



“十一五国家重点图书”

《当代科学技术基础理论与前沿问题研究丛书》

《中国科学技术大学校友文库》

**ISBN: 978-7-312-02253-1**

中国科学技术大学出版社出版，2010年

《自然语言处理的形式模型》

冯志伟 著

## 内容简介

《自然语言处理的形式模型》对自然语言处理中的各种形式模型进行了系统的梳理，分别讨论了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、语用自动处理的形式模型、隐马尔可夫模型、统计机器翻译的形式模型。

《自然语言处理的形式模型》说理透彻、语言流畅、实例丰富、深入浅出，适合于从事自然语言处理教学和研究的科研人员、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。

## 序言

大学最重要的功能是向社会输送人才，大学对于一个国家、民族乃至世界的重要性和贡献度，很大程度上是通过毕业生在社会各领域所取得的成就来体现的。

中国科学技术大学建校只有短短的五十年，之所以迅速成为享有较高国际声誉的著名大学之一，主要就是因为她培养出了一大批德才兼备的优秀毕业生。他们志向高远、基础扎实、综合素质高、创新能力强，在国内外科技、经济、教育等领域做出了杰出的贡献，为中国科大赢得了“科技英才的摇篮”的美誉。

2008年9月，胡锦涛总书记为中国科大建校五十周年发来贺信，信中称赞说：半个世纪以来，中国科学技术大学依托中国科学院，按照全院办校、所系结合的方针，弘扬红专并进、理实交融的校风，努力推进教学和科研工作的改革创新，为党和国家培养了一大批科技人才，取得了一系列具有世界先进水平的原创性科技成果，为推动我国科教事业发展和社会主义现代化建设做出了重要贡献。

据统计，中国科大迄今已毕业的5万人中，已有42人当选中国科学院和中国工程院院士，是同期（自1963年以来）毕业生中当选院士数最多的高校之一。其中，本科毕业生中平均每1000人就产生1名院士和700多名硕士、博士，比例位居全国高校之首。还有众多的中青年才俊成为我国科技、企业、教育等领域的领军人物和骨干。在历年评选的“中国青年五四奖章”获得者中，作为科技界、科技创新型企业界青年才俊代表，科大毕业生已连续多年榜上有名，获奖总人数位居全国高校前列。鲜为人知的是，有数千名优秀毕业生踏上国防战线，为科技强军做出了重要贡献，涌现出20多名科技将军和一大批国防科技中坚。

.....

## 作者前言

采用计算机技术来研究和处理自然语言是 20 世纪 40 年代末期和 50 年代才开始的,50 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,简称 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,成为了文科和理科紧密结合的一门典型的交叉学科。

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、隐马尔可夫模型(HMM)、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 年代在北京大学求学时,我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,后来当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,我被迫改行,我们这些北京大学的研究生成为了“旧教育制度的牺牲品,新教育制度的实验品,社会上的处理品”(当时叫做“三品学生”),我也不得不离开了北京大学,到边疆当一名中学物理教员。

1978 年文化大革命结束,高考制度恢复,我本想回到北京大学继续从事在我过去当研究生时未完成的研究工作,但根据毛泽东生前的指示,文科暂时不招生,我没法回到北京大学去研究语言学,只好报考理工科,考入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沃古瓦(B.Vauquois)教授,系统地学习计算机科学和数学的知识,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,使我有可能是我毕生钟爱的这个学科尽自己绵薄之力,我永远也忘不了中国科学技术大学对我的恩情。在中国科学技术大学建立 50 周年的时候,我以此书作为对于母校 50 周年的献礼。

在母校成立 50 周年的时候,我自己从事自然语言处理也恰巧有 50 年的日子了。50 年前,我还是一个不谙世事的 19 岁的小青年,现在,我已经是年过 70 岁的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，她还非常年青，充满了青春的活力，尽管她还比较幼稚娇嫩，还不够成熟，但是她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常青的参天大树相比较，显得多么地微不足道，有如沧海之一粟。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，我曾参考过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，我对他们表示由衷的感谢，就不一一列名道谢了。

