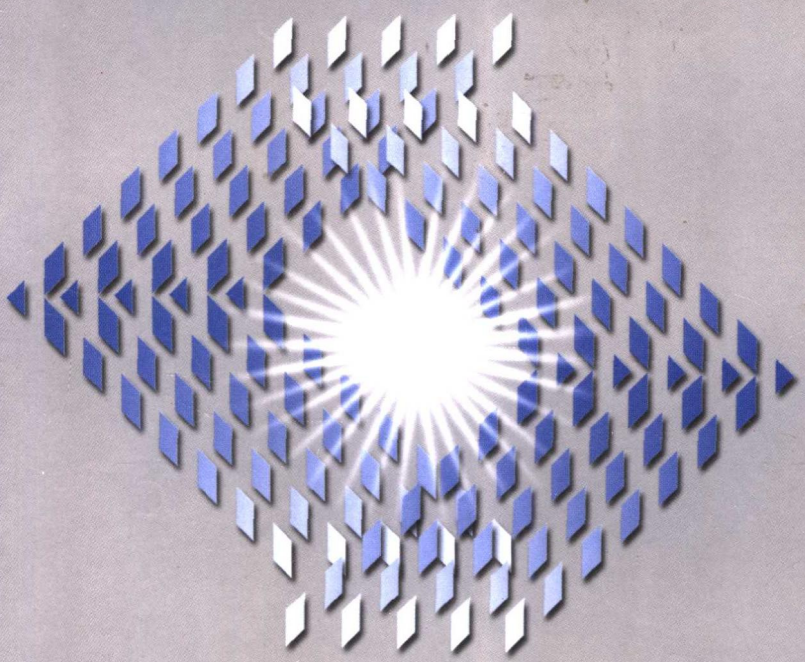
21世纪信息通信系列教材

自然语言处理技术基础

ZIRAN YUYAN CHULI JISHU JICHU

王小捷 常宝宝 编著



北京邮电大学出版社
www.buptpress.com

自然语言处理技术基础

王小捷　常宝宝　编著

北京邮电大学出版社
·北京·

内 容 简 介

本书包括了三个方面的内容。第一部分介绍基于规则的自然语言处理技术，分别从语法和语义两个层面入手。首先介绍了几种语法系统的形式化表示方案，在此基础上，介绍了几种典型的上下文无关句法分析和基于复杂特征的句法分析方法。在语义层面，分别从词义和句义两个层次介绍了进行词义和句义分析的方法。第二部分介绍基于统计的自然语言处理技术，包括词汇层的一些统计语言模型以及在句法层的概率上下文无关语法。第三部分介绍一种重要的应用——机器翻译，分别从规则和统计两个方面来介绍它的理论和实现。

图书在版编目(CIP)数据

自然语言处理技术基础/王小捷等编著. —北京：北京邮电大学出版社，2001.6

ISBN 7-5635-0527-X

Ⅰ.自… Ⅱ.王… Ⅲ.自然语言处理 Ⅳ.TP391

中国版本图书馆 CIP 数据核字(2001)第 084948 号

出 版 者：北京邮电大学出版社(北京市海淀区西土城路 10 号)
邮编：100876 电话：62282185(发行部) 62283578(传真)
网址：http://www.buptpress.com

经 销：各地新华书店

印 刷：北京忠信诚胶印厂

印 数：4 000 册

开 本：787 mm × 1 092 mm 1/16

印 张：9.75

字 数：229 千字

版 次：2002 年 12 月第 1 版 2002 年 12 月第 1 次印刷

书 号：ISBN 7-5635-0527-X/TP·54

定 价：19.00 元

序　言

自然语言处理:自然的人机交互

随着计算技术的飞速发展,计算机已成为辅助人类认识和改造世界最为强大的工具之一,自出现那一天起至今,帮助人类完成了许多自身难以完成的工作,使人类社会在这一段时期里获得了比以往任何时期都要快的发展。相信在可以预见的未来,计算机对人类发展的重要辅助作用还将持续。

为了让计算机能完成人类所赋予的各项任务,一个首要的问题就是人和计算机的通信问题,即如何把人类希望计算机完成的任务告诉计算机,以及计算机在完成任务后又如何把结果告诉人们。

