“十三五”国家重点图书出版规划项目
当代科学技术基础理论与前沿问题研究丛书


# 自然语言
## 计算机形式分析的理论与方法

Theory and Method for Formal Analysis
of Natural Language by Computer

冯志伟 著

中国科学技术大学出版社

国家出版基金项目
NATIONAL PUBLICATION FOUNDATION

“十三五”国家重点图书出版规划项
当代科学技术基础理论与前沿问题研究丛书

# 自然语言
## 计算机形式分析的理论与方法

Theory and Method for Formal Analysis
of Natural Language by Computer

中国科学技术大学出版社

冯志伟 著

## 内 容 简 介

自然语言计算机形式分析是横跨语言学、计算机科学和数学的一个交叉研究领域，是自然语言计算机处理的关键。自然语言是信息最主要的负荷者，在当今信息网络时代，计算机已经日益普及，普通计算机用户可以使用的语言资源正以惊人的速度飞快增长。互联网主要是由自然语言构成的，它已经成为了极为丰富的语言信息资源；移动通信也是以自然语言为媒介的，它已经渗透到日常生活的各个领域。因此，自然语言计算机形式分析对于国家的信息化建设，对于互联网和移动通信的安全具有重要作用。

本书对自然语言处理中的各种理论和方法进行了系统的总结和梳理。首先讨论了自然语言处理的学科定位；接着介绍了语言计算的一些先驱研究；然后以主要的篇幅讨论自然语言处理中的各种形式模型，包括基于短语结构语法的形式模型、基于合一运算的形式模型、基于依存和配价的形式模型、基于格语法的形式模型、基于词汇主义的形式模型、语义自动处理的形式模型、系统功能语法、语用自动处理的形式模型、概率语法、Bayes 公式与动态规划算法、$N$ 元语法和数据平滑、隐 Markov 模型(HMM)、语音自动处理的形式模型、统计机器翻译的形式模型；同时还讨论了自然语言处理系统的评测问题；最后从哲学的角度讨论了自然语言处理中的理性主义和经验主义，探索理性主义方法和经验主义方法相结合的途径。

本书说理透彻、语言流畅、实例丰富、深入浅出，适合从事自然语言处理研究的科研人员、大学师生阅读，也可以作为人工智能、计算语言学等课程的教学参考书。

**图书在版编目(CIP)数据**

自然语言计算机形式分析的理论与方法/冯志伟著. —合肥：中国科学技术大学出版社，2017.1

(当代科学技术基础理论与前沿问题研究丛书)

“十三五”国家重点图书出版规划项目

国家出版基金项目

ISBN 978-7-312-04130-3

Ⅰ.自… Ⅱ.冯… Ⅲ.自然语言处理—研究 Ⅳ.TP391

中国版本图书馆 CIP 数据核字(2016)第 325329 号

**出版** 中国科学技术大学出版社
安徽省合肥市金寨路 96 号，230026
http://press.ustc.edu.cn

**印刷** 安徽联众印刷有限公司

**发行** 中国科学技术大学出版社

**经销** 全国新华书店

**开本** 787 mm×1092 mm 1/16

**印张** 53.5

**字数** 1135 千

**版次** 2017 年 1 月第 1 版

**印次** 2017 年 1 月第 1 次印刷

**印数** 1—1000 册

**定价** 198.00 元

# 序

采用计算机技术来研究和处理自然语言是20世纪40年代末期才开始的，六十多年来，这项研究取得了长足的进展，形成了当代计算机科学中一门重要的新兴学科——自然语言处理(Natural Language Processing，NLP)。在信息网络时代，自然语言处理引起了包括计算机专家和语言学家在内的越来越多的学者的重视，成为了一门文科和理科紧密结合的典型的交叉学科。

由于现实的自然语言极为复杂，不可能直接作为计算机的处理对象，为了使现实的自然语言成为可以由计算机直接处理的对象，在自然语言处理的各个应用领域中，我们都需要根据处理的要求，把自然语言处理抽象为一个“问题”(problem)，再把这个问题在语言学上加以“形式化”(formalism)，建立语言的“形式模型”(formal model)，使之能以一定的数学形式，严密而规整地表示出来，并且把这种严密而规整的数学形式表示为“算法”(algorithm)，建立自然语言处理的“计算模型”(computational model)，使之能够在计算机上实现。在自然语言处理中，算法取决于形式模型。形式模型是自然语言计算机处理的本质，而算法只不过是实现形式模型的手段而已。这种建立语言形式模型的研究是非常重要的，它应当属于自然语言处理的基础理论和方法研究。

