

第3章 传统自然语言 处理分析



目录

- 一、形式文法-乔姆斯基文法体系
- 二、句法分析方法
- 三、词法分析方法
- 四、语义分析与计算



一、形式文法-乔姆斯基文法体系



3.1 形式文法-Chomsky语法体系

如果一个语言的词汇集是一个有限集 V ，对 V 中的元素毗连计算可以得到符号串集合 V^* ，那 V^* 就是由 V 构成的句子。

而通过毗连计算得到的字符串如果一个语言的词汇集是一并不一定都是某种语言中的句子。例如，“the man saw the ball”（“人看球”）在英语中是正确的，而由同样符号构成的“the saw the man ball”在英语中却是不正确的。我们把前者叫作成立的句子，后者叫作不成立的句子，而要区别一种语言中的成立的句子和



3.1 形式文法-Chomsky文法体系

不成立的句子，就必须采用某些办法把语言刻画出来，从而说明，在这一种语言中，什么样的句子是成立的什么样的句子是不成立的。我们可以采用三种办法来刻画语言。

(1) 穷举法 — 把语言中全部成立的句子穷尽枚举出来。只适合句子数目有限的语言。

(2) 语法描述 — 制定有限数量的规则来生成语言中无限个数的句子，这些句子是语言中合格的句子。这种能够刻画语言的有限个数的规则称为文法。记为G。



3.1 形式文法-Chomsky文法体系

(3) 语言识别程序自动机 — 设计一种装置来检验输入符号串，来识别该符号串是不是语言L中成立的句子，如果是，这个装置就接收，如果不是语言中成立的句子，这个装置就不接收。

由此可见，刻画某类语言的有效手段是文法和自动机，文法用于生成语言，而自动机则用于识别语言。

美国著名语言学家乔姆斯基（N.Chomsky）将文法抽象成一个四元组，称为短语结构文法或短语结构语法。

3.1 形式文法-Chomsky文法体系

3.1.1 短语结构语法理论

一种语言就是一个句子集，它包含了属于这种语言的全部句子，而语法是对这些句子的一种有限的形式化描述。可以利用一种基于产生式的形式化工具对某种语言的语法进行描述。

一部短语结构语法G可以用一个四元组来定义：

$$G = (V_t, V_n, P, S)$$

V_t: 终结符集合，终结符是指被定义的那个语言的词或符号；

V_n: 非终结符的集合，这些符号不能出现在最终生成的句子中，是专门用来描述语法的。V_t和V_n的并(\cup)构成了符号集

3.1 形式文法-Chomsky语法体系

V , 称为总词汇表, 且 V_t 和 V_n 不相交, 因此有: $V = V_t \cup V_n$,
 $V_t \cap V_n = \varnothing$ (\varnothing 表示空集);

P: 有穷产生式集: $\alpha \rightarrow \beta$

式中 $\alpha \in V^* V_n V^*$, $\beta \in V^*$, $*$ 表示它前面的字符可以出现任意次;

S: 非终结符表 V_n 的一个元素, 称为起始符。

下面就是采用短语结构语法对一个英语子集 (受限英语) 的语法的描述:

$$G = (V_t, V_n, P, S)$$

$$V_n = \{S, NP, VP, Det, N, V, Prep, PP\}$$

$$V_t = \{the, girl, letter, pencil, write, with, a\}$$

3.1 形式文法-Chomsky语法体系

S=S

P: S → NP VP

NP → Det N

VP → V NP

VP → VP PP

PP → Prep NP

Det → the | a

N → girl | letter | pencil

V → write

Prep → with

这一语法所描述的英语子集中，只有the、girl、Letter、pencil、write、with、a几个单词。

3.1 形式文法-Chomsky文法体系

形式文法的直观意义

形式文法是用来精确地描述语言（包括人工语言和自然语言）及其结构的手段。形式语言学 也称 代数语言学。

以重写规则 $\alpha \rightarrow \beta$ 的形式表示，其中， α ， β 均为字符串。顾名思义：字符串 α 可以被改写成 β 。一个初步的字符串通过不断地运用重写规则，就可以得到另一个字符串。通过选择不同的规则并以不同的顺序来运用这些规则，就可以得到不同的新字符串。

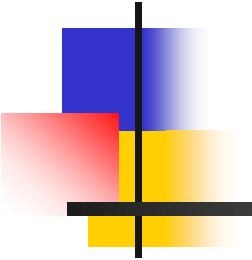
3.1 形式文法-Chomsky文法体系

3.1.2 约束的短语结构语法——乔姆斯基语法体系

短语结构语法是用于描述语言特性的一种形式体系。对于一种形式体系，如果它能定义的语言类型越多，就说它的描述能力越强。例如，如果形式体系T1可以定义5种语言，而形式体系T2可以定义10种语言，而且包含了所有可以被T1定义的5种语言，就说T2比T1具有更强的描述能力。

理论语言学家将语言分成两类：**递归语言**和**可递归枚举语言**。

对于一种语言，若能编写一部程序，使其能以某种顺序逐个地输出该语言的全部句子，就称该语言是**可递归枚举的(生成)**；



如果能编一部程序，使其能在读入一个符号串后，可以判断该符号串是否为该语言的句子，就称该语言是递归的（可识别）。

一种语言可以是可递归枚举的，但却不一定是递归的，因为对给定的一个符号串，可能无法判断它是否是该语言的一个句子。

乔姆斯基语法体系是一组受限的短语结构语法，降低它的描述能力。他定义了四种语法：0型语法、1型语法、2型语法和3型语法。

3.1 形式文法-Chomsky文法体系

0型语法：是一种无约束的短语结构语法，也就是前面已经介绍的短语结构语法。

1型语法：也称做上下文有关语法，是一种满足下列约束条件的短语结构语法：

对于每一条形式为

$$x \rightarrow y$$

的产生式，符号串 y 中所包含的字符个数不少于字符串 x 中所包含的字符个数，而且 $x, y \in V^*$ 。

3.1 形式文法-Chomsky文法体系

2型语法：也称做上下文无关语法，是一种满足下列约束条件的短语结构语法：

对于每一条形式为

$$A \rightarrow x$$

的产生式，其左侧必须是一个单独的非终结符，而右侧则是任意的符号串，即 $A \in V_n, x \in V^*$ 。在这种语法中，由于产生式规则的应用不依赖于符号A所处的上下文，因此，称为上下文无关语法。

