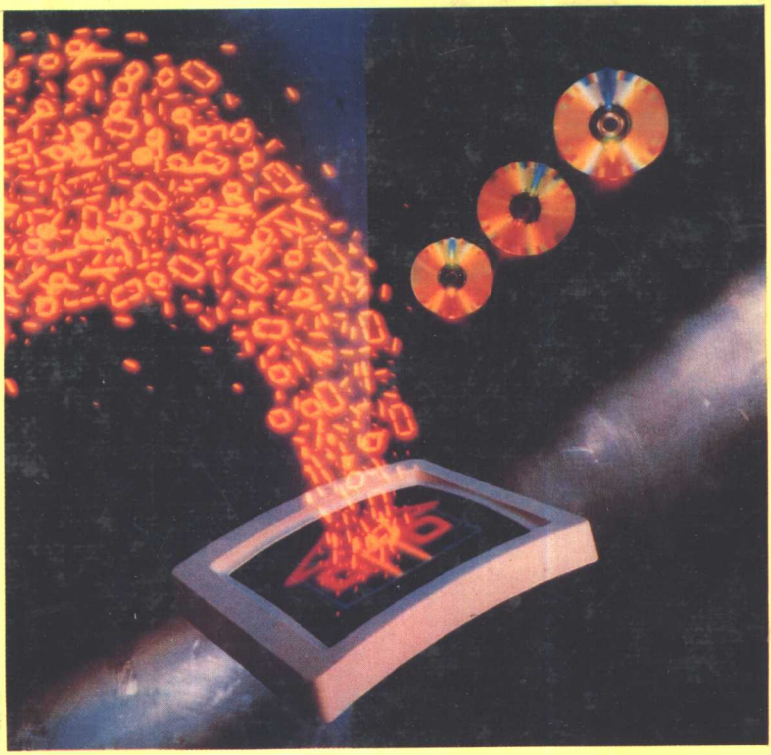
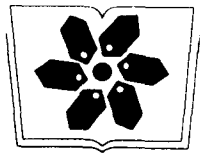
语言学知识的计算机辅助发现



白 硕 著

科 学 出 版 社



中国科学院科学出版基金资助项目

语言学知识的计算机辅助发现

白　硕 著

科 学 出 版 社

（京）新登字092号

内容简介

本书系统地介绍了作者利用机器学习、机器发现方面的新成果，在探索从语料中直接获取面向汉语的语言学知识方面的研究成果。作者选择了几个关键性的突破口进行了深入的研究，获得了有指导意义的理论成果，研制出了具体的发现工具，并用这些发现工具做出了有语言学价值的发现。

本书可供从事理论语言学、计算语言学、自然语言处理、机器学习与机器发现方面的研究人员参考。从事心理语言学特别是儿童语言习得的研究人员及研究科学发现方法论的学者也可以从本书中获益。

语言学知识的计算机辅助发现

白 硕 著

责任编辑 刘晓融 留 霞

科学出版社出版

北京东黄城根北街16号

邮政编码：100717

中国科学院印刷厂印刷

新华书店北京发行所发行 各地新华书店经售

*

1995年7月第 一 版 开本：850×1168 1/32

1995年7月第一次印刷 印张：4 插页：2

印数：1—2 000 字数：102 000

ISBN 7-03-004527-0/TP·412

定价：9.80 元

序

(一)

能把这样一本书贡献给读者，是笔者感到无限欣慰的事.

多年来，大家提倡计算语言学，多是从开发计算机应用领域着眼，也就是把语言学的成果拿来应用. 这种工作常被称为语言工程. 然而反过来如何呢?能不能把计算机科学的概念和方法用于研究语言呢?这是语言学特别关心的问题,也是笔者多年来想回答的问题.

不入虎穴，焉得虎子. 为了回答上述问题，我和包括本书作者的许多人一起，深入语言学领域，从事语言学家的工作（写论文，做演讲，参加学术会议，等等). 工作做得好坏当然应请别人评论. 我们自己则可以毫无愧色地说，我们摸到了语言学的脉搏——它的理想、追求和困惑.

