王宗炎 主编

现代语言学丛书

自然语言的计算机处理

冯志伟 著

上海外语教育出版社

现代语言学丛书

自然语言的计算机处理

冯志伟 著

上海外语教育出版社

自然语言的计算机处理

冯志伟 著

上海外语教育出版社出版发行
（上海外国语大学内）
上海信老印刷厂印刷
新华书店上海发行所经销

开本 850×1168 1/32 16.75 印张 374 千字
1996 年 10 月第 1 版 1996 年 10 月第 1 次印刷
印数：1－2 000

ISBN 7－81046－036－6

T·006 定价：20.00 元

总　序

为什么出版《现代语言学丛书》？

因为我们感到，中国现代化包括许多方面的工作，其中之一是语言学研究的现代化。我们希望这一套丛书的出版，会有助于这一工作的开展。

近几十年来，国外语言学的研究进展很快。一方面，关于语言的内部结构，出现了各种理论和模式；另一方面，从各种不同的学科去研究语言，产生了诸如人类语言学、社会语言学、心理语言学、神经语言学、计算语言学等多科性研究。了解和介绍这两方面的理论、模式、实验和数据，供我国语言研究者参考，从而为语言学研究的现代化出一点力，这是我们的希望。

要做到语言学研究的现代化是不容易的。首先要对国外新的语言学理论加以分析和比较，作出我们自己的判断；更重要的是要结合汉语的研究加以验证，写出结合中国实际的论著。我们这里先做第一步工作。

中国语言学史上，不乏利用外国的语言理论，为汉语研究开辟新路的例子。郑樵说："初韵之学，起自西域。"马建忠以拉丁文法为范式，写出了《马氏文通》。赵元任、罗常培等前辈先生运用描写语言学的方法，为我国方言调查做出了典范。近时汉语语法学家利用国外语言学的研究方法，使语法现象的分类和范畴的描写更有理据，更为精确。先行者研究外国语言理

论的态度，永远是值得我们学习的。

作为第一步，我们打算出版15至20种书。以普及为主，逐步提高，以引进为主，同时注意结合我国的实际。我们希望和国内语言学界同志共同努力，填补我国语言学科中的一些空白点。

我们心目中的读者，是高等学校中文、外文和其他文史专业的师生，翻译界、新闻出版界人士，中学语文教师，以及一般语文工作者和爱好者。我们将力求用明白易懂的语言介绍新的学说和理论。

我们将注意国外新出的语言学文献，为中国的语言学的现代化尽快提供信息。我们的力量还很薄弱，我们要努力去做，并热诚希望国内语言学者和语文工作者给予指导、批评和支持。

《现代语言学丛书》编委会

一九八二年十一月

前　言

自然语言处理(Natural Language Processing，简称NLP)就是利用电子计算机为工具对人类特有的书面形式和口头形式的自然语言的信息进行各种类型处理和加工的技术，这种技术现在已经形成一门专门的边缘性交叉性学科，它涉及语言学、数学和计算机科学，横跨文科、理科和工科三大知识领域。自然语言处理的目的在于建立各种自然语言处理系统，如机器翻译系统、自然语言理解系统、情报自动检索系统、电子词典和术语数据库系统、计算机辅助教学系统、语音自动识别系统、语音自动合成系统、文字自动识别系统等。由于自然语言处理离不开电子计算机，因此，自然语言处理又可以叫做“自然语言的计算机处理”(Natural Language Processing by Computers)，以强调电子计算机对自然语言处理的作用。

自然语言处理又是语言文字应用的一个新课题，从语言学的观点来看，我们可以把它作为应用语言学的一个分支。

自然语言处理又是人工智能(Artificial Intelligence，简称AI)的一个主要内容，它是电子计算机模拟人类智能的一个重要方面。因此，自然语言处理还是研制智能化的电子计算机的一项基础性工作。目前，科学技术的发展突飞猛进，信息的数量与日俱增，电子计算机技术得到越来越广泛的运用，正在联成世界性的网络，并向更高的层次迈进，向智能化的方向

发展。智能化的电子计算机已经不是十分遥远或虚无缥缈的幻想，而是近在眼前、指日可待的现实。当前，美国、英国、日本等发达国家，都投入大量的人力、物力和财力，把智能化电子计算机的研制放在十分突出的地位。预计智能化计算机将会在本世纪末问世，并对人类社会产生不可估量的影响。它同过去人类历史上语言的出现、文字的创造、造纸技术的发明以及印刷技术的发明一样，将成为人类文明史上的又一件大事。

自然语言是人类区别于其它动物的重要标志之一。人借助于自然语言交流思想，达到互相了解，组成人类社会生活；人还借助于自然语言进行思维活动，认识事物的本质和规律，创造了人类的物质文明和精神文明。