人机通信经过了几个时期,编写二进制代码、汇编代码、高级语言、第四代语言,人类为了与计算机进行通信,创造了一系列人工语言。为了和计算机进行通信,人类付出了许多的努力。在人类使用工具的历史长河中,人类还从来没有为了和自己创造的工具进行交流而如此屈尊过;如此为了使用这种工具而使自己向这种工具靠拢。人机的矛盾、人因为工具而产生的异化在这里表现得十分突出。一些哲学家早就注意到这个问题,提出了哲学和社会学上的解决方案。

但是,也可以看出,所有这些不断发展新人工语言的努力,正在让人类在使用计算机时离计算机远一些,而离人类本身更近一些。然而,我们知道,人类表达自己思想最方便、最自然的方式是利用人类自身的语言——各种自然语言;人与人之间交流观点、传播消息最方便、最自然的方式也是利用自然语言。因此,最自然的人机通信不应该是任何人工语言,而应该是自然语言。

要使计算机与人能通过自然语言进行通信,就要使计算机能够理解和运用自然语言。早在计算机发明不久,人们就开始了这个方面的尝试,自然语言处理[①]技术就是几十年来人们在这个方向不断努力的产物。

从某种意义上来说,自然语言处理技术提供了一个解决人机异化问题的技术上的解决方案:计算机直接处理自然语言,无需人去适应机器。这将是一个更自然、消除了异化的人机环境,计算机将能帮助人类完成更多的工作。

为了让计算机能很好地进行自然语言处理,一个有益的工作是考察人类的自然语言运用方式,虽然计算机进行自然语言处理的方式很可能与人类不同,但是,毕竟到目前为止,人类的自然语言运用是自然语言处理的唯一原型。遗憾的是,迄今为止,人类对自身运用自然语言的机制还不甚了解,更多的研究还集中在外在的语言本身上。

---

① 自然语言处理也称为计算语言学,它们所指的是同一个研究领域,只是在使用时稍有不同。通常的使用习惯是,在偏重本研究领域的理论时,使用计算语言学这一术语;而偏重于本研究领域的应用方面时,常使用自然语言处理。

语言学:经验材料和理性规则

最简单地说,人类对于自身所使用的自然语言的研究称为语言学。这种研究从很早以来就一直没有终止过,通常分为几个交错的阶段。

最早的研究是由希腊人创立的所谓"语法",并在法国人波尔·洛瓦雅尔的"唯理普遍语法"中得到了显著的体现。其特征是以逻辑为基础,制订出一些规则,用以区别正确的语言形式和非正确的语言形式。其对于语言材料本身缺乏科学的观察。

其后出现了语文学,其首要任务是确定、解释和评注各种文字的文献,通过比较不同时代的文献,确定每个作家的特殊语言,解读和说明用某种古代的或晦涩难懂的语文写出的碑铭。

随后,人们发现不仅可以进行这种比较,还可以进行不同语种间的比较,用一种语言阐明另一种语言,用一种语言的形式解释另一种语言的形式。这就是语言学的第三个阶段——历史比较语言学。

在这样一些研究的基础上,德·索绪尔建立了普通语言学,以此为界,标志着现代语言学的开始。在索绪尔那里,语言的研究重心转向共时语言学,研究语言体系的内部结构。这成为了结构主义语言研究的开始。在结构主义语言学派中,美国的描写派是最有影响的流派之一,他们注重记录实际语言,注重语言中各种单位的分布,基于分布信息的基础上对语言各单位进行切分、归并分类和组合。这时的语言学研究重视语言材料,具有很强的经验主义色彩。其主要原因是,美国语言学家十分强烈地受到一种需求的影响,这就是要把多达几百种以往没有文字记载的北美语言尽可能多地描写出来。最初的代表人物是弗朗兹·博厄斯(1858~1942年),他认为每一种语言都有其独特的语法结构,语言学家的任务就是要为每一种语言找到适合于该语言的描写范畴。其后,从1924年美国语言学会成立到第二次世界大战开始这段时间内的重要代表人物之一是伦纳德·布隆菲尔德(1887~1949年),他明确采用行为主义作为语言描写的框架。为了按照他所理解的"科学性"来描写语言,他认为应排除一切不能直接观察到的、也不能进行物理测量的素材,因此,语义的研究并不属于正规的语言学研究范围。这些观点,直到20世纪60年代,由美国后布隆菲尔德学派的结构主义语言学家齐格律·哈里斯(Zellig Harris)进一步继承。