我们特别注意从方法的角度观察自己的（语言学）工作. 我们看到许多语言学家都要做卡片，制表格，做许多机械的枯燥的工作. 因此，我们相信，语言学工作同许多别的学科一样，总有大量的可以严格化、形式化、算法化的部分. 一旦把这些东西界定出来并用计算机实现，语言学的工作就会效率更高，差错更少.

早就有人在这方面动过脑筋了. 早年的统计语言学、语言信息的研究，近年来的人工智能和语言工程，都在不同程度上涉及这个问题. 但是许多研究者都想撇开语言学家的工作，独出心裁，另搞一套. 因此，也没有被语言学家接受：你走你的阳关道，我走我的独木桥. 真正的问题是：语言学界真正关心的东西恰在河对岸——阳关道到达不了的地方.

我们不认为语言学工作可以用计算机来做，也不认为有什么

计算机上的绝技可以取代语言学. 但我们相信，只要用理论计算机科学的观点剖析当代语言学方法，一定可以搞出有用的东西来. 不是另铺一条阳关道，而是改造那座独木桥，使语言学得以驱车而过.

经过几年的积累，终于有了足够的基础，可以开始做实验了. 实验工作从模仿语言学家入手. 本书作者以极大的热情投入这项工作. 他和他编制的软件组成一个人机系统，模仿语言学家朱德熙做的一项工作，得到了令人振奋的结果.

我们仔细分析了这个实验，确定了进一步的目标，包括理论方面的整理和发展，也包括对一些语言学家不知道答案的问题用我们的办法做出解答. 本书就是这些工作的结晶.

本书从某一个角度勾画出了计算语言学的一个蓝图，使读者能对我们的理想和实现这个理想的手段有所了解. 这就是笔者所以感到欣慰的原因.

(二)

在开创计算机的新应用时，人们注意到基于演绎法和基于归纳法的思维活动的深刻区别.

演绎法的出发点是多少已经抽象化、形式化的前提. 从一些前提出发，演绎出种种结论来，是计算机一定可以胜任的事（注意：我们这里不是说机器证明，那包括了一个判定问题，即回答某一命题是否可以在演绎过程中得到). 只要前提含有可以互相消解（resolve）的对象，就可以衍生出新的命题来. 但归纳法的出发点常常是未充分抽象化、形式化的大量个别事例，希望从中抽象出有用的概念、模式、定律来. 这种工作能不能用计算机来完成呢?

思维活动总是有目的的. 演绎法作为一个手段虽无目的可言，但是基于演绎法的思维活动比如证明（或反驳）还是有目的的. 只是判断目的是否达到十分简单——只要演绎出所要的命题（或它

的逆命题）就达到了目的.

使用归纳法的时候（例如划分词类、发现句法模式等）遇到的情况大不相同.因为在达到目的之前往往说不清目的是什么，到了目的地，也未必一下子就能判断出来. 只有反复地尝试、失败，目的才渐渐明晰，手段才逐步建立起来.

我们认为，要单独由计算机来完成这个过程是不可能的，至少在当代是如此. 需要人机共生系统来做. 其中的人负责设定目标手段，机负责实现这种手段而不管目标是什么. 有了这样的系统，可以大大提高工作的效率和质量.

如果认为这种看法是有道理的，就应该进一步研究归纳的手段. 本书主要就是以语言学为背景提出许多概念和方法，做为人机共生系统的基础，作者在实验中的成绩，应能成为这种看法的例证. 笔者希望看到这种看法与做法能在计算语言学领域中得到更加深入的应用，更希望它能在更多的领域中得到发扬.

（三）

本书以语言学为大系，对每个问题分析已知研究工作的得失，设定自己的目标，并用理论计算机科学的方法做出解答. 然后，又回到语言学中诠释所得到的结论. 单纯从语言学角度阅读本书的读者也一定会感到兴趣.