本书对自然语言处理中的各种理论和方法进行了系统的总结和梳理。首先讨论了自然语言处理的学科定位；接着介绍了语言计算的一些先驱研究；然后以主要的篇幅讨论自然语言处理中的各种形式模型，包括基于短语结构语法的形式模型、基于合一运算的形式模型、基于依存和配价的形式模型、基于格语法的形式模型、基于词汇主义的形式模型、语义自动处理的形式模型、系统功能语法、语用自动处理的形式模型、概率语法、Bayes公式与动态规划算法、$N$元语法和数据平滑、隐Markov模型(HMM)、语音自动处理的形式模型、统计机器翻译的形式模型；同时还讨论了自然语言处理系统的评测问题；最后从哲学的角度讨论了自然语言处理中的理性主义和经验主义，探索理性主义方法和经验主义方法相结合的途径。

早在20世纪50年代我在北京大学求学时，就对自然语言的数学模型研究产生了兴趣，毅然从理科转到文科，师从王力、岑麒祥、朱德熙等著名语言学家学习语言学，探讨语言研究中的数学方法；后来考上了理论语言学的研究生，试图从理论上探

讨自然语言处理的形式模型。可惜不久就发生了“文化大革命”，我被分配到边疆当了一名中学物理教员。

1978年高考制度恢复，我考入了中国科学技术大学研究生院。入学之后，学校公派我到法国格勒诺布尔理科医科大学应用数学研究所（IMAG）自动翻译中心（CETA）留学，师从法国著名数学家、国际计算语言学委员会主席 B. Vauquois（沃古瓦）教授，系统地学习计算机科学知识和数学知识，并专门研究文理交叉的自然语言处理问题，把语言学、计算机科学和数学紧密地结合起来。

中国科学技术大学使我有机会重新回到自然语言处理的队伍，使我有可能为我毕生钟爱的这个学科尽自己的绵薄之力。现在，中国科学学技术大学出版社又决定出版我的专著，并为此申请了国家出版基金，使我有机会系统地总结自己研究自然语言形式分析技术的经验和教训。我永远也忘不了中国科学技术大学对我的恩情。

我从事自然语言处理已经五十多年了。五十多年前，我还是一个不谙世事的小青年，现在，我已经是年过花甲、白发苍苍的古稀老人了。我们这一代人正在一天天地变老；然而，我们如痴如醉地钟爱着的自然语言处理事业却是一门新兴的学科，它还非常年轻，充满了青春的活力，尽管它还比较稚嫩，还不够成熟，但是它无疑地有着光辉的发展前景。我们个人的生命是有限的，而科学知识的探讨和研究却是无限的。我们个人渺小的生命与科学事业这棵长青的参天大树相比较，显得多么微不足道，如沧海之一粟。想到这些，怎不令我们感慨万千！“书山有路勤为径，学海无涯苦作舟。”我们应当勤苦地工作，把个人的有限的生命投入到无限的科学知识的探讨和研究中去，从而实现人生的价值。

在本书的写作过程中，我参考了国内外时贤著作多种，没有他们丰厚的研究成果，本书是不可能写出来的。在此，我对他们表示诚挚的谢意，就不一一列名道谢了。

本书涉及语言学、计算机科学、数学等多个领域的知识。我自己水平有限，错误在所难免，敬请广大读者提出宝贵的意见。

冯志伟

2016年1月于杭州下沙钱塘江畔

# 目 次

# 第 1 章
# 自然语言处理的学科定位

采用计算机技术来研究和处理自然语言是 20 世纪 40 年代末才开始的，六十多年来，这项研究取得了长足的进展，成为了当代计算机科学中一门重要的新兴学科——自然语言处理(Natural Language Processing，NLP)。在信息网络时代，自然语言处理引起了越来越多的学者的重视，成为一门“显学”，人们提出了各种不同的理论和方法。