3.1 形式文法-Chomsky文法体系

3型语法：也称做正则语法，分左线性语法和右线性语法两种形式。在左线性语法中，每一条产生式的形式为

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

而在右线性语法中，每一条产生式的形式为

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

这里，A和B都是单独的非终结符，t是单独的终结符，即A, $B \in V_n$, $t \in V_t$ 。

在这四种语法中，型号越高，所受到的约束就越多，其生成语言的能力就越弱，因而生成的语言集就越小，也更易于对其生成的语言进行计算机自动分析。

3.1 形式文法-Chomsky语法体系

3.1.3 句法分析树

在对一个句子进行分析的过程中，如果把分析句子各成份间关系的推导过程用树形图表示出来的话，那么，这种图称做句法分析树。

图3.1就是依据上述定义的语法对语句
The girl writes the letter with a pencil
进行句法分析时建立的句法分析树。

在句法分析树中，起始符总是出现在树的根上，终结符则出现在树的叶子上。

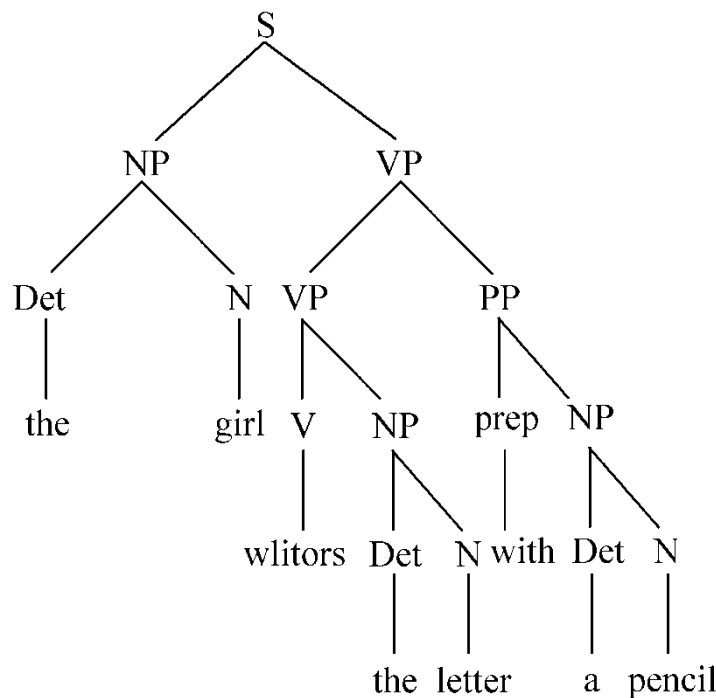


图3.1 句法分析树示例





二、句法分析方法

3.2 基于短语结构的句法分析

基于短语结构语法的自动分析算法主要有自顶向下回溯算法、自底向上并行算法、富田算法、左角分析算法和CYK算法等等。今天我们先介绍自顶向下回溯算法和自底向上并行算法。

3.2.1 自顶向下回溯算法

自顶向下分析算法的思想就是从起始符开始向着被分析的句子进行推导，推导过程的语法树建立从根节点开始，自上而下进行。每次推导只选择一种路径进行尝试，并保留其他可选择的路径，当推导失败时，进行回溯，尝试另一种推导路径。

3.2 基于短语结构的句法分析

例如，我们定义下面的一个语法：

$G=(V_t, V_n, P, S)$

$V_n=\{S, NP, VP, Det, N, V, Prep, PP\}$

$V_t=\{the, girl, letter, pencil, writes, with, a\}$

$S=S$

$P: S \rightarrow NP VP \quad (a)$

$NP \rightarrow Det N \quad (b)$

$VP \rightarrow V NP \quad (c)$

$VP \rightarrow VP PP \quad (d)$

$PP \rightarrow Prep NP \quad (e)$

$Det \rightarrow the \mid a \quad (f)$

$N \rightarrow girl \mid letter \mid pencil \quad (g)$

$V \rightarrow writes \quad (h)$

$Prep \rightarrow with \quad (i)$

3.2 基于短语结构的句法分析

应用定义的语法对句子 “the girl writes the letter with a pencil” 的分析过程。

搜索步骤	搜索对象	所使用的规则	输入句子中遗留部分
(1)	S	(a)	the girl writes the letter with a pencil
(2)	NP VP	(b)	the girl writes the letter with a pencil
(3)	Det N VP	(f)	the girl writes the letter with a pencil
(4)	the N VP		the girl writes the letter with a pencil
(5)	N VP	(g)	girl writes the letter with a pencil
(6)	girl VP		girl writes the letter with a pencil
(7)	VP	(c)	writes the letter with a pencil
(8)	V NP	(h)	writes the letter with a pencil
(9)	writes NP		writes the letter with a pencil
(10)	NP	(b)	the letter with a pencil
(11)	Det N	(f)	the letter with a pencil
(12)	the N		the letter with a pencil
(13)	N	(g)	letter with a pencil
(14)	letter		letter with a pencil
(15)			with a pencil

3.2 基于短语结构的句法分析

这时，句子中还有遗留部分，但搜索对象中却已变空，分析过程已无法继续，只得回溯。回溯到第（7）步，看看是否还能利用别的规则进行分析。

(7')	VP	(d)	writes the letter with a pencil
(16)	VP PP	(c)	writes the letter with a pencil
(17)	V NP PP	(h)	writes the letter with a pencil
(18)	writes NP PP		writes the letter with a pencil
(19)	NP PP	(b)	the letter with a pencil
(20)	Det N PP	(f)	the letter with a pencil
(21)	the N PP		the letter with a pencil
(22)	N PP	(g)	letter with a pencil
(23)	letter PP		letter with a pencil
(24)	PP	(e)	with a pencil
(25)	Prep NP	(i)	with a pencil
(26)	with NP		with a pencil

3.2 基于短语结构的句法分析

(27)	NP	(b)	a pencil
(28)	Det N	(f)	a pencil
(29)	a N		a pencil
(30)	N	(g)	pencil
(31)	pencil		pencil
(32)	NIL		NIL

在应用规则(f)和(g)对搜索对象进行替换时，由于规则的右边有多个单词可供选择，这时，可根据句子遗留部分的第一个单词确定。

3.2 基于短语结构的句法分析

3.2.2 自底向上并行算法

自底向上分析算法是从输入句子的句首开始依次取词向前移进，并应用合适的语法规则逐级向上归约（产生式倒过来用），直到构造出表示句子结构的整个推导树为止。换句话说，句法树的建立从树底部的叶节点（即词和词类）开始，直到根部。

本算法实际上分**移进**、**归约**两个步骤。所谓**移进**，就是把一个尚未处理过的符号移入栈顶，并等待更多的信息到来之后再决定；所谓**归约**，就是对栈顶的那些与某一语法规则右边相匹配的符号，用该语法规则左边的符号来取代。

3.2 基于短语结构的句法分析

在移进-归约过程中，可能会出现有多条语法规则符合归结条件，这种情况称为“归约-归约”冲突；

也可能出现既符合移进条件又符合归约条件的情况，在这种情况下是移进还是归约呢？这称做“移进-归约”冲突。

解决这两种冲突是移进-归约算法的中心问题。

下面以对句子“**the girl writes the letter with a pencil**”的分析为例，说明采用移进-归约算法进行自底向上分析的过程。

3.2 基于短语结构的句法分析

步骤	栈	操作	输入句子中的遗留部分
(1)			the girl writes the letter with a pencil
(2)	the	移进	girl writes the letter with a pencil
(3)	Det	用规则(f)归约	girl writes the letter with a pencil
(4)	Det girl	移进	writes the letter with a pencil
(5)	Det N	用规则(g)归约	writes the letter with a pencil
(6)	NP	用规则(b)归约	writes the letter with a pencil
(7)	NP writes	移进	the letter with a pencil
(8)	NP V	用规则(h)归约	the letter with a pencil
(9)	NP V the	移进	letter with a pencil
(10)	NP V Det	用规则(f)归约	letter with a pencil
(11)	NP V Det letter	移进	with a pencil
(12)	NP V Det N	用规则(g)归约	with a pencil

3.2 基于短语结构的句法分析

- | | | |
|------------------------|----------|---------------|
| (13) NP V NP | 用规则(b)归约 | with a pencil |
| (14) NP VP | 用规则(c)归约 | with a pencil |
| (15) S | 用规则(a)归约 | with a pencil |
| (16) S with | 移进 | a pencil |
| (17) S Prep | 用规则(i)归约 | a pencil |
| (18) S Prep a | 移进 | pencil |
| (19) S Prep Det | 用规则(f)归约 | pencil |
| (20) S Prep Det pencil | 移进 | |
| (21) S Prep Det N | 用规则(g)归约 | |
| (22) S Prep NP | 用规则(b)归约 | |
| (23) S PP | 用规则(e)归约 | |