本书涉及到语言学、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

2008年

## 目录

### 第一章 自然语言处理的学科定位

第一节 从自然语言处理的过程来考察其学科定位

第二节 从自然语言处理的范围来考察其学科定位

第三节 从自然语言处理的历史来考察其学科定位

第四节 当前自然语言处理发展的几个特点

### 第二章 语言计算研究的先驱

第一节 Markov 链

第二节 Zipf 定律

第三节 Shannon 关于“熵”的研究

第四节 Bar-Hillel 的范畴语法

第五节 Harris 的语言串分析法

第六节 О. С. Кулагина 的语言集合论模型

### 第三章 基于短语结构语法的形式模型

第一节 语法的 Chomsky 层级

第二节 有限状态语法和它的局限性

第三节 短语结构语法

第四节 递归转移网络和扩充转移网络

第五节 自底向上分析与自顶向下分析

第六节 通用句法生成器和线图分析法

第七节 Earley 算法

第八节 左角分析法

第九节 CYK 算法

第十节 Tomita 算法

第十一节 管辖-约束理论与最简方案

第十二节 Joshi 的树邻接语法

第十三节 汉字结构的形式描述

#### 第四章 基于合一运算的形式模型

第一节 中文信息 MMT 模型

第二节 Kaplan 的词汇功能语法

第三节 Kay 的功能合一语法

第四节 Gazdar 的广义短语结构语法

第五节 Shieber 的 PATR

第六节 Pollard 的中心语驱动的短语结构语法

第七节 Pereira 的定子句语法

#### 第五章 基于依存和配价的形式模型

第一节 依存和配价观念的起源

第二节 Tesnière 的依存语法

第三节 依存语法在自然语言处理中的应用

第四节 配价语法

第五节 配价语法在自然语言处理中的应用

#### 第六章 基于格语法的形式模型

第一节 Fillmore 的格语法

第二节 Fillmore 的框架网络

#### 第七章 基于词汇主义的形式模型

第一节 Gross 的词汇语法

第二节 Sleator 和 Temperley 的链语法

第三节 词汇语义学

第四节 知识本体

第五节 词网 WordNet

第六节 知网 HowNet

#### 第八章 语义自动处理的形式模型

第一节 义素分析法

第二节 语义场

第三节 语义网络

第四节 Montague 的蒙塔鸠语法

第五节 Wilks 的优选语义学

第六节 Schank 的概念依存理论

	第七节	Mel'chuk 的意义-文本理论
	第八节	词义排歧方法
第九章	系统功能语法	
	第一节	系统功能语法的基本概念
	第二节	系统功能语法对于自然语言处理的作用
第十章	语用自动处理的形式模型	
	第一节	Mann 和 Thompson 的修辞结构理论
	第二节	文本连贯中的常识推理技术
第十一章	概率语法	
	第一节	概率上下文无关语法与句子歧义
	第二节	概率上下文无关语法的基本原理
	第三节	概率上下文无关语法的三个假设
	第四节	词汇化的概率上下文无关语法
第十二章	Bayes 公式与动态规划算法	
	第一节	拼写错误的检查与更正
	第二节	Bayes 公式与噪声信道模型
	第三节	最小编辑距离算法
	第四节	发音问题研究中的 Bayes 方法
	第五节	发音变异的决策树模型
	第六节	加权自动机
	第七节	向前算法
	第八节	Viterbi 算法
	本章附录	
第十三章	N-元语法和数据平滑	
	第一节	N-元语法
	第二节	数据平滑
第十四章	隐马尔可夫模型 (HMM)	
	第一节	HMM 模型概述
	第二节	HMM 模型在语音识别中的应用
第十五章	统计机器翻译的形式模型	
	第一节	机器翻译与噪声信道模型
	第二节	基于平行概率语法的形式模型

- 第三节 最大熵模型
- 第四节 结合短语知识的统计机器翻译模型
- 第五节 结合句法知识的统计机器翻译

## 第十六章 自然语言处理系统的评测

- 第一节 评测的原则和方法
- 第二节 文语转换和语音合成系统的评测
- 第三节 机器翻译的评测
- 第四节 语料库系统的评测

## 第十七章 自然语言处理中的理性主义与经验主义

- 第一节 哲学中的理性主义与经验主义
- 第二节 自然语言处理中理性主义与经验主义的消长
- 第三节 理性主义和经验主义的利弊得失
- 第四节 探索理性主义方法与经验主义方法结合的途径