其实，作者提出的概念和方法，都是基于语言学研究成果，进行抽象，使之升华，然后结晶，回到语言学中. 许多章节后面，作者给出了浅显的例子. 这些例子说明了其成果的价值，更是对语言学研究方法的一种洞察. 这些例子非常精彩，有时看来奇怪，值得仔细玩味. 比如说，从一个语言的两个例句

（1）我是学生.

（2）你是我学生

出发，能对这个语言做出什么样的模型来呢？

本书第八、九章尤其值得注意. 因为本书宗旨之一是得到有价

值的语言学成果. 这两章报告了一个具有一定规模的实验并分析了所得到的结果. 用语言学的尺度来分析这些结果，也许还嫌粗糙. 但在现代汉语学界中这种语法、语义、语用三位一体的研究成果还很少见. 相信一定会引起读者的兴趣 . 至于笔者说的粗糙，并不是说还要分许多小类，而是说，在分类方面要利用不同层次、不同角度，交织而成，以便以纲带目，便于把握. 这就要做进一步的归纳，比如那些可以进入双宾语句式的动词有什么语义上的共性. 还需要用稍大一点的语境来证实或发展这种分类，比如

一 VP……

等 VP 再……

我是 VP 才……的.

不用说，还应该用更多的动词来丰富这项研究.

（四）

计算机科学与语言学有不解之缘. 然而，语言学一直是贡献者，极少得到回报. 形式语言理论的创造人之一 Chomsky 本人就是一位语言学大师. 更不消说语言学在语言工程方面的贡献了（计算机的应用给语言学家在收集资料和统计分析中提供了便利. 但那毕竟不能算是语言学家的工作，而是助手的工作）. 本书作者的研究可说开创了一个新的局面，使语言学家可以利用计算机来做自己感兴趣的事.

德雷福斯在“计算机不能做什么”一书中说过，如果有一位高级棋手和计算机搜索能力有效地结合在一起，可以胜过任何人间棋手和计算机棋手. 我们相信，语言学与计算机科学的结合会在两个领域中都开创出新的天地.

马希文

一九九五年六月

目　录

第一章 引 论

1.1 问题的提出

从一组事例中发现一般性规则，是人类认知活动的基本形式之一，也是机器学习的中心课题. 这种认知活动在逻辑上被描述成一种“归纳”. 由于已知前提（包括事例与背景知识）的不完全性，归纳过程往往是非单调的，就是说，已经学到的规则总有可能被后来的事实证明是不正确的、需要修改的，然而在没有遇到这样的事实时，这些规则又可以认为是近似正确的、不妨试用的. 动态地看，这与科学理论的进化有很多相似之处.

上述“归纳过程”的一个具体而重要的例子是从一个由例句组成的语料库中发现特定自然语言的规律. 这就是所谓“语言学知识的发现”. 这本来是一种不折不扣的“归纳”，然而这方面的研究同逻辑一直是相脱节的. 尽管学术界已提出过一些算法，证明过一些涉及可学习性和复杂性方面的结果，最近又在基于统计的语料库语言学方面取得了给人以深刻印象的进展，但关于语言学知识获取的逻辑实质，一直没有见到相应的报道.

这就难免给人一个印象，似乎对于归纳的逻辑分析在语言学规则这种特殊形式的知识的发现上无所作为. 实际上，完全不是这么回事. 本书将展示与语言学知识发现有关的各个层次上的形式化机制——从数学建模、逻辑分析、算法描述、具体实现直到结果的语言学解释.