在工业革命时代，人类需要探索物质世界的奥秘。由于物质世界是由原子和各种基本粒子构成的，因此，研究原子和各种基本粒子的物理学成了非常重要的学科；在信息网络时代，由于信息网络主要是由语言构成的，因此，我们可以预见，在不久的将来，研究语言结构的自然语言处理必定也会成为像物理学一样非常重要的学科。物理学研究物质世界中各种物理运动的规律，而自然语言处理则研究信息网络世界中语言载体的规律。自然语言处理的重要性完全可以与物理学媲美，它们将成为未来科学世界中举足轻重的双璧。这是我们在直觉上的一种估计，我们坚信这样的估计将会成为活生生的现实。

在这样的情况下，如何对自然语言处理进行正确的学科定位，使我们认识到自然语言处理在整个学科体系中的位置，从而自觉地推动自然语言处理的发展，是一个至关重要的问题。

我们可以从自然语言处理的过程、自然语言处理的范围以及自然语言处理的历史三个角度来考察自然语言处理的学科定位问题。从自然语言处理的过程来考察它的学科定位，是从纵的角度来讨论；从自然语言处理的范围来考察它的学科定位，是从横的角度来讨论。纵横交错，我们对于自然语言处理的学科定位就可以在共时的平面上得到比较清晰的认识。最后，我们再从自然语言处理的历史来考察，也就是从发展的角度来讨论。这样，我们对于自然语言处理的学科定位就可以在历时的平面上得到比较清晰的认识。

## 1.1 从自然语言处理的过程来考察其学科定位

首先，我们从自然语言处理的过程，也就是从纵的角度来讨论这个问题。

我们认为，计算机对自然语言的研究和处理，一般应经过如下四个方面的过程：

- 把需要研究的问题在语言学上加以形式化，建立语言的形式化模型，使之能以一定的

数学形式，严密而规整地表示出来。这个过程可以叫作“形式化”。

● 把这种严密而规整的数学形式表示为算法。这个过程可以叫作“算法化”。

● 根据算法编写计算机程序，使之在计算机上实现，建立各种实用的自然语言处理系统。这个过程可以叫作“程序化”。

● 对于所建立的自然语言处理系统进行评测，不断地改进其质量和性能，以满足用户的要求。这个过程可以叫作“实用化”。

美国计算机科学家 Bill Manaris(马纳利斯)在 1999 年出版的《计算机进展》(*Advances in Computers*)第 47 卷的《从人-机交互的角度看自然语言处理》一文中曾经对自然语言处理提出了如下的定义：

“自然语言处理可以定义为研究在人与人交互中以及在人与计算机交互中的语言问题的一门学科。自然语言处理要研制表示语言能力(linguistic competence)和语言应用(linguistic performance)的模型，建立计算框架来实现这样的语言模型，提出相应的方法来不断地完善这样的语言模型，根据这样的语言模型设计各种实用系统，并探讨这些实用系统的评测技术。”这个定义的英文如下：“NLP could be defined as the discipline that studies the linguistic aspects of human-human and human-machine communication, develops models of linguistic competence and performance, employs computational frameworks to implement process incorporating such models, identifies methodologies for iterative refinement of such processes/models, and investigates techniques for evaluating the result systems.”

Manaris 关于自然语言处理的这个定义，比较全面地表达了计算机对自然语言的研究和处理的上述四个方面的过程。我们认同这样的定义。

在 2001 年的美国电影《太空漫游》(*A Space Odyssey*, Stanley Kubrick 和 Arthur C. Charke 编)中机器人 HAL 和 Dave 进行了如下对话：

Dave Bownman: Open the pod bay doors, HAL.

HAL: I'm sorry Dave, I'am afraid I can't do that.