20世纪50年代中后期,诺姆·乔姆斯基(Noam Chomsky,1928~)提出了转换生成语法,他秉承波尔·洛瓦雅尔"唯理普遍语法"的衣钵,重新确立了理性主义在语言研究中的地位。他认为:语言描写和分析的目的不在于分类,而在于建立一种理论,研究人的语言生成能力,即怎样用有限的成分和规则生成无限的句子,其目标是提出一个能产生所有句子的语法系统。他认为:人存在着先天语言能力,语言的结构是由人类的心理结构决定的,而语言的某些特征所具有的普遍性也证明了人类天性的这一部分为全体成员所共有,不论其种族或阶级如何,也不论其智力、性格和体质方面所显然具有的区别。乔姆斯基的理性主义观点曾经在语言学研究中占据着主导地位,时至今日,依然有着重要的影响。

与此同时,注重语言材料的语料库语言学仍然是一个重要的分支,并在80年代随着计算机计算能力的迅猛发展得到越来越多的重视。在80年代,一些语言学家、哲学家还发展了把语言纳入认知范畴来研究的认知语言学。

可以看到，在语言学发展的过程中，存在着经验主义（注重语言材料）和理性主义（注重语言机制）的交替发展。这种情形也出现在了计算语言学的发展过程中。

从语言学到计算语言学

计算语言学诞生之日正值乔姆斯基学派的理论大行其道之时，自然语言处理的主流技术是基于规则的，从各种句法分析技术到句法语义分析技术，利用规则来描述语言现象使之能为计算机所处理是计算语言学的主导方法。

20 世纪 80 年代末和 90 年代初，由于大量联机语料的出现以及计算机处理能力的大幅度提高，也由于规则方法迟迟未能达到人们预期的目标，统计自然语言处理逐渐兴起，成为自然语言处理中与规则方法比肩发展的两个方向。

在统计方法开始盛行之初，规则方法和统计方法存在着很多的对立，但是不久，人们便认识到二者并不是不可调和的两个对立面，而是互为补充的。詹姆士·艾伦（James Allen）在他的《Natural Language Understanding》（第二版）一书中，在保留规则方法的同时，增加了一些统计方法的内容，在序言中，他谈到，老方法（基于规则）和新方法（基于统计）是互为补充的，谁也不能替代谁。

全 书 安 排

全书分为三个部分。

第一部分用来介绍一些重要的基于规则的自然语言处理技术，这部分是从第一章开始直到第五章。其中，第一章介绍面向计算机处理的上下文无关语法及其形式化表示方式；第二章介绍了几种基于上下文无关语法的句法分析算法；第三章介绍基于特征的增强上下文无关语法以及基于该类语法的句法分析方法。后面两章介绍语义层面，其中，第四章介绍词汇语义的表示和处理；第五章介绍句义的表示和处理。

第二部分从第六章到第八章，介绍一些基于统计的自然语言处理技术。其中，第六章介绍 n 元语言模型；第七章介绍隐马尔科夫模型；第八章介绍概率上下文无关语法。（王伟、孙健两位博士参与了第六和第七章的选材和编写。）

在第三部分介绍一个典型的自然语言处理的应用——机器翻译，为本书的第九章。这部分主要从技术的角度来考察、分析各种机器翻译系统在规则和统计技术下是如何来实现的，而不过多地介绍某个具体的系统。

作者

2002 年 1 月

目　录

第一章　上下文无关语法

说到语法(在本书中,语法均在句子层面使用,因此将不仔细区分语法和句法两个词的使用),人们可能首先会想到语言学课程。在语言学的教科书中,语法是一个主要的内容。在那些语法中,规定了如何用词构造句子,何种用法是不允许的等等。通常,语法可以用来辅助人们完成两件事情,其一是作为判定一个句子构造得是否合适的重要依据,也即,一个句子是否合乎语法;其二,依据语法来分析句子的结构,帮助人们理解句子内容,这一过程在人们学习外语时是尤为明显和重要的。(由此也可见,利用语法来进行句子结构分析对于进行自然语言理解是有一定认知依据的。)

对于计算机自然语言处理,利用认知依据来建立计算模型是一种可行的途径。因而,让计算机能够利用语法来分析句子是进行自然语言处理的一个重要阶段。与人类使用语法相同,计算机利用语法来分析句子也可以有两个层次:其一是识别一个句子是否合乎语法。通常把能完成该任务的计算机程序称为句子识别器。其二是分析句子的内部结构,确定句子的语法成分,为进一步的句子分析和理解提供足够的基础。通常把能完成第二个任务的计算机程序称为句法分析器。显然可以看出,句法分析器比识别器具有更强的能力。