另一方面，归纳和类比的逻辑实质目前研究得还不十分透彻，无论在 AI 逻辑还是哲学逻辑中都是如此. 这类“非演绎逻辑”一般都是可错的. 它们允许某种“逻辑跳跃”来达到一些好

的猜测，而这种机制决不仅仅局限于语言学知识的发现．因此，我们的研究也将为进一步深化对于归纳和类比的认识、探索知识发现的一般途径有所贡献．

计算机在中国的普及，离不开汉字环境．而汉字输入的困难，又极严重地限制了普通用户对计算机的使用．近年来的一些研究表明，要想进一步提高汉字输入的效率和质量，需要许多层次的语言学知识．另外，还有许多应用领域也都不同程度地对面向汉语的自然语言处理技术提出了需求．一些自然语言处理系统在商业上的成功，使这种需求变得更加迫切了．这样，以自然语言为主要研究对象的语言学，理所当然地受到计算机浪潮的猛烈冲击．像计算语言学这样的新兴边缘学科应运而生，并通过与中文信息处理技术的结合，在中国取得了很大的发展．

计算语言学旨在以自然语言处理（包括理解、生成、人机对话、机器翻译以及语音/文字输入的后处理等）为技术背景，揭示自然语言的词法、句法、语义、语用诸平面及其相互作用的计算结构，把语言学知识重塑成可以转化成产品的计算模型．计算语言学的出现，给语言学增添了新的活力，为当代语言学研究的导向起了极为关键的作用．

然而，长期以来，绝大多数语言学家，包括一些计算语言学家，主要还是以手工方式从事语言学研究、发现语言学知识，其规模与效率同实际需要相比极不匹配．随着时代的发展，机器可读的真实语料正在以惊人的速度积累．计算机这样一个强有力的信息处理工具，如果不能利用这些宝贵的语料资源来辅助进行语言学知识的整理和发现，那真是莫大的浪费！

面对这样一种强烈的反差，一些研究者提出了“语料库语言学”的思想，以统计学的方法从语料中发现统计性的语言模型．这一方法在许多方面得到了成功的应用，但也有明显的不足，即如何从语料中发现确定性的语言学规则的问题，仅靠统计学是无法解决的．

因此，我们在本书中采纳了“分布分析”的合理思想，并针对

真实语料的种种特点，结合汉语的实际，从数学、逻辑、算法和实现等各个角度，全面阐述了从语料中发现确定性语言学知识的理论和方法．其中大部分工作是在作者攻读博士学位期间在马希文教授指导下完成的，其余部分则是作者后来的补充和发展．

这项工作的核心任务是研制具有实用价值的确定性语言学知识辅助发现工具．我们知道，从大量语料中发现带有规律性的语言学知识，是语言学家一种典型的工作方式，其中有大量的可形式化的、可借助于计算机来完成的操作．在这些环节上，如能用计算机代替语言学家，不仅可以扩大研究规模、提高研究效率，而且可以把语言学家从繁琐的事务中解放出来，更充分地发挥他们的语言直觉在语言学研究中所起的计算机无法替代的作用．

这项工作同时也是面向自然语言处理系统的开发、维护自动化的．一个自然语言处理系统要想具有接近实用的处理能力，必须以大量的语言学知识及非语言学知识为后盾．因此在其开发和维护过程中，语言学知识的获取是一个非常关键的环节．在目前的技术条件下，这一环节仍然是所谓的“瓶颈”．于是，一个具有实用价值的语言学知识辅助发现工具，实际上就相当于一个使自然语言处理系统具有自扩充、自维护功能的高级开发手段，其实用价值是不言而喻的．

心理语言学家也在从一个独特的角度进行着与这项工作有关的研究．儿童语言习得是近年来心理语言学中比较活跃的研究领域．研究语言习得模型，不能不涉及到人类发现语言学知识的机制问题．在这个问题上，计算机既是一个合适的参照物，又是一个检验模型动态特性的绝好的实验工具．事实上，有一些学者，如麻省理工学院的 Berwick，就是同时研究人类语言习得和计算机辅助语言学知识发现的心理语言学家兼计算语言学家．

此外，在中国目前情况下开展这项研究，还有其更为重要的原因．

当关于自然语言处理的研究起步之时，像英语这样的印欧语言的语法研究已经相当成熟和深入．这使得从事这些语言的计算

机处理方面的研究人员少走了不少弯路. 他们可以在语法方面少花一些时间，把更多的精力集中于语义、语用方面的课题. 这是语言学史留给他们的一份得天独厚的幸运礼物.