自然语言是人脑的高级功能之一。心理学研究表明，人脑的语言功能具有一侧化的性质。它主要定位在大脑左半球，由大脑左半球所控制。因此，自然语言是人类特有的一种最重要的智能。智能化电子计算机的研究离不开自然语言处理，因自然语言处理的研究水平，在智能化计算机的研制中，起着举足轻重的作用。我们中国的自然语言处理工作者，应该站在电子计算机智能化这样的高度，以战略的眼光来看待自然语言处理技术的研究，把我国的自然语言处理提高到一个新的水平。

在电子计算机软件中，早已设计了许多人工语言，如 BASIC，PASCAL，COBOL，PROLOG，LISP 等程序设计语言。这些人工语言与自然语言一样，都遵循着形式语言的规律和法则。美国语言学家乔姆斯基(N. Chomsky)的形式语言理论，既适用于人工语言，也适用于自然语言。这有力地说明，自然语言与人工语言之间，在形式描述方面，确实存在着某些共同的性质。正如美国著名的逻辑学家蒙德鸠(R. H. Montague)在《英语作为一种形式语言》一文中所说的："我并不认

为形式语言和自然语言之间在理论上存在着重要的区别。”

但是，自然语言毕竟是人类历史长期发展而约定俗成的产物。它带着几千年人类历史的痕迹，比人工语言要复杂得多，因而用计算机处理起来也就困难得多。

自然语言起码在下面4个方面与人工语言大相径庭：

1. 自然语言中充满着歧义，而人工语言中的歧义则是可以控制的；

2. 自然语言的结构复杂多样，而人工语言的结构则相对简单；

3. 自然语言的语义表达千变万化，迄今还没有一种简单而通用的途径来描述它，而人工语言的语义则可以由人来直接定义；

4. 自然语言的结构和语义之间有着千丝万缕的、错综复杂的联系，一般不存在一一对应的同构关系；而人工语言则常常可以把结构和语义分别进行处理，人工语言的结构和语义之间有着整齐的一一对应的同构关系。

由于自然语言的这些独特性质，使得自然语言处理成为人工智能的一大难题。自然语言处理的种种难题常常使研究者们感到心力交瘁，进退维谷，往往使他们陷入束手无策、一筹莫展的困境中。然而，恰恰因为自然语言处理的这些困难，却吸引了一大批敢于迎着困难上的、毫无畏惧的探索者。他们以战胜困难为乐，以克服困难为荣，每当他们有所前进的时候，就会产生“山穷水尽疑无路，柳暗花明又一村”的清新之感，体验到胜利者的欢乐。有志于自然语言处理的探索者就像科学战线上的侦察兵。对于侦察兵来说，没有道路的路，才是世界上最好的路。自然语言处理有如一条充满艰险的荆棘之路，这条荆棘之路一旦被勇于探索的侦察兵披荆斩棘之后开

通了，前面就是一马平川的坦途。正是这种对未来的坚强信念，从本世纪50年代以来，国内外学者在这个新的学科领域进行了不屈不挠的探索，历时40余年，现在已经取得了可喜的成绩。

自然语言处理有时也叫做“计算语言学”(Computational Linguistics)。计算语言学这个术语，最早是在美国的语言自动处理咨询委员会(Automatic Language Advisory Committee，简称ALPAC)1966年11月的ALPAC报告中出现的。据说这个术语的提出人就是ALPAC报告的编委之一海斯(D. G. Hays)。他曾经提出运用从属理论(Dependency Theory)来进行机器翻译的自动句法分析，是机器翻译的最早的研究者之一。他当时提出计算语言学这个术语，是希望学者们花更多的精力去研究自然语言计算机处理的基本理论。因此，计算语言学也可以看成是自然语言处理的同义词。

本书着重论述自然语言处理的方法，当涉及到自然语言处理的基本理论的时候，我们才使用计算语言学这个术语，也就是说，自然语言处理这个术语主要用于说明方法，计算语言学这个术语主要用于说明理论。两个术语各有分工，以体现它们各自的特点。