(Dave Bownman: HAL，请你打开太空舱的分离舱门。

HAL: 对不起，Dave，我不能这样做。)

HAL 实际上是一台名为“9000”的电子计算机，这台计算机具有 20 世纪最受人们认可的一些特征。HAL 实际上是一个具有高级的语言处理能力并且能够说英语和理解英语的智能机器人(artificial agent)，在影片情节的关键时刻，HAL 甚至能够进行唇读(reading lip)。上面就是电影中的角色 Dave 先生请求智能机器人 HAL 打开宇宙飞船的分离舱门时与 HAL 之间的一段对话。作者 Arthur C. Charke 曾经乐观地预言，到一定的时候，我们就可以制造出像 HAL 这样的智能机器人。但是，现在我们离这样的预言还有多远呢？为了让 HAL 具有与语言相关的能力，我们究竟还应该做些什么呢？

我们认为，像 HAL 这样的机器人至少应该能通过语言与人类进行交流。其中包括通过语音识别(speech recognition)和自然语言理解(natural language understanding，当然包括唇读)来与人类沟通，通过自然语言生成(natural language generation)和语音合成(speech synthesis)来与人类交互。HAL 也应该能够做信息检索(information retrieval，发现它所需要的文本资源在哪里)和信息抽取(information extraction，从文本资源中抽取它所需要的信息)，并且进行知识推理(reference，根据已知的事实推出结论)。

尽管这些问题现在还远远没有完全解决，但 HAL 需要的一些与语言相关的技术现在已经研发出来了，并且有一部分技术已经商品化。解决这样的问题以及其他类似的问题，是自然语言处理、计算语言学、语音识别与语音合成的主要研究内容。我们把它们统称为语音与语言的计算机处理，或者简单地称为自然语言处理，因此，自然语言处理也同时包括了语音处理的内容。

像 HAL 这样有复杂的语言能力的智能机器人将要求具有非常广泛和深刻的语言知识。我们只要读一读前面 HAL 和 Dave 之间进行的对话，就可以了解到这样复杂的应用所需要的语言知识的范围和种类。

为了确定 Dave 讲什么，HAL 必须能够分析它所接收的声音信号，并且把 Dave 的这些信号复原成词的系列。与此相似，为了生成回答，HAL 必须把它的回答组织成词的系列，并且生成 Dave 能够识别的声音信号。要完成这两方面的任务，需要语音学(phonetics)和音系学(phonology)的知识，这样的知识可以帮助我们建立词如何在话语中发音的模型。

值得注意的是，HAL 还能够说出如 I'm 和 can't 这样的缩约形式。HAL 必须把它们分别还原为 I am 和 can not，才能在它的词库中找到这些单词的对应物，从而明白这些缩约形式究竟代表什么样的语言成分。HAL 还要能够产生并且识别单词的这样或那样的变体(例如，识别 doors 是复数)。这些都要求 HAL 具有形态学方面的知识，这些知识能够反映关于上下文中词的形态和行为的有关信息。

除了处理一个一个的单词之外，HAL 还应该知道怎样分析 Dave 所提出的请求的结构。这样的分析能够使 HAL 确定，Dave 说的话是关于要 HAL 采取某种行动的一个请求，这样的请求不同于下面关于陈述客观世界的简单命题，也不同于下面关于 door 的问话，它们是 Dave 请求的不同变体：

HAL, the pod bay door is open. (HAL，分离舱的门是开着的。)

HAL, is the pod bay door open? (HAL，分离舱的门是开着的吗?)

此外，HAL 还必须使用类似的结构知识把一个个的单词组织成为符号串，构成它的回答。例如，HAL 必须知道，下面的单词序列对于 Dave 是没有意义的，尽管这个单词系列所包含的单词与它原来的回答中所包含的单词完全一样：

I'm I do, sorry that afraid Dave I'm can't.

这里所说的关于组词成句的知识，叫作句法(syntax)。

显而易见，如果只是知道 Dave 所说的话语的各个单词以及句法结构，并不能使 HAL 了解 Dave 提出的请求的实质。为了理解 Dave 的请求事实上是关于要求打开 pod bay door(分离舱门)的一个命令，而不是讲关于当天中饭的菜单的事情，就要有复合词的语义的知识、词汇语义学(lexical semantics)的知识以及如何把这样的复合词组成更大意义的知识，即关于组合语义学(compositional semantics)的知识。pod bay door 按照字面逐词翻译是"豆荚-海湾-门"，但是它们组合成的意思却是"分离舱门"。这是关于科学技术术语(terminology)的知识。