3.2 基于短语结构的句法分析

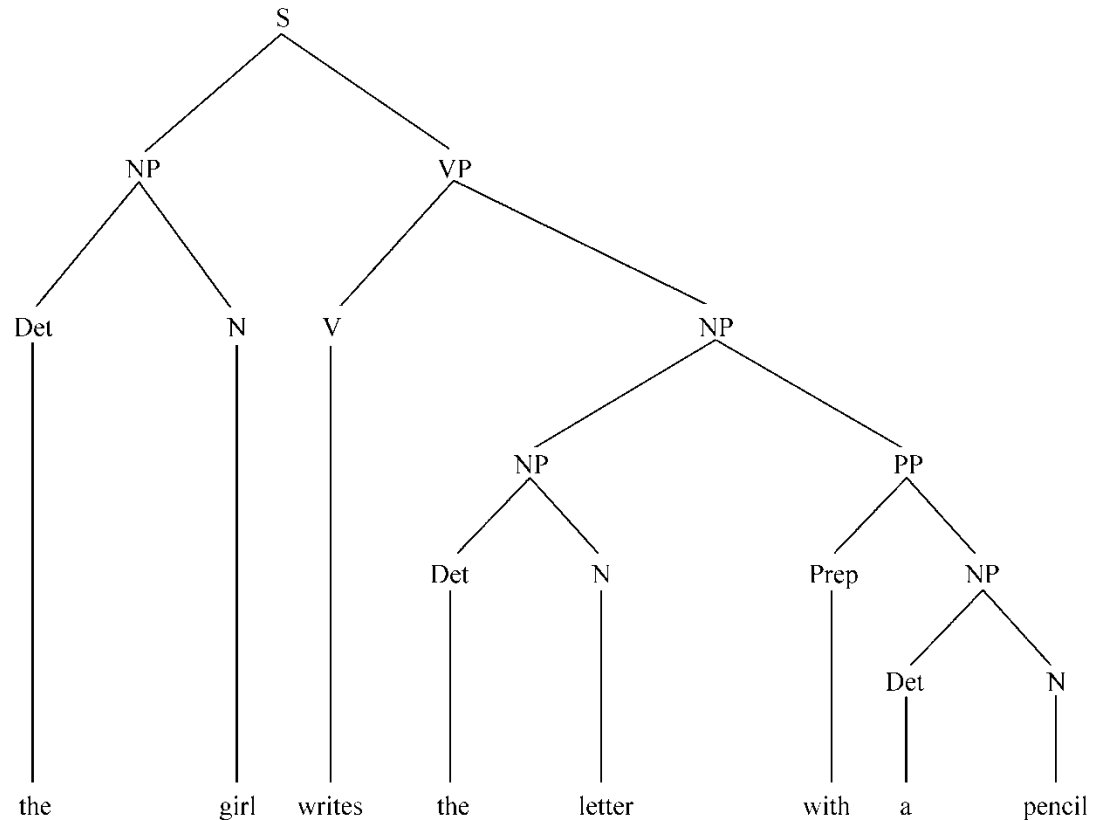
这时，输入句子串已空，但栈中却不是只有起始符S，况且语法中已无合适的规则可用来归约。进行回溯，返回到(14)，在这一步，先不采用规则(a)对其进行归约，而是移进下一个单词with，再使用规则(i)归约。

(14') NP VP	回溯	with a pencil
(24) NP VP with	移进	a pencil
(25) NP VP Prep	用规则(i)归约	a pencil
(26) NP VP Prep a	移进	pencil
(27) NP VP Prep Det	用规则(f)归约	pencil
(28) NP VP Prep Det pencil	移进	
(29) NP VP Prep Det N	用规则(g)归约	
(30) NP VP Prep NP	用规则(b)归约	
(31) NP VP PP	用规则(e)归约	
(32) NP VP	用规则(d)归约	
(33) S	用规则(a)归约	

这时，输入句子串已空，且栈中只剩下起始符S，该句子被接受，分析成功。

3.2 基于短语结构的句法分析

利用自底向上分析算法对句子 “the girl writes the letter with a pencil” 的分析句法树如下图所示



3.3 递归转移网络与扩充转移网络

扩充转移网络 (Augmented Transition NetWorks, ATN) 属于一种增强的上下文无关语法, 其基本思想是采用上下文无关语法来描写句子的成分结构, 但对语法中的个别产生式增添了某些功能, 主要是描写某些必要的语法限制, 并建立句子的深层结构。

ATN是在递归转移网络 (Recursive Transition NetWorks, RTN) 上附加若干控制条件所形成的网络, 而递归转移网络又是扩展的有限状态转移图 (Transition NetWorks, TN) 。所以本节先介绍有限状态转移网络和递归转移网络, 最后再介绍扩充转移网络。

3.3 递归转移网络与扩充转移网络

3.3.1 有限状态转移网络

有限状态转移网络（TN）只能用来生成和识别正则语言。

一个有限状态转移网络由一组状态（即结点）和一组弧组成：

(1) 其中的一个状态被指定为起始状态。

(2) 在每条弧上都标注着该语法的终结符（词或词类）。表明在句子分析和识别时状态转移的条件和转移的方向。必须在输入句子中找到符合该弧上标注的词，才可以进行这条弧所规定的转移。

(3) 状态集中有一个名为结束状态的子集。如果输入句子的头从起始状态开始，经过一系列的转移，句尾恰好到达结束状态，就说这个句子被这个转移网络所接受（或识别）。

3.3 递归转移网络与扩充转移网络

TN的工作过程为：输入某一个句子（句子定义为终结符连接成的串），从起始状态出发，按有限状态转移网络中箭头所指方向，依次扫描输入词，观察所输入词与相应状态弧上的标记是否匹配，匹配的话即通过该弧，进入下一个状态。如果扫描到句子的终点，有限状态转移网络也进入了结束状态，就说这个句子被这个转移网络所接受（或识别）。

例1. 用转移网络来识别句子The small black ducks swallow flies的过程如表1.1（这里忽略了词法分析），转移网络如图3.1所示。

3.3 递归转移网络与扩充转移网络

词典

ducks	<u>noun,verb</u> (躲避、低下头、弯下腰)
<u>flies</u>	<u>noun,verb</u>
<u>small</u>	adj.
<u>black</u>	<u>adj.,noun</u>
<u>swallow</u>	<u>noun,verb</u>
<u>the</u>	det.