为了实现句子识别器或句法分析器,需要预先赋予计算机两个东西。

第一个是语法:通常语言学教材中的语法是面向人的,为了让机器分析句子,需要让机器知道这些语法,这种面向机器处理的语法也称为形式语法,它是规定语言中允许出现的结构的形式化说明。其中很重要的是如何表示形式语法,即形式语法的表示方式。本章将介绍两种表示方式:重写规则和转移网络。

第二个是语法分析算法:机器依据形式语法来识别和分析句子并决定其结构的方式。在计算机自然语言处理中,我们更多地关心句法分析器的算法,因为句法分析器比识别器具有更强的能力,能够提供更多的信息。句法分析算法还应包括其中采用的数据结构的构造,在分析之后如何表示句子的句法结构等各个方面。在通常的人类自然语言中,未经分析的句子是线性的符号串表示。本章将介绍在经过分析后产生的句子结构的树形表示,以及两种表示对于理解句子所带来的差异,也即句子的结构歧义问题。

本章主要明确两个方面的内容,其一是形式语法的表示;其二是句子结构的表示。各部分是这样安排的:1.1 节一般性介绍形式语法的描述问题;1.2 节利用重写规则描述上下文无关语法;在 1.3 节介绍用转移网络和递归转移网络来描述上下文无关语法;在 1.4 节介绍句子在经过句法分析后产生的句法结构的树形表示;最后是对本章的小结。

1.1 形式语法描述

最简单的描述语法的方式是把一种语言中所有可能的句子都列举出来作为这种语言的语法。

这种描述语法的方式其问题是明显的，可以从以下两个方面来看。

第一，在这种语法描述方式下，为了要完成句子识别的任务，即判断一个句子是否符合该语法，也即判断该句子是否是这种语言中的一个合法的句子，就需要列出这种语言中所有可能的句子，这样要判断一个句子是否合乎语法，只需要把该句子和这种语言中的句子逐一比较，看看是否有和该句子完全相同的句子。而通常，我们所使用的语言其句子是无穷多的，无论是对于计算机还是对于人，穷举都是不可能的，对于计算机处理，一个可行的方案是编制一个程序来按某种算法生成并输出这种语言的所有句子，显然，对于有无穷个可能句子的语言而言，这个输出过程是无限的。对于这类语言，有如下的定义：

对于一种语言，如果能编写一部程序，使得能按某种次序输出该语言的所有句子，则称该语言是可递归枚举的。形式语言理论的一个结论是，可递归枚举语言是一种很强的语言，对它的句子进行是否合乎语法的判断并不一定能完全实现。假设给定某种可递归枚举语言，并编写出了一部程序能生成其所有的句子，现在来判断一个句子是否合乎语法，即该句子是否能和程序输出的某个句子完全匹配，如果找到一个完全匹配的句子，那么，可以说该句子是合乎语法的。但是，如果一直没有找到匹配的句子，也不能断定该句子不合乎语法，因为它还可能与后面输出的句子相匹配。由于在句子个数无限多时程序的输出过程是不会终止的，因而它与后面输出的句子相匹配的可能性就一直存在，也即是说，对句子的合法性判断可能不会在有限步骤结束，这对于计算机处理而言，是不可实现的。

可用计算机实现合法性判定的语言应该如下定义：

如果对于一种语言，能编写一个程序在有限步骤内完成上述判断，则该语言称为是可递归的。一种语言是可递归枚举的，却不一定是可递归的。

可见，自然语言句法分析的任务只能在可递归语言上实现，因此，相应的语法描述也应该是可递归的。

第二，如果用列举所有句子的方法作为语法描述，那么这种描述是无助于对新句子进行结构分析的，而只能实现新句子的合法性识别。其方法是通过把新句子与该语法所列举出的句子进行匹配来判定新句子是否来自于这些合法句子的集合中，即是否是一个合法的句子。除此之外，不能得出关于句子结构的进一步信息。

从上述的两点可以看出，用列举句子的方法作为语法描述难以完成对句子结构进行分析的任务。

另外，用列举句子的方法作为语法描述对于解释人类语言构造的方式没有任何帮助，好的语法应该具有推广能力，能够抓住语言现象中的共同点，这对于理解语言、发掘语言运用的认知原理，进而发掘人类思维的本质都具有重要意义。更具现实意义的是，从计算资源的角度来看，具有推广能力的语法更能节约存储空间。