汉语的情况就完全不同了. 一百年来，几代语言学家为了寻求汉语语法的合理描述，仿照、借鉴了多种西方语言学理论，但迄今为止只有部分成功. 相反，汉语不同于印欧语的奇妙特性正在逐渐暴露出来. 在汉语里，不存在可以从形态上加以区别的“词类”，但那些可以从分布上加以区别的“词类”，却在句式的描述和歧义的甄别中占有重要的地位. 汉语的短语（乃至词）往往可以不加标记地被另外的成分穿插分割或被变更语序，但仍不失为一个表义整体……. 这些现象，给面向汉语的自然语言处理研究者们带来了许多困难.

这就是这些研究者们不满足于汉语研究现状的原因. 他们没有时间去等待一个现成的、令他们满意的语法理论，他们中的一些人只好亲自去从事汉语的研究. 现在中国的计算语言学界中，属于这种情况的人为数不少.

我们相信，这种状况绝不是汉语所独有的. 像英语那样语法理论非常成熟的语种毕竟是少数. 大多数语种的语法研究现状与其自然语言处理的要求相比还有不少差距，其语言的许多独特性质有待进一步的揭示. 因而，从词类和句法做起的一整套计算机辅助发现工具的研制，在汉语的故乡——中国，更有其必要性和紧迫性.

此外，语言学知识的计算机辅助发现问题，不仅是一个技术问题，还是一个理论问题. 计算机只能在一定的层次上充任发现工具. 以句法知识为例，计算机只能“发现”一条一条具体的句法知识，但句法知识的一般形态却必须由人来设计. 为句法知识规定什么样的一般形态，其背后是有一定的语言学、元语言学理论作指导的. 一般形态的选择将直接影响到发现工具的质量. 因此，在关键问题上，我们无法回避语言学界的学术争论，而只能在博采众长的基础上，直言我们的观点和主张. 另一方面，从语料中

发现语言学知识，其逻辑实质是归纳，而归纳推理如前所述是可错的.如何保证一个发现工具能随着语料的增加不断地纠正错误，不断地逼近自然语言某一个侧面的真实情况呢？这就需要研究语言学知识体系随着语料的增加进行动态调整和演化的规律，即一种面向自然语言的知识进化理论，一种不拘泥于统计方法的“广义”语料库语言学. 由此可见，研究语言学知识的计算机辅助发现，不只是工具的研制，还包括了与此有关的语言学、元语言学、心理语言学、人工智能中的学习与推理乃至数学上的全面深刻的理论探讨，是一项跨学科、综合性的研究.

由此可见，研究确定性语言学知识的计算机辅助发现，具有十分重要的理论意义和应用前景.

1.2 语言学知识及其形式化

从传统的观点看，一般把语言学知识分成语音、词汇、句法、语义、语用等几大块. 从发现的角度看，语言学知识又可分为以下三种类型：范畴、规则、策略.

范畴 指具有某种共性的语言学对象的集合，如语音学中的韵部、语法学中的词类（包括短语的类）、语义学中的格等等. 其中，词类是书写句法规则的基础，在像汉语这样形态很不发达的语言中如何划分词类，现在还是一个有争议的问题. 本书将花费较多篇幅来讨论词类问题. 粗略地说，划分词类的标准可以归结为形态标准、意义标准和语法功能分布标准三派. 其中，主张语法功能分布标准的意见最为客观，也最容易用计算机辅助实现词类的划分，但由于其理论上目前尚不完善，遭到了一些批评. 我们在本书中的一部分工作，就是对分布分析理论的完善和推广.