表 3.1 句子识别过程

词	当前状态	弧	新状态
the	a	a $\xrightarrow{\text{det}}$ b	b
small	b	b $\xrightarrow{\text{adj.}}$ <u>b</u>	b
black	b	b $\xrightarrow{\text{adj.}}$ <u>b</u>	b
ducks	b	b $\xrightarrow{\text{noun}}$ c	c
swallow	c	c $\xrightarrow{\text{verb}}$ e	e
flies	e	e $\xrightarrow{\text{noun}}$ f	f(识别)

3.3 递归转移网络与扩充转移网络

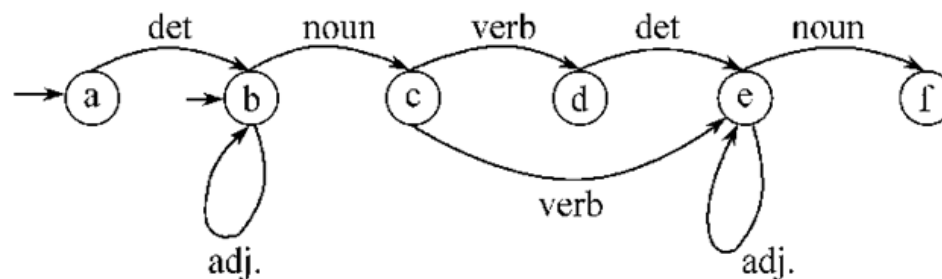


图 3.1 转移网络实例

识别过程到达 f 状态（终态），所以该句子被成功地识别了。分析结果如图 3.2 所示。

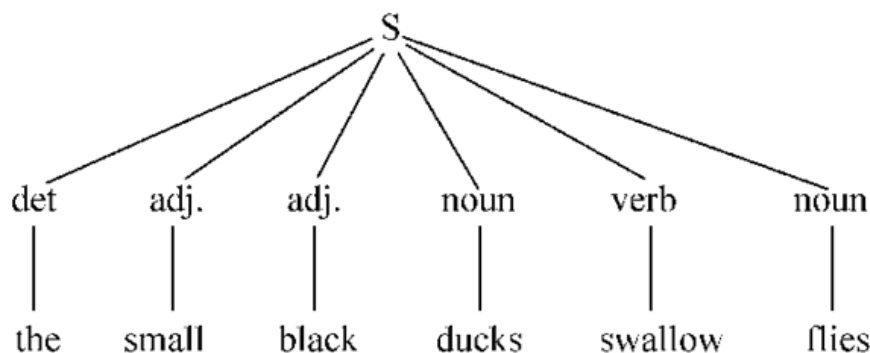


图 3.2 TN 分析树

3.3 递归转移网络与扩充转移网络

从上可以看出，这个句子还可以在网络中走其他弧，如词ducks也可以走弧c d，但接下来的swallow就找不到合适的弧了，对应于这个路径，该句子就被拒识了。由此看出，网络识别的过程中应找出各种可能的路径，因此算法要采用并行或回溯机制。

并行算法。并行算法的关键是在任何一个状态都要选择所有可以到达下一个状态的弧，同时进行试验。

回溯算法。在所有可以通过的弧中选一条往下走，并保留其他的可能性，以便必要时回过来选择之。这种方法需要一个堆栈结构。

TN只能识别正则语言，实际上任何一个有限状态转移网络都对应一部正则语法。所以，用有限状态转移网络表达自然语言是远远不够的。为提高TN的识别能力，提出了递归转移网络RTN。

3.3 递归转移网络与扩充转移网络

3.3.2 递归转移网络

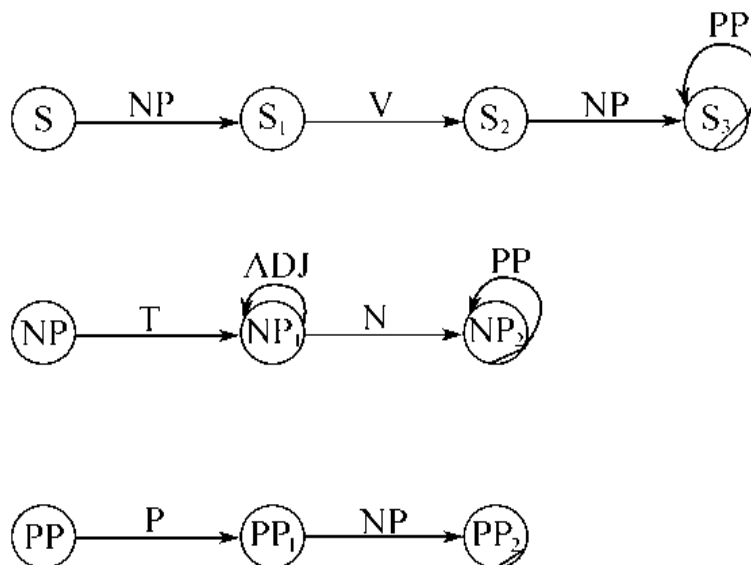
递归转移网络 (Recursive Transition Networks, RTN) 是对有限状态转移网络 (TN) 的一种扩展, 在RTN中每条弧的标注不仅可以是一个终结符 (词或词类) 而且可以是一个用来指明另一个网络名字的非终结符。

例3.2下面是一部上下文无关语法:

$$S \rightarrow NP \ V \ NP \ PP^*$$
$$NP \rightarrow T \ ADJ^* \ N \ PP^*$$
$$PP \rightarrow P \ NP$$

其中 X^* 表示符号 X 可以出现零次或多次。这三条语法规则可以图3.3所示的递归网络来表示。

3.3 递归转移网络与扩充转移网络



在递归转移网络中，任何一个子网络都可以调用包括它自己在内的任何其他子网络。在图3.3中，表示名词短语NP的子网络中包含了介词短语PP，而在表示PP的子网络中又包括了NP。这种在NP的定义中包含了NP自身的定义叫做递归定义。相应的状态转移网络叫做递归转移网络。

3.3 递归转移网络与扩充转移网络

从生成能力上看，递归转移网络等价于上下文无关语法。但是要用它来分析自然语言，还必须在功能上予以增强，以便它可以描写各式各样的语法限制以及在识别过程中同时构造输入句子的句法结构。经过增强的递归转移网络就是下面要介绍的**扩充转移网络**。

3.3 递归转移网络与扩充转移网络

3.3.3 扩充转移网络

扩充转移网络(Augmented Transition Networks,简称 ATN)是由一组网络构成的递归转移网络,每个网络都有一个网络名,它在以下三个方面对RTN进行了扩充:

(1) 增加了一组寄存器,用以存储分析过程中得到的中间结果和有关信息。

(2) 每条某些弧上除了用句法范畴(如词类和短语标记)来标注外,可以附加任意的测试,只有当弧上的这种测试成功之后才能通过这条弧。

(3) 每条弧上还可以附加操作,当通过一条弧时,相应的动作便被依次执行,这些动作主要用来设置或修改寄存器的内容。

3.3 递归转移网络与扩充转移网络

ATN的每个寄存器由两部分构成：句法特征句法功能寄存器。

(1) 特征寄存器中，包含着许多维的特征，每一维特征都由一个特征名和一组特征值以及一个缺省值来表示。例如：

“数”：单数，复数。缺省:空）。

可以使用一维特征值来表示英语中动词的各种形式。

例如，对动词Work，可以使用下面的一维特征值来表示它的各种形式：

Work: present, past, present-participle, past-participle.

Default: present.