从上面的分析可以看到改进语法描述的某些端倪：它描述的应该是可递归语言，并且

它描述的句子应该是有内部结构的，而且这种内部结构是具有共性的，因而这种语法是有推广能力的。

一个让语法具有推广能力的方法是首先建立一些语法范畴（也常被称为语法类别、词性等），把具有相似语法行为的词归入一个相同的语法类别；然后，描述这些语法类别如何进一步组合的语法行为，这样构造的任何一个语法行为对于具有相似语法范畴的词都具有推广能力。

这时，语法描述就是列出语法类别的所有可能的组合模式。例如：（本章使用几种常用的语法类别，包括 ART（冠词）、N（名词）、V（动词）、ADJ（形容词）、ADV（副词）和 PRON（代词）等。）

$$\mathrm{ART} + \mathrm{N}$$

$$\mathrm{ART} + \mathrm{N} + \mathrm{V}$$

$$\mathrm{ART} + \mathrm{ADJ} + \mathrm{N} + \mathrm{V}$$

就是几个在英语中允许的语言模式，这种以语法类别为单元的模式就比以词本身为单元的列举具有更强的推广能力。例如，上述的

$$\mathrm{ART} + \mathrm{N}$$

模式就可以用来描述诸如：a book，the sentence 等等很多个词串。

但是，如果模式有限，每个语法类别中的词有限，则这样的语法可以生成的句子是有限的。然而通过引入几个记号可以大大扩展上述模式的描述能力。

（1）Kleene 星，记为 *，例如：

$$\mathrm{ART} + \mathrm{ADJ} + \mathrm{ADJ}^{*} + \mathrm{N} \tag{1-1-1}$$

* 号出现在 ADJ 的右上角，表示 ADJ 可以出现 0 次或 0 次以上。这样，可以描述在一个冠词和名词之间插有多个形容词的语言模式。

（2）Kleene 加，记为 +，例如：

$$\mathrm{ART} + \mathrm{ADJ}^{+} + \mathrm{N} \tag{1-1-2}$$

+ 号出现在 ADJ 的右上角，表示 ADJ 可以出现 1 次或 1 次以上。这个式子描述的内容与模式（1-1-1）是相同的。

（3）圆括号，记为（ ），例如：

$$\mathrm{ART} + (\mathrm{ADJ}) + \mathrm{N} \tag{1-1-3}$$

ADJ 外加一个圆括号表示 ADJ 可以出现 1 次，也可以 1 次也不出现。也就是说，ADJ 是可选的。

（4）垂直线，记为 |，例如：

$$\mathrm{N} \mid \mathrm{PRON} + \mathrm{V} \tag{1-1-4}$$

N 和 PRON 中间的直线表示可以是 N，也可以是 PRON，它们都可以与后面的 V 组成这个模式，但二者不能同时出现。

在引入了这几个记号后，基于有限个语法类别的组合模式就可以构造无限多个句子，比如，在模式

$$\mathrm{ART} + \mathrm{ADJ}^{+} + \mathrm{N} + \mathrm{V}$$

中，可以通过无限次重复出现 ADJ 而产生无限多个句子。

这样形成的上述语法描述称为正则表达式。与第一种用列举的方法来作为语法描述

相比，利用正则表达式来描述语法具有了一定的推广能力，并可以利用有限的语法类别的组合模式来生成无限的句子。但是，这种语法描述所表现出的推广能力还是远远不够的，还有进一步改进的余地。例如，上述的模式（1-1-1）和模式（1-1-3）通常会在句子中处于类似的位置，起着类似的作用，因此，应该可以进入更高一层次的抽象。这可以通过简单地定义一个比语法类别具有更高抽象的概念——短语来实现。简单地说，短语是经常反复出现的符号串，它构成确定的语法成分。例如，上述的两个字符串：模式(1-1-1)和模式(1-1-3)可以进一步用一个短语名称 NP（名词短语）来概括。

在引入短语概念的基础上，可以进一步扩展正则表达式的描述能力。如果在正则表达式中，除了可以包含原有的语法类别，还可以包含短语，就可以形成一种描述语言的语法——短语结构语法。

1.2 短语结构语法

为描述短语结构语法，需要先介绍重写规则。重写规则是一种形式化表示方式，可以用来描述规则，例如：

$$S \rightarrow NP\ VP$$

就是一个重写规则。其中，S 代表一个句子；NP，VP 表示两个短语，NP 表示一个名词短语，VP 表示一个动词短语。该规则的意思是说左边的符号 S 所代表的项可以被合乎语法地替换成右边符号所代表的两个项，即被重写为右边两项的组合。