规则 指某些语言学对象之间带有一定普遍性的变换或推演关系. 我们可以举出种种语言学规则的类型来：

（1）语音规则 在某些用拼音文字记录的语言中，词的拼

写和实际发音（音位序列）之间不存在一对一的简单对应，而是有一套比较复杂的发音规则支配着拼写与发音之间的转换．在汉语中，连读变调、协同发音等也有其独特的规则．在语音合成等研究中，这一套规则是不可缺少的．

（2）词法规则　在某些有词尾变化的语言中，词的形态变化受词法规则的影响．特别是像英语这样的语言，经过长期演变，其规则中有例外，例外中又有进一步的规则．在自然语言理解系统中，首先要把文本中经过形态变化的词还原成其“基本形态”，才能从机器字典中查到词义．因此，对这类语言来说，词法规则是一类很重要的形式规则．汉语虽不存在词尾变化问题，但是汉语中有一类可以扩展的词——离合词．如何把扩展形式的离合词还原成其“基本形态”，也是汉语理解中同样要遇到的问题．汉语离合词的扩展规则及其逆方向的还原规则，已经有很强的“句法”特性了，可以说是介于词法和句法之间的规则．“词”这一范畴的边界模糊，是汉语的一个重要特点．本书第四章对此有详细的讨论和处理方法．

（3）句法规则　在不同的语言学流派看来，“句法规则”是很不相同的一些东西．传统的中心词分析法认为，句法规则就是“句成分”的组成规则；结构主义语言学派认为，句法规则就是“直接成分”的组合规则；转换生成学派认为，句法规则就是“生成规则”（句型）加转换规则（句型推演规则）．更新一些的学派还有新的说法，比如“合一规则”等等．本文又提出了“环境推演规则”及其发现的问题，详细论证参见第三章．

（4）语义规则　在计算语言学中，语义总是用一定的形式系统，如内涵逻辑公式、语义网、Frame（框架）、Script（脚本）等等“深层结构”来表示的．这时，语义规则就表现为“表层结构”（即语言形式）到这些“深层结构”的对应规则（所谓“投射规则”），或是反向的由“深层结构”到“表层结构”的对应规则（所谓“表达规则”）．现代自然语言处理技术的发展，对语义规则的要求日益迫切，这对语言学界是一个重要的刺激源．

(5)翻译规则　在有共同的深层结构作为中介的情况下，双语翻译规则实际上可以拆成甲语言的表层结构与公共的深层结构间对应的语义规则和乙语言的表层结构与公共的深层结构间对应的语义规则两部分．在没有共同的深层结构的情况下，双语对译规则实际上是两种语言的表层结构之间极其复杂的、在人看来是保持意义等价的对应关系．现代机器翻译界日益重视所谓“逻辑语义平面”的作用，看来前一种形式的双语对译规则的发现更加重要一些．

策略　以上提到的种种规则，都是具有显式的形式化表示并具有可操作性的语言学知识，我们称之为确定性语言学知识．它们只是庞大的语言学知识体系的一部分．还有很多很重要的语言现象，例如绝大多数修辞现象和文学现象，其背后确实有某种机制在起作用，可是目前它们还只是一种尚未形式化的直觉知识．属于这一情况的还有使用形式规则的“策略”．一个自然语言处理系统经常会遇到这样的情况：从局部看，几个规则执行的前提条件都满足，因而都可执行；但从全局看，其中只有一个可以导向预期的目标．我们人类显然有着一套从若干个候选规则中选择“最可能导向预期目标的”那条规则的复杂的决策机制，它也是一种尚未形式化的直觉知识．我们说这类机制是“尚未形式化的直觉知识”，是指它们不具备显式的形式描述，没有可操作性，更准确地说它们不具备显式地可描述的保证其施用有效的疆域．这一类直觉知识我们统称为一种广义的“策略”．它们一般有可能借助于对特定的人工神经网络（artificial neural network）或某种统计模型的“训练”，体现为相应的“权值”（weights）或“概率分布”．但训练后它们仍无法严格保证施用的有效性，只能算做一种近似的、统计意义上的“发现”．这种“发现”的机制，与范畴和规则的发现机制相比，是有着很大差别的，但无疑也是非常重要的．