这里work就是特征名， present, past, present-participle等则是它的一组特征值。

3.3 递归转移网络与扩充转移网络

(2) 功能寄存器则反映了句法成分之间的关系和功能。

分析树的每个结点都有一个寄存器，寄存器的上半部分是特征寄存器，下半部分是功能寄存器。

图3.4所示是一个简单的名词短语（NP）的扩充转移网络，网络中弧上的条件和操作如下：

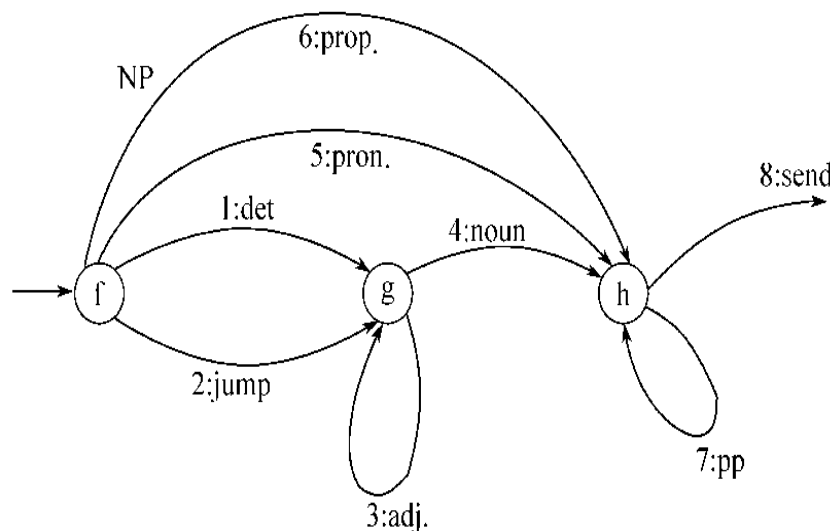


图3.4 名词短语（NP）的扩充转移网络

3.3 递归转移网络与扩充转移网络

NP-1: $\underline{f} \xrightarrow{\text{det}} g \quad \downarrow$

A: Number \leftarrow *.Number \downarrow

NP-4: $g \xrightarrow{\text{Noun}} h \quad \downarrow$

C: Number = *.Number or $\phi \quad \downarrow$

A: Number \leftarrow *.Number \downarrow

NP-5: $f \xrightarrow{\text{pronoun}} h \quad \downarrow$

A: Number \leftarrow *.Number \downarrow

NP-6: $f \xrightarrow{\text{proper}} h \quad \downarrow$

A: Number \leftarrow *.Number \downarrow

3.3 递归转移网络与扩充转移网络

上面的网络主要是用来检查NP中的数的一致值问题。

其中用到的特征是Number(数)，它有两个值：Singular(单数)和plural(复数)，缺省值是 \varnothing （空）。

C是弧上的条件，A是弧上的操作，*是当前词，proper是专用名词，Det是限定词，PP是介词短语，*.Number表示当前词的值。

NP是该扩充转移网络的网络名。网络NP可以是其他网络的子网络，也可包含其他网络，如其中的PP就是一个子网络，这就是网络的递归性。

3.3 递归转移网络与扩充转移网络

弧NP-1将当前词的Number放入当前NP的Number中，而弧NP-4则要求当前noun的Number与NP的Number是相同时，或者NP的Number为空时，将noun作为NP的Number，这就要求det的数和noun的数是一致的。因此，this book, the book, the books, these books 都可顺利通过这一网络，但是this books 或these book就无法通过。如果当前NP是一个代词（Pron.）或者专有名词（Proper），那么网络就从NP-5或NP-6通过，这时NP的数就是代词或专用名词的数。PP是一个修饰前面名词的介词短语，一旦到达PP弧就马上转入子网络PP。

3.3 递归转移网络与扩充转移网络

ATN方法是一个比较复杂的方法，尽管在自然语言理解的研究中得到了广泛的应用，但在实现过程中，还有许多问题，如非确定性分析、弧的顺序、非直接支配关系的处理等需要进一步的研究。

3.4 词汇功能语法

词汇功能语法是由 **J.Bresnan** 和 **R.M.Kaplan** 在 **1982**年提出的，它是一种功能语法，但是更加强调词汇的作用。上面介绍的扩充转移网络（**ATN**语法）是有方向性的，也就是说，**ATN**语法的条件和操作要求语法的使用是有方向的，因为只有在寄存器被设置过之后才可被访问。而词汇功能语法（**LFG**）试图通过互不矛盾的多层描述来消除这种有序性限制，它利用一种结构来表达特征、功能、词汇和成份的顺序。

3.4 词汇功能语法

在**LFG**中，对句子的描述包括两部分：一个直接成分结构（**C-structure**）和一个功能结构（**F-structure**）。直接成分结构（**C-structure**）是由上下文无关语法产生的，用来描述表层句子的层次结构。功能结构（**F-structure**）则是通过附加到语法规则和词条定义上的功能方程来生成，其作用是表示句子的结构功能。

LFG采用了两种规则，一种是带有功能方程式的上下文无关语法规则，一种是词汇规则。表8.2给出了词汇功能语法（**LFG**）的语法规则，是带有功能方程式的上下文无关文法。

3.4 词汇功能语法

表 7.2 LFG 的语法规则

-
- (1) $S \rightarrow NP \quad VP$
 $(\uparrow \text{ Subject}) = \downarrow \quad \uparrow = \downarrow$
- (2) $NP \rightarrow \text{Determiner Noun}$
- (3) $VP \rightarrow \text{Verb} \quad NP \quad NP$
 $\uparrow = \downarrow \quad (\uparrow \text{ Object}) = \downarrow \quad (\uparrow \text{ Object2}) = \downarrow$
-

3.4 词汇功能语法

其中符号 \uparrow 和 \downarrow 称作元变量。 \uparrow 表示当前成分的上一层次的直接成分，如规则中NP的 \uparrow 就是S，VP的 \uparrow 也是S； \downarrow 则表示当前成分。因此，规则（1）中的第一个方程式 $(\uparrow\text{Subject})=\downarrow$ 就可解释为把NP的属性传递给S的Subject特征。第二个方程式 $\uparrow=\downarrow$ 表示将VP的所有属性传递给它的上一层成分S。

LFG的分析还依赖于句子中的词汇，词汇也带有功能方程式。例如，表8.3就是给出了一些词汇的LFG规则。

3.4 词汇功能语法

表 7.3 LFG 的词汇规则

handed	Verb	(↑ Tense)=Past (↑ Predicate) = ‘Hand<(↑ Subject),(↑ Object),(↑ Object2)>’
girl	Noun	(↑ Number)= Singular (↑ Predicate)=‘Girl’
baby	Noun	(↑ Number)= Singular (↑ Predicate)=‘Baby’
toys	Noun	(↑ Number)=Plural (↑ Predicate)= ‘Toy’
the	Determiner	(↑ Definiteness)=Definite
A	Determiner	(↑ Definiteness)=Indefinite (↑ Number)=Singular

3.4 词汇功能语法

其中，在动词的词条中，通过功能方程式定义了从语法功能到谓词—变元关系的映射。“<>”中表达的是句法模式，
 $\text{Hand} = \langle (\uparrow \text{Subject}), (\uparrow \text{Object}), (\uparrow \text{Object2}) \rangle$
，表示谓语动词hand要有一个主语，一个直接宾语和一个间接宾语。

3.4 词汇功能语法

用LGF语法对句子进行分析的过程如下：

(1) 用上下文无关语法分析获得C-structure，不考虑语法中的功能方程式；该C-structure就是一棵直接成分树。

(2) 将各个非叶节点定义为变量，并用这些变量置换词汇规则和语法规则中功能方程式的元变量（ \uparrow 或 \downarrow ），建立功能描述，这一描述实际上就是一组功能方程式。

对方程式作代数变换，求出各个变量，获得功能结构F-structure。

利用词汇功能语法对句子 “A girl handed the baby the toys” 进行分析的过程请参阅教材中的例8.4

3.5 依存句法分析

□ 依存句法理论

现代依存语法理论的创立者是法国语言学家Lucien Tesnière(1893-1954)。其思想主要反映在他1959年出版的《结构句法基础》。

3.5 依存句法分析

L. Tesnière 的理论认为:

一切结构句法现象可以概括为关联(**connexion**)、组合(**jonction**)和转位(**translation**)这三大核心。句法关联建立起词与词之间的从属关系,这种从属关系是由支配词和从属词联结而成;动词是句子的中心并支配别的成分,它本身不受其他任何成分支配。