本书共分八章。第一章至第七章讲述基本知识，第八章讲述应用系统。各章内容简述如下：

第一章讲述自然语言处理与理论语言学的关系，说明自然语言处理对语言学各个方面的深刻影响。

第二章讲述自动词法分析，以有限状态转移网络为工具，说明黏着型语言和分析型语言的自动词法分析方法，并介绍了书面汉语的自动切词与文本自动标注的方法。

第三、第四、第五章讲述自动句法分析。以递归转移网络

和扩充转移网络为工具，说明基本的剖析技术，提出了“潜在歧义论”，分析了科技术语和日常语言中的潜在歧义，并介绍了良构子串表和线图分析法。

第六章讲述复杂特征理论以及合一运算方法（这是当前自然语言处理中最有影响的方法），并介绍了中文信息处理中的多叉多标记树模型。

第七章讲述自动语义分析，介绍了义素分析法、语义场、语义网络等语义分析方法。

第八章讲述各种自然语言处理系统，使读者进一步了解到自然语言处理研究的实用价值。

从本书内容安排可以看出，本书的重点是自然语言处理的方法，而不是理论。对于自然语言处理的许多理论（如广义短语结构语法、词汇功能语法、功能合一语法等），仅在说明方法时加以简要的介绍，不作详尽的叙述，以便提高本书的通俗性和实用性。有中等文化程度的广大读者，理解本书的内容将不会有很大的困难。

本书特别注意介绍自然语言处理中的新方法，尽可能深入地、具体地描述每一种方法的操作过程。对于自然语言处理中的一些传统方法，请读者参阅笔者的《数理语言学》、《自动翻译》、《中文信息处理与汉语研究》、《现代汉字和计算机》、《数学与语言》等著作，本书不再作介绍。

本书在写作时，还尽量考虑到不同学科读者的需要，使语言学工作者可以从中了解到计算机处理自然语言的有关技术，使计算机工作者可以从中了解到现代语言学的有关知识。希望本书的出版，对于语言学工作者和计算机工作者在自然语言处理这个学科中的进一步合作，能够有所裨益。

我曾于 1979—1981 年在法国格勒诺布尔大学自动翻译

中心(GETA)学习，师从当时的国际计算语言学委员会主席沃古瓦(B. Vauquois)教授，进行了汉外多语言机器翻译试验。1986—1988年我又到德国夫琅禾费研究院新信息技术与通讯系统研究所担任客座研究员，进行了术语数据库的开发研究。1990—1993年我在德国特里尔大学担任客座教授，讲授中文信息处理和机器翻译等课程。在前后几次出国期间，我有机会直接阅读到国外自然语言处理研究的最新文献，亲自了解到国外在这个领域中的最新成果，分别拜访了好几位国外在这个领域中卓有建树的专家学者。所有这些使我对于自然语言处理有了更深的认识，耳目为之一新。我在本书中，力图把在国外学习和研究的所得反映出来，本书在写法上以及章节的安排上，受到了国外有关自然语言处理著作的启发和影响。当然，本书的写作也参考过国内时贤的论文和著作多种。如果没有国内外学者的出色工作，没有他们极为宝贵的研究成果，本书是写不出来的。本书在每章末均列出有关的参考文献，在本书出版之际，谨向他们表示衷心的感谢。

中山大学外语系王宗炎教授始终鼓励我写作本书，并给了我极大的支持和热情的帮助。上海外语教育出版社顾霞君教授对于本书的内容、体例和写法提出过十分有益的建议，特在此致谢。

在本书写作过程中，笔者常为自己的学识不足而苦恼。自然语言处理作为一门交叉性边缘性学科，涉及到文科、理科、工科各个方面的知识，笔者学识浅陋，总有绠短汲深之感。论述之中，倘有不当，恳请海内外读者批评指正。

冯志伟

1994年9月30日于北京

目　录

第一章

自然语言处理与理论语言学

采用计算机技术来研究和处理自然语言是本世纪50年代才开始的。40年来，这项研究取得了长足的进展，成为了一门重要的新兴学科——自然语言处理。在这一章中，我们将说明自然语言处理在语言学以及现代科学体系中的地位及其对语言研究各个方面的深刻影响。

