP 附着问题：在英语句子中，介词短语 PP 可以做中心动词短语 VP 的状语，也可以做它前面名词短语 NP 的修饰语，究竟是附着于 VP，还是附着于 NP，这就是所谓"PP-

附着”（PP-attachment）问题。PP-附着与词汇有着密切的关系。

例如，在句子 “Washington sent more than 10,000 soldiers into Afghanistan”中, 介词短语（PP） “into Afghanistan”或者附着于名词短语（NP） “more than 10,000 soldiers”，或者附着于动词短语（VP）“sent”(单独的动词也可以看成一个动词短语)。这里存在 PP-附着问题。

在概率上下文无关语法中，这种 PP-附着的判定要在下面的规则之间进行选择：

NP → NP PP　　(PP 附着于 NP)

和　　VP → VP PP　　(PP 附着于 VP)

这两个规则的概率依赖于训练语料库。在训练语料库中，NP 附着和 VP 附着的统计结果如下：

语料库	PP 附着于 NP	PP 附着于 VP
AP Newswire (13 00 万词)	67%	33%
Wall Street Journal & IBM manuals	52%	48%

可以看出，在两个训练语料库中，“PP 附着于 NP”都处于优先地位。根据这样的统计结果，我们应该选择 PP 附着于 NP，也就是选择 PP “into Afghanistan”附着于 NP “more than 10,000 soldiers”这个结果。但是，在我们上面的句子中，介词短语“into Afghanistan”的正确附着却应该是附着于动词短语 VP(“sent”)，这是因为这个 VP“sent”往往要求一个表示方向的介词短语 PP，而介词短语“into Afghanistan”正好满足了这个要求。概率上下文无关语法显然不能处理这样的词汇依存问题。

⑵　并列结构的歧义：

句子“dogs in houses and cats”是有结构歧义的：

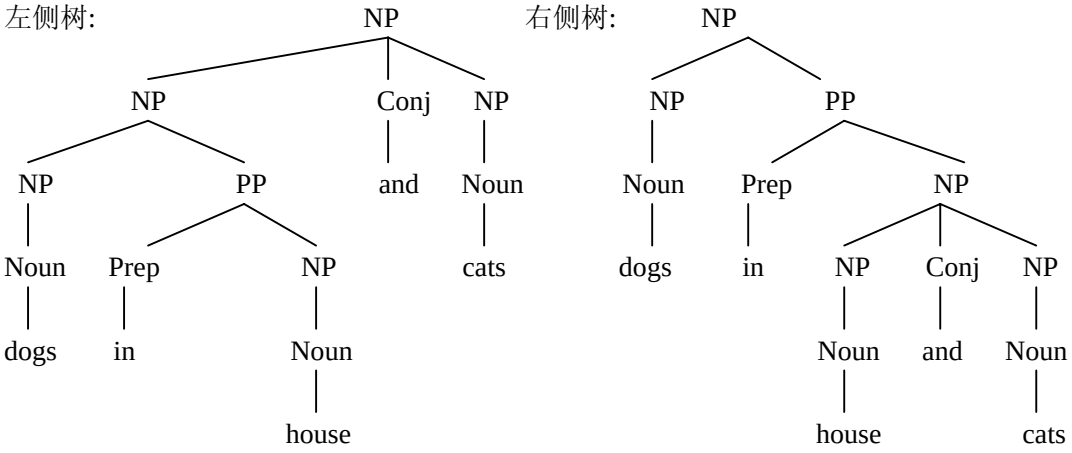

图 2　并列结构歧义

尽管在直觉上我们认为图 2 中左侧树是正确的，但是，左右两侧的树所使用的规则却是完全一样的。这些规则如下：

NP → NP Conj NP

NP → NP PP

NP → Noun

PP → Prep NP

Noun → dogs

Noun →house

Noun → cats

Prep → in

Conj → and

根据概率上下文无关语法的无关性假设，由于规则完全相同，使用这些规则的概率相乘而计算出来的两个树形图的概率也应该是一样的。在这种情况下，概率上下文无关语法将指派这两个树形图以相同的概率，也就是说，概率上下文无关语法无法判定这个句子的歧义。