3.5 依存句法分析

欧洲传统的语言学突出一个句子中主语的地位，句中其它成分称为“谓语”。依存语法打破了这种主谓关系，认为“谓语”中的动词是一个句子的中心，其他成分与动词直接或间接地产生联系。

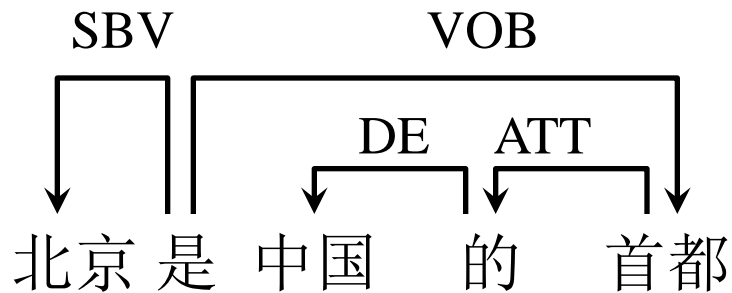
Tesnière 还将化学中“价”的概念引入依存语法，一个动词所能支配的行动元（名词词组）的个数即为该动词的价数。

3.5 依存句法分析

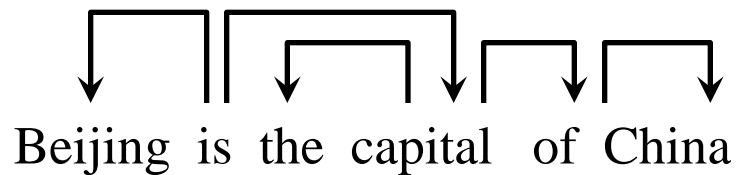
依存语法：用词与词之间的依存关系来描述语言结构的框架被称为依存语法，又称从属关系语法。

在依存语法理论中，“依存”就是指词与词之间支配与被支配的关系，这种关系不是对等的，而是有方向的。处于支配地位的成分称为支配者(governor, regent, head)，而处于被支配地位的成分称为从属者(modifier, subordinate, dependency)。

3.5 依存句法分析



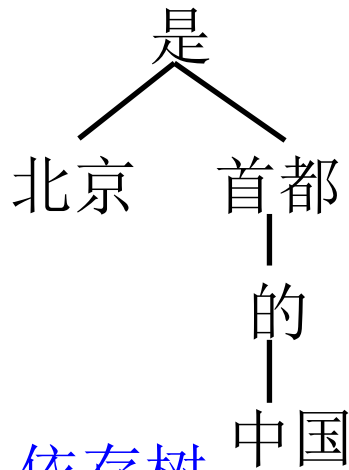
(e) 有向图-1



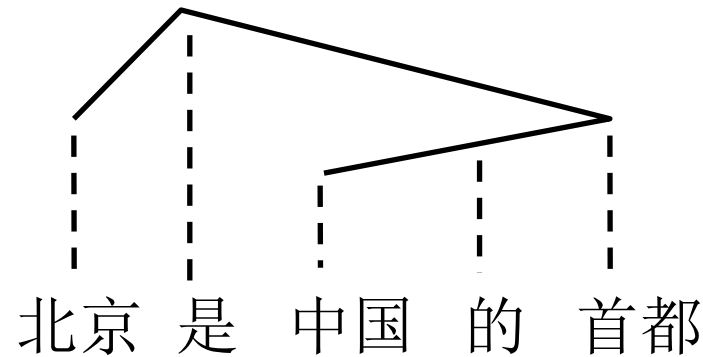
(e) 有向图-2

两个有向图用带有方向的弧(或称边, edge)来表示两个成分之间的依存关系, 支配者在有向弧的发出端, 被支配者在箭头端, 我们通常说被支配者依存于支配者。

3.5 依存句法分析



(f) 依存树



(g) 依存投射树

图(f)是用树表示的依存结构，树中子节点依存于该节点的父节点。

图(g)是带有投射线的树结构，实线表示依存联结关系，位置低的成份依存于位置高的成份，虚线为投射线。

3.5 依存句法分析

1970年计算语言学家J. Robinson在论文《依存结构和转换规则》中提出了依存语法的四条公理：

- (1) 一个句子只有一个独立的成分；
- (2) 句子的其他成分都从属于某一成分；
- (3) 任何一成分都不能依存于两个或多个成分；
- (4) 如果成分A直接从属于成分B，而成分C在句子中位于A和B之间，那么，成分C或者从属于A，或者从属于B，或者从属于A和B之间的某一成分。

3.5 依存句法分析

这四条公理相当于对依存图和依存树的形式约束为：

单一父结点(single headed)、连通(connective)、无环(acyclic)和可投射(projective)，并由此来保证句子的依存分析结果是一棵有“根(root)”的树结构。

3.5 依存句法分析

在处理中文信息的研究中，中国学者提出了依存关系的第五条公理，如下：

(5)中心成分左右两面的其它成分相互不发生关系

。

3.5 依存句法分析

□ 依存句法分析

建立一个依存句法分析器一般需要完成以下三部分工作：

- (1) 依存句法结构描述
- (2) 分析算法设计与实现
- (3) 语法规则或参数学习

3.5 依存句法分析

目前依存句法结构描述一般采用有向图方法或依存树方法，所采用的句法分析算法可大致归为以下四类：

- 生成式的分析方法(Generative parsing)
- 判别式的分析方法(Discriminative parsing)
- 决策式的分析方法(Deterministic parsing)
- 基于约束满足的分析方法(Constraint satisfaction parsing)

A decorative graphic on the left side of the slide, consisting of overlapping blue, red, and yellow squares with a black crosshair.

3.5 依存句法分析

(1) 生成式的分析方法

生成式的句法分析方法采用联合概率模型生成一系列依存句法树并赋予其概率分值，然后采用相关算法找到概率打分最高的分析结果作为最后输出。这是一种完全句法分析方法，它搜索整个概率空间，得到整个句子的依存分析结果。

3.5 依存句法分析

- 二元文法的词汇关系模型
(Bigram lexical affinities)

$$\Pr(words, tags, links) \approx \prod_{1 \leq i \leq n} \Pr(tag(i) | tag(i+1), tag(i+2)) \cdot \Pr(word(i) | tag(i)) \cdot \prod_{1 \leq i, j \leq n} \Pr(L_{ij} | tword(i), tword(j))$$

其中， $tword(i)$ 表示符号 i 的标记 ($tag(i)$) 和词本身 ($word(i)$)； L_{ij} 是取值0或1的二值函数， $L_{ij}=1$ 表示 i 和 j 具有依存关系， $L_{ij}=0$ 表示 i 和 j 不具有依存关系； n 是句子长度。

3.5 依存句法分析

一个标记序列(tags)由马尔柯夫(Markov)过程产生, 某一个标记由该标记前面的两个标记决定, 词由标记决定, 观察每一对词(words)是否可以构成链接关系(link)的决策依赖于[tags, words], 即 link 对词汇是敏感的。最终生成words, tags, links 的联合概率模型。

3.5 依存句法分析

(2) 判别式的分析方法

判别式句法分析方法采用条件概率模型，避开了联合概率模型所要求的独立性假设。

- 最大跨度树模型

(Maximum Spanning Trees, Mst)

定义整棵句法树的打分是树中各条边打分的加权和：

$$s(\mathbf{x}, \mathbf{y}) = \sum_{(i,j) \in \mathbf{y}} s(i, j) = \sum_{(i,j) \in \mathbf{y}} \mathbf{w} \cdot \mathbf{f}(i, j)$$