另外，尽管智能机器人 HAL 的行为还不十分熟练，但它也应该充分地懂得如何对 Dave 表示礼貌。例如，它不要简单地回答 No 或者 No，I won't open the door。HAL 首先用表示客气的话(I'm sorry 和 I'm afraid)回答，然后委婉地说 I can't，而不是直截了当地说 I won't。这种礼貌和委婉语言的用法属于语用学(pragmatics)的研究领域。

最后，HAL 不是简单地无视 Dave 的请求，让门继续关着，而是对于 Dave 开始的请求，选择结构会话的方式来对待。HAL 在它给 Dave 的回答中，正确地使用单词 that 来简单地表示会话中话段之间的共同部分。正确地把这样的会话组织成结构，需要话语规约(discourse convention)的知识。

因此，我们认为，建立自然语言处理模型需要如下九个不同平面的知识：

- 声学和韵律学的知识：描述语言的节奏、语调和声调的规律，说明语音怎样形成音位。
- 音位学的知识：描述音位的结合规律，说明音位怎样形成语素。
- 形态学的知识：描述语素的结合规律，说明语素怎样形成单词。
- 词汇学的知识：描述词汇系统的规律，说明单词本身固有的语义特性和语法特性。
- 句法学的知识：描述单词(或词组)之间的结构规则，说明单词(或词组)怎样形成句子。
- 语义学的知识：描述句子中各个成分之间的语义关系，这样的语义关系是与情景无关的，说明怎样从构成句子的各个成分中推导出整个句子的语义。
- 话语分析的知识：描述句子与句子之间的结构规律，说明怎样由句子形成话语或对话。
- 语用学的知识：描述与情景有关的情景语义，说明怎样推导出句子具有的与周围话语有关的各种含义。
- 外界世界的常识性知识：描述关于语言使用者和语言使用环境的一般性常识，例如语言使用者的信念和目的，说明怎样推导出这样的信念和目的内在的结构。

当然，关于自然语言处理所涉及的知识平面还有不同的看法，不过，一般而言，大多数的自然语言处理研究人员认为，这些语言学知识至少可以分为词汇学知识、句法学知识、语义学知识和语用学知识等平面。每一个平面传达信息的方式各不相同。例如，词汇学平面可能涉及具体的单词的构成成分(例如语素)以及它们的屈折变化形式的知识；句法学平面可能涉及在具体的语言中单词或词组怎样结合成句子的知识；语义学平面可能涉及怎样给具体的单词

或句子指派意义的知识；语用学平面可能涉及在对话中话语焦点的转移以及在给定的上下文中怎样解释句子含义的知识。

下面我们具体说明在自然语言处理中这些知识平面的一般情况。如果我们对计算机发一个口头的指令“Delete file x”（删除文件 x），我们要通过自然语言处理系统让计算机理解这个指令的含义，并且执行这个指令，一般来说需要经过处理，过程如图 1.1 所示。