由此可见，尽管我们使用数学，使用概率的方法，概率上下文无关语法在遇到词汇依存问题的时候就显得捉襟见肘、无能为力了，我们还需要探索其他的途径来进一步提升概率上下文无关语法的功能，其中的一个有效的途径，就是在概率上下文无关语法中引入词汇信息，采用词汇中心语概率表示法，把概率上下文无关语法提升为概率词汇化上下文无关语法。

在理论语言学中，N. Chomsky 最近提出了“最简方案”，所有重要的语法原则直接运用于表层，把具体的规则减少到最低限度，不同语言之间的差异由词汇来处理，也非常重视词汇的作用。在语言学中出现了“词汇主义”（lexicalism）的倾向。在自然语言处理中，词汇知识库的建造成为了普遍关注的问题。美国的 WordNet，FrameNet 以及我国各种语法知识库和语义知识库的建设，都反映了这种强烈的“词汇主义”的倾向。

在这样的新形势下，自然语言处理这个学科的交叉性和边缘性显得更加突出了，我们自然语言处理的研究者如果只是局限于自己原有的某一个专业的狭窄领域而不从其他相关的学科吸取营养来丰富自己的知识，在自然语言处理的研究中必将一筹莫展、处处碰壁。面对这样的形势我们应该怎么做？是抱残守缺，继续把自己蜷缩在某一个专业的狭窄领域之内孤芳自赏，还是与时俱进，迎头赶上，努力学习新的知识，以适应学科交叉性和边缘性的要求？这是我国自然语言处理工作者必须考虑的大问题。

我国自然语言处理虽然已经取得不少成绩，但是，与国际水平相比，差距还很大。自然语言处理是国际性的学科，我们不能闭门造车，而应该参与到国际自然语言处理的研究中去，用国际的水平和国际的学术规范来要求我们的研究。近年来，我国的自然语言处理工作者也到国外参加过一些第一流的自然语言处理国际会议，如 COLING，ACL，LREC 等，但是，在这些国际会议上，我国学者几乎从来也没有被邀请做代表当前最高研究水平并且引导计算语言学发展潮流的“主题报告”，我们只能做代表一般水平的发言，或者在分组会议上讲一讲我们的成绩和体会。这种情况说明，我国的自然语言处理研究，不论在理论上还是在应用系统的开发上，基本上还没有什么重大的创新，尽管我们的自我感觉良好，但实在还没有什么特别值得称道的突破，我们的研究，基本上还是跟踪性的研究，很少有创造性的研究，当然更没有具有原创思想的研究了。因此，我们不能夜郎自大，不能坐井观天，我们只有努力学习国外的先进成果，赶上并超过国际的先进水平，使我国的自然语言处理在国际的先进行列中占有一席之地，以无愧于我国这个国际大国的地位。

参考文献

[1] J. A. Fodor, Representations, MIT Press, 1980.

[2] J. R. Searle, Minds, Brains and Programmes, 载 *Behavioral and Brain Sciences*, Vol. 3, 1980.

[3] 冯志伟，机器翻译研究，中国对外翻译出版公司，2004 年。

[4] D. Jurafsky, J. Martin,（冯志伟 孙乐 译），自然语言处理综论，电子工业出版社，2005 年。

冯志伟，1939 年生，云南昆明人，教育部语言文字应用研究所研究员，中国传媒大学博士生导师，韩国科学技术院电子工程与计算机科学系教授，《中国语文》《语言科学》《语言文字应用》《中文信息学报》编委，主要研究方向为计算语言学和应用语言学，发表中外文专著 20 部，论文 200 余篇。

Four Features of Recent Development of Natural Language Processing

Feng Zhiwei

(Institute of Applied Linguistics, the Ministry of Education)

Abstract: Four features of recent development of natural language processing (NLP) are discussed in this paper: 1. As growing up of corpus linguistics, the syntactic-semantic rules-based rationalism falls in question, large-scale authentic text processing becomes main goal of NLP. 2. The automatic machine learning becomes main approach for acquiring language knowledge. 3. Statistical mathematical approach becomes more and more important. 4. Lexicalism becomes main tendency in NLP.

Keywords: Natural Language Processing (NLP), corpus, automatic machine leaning, statistical mathematics, lexicalism.