一个形式语法可以包含若干条重写规则。通常一些重写规则的集合用 P 来表示。除此之外，组成一个完整的形式语法还有另外几个要素：其一是所谓终结符号集合，用 T 来表示，一个终结符号代表一个这样的项，它在此语法中不能再被重写为其他项的组合，通常是该形式语法所描述的语言中的词汇的语法类别（如 N，V 等等），或者就是该语言中使用的词汇（如英语中的单词 a，boy 等等）；其二是非终结符号集合，用 NT 来表示，一个非终结符号代表一个这样的项，它在此语法中可能再被重写为其他项的组合，如果上述终结符号指的是语言中的词汇本身，那么非终结符号也包括词的语法类别；其三是一个特殊的非终结符号 S，表示句子。因为句法分析针对的单位均为句子，因而 S 就十分重要，它通常是对句子进行语法分析的开始或结束符号。

这样，一个完整的用来描述一种语言的形式语法就可以表示为四元组 (T，NT，S ，P)，且 $T \cap NT = \phi$，即一个符号不能同时既是终结符号又是非终结符号。令 $V = T \cup NT$，V^* 表示由 V 中的符号所构成的全部符号串（包括空符号串 ϕ），而 V^+ 表示 V^* 中除 ϕ 之外的一切符号串的集合。P 中的每条规则形如：

$$a \rightarrow b$$

其中，$a \in V^+$，$b \in V^*$，且 $a \neq b$。

一个简单的例子：

$$NT = \{S, NP, VP, ART, N, V\}$$

$$T = \{the,\ a,\ boy,\ sees,\ cat,\ dirty\}$$

T 中的符号串均为英语单词。

P 中包含如下几条重写规则：

$$S \rightarrow NP\ VP \tag{1-2-1}$$

$$NP \rightarrow ART\ N \tag{1-2-2}$$

$$NP \rightarrow ART\ ADJ\ N \tag{1-2-3}$$

$$VP \rightarrow V\ NP \tag{1-2-4}$$

$$ART \rightarrow the \mid a \tag{1-2-5}$$

$$N \rightarrow boy \mid cat \tag{1-2-6}$$

$$V \rightarrow saw \tag{1-2-7}$$

$$ADJ \rightarrow dirty \tag{1-2-8}$$

其中，(1-2-5) ~ (1-2-8)表明非终结符号的所属语法类别，四元组 (S,NT,T,P) 就表示了一个语法。

利用上述语法，不仅可以进行句子的合法性识别，还可以对一些句子进行结构分析。下面对这种合乎语法的可识别性以及结构分析的内容进行定义。

导出：某个句子被称为由一个语法导出的(一个语法导出了某个句子)，如果能由 S 开始依据语法中的一系列重写规则重写出该句子。如果一个句子能由某个语法导出，则称这个句子是合乎该语法的。

看看下面的句子：

$$\text{The boy saw a cat.} \tag{1-2-9}$$

是否合乎上述语法，如果合乎语法，其结构又是怎样的。

(1) 由规则(1-2-1)，合乎该语法的句子都是应该能被重写为一个名词短语加一个动词短语，因此，只需看句子(1-2-9)是否是由这两部分组成的。名词短语和动词短语是非终结符号，还可以进一步分解。

(2) 名词短语是应该能被重写为由一个冠词加一个名词组成的(规则(1-2-2))，或者由一个冠词加一个形容词再加一个名词组成的(规则(1-2-3))；而动词短语是应该能被重写为由一个动词加一个名词短语组成的(规则(1-2-4))。显然，句子(1-2-9)中 the boy 可以组成一个名词短语，see a cat 可以组成一个动词短语，而其中 a cat 又是一个名词短语。

(3) 由上述两条，可以看到句子(1-2-9)是能够依据一系列重写规则导出的，规则的使用次序可以如下：

规则(1-2-1)，(1-2-2)，(1-2-5)，(1-2-6)，(1-2-4)，(1-2-7)，(1-2-2)，(1-2-5)，(1-2-6)

同时，得到其组成结构：

(NP1 (The boy) VP1(saw NP2(a cat)))

括号由内向外，反映了句子的组成，The 和 boy 组成一个名词短语 NP1，a 和 cat 组成一个名词短语 NP2，saw 和 NP2 组成一个动词短语 VP1，NP1 和 VP1 组成句子 S。