3.5 依存句法分析

其中， s 表示打分值， y 是句子 x 的一棵依存树， (x, y) 是 y 中的结点对。 $f(\bullet)$ 是取值为1或0的高维二元特征函数向量，表示结点 x_i 和 x_j 之间的依存关系，如果一棵依存树中两个词“打”和“球”存在依存关系，则：

$$f(i, j) = \begin{cases} 1 & \text{如果 } x_i = \text{'打'} \text{ and } x_j = \text{'球'} \\ 0 & \text{其他} \end{cases}$$

w 是特征 $f(i, j)$ 的权值向量， w 在确定了特征后由样本训练得到。

3.5 依存句法分析

该方法基本思想就是，在点和边组成的跨度树(spanning tree)中找到加权和分值最高的边的组合。跨度树中任意两个由词表示的节点之间都有边，根据特征和权值为每条边打分，求解最佳分析结果转化为搜索打分最高的最大跨度树问题。

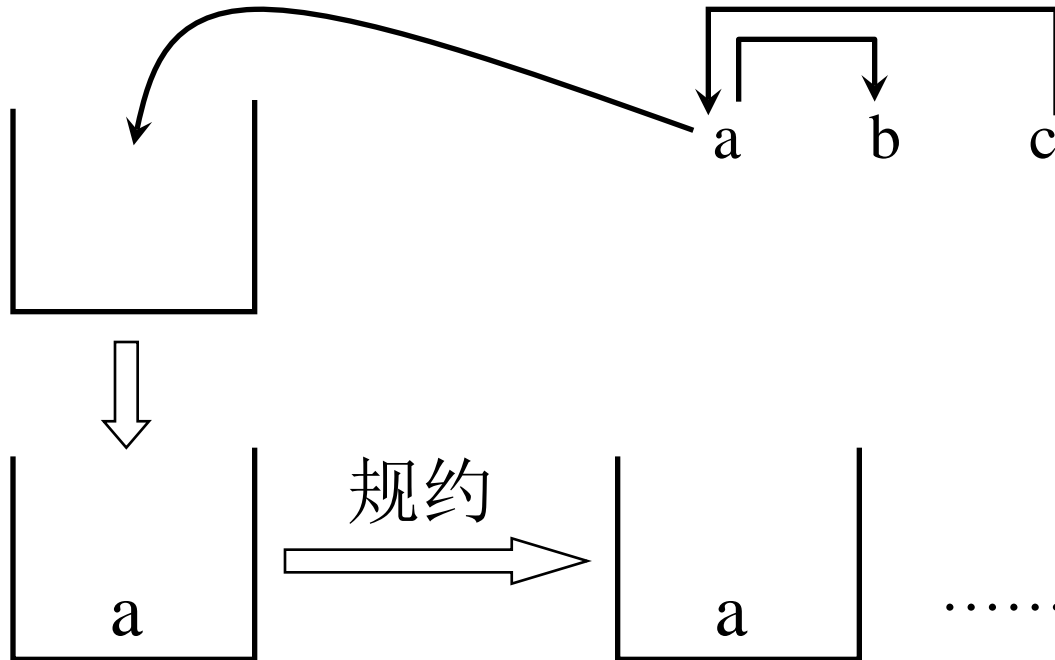
3.5 依存句法分析

(3) 决策式的分析方法

J. Nivre等(2003)提出的由左向右、自底向上的分析算法(移进-归约算法):

分析结构是一个三元组 $\langle S, I, A \rangle$ 。 S 是栈(stack), I 是待分析(剩余)的符号(token)序列, A 是当前已有的依存关系集合。决策时有三种转换操作(transition)可供选择: Left-reduce, Right-reduce和Shift。分析器根据规则判断当前栈顶符号(token)与下一个输入符号(token)是否存在依存关系, 如果存在, 则将这一依存关系添加到集合 A 中, 然后归约(Reduce)处于从属地位的符号, 否则, 移进Shift。

3.5 依存句法分析



将a、b依存关系添加到集合A中，

3.5 依存句法分析

(4) 基于约束满足的分析方法

基于约束满足的依存句法分析方法采用约束依存语法(Constraint Dependency Grammar, CDG), 该方法将依存句法分析过程看作可以用约束满足问题(Constraint satisfaction problem, CSP)来描述的有限构造问题(finite configuration problem)。

3.5 依存句法分析

判别式方法将寻求**最佳依存分析**转化为**最优路径搜索**问题，使得诸多机器学习方法和运筹学的方法得以应用，在可计算性上具有优势，该方法的大部分精力放在如何降低算法复杂度上。

决策式方法的提出是为了提高依存句法分析的有效性即降低算法复杂度，分析的每一步都不需要保留多个可能的结果，而只给出一个确定的结果。这种算法属于贪婪(Greedy)算法，在准确率上没有优势，但算法复杂度一般是线性的。



3.6 格语法

格语法（Case Grammar）是美国语言学家菲尔墨（C.J.Fillmore）在60年代中期提出来的着重探讨句法结构与语义之间关系的一种语法理论和语义学理论。

3.6 格语法

□ 格语法的来源

乔姆斯基在1957年出版的第一本书《句法结构》中提出了三大规则：短语结构规则、转换规则、语素音位规则。其短语结构规则（ $S \rightarrow NP + VP; V + NP$ ）的目标是生成所有的句子。结果，生成所有句子的目标虽然达到了，但是在生成正确句子（“约翰喝酒”）的同时，也生成出错误的句子（“洒喝约翰”）。这说明动词和名词之间要有一种语义限制。



3.6 格语法

乔姆斯基针对他第一本书存在的问题，于1965年出版了第二本书《语法理论的各方面》（The Aspects of the Theory of Yourself），主要是对第一本书的规则加以语义限制。但第二本书出版后不到一年又发现有新的问题。首先起来反对的是乔姆斯基的学生菲尔墨，他认为用各类格框架分析句法结构要比乔姆斯基的转换规则方便精密得多。

3.6 格语法

为了从语义的角度弥补转换生成语法的不足，菲尔墨1966年发表了《关于现代的格理论》（Toward a Modern Theory of Case），1968年发表了《格辨》（The Case for Case），1971年发表了《格语法的某些问题》（Some Problem for Case Grammar），1977年发表了《再论格辨》（The Case for Case Reopened）。其中的《格辨》是代表性论文。菲尔墨以上这些系列论文形成了一个语法学派，即所谓格语法，它实际上是转换生成语法发展出来的一个分支。

3.6 格语法

□ 格的含义

在传统语法中，“格”是指某些屈折语法中用于表示词间语法关系的名词和代词的形态变化，这种格必定有显性的形态标记，即以表层的词形变化为依据。如德语的四格。在汉语中，名词和代词没有形态变化，所以没有格。

3.6 格语法

传统语言学中的格只是表层格，其形式标志是词尾变化或者词干音变，这是某些屈折语的特有现象。格语法中的“格”是“深层格”，它是句子中体词（名词，代词等）和谓词（动词，形容词等）之间的及物性关系（transitivity），这些关系是语义关系，它是一切语言中普遍存在的现象。

这种格是在底层结构中依据名词与动词之间的句法语义关系来确定的，这种关系一经确定就固定不变，不管它们经过什么转换操作，在表层结构中处于什么位置，与动词形成什么语法关系，底层上的格与任何具体语言中的表层结构上的语法概念没有对应关系。