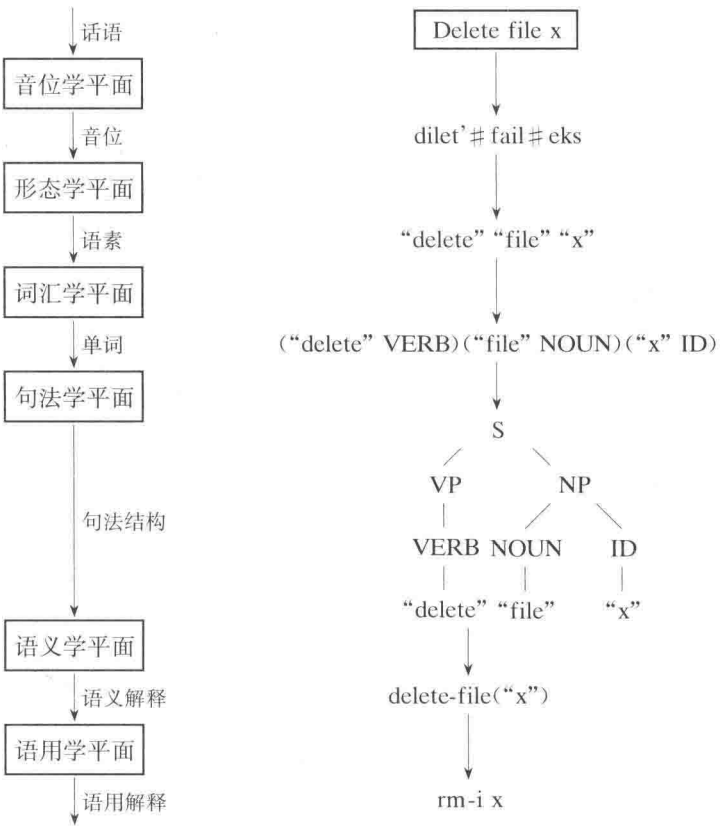


图 1.1　自然语言处理系统中的知识平面

从图 1.1 中可以看出，自然语言处理系统首先把指令“Delete file x”在音位学平面转化成音位系列“dilet'＃fail＃eks”；然后在形态学平面把这个音位系列转化为语素系列“delete” “file” “x”；接着在词汇学平面把这个语素系列转化为单词系列并标注相应的词性：(“delete” VERB) (“file”NOUN) (“x”ID)；在句法学平面进行句法分析，得到这个单词系列的句法结构，用树形图表示；在语义学平面得到这个句法结构的语义解释：delete-file (“x”)；在语用学平面得到这个指令的语用解释“rm-i x”，最后让计算机执行这个指令。

这个例子来自美国自然语言处理学者 Wilensky（威林斯基）为 UNIX 设计的一个语音理解界面，叫作 UNIX Consultant。这个语音理解界面使用了上述的第 1 至第 6 个平面的知识，得到口头指令“Delete file x”的语义解释：delete-file (“x”)；然后使用第 8 个平面的语用学知

识把这个语义解释转化为计算机的指令语言“rm-i x”，让计算机执行这个指令，这样便可以使用口头指令来指挥计算机的运行了。

不同的自然语言处理系统需要的知识平面可能与 UNIX Consultant 不一样，根据实际应用的不同要求，很多自然语言处理系统只需要使用上述九个平面中的部分平面的知识就行了。例如，书面语言的机器翻译系统只需要第 3 至第 7 个平面的知识，个别的机器翻译系统还需要第 8 个平面的知识，语音识别系统只需要第 1 至第 5 个平面的知识。

上述九个平面的知识主要涉及的是语言学知识，所以我们认为自然语言处理原则上是一个语言学问题。除了语言学之外，自然语言处理还涉及如下的知识领域：

- 计算机科学：给自然语言处理提供模型表示、算法设计和计算机实现的技术。
- 数学：给自然语言处理提供形式化的数学模型和形式化的数学方法。
- 心理学：给自然语言处理提供人类言语行为的心理模型和理论。
- 哲学：给自然语言处理提供关于人类的思维和语言的更深层次的理论。
- 统计学：给自然语言处理提供基于样本数据来预测统计事件的技术。
- 电子工程：给自然语言处理提供信息论的理论基础和语言信号处理技术。
- 生物学：给自然语言处理提供大脑中人类语言行为机制的理论。

因此，自然语言处理是一个多边缘的交叉学科。自然语言处理的研究，应该把这些学科的知识结合起来。每一个从事自然语言处理研究的人，都应该尽量使自己成为文理兼通、博学多识的人。

## 1.2 从自然语言处理的范围来考察其学科定位

上面，我们从自然语言处理的过程，也就是从纵的角度，考察了自然语言处理的学科定位。下面，我们换一个角度，从自然语言处理的范围，也就是从横的角度来考察自然语言处理的学科定位。

自然语言处理的范围涉及众多的领域，如语音的自动识别与合成、机器翻译、自然语言理解、人机对话、信息检索、文本分类、自动文摘等等。我们认为，这些领域可以归纳为如下四个大的方向：

- 语言学方向：把自然语言处理作为语言学的分支来研究，它只研究语言及语言处理与计算相关的方面，而不管其在计算机上的具体实现。这个研究方向的最重要的研究领域是语法形式化理论和自然语言处理的数学理论。
- 数据处理方向：把自然语言处理作为开发语言研究相关程序以及语言数据处理的学科