显然，很多自然语言的句子在上述几条规则(语法)下是不合法的，即有很多自然语言的句子不能由上述语法导出。通过增加重写规则、增加(非)终结符号都可以增加语法能导出的句子的数量。一般地，我们称能生成更多个句子的语法具有更强的生成能力，显然，对语法具有较少约束的语法具有更强的生成能力。

按照上一节定义的可递归枚举语言与可递归语言来分，上述的一般的短语结构语法是可以描述可递归枚举语言的，即某些短语结构语法导出的语言是可递归枚举的而不是可递归的。

通过对一般的短语结构语法进行限制，可以得到被称为乔姆斯基体系的4类语法。下面，按照生成能力由弱到强（约束由多到少）的次序分别简单介绍。

1. 正则语法（3型语法）

正则语法分为左线性语法和右线性语法。

在左线性语法中，所有重写规则必须采用如下的形式：

$$A \rightarrow Bt \text{ 或 } A \rightarrow t$$

其中，A，B是非终结符号；t而为终结符号。

而在右线性语法中，所有重写规则必须采用如下的形式：

$$A \rightarrow tB \text{ 或 } A \rightarrow t$$

正则语法是乔姆斯基体系中生成能力最弱的一个，一些常见的语言现象都不能用正则语法来生成。一个简单的例子是任意符号“x”两边成对匹配添加括号，通过不断嵌套的方式可以实现一系列句子：

$$x,(x),((x)),(((x))),((((x)))),\cdots$$

为了生成这种语言的句子，当生成到“x”时必须知道前面已经生成了多少个“(”，以便能生成同样数量的“)”相匹配。而对于正则语法，无论是左线性语法还是右线性语法，都只能独立地生成“x”某一侧的符号，无法进行匹配。

在自然语言中，也存在着类似的匹配模式。例如：“如果A那么……”、“因为A所以……”等句子结构（其中A表示一个符号串），通常都需要匹配出现，这种模式也可以进行不断地嵌套形成复杂句子：

如果A那么……，如果……如果A那么……那么……，……

同样，当生成到A时，也必须知道前面已经生成了多少个“如果”，以便能生成同样数量的“那么”相匹配。

2. 上下文无关语法（2型语法）

在上下文无关语法中，每一条规则都采用如下的形式：

$$A \rightarrow x$$

其中，A是非终结符号，$x \in V^*$。这种规则的应用不依赖于A出现在什么上下文环境中，因此称为上下文无关语法。

上下文无关语法比正则语法具有更强的生成能力，能反映更多的自然语言现象。但是，还有一些自然语言现象并不能由上下文无关语法来描述，有些情况下，一条重写规则的应用是受上下文制约的。

3. 上下文有关语法（1型语法）

在上下文有关语法中，每一条重写规则都是这样的：

$$x \rightarrow y$$

其中，$x,y \in V^*$，且y的长度（即符号串y中的符号个数）总是大于或等于x的长度。

上下文有关语法的重写规则也可以这样来表示：

$$A \rightarrow y / x\_z$$

其中，A 是非终结符号；$y \in V^+$；$x, z \in V^*$，在这种表示中，可以很明显地看出所谓上下文有关的含义来：如果 A 出现在上下文 x_z 中，即前面紧挨着符号串 x，后面紧挨着符号串 z，则 A 可重写为 y，可以看到 A 可重写为 y 是有上下文约束的。

4. 无约束短语结构语法(0 型语法)

0 型语法对规则没有任何约束，其定义的语言可能不是递归的，因而就不可能设计一个程序来判别一个输入的符号串是否是 0 型语言中的一个句子，所以 0 型语言很少被用来处理自然语言。

总而言之，在乔姆斯基体系中，如果一种语言可以被一部 i（$i = 0, 1, 2, 3$）型语法所生成，就称它为 i 型语言。

由于在乔姆斯基体系中，语法的型号越高，对重写规则所附加的限制也越多，所以 3 型语言是 2 型语言的一个子集，2 型语言是 1 型语言的一个子集，依此类推，有：0 型语言 ⊇1 型语言 ⊇2 型语言 ⊇3 型语言。从语法的生成能力看，0 型语言最强，1 型到 3 型依次递减，3 型最弱。