3.6 格语法

格语法有三部分组成：基本规则，词汇部分和转换部分。

□格语法基本规则

最基本的有三条规则：

(1) $S \rightarrow M + P$

(2) $P \rightarrow V + C_1 + C_2 + \dots + C_n$

(3) $C \rightarrow K + NP$

3.6 格语法

●格表

底层格的概念相当于人类对周围发生的事情所作出的判断。

菲尔墨在1996年认为命题中需用的格包括6种：

（1）施事格，（2）工具格，（3）承受格，（4）使成格，（5）方位格，（6）客体格。

后来，他在语言分析时又加了一些格：

（7）受益格，（8）源点格，（9）终点格，（10）伴随格。

3.6 格语法

□ 词汇部分

● 词库

词库是语言中词汇的集合。在词库中除了要标明每一个词条在句法、语义和语音方面的特征外，还需标明它们的底层格的特征。

● 词汇插入

格语法中词汇插入问题主要是名词和动词的选择问题。对于名词来说，把词库中每一个名词的特征与格范畴联系起来。

3.6 格语法

□转换部分

格的转换部分操作与转换生成语法大同小异，大致采用移动、删除、插入、复写等方法。

菲尔墨主要研究了有关格的形式和主语确定的转换规则。他认为深层格所体现的语义关系是一个固定而统一的概念，而在表层结构中的表现形式则因语言而异。有些语言主要通过介词来表现，有些语言用屈折变化和词汇变化来表现，有些语言则主要采用次序来表现，有些语言综合采用上述各种形式

3.6 格语法

□使用格语法进行语言分析-格框架约束分析技术

●分析结果可以使用“格框架”来表示

在格框架中，不仅可以有语法信息，而且还有许多语义信息，语言信息是整个格框架的最基本的部分。一个格框架可由一个主要概念和一组辅助概念组成，这些辅助概念以一种适当定义的方式同主要概念相联系。在实际使用中，主要概念可以理解为动词，辅助概念理解为施事格，受事格，处所格，工具格，工具格等语义深层格。

3.6 格语法

- 使用格语法进行语义分析的内容

把格框架中的格映射到输入句中找到的短语上。

- 分析基础

词典中要记录动词的格框架和名词的语义信息。

- 分析步骤

(1) 判断待分析词序列中主要动词，如果判断出，则在动词词典中找出该词的格框架。否则，对于待分析的词序列，查找带有格框架的动词词典。

(2) 识别必备格

3.6 格语法

(3) 按照与(2)相似的方法识别可选格

(4) 根据句子中出现的标志判断句子的情态Modal

如果处理完(2)、(3)和(4)后,分析词序列中还有未识别的成分,则或者分析出错,或者待分析的词序列不合法,或者动词的格框架,名词的语义信息不正确。如果分析成功,则得到待分析的词序列的格框架。

3.6 格语法

□ 格语法描写汉语的局限性

汉语的一些流水句、无动句。连动、紧缩、动补、省略等结构，无法或不必用一个动词统率一个句子的模式来描述。其中连动句和兼语句尤为突出。

。

3.7 概率上下文无关文法

□ PCFG 规则

形式: $A \rightarrow \alpha, P$

约束: $\sum_{\alpha} P(A \rightarrow \alpha) = 1$

例如: $\left. \begin{array}{l} NP \rightarrow NN \ NN, \ 0.60 \\ NP \rightarrow NN \ CC \ NN, \ 0.40 \end{array} \right\} \sum p = 1$

$\left. \begin{array}{l} CD \rightarrow QP, \ 0.99 \\ CD \rightarrow LST, \ 0.01 \end{array} \right\} \sum p = 1$

3.7 概率上下文无关文法

◆例-1: $S \rightarrow NP VP, 1.00$ $NP \rightarrow NP PP, 0.40$

$NP \rightarrow \text{astronomers}, 0.10$

$NP \rightarrow \text{ears}, 0.18$

$NP \rightarrow \text{saw}, 0.04$

$NP \rightarrow \text{stars}, 0.18$

$NP \rightarrow \text{telescopes}, 0.1$

$PP \rightarrow P NP, 1.00$

$P \rightarrow \text{with}, 1.00$

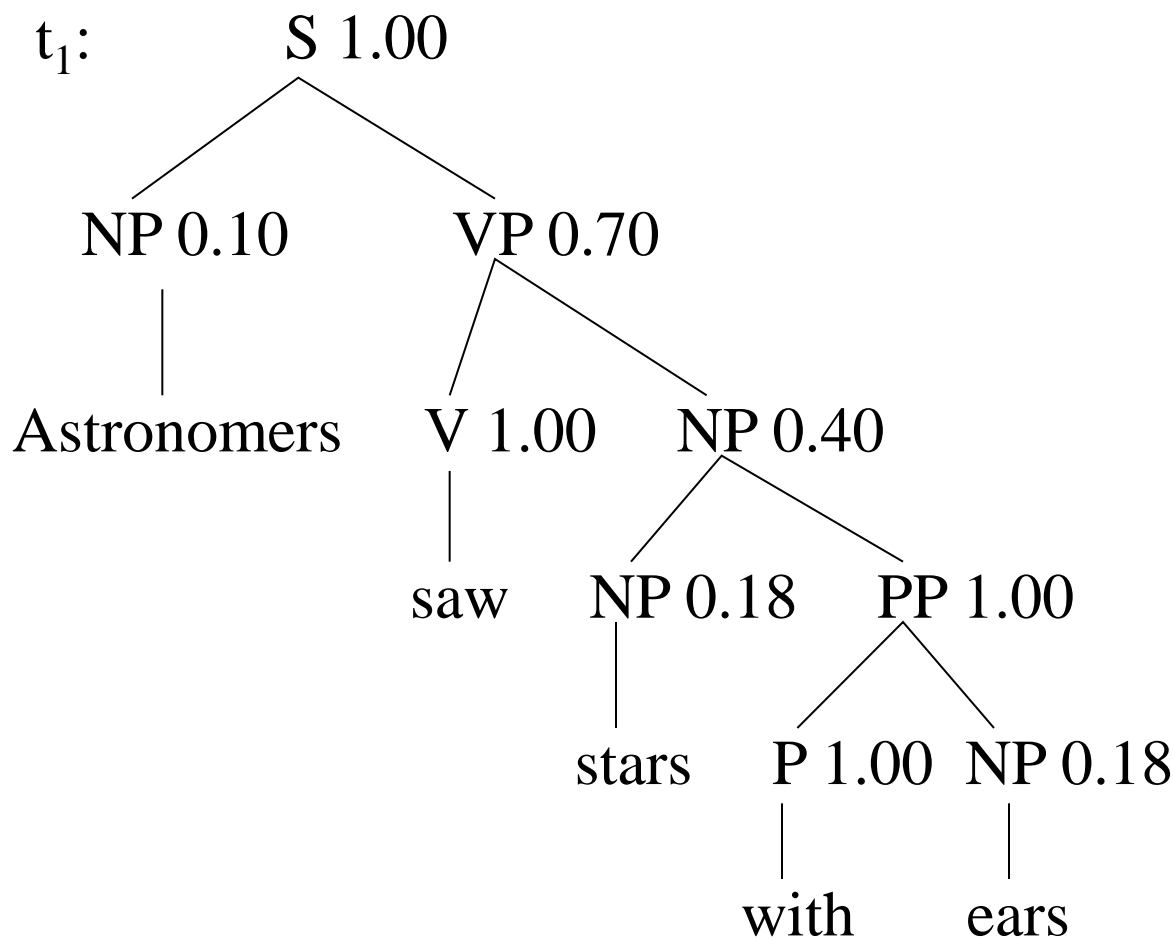
$VP \rightarrow V NP, 0.70$

$VP \rightarrow VP PP, 0.30$