计算机对自然语言的研究和处理，一般应经过如下3个方面的过程：

第一，把需要研究的问题在语言学上加以形式化(linguistic formalism)，使之能以一定的数学形式，严密而规整地表示出来；

第二，把这种严密而规整的数学形式表示为算法(algorithm)，使之在计算上形式化(computational formalism)；

第三，根据算法编写计算机程序，使之在计算机上加以实现(computer implementation)。

因此，为了研究自然语言处理，我们不仅要有语言学方面的知识，而且，还要有数学和计算机科学方面的知识，这样自然语言处理就成为了一门界乎于语言学、数学和计算机科学

之间的边缘性的交叉学科，它同时涉及到文科、理科和工科三大领域。

早在计算机出现以前，英国数学家图灵(A. M. Turing)就预见到未来的计算机将会对自然语言研究提出新的问题。他在《机器能思维吗》一文中指出："我们可以期待，总有一天机器会同人在一切的智能领域里竞争起来。但是，以哪一点作为竞争的出发点呢？这是一个很难决定的问题。许多人以为可以把下棋之类的极为抽象的活动作为最好的出发点，不过，我更倾向于另一种主张，这种主张认为，最好的出发点是制造出一种具有智能的、可用钱买到的机器，然后，教这种机器理解英语并且说英语。这个过程可以仿效小孩子说话的那种办法来进行。"图灵提出，检验计算机智能高低的最好办法是让计算机来讲英语和理解英语，他天才地预见到计算机和自然语言将会结下不解之缘。美国语言学家乔姆斯基(N. Chomsky)在计算机出现的初期也不约而同地把计算机程序设计语言与自然语言置于相同的平面上，用统一的观点进行研究和界说。他在《自然语言形式分析导论》一文中，从数学的角度给语言提出了新的定义，指出："这个定义既适用于自然语言，又适用于逻辑和计算机程序设计理论中的人造语言。"在《语法的形式特性》一文中，他专门用了一节的篇幅来论述程序设计语言，讨论了有关程序设计语言的编译程序问题，这些问题，是作为"组成成分结构的语法的形式研究"，从数学的角度提出来，并从计算机科学理论的角度来探讨的。他在《上下文无关语言的代数理论》一文中提出："我们这里要考虑的是各种生成句子的装置，它们又以各种各样的方式，同自然语言的语法和各种人造语言的语法二者都有着密切的联系。我们将把语言直接地看成在符号的某一有限集合 V 中的符号

串的集合，而 V 就叫做该语言的词汇……我们把语法看成是对程序设计语言的详细说明，而把符号串看成是程序。”在这里乔姆斯基把自然语言和程序设计语言放在同一平面上，从数学和计算机科学的角度，用统一的观点来加以考察，对“语言”、“词汇”等语言学中的基本概念，获得了高度抽象化的认识。

图灵和乔姆斯基都是当代第一流的学者。图灵是现代计算机科学理论的奠基人，而乔姆斯基则是转换生成语法学派的奠基人。他们以学术大师特有的远见卓识，指出了计算机与自然语言的密切联系，他们的思想成为了尔后自然语言处理取之不尽的源泉。

自然语言处理的出现，使得语言学在现代科学体系中的地位有了明显的变化，使语言学由一门基础科学变成了带头科学，获得了与数学、哲学同等的地位，语言学将成为人文科学发展的突破点和生长点，它的重要意义已经为越来越多的人所认识。

自然语言处理的研究首先是从机器翻译开始的。1946 年电子计算机刚一问世，人们在把计算机广泛地应用于数值运算的同时，也想到了利用计算机把一种或几种语言翻译成另外一种语言或另外几种语言。从 50 年代初期到 60 年代中期，机器翻译一直是自然语言处理研究的中心课题，当时采用的主要是“词对词”翻译方式，这种不是建立在对自然语言理解的基础上的简单技术，没有得到预期的翻译效果。60 年代中期，人们开始转入对自然语言的语法、语义和语用等基本问题的研究，并尝试着让计算机来理解自然语言。许多学者认为，断定计算机是否理解了自然语言的最直观的方法，就是让人们同计算机对话，如果计算机对人用自然语言提出的问题能

作出回答，就证明计算机已经理解了自然语言，这样，就出现了“人机对话”(或“自然语言理解”)的研究。自然语言处理的理论和方法也就在这些具体的研究中逐渐形成、成熟并完善起来。目前，除了机器翻译和自然语言理解之外，自然语言处理的研究领域还扩展到了自然语言人机接口、情报自动检索、术语数据库、语料库、计算机辅助教学、语音自动识别与合成、文字自动识别、言语统计、词典编纂、风格学研究等领域。自然语言处理已经成为现代科学技术的一个研究热点。

自然语言处理的研究与计算语言学的研究是密切不可分的。我们在前言中说过，计算语言学可以看成是自然语言处理的同义词，当我们主要涉及方法的时候，用自然语言处理这个术语；当我们主要涉及理论的时候，用计算语言学这个术语。因此，在我们讨论自然语言处理的各种问题时，也不可避免地会讨论到计算语言学的问题，用到计算语言学这个术语。

1962 年美国成立了计算语言学学会，每年开一次年会，并且出版学术季刊《美国计算语言学杂志》(American Journal of Computational Linguistics)，后改名为《国际计算语言学杂志》(International Journal of Computational Linguistics)。1965 年在美国纽约成立了国际计算语言学委员会(International Committee of Computational Linguistics，简称 ICCL)，每两年召开一次国际会议，叫做 COLING，现已召开了 15 届，我国学者从 1982 年起就参加了 COLING 的活动，并在学术会议上发表论文。近年来，我国的自然语言处理研究很活跃，我国于 1983 年 5 月由中国中文信息学会组建了自然语言处理专业委员会，该专业委员会主要研究机器翻译，中国中文信息学会又于 1987 年 6 月组建了计算语言学专业委员会，接着，于 1988 年 6 月召开了首届计算语言学学术会议，