在上述乔姆斯基体系的四种语法中，上下文无关语法是计算语言学的重要研究对象。由于其描述能力强，足以描述自然语言中的大部分结构，同时又是可递归的，可以构造有效的句法分析器来进行句子的分析，因此，目前大多数计算机处理用的语法都是基于上下文无关语法的。

1.3 转移网络

除了用重写规则来描述语法之外，转移网络也是一种方式。

转移网络是一个图，图由结点集合和边集合组成，每条边都是带标记的。结点集合中有一个结点是初始状态或称开始状态，还有一个或多个终止状态。

有限状态转移网络是一种较为简单的转移网络。上述乔姆斯基体系中的正则语法可以用有限状态转移网络来等价地描述。图 1-1 是一个有限状态转移网络。

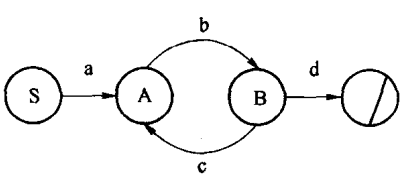


图 1-1 有限状态网络

图中标为 S 的结点是句子分析的起始点，从结点 S 有一条指向结点 A 的有向弧，弧上有一个语法类别标记，意为句子的第一个词如果具有语法类别 a，则 a 可以转移到结点 A；然后如果第二个词具有语法类别 b，则可转移到结点 B；在结点 B 时，有两个有向弧分别指向两个结点，其中标有 d 的有向弧指向的是终止结点(终止结点中标有一斜划线)。如果一个句子能沿着有向弧最终达到终止结点，同时句子也到最后的话，就可以断定该句子是符合该有限状态转移网络所描述的语法。

显然，上述转移网络能用来等价地描述如下的一部正则语法：

S→aA

A→bB

B→cA

B→d

因此，正则语法也可称之为有限状态语法。

对有限状态转移网络进行扩展，可以建立具有更强描述能力的递归转移网络。递归转移网络的描述能力与上述乔姆斯基体系中的上下文无关语法等价。例如，如下一个上下文无关语法：

S→NP VP

NP→ART ADJ* N (PP)

VP→V (NP)

PP→PREP (NP)

可等价地用如图 1-2 中的递归转移网络来描述(PP 表示介词短语)，图中每一个单独的网络与上述的一个规则相对应。

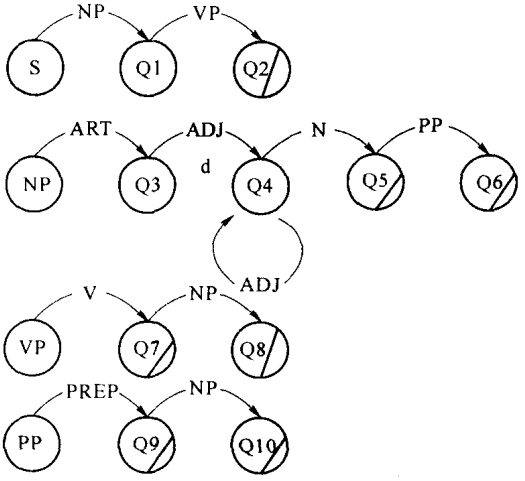


图 1-2　递归转移网络

可以看到，在递归转移网络中，有向弧上的标记不仅可以是某个语法类别，还可以是某个短语结构，这是递归转移网络与有限状态转移网络的一个重要区别所在，这使得递归转移网络能支持递归结构。例如：在图 1-2 的四个网络所构成的递归网络描述中，主网络是起始结点为 S 的网络，一旦在这个网络中沿着有向弧能到达终止结点，则可判定相应的句串是合乎该网络所描述的语法的。而为了从 S 结点转移到终止结点 Q2，首先要转移到 Q1 结点。而为了转移到 Q1 结点，首先要判定句子前面是否有一个有向弧上标的结构：NP。而为了判定句子中的 NP 结构，需要调用第二个网络，即 NP 网络。在 NP 网络中有两个终止结点，如果在第二个终止结点终止的话，将意味着在 N 后面有一个 PP 结构。为判定 N 后面是否有一个 PP 结构，需要调用第四个网络，即 PP 网络。在 PP 网络中若要达到终止结点，首先要找到一个 PREP，而后要判断后面是否有一个 NP 结构，这样反过来又需要调用第二个 NP 网络，如此形成递归调用。

1.4　短语结构与句法树

自然语言，无论是语音形式还是文本形式都表现为线性的，语音在时间上顺序出现，文本在空间上顺序出现。例如，下面以文本形式出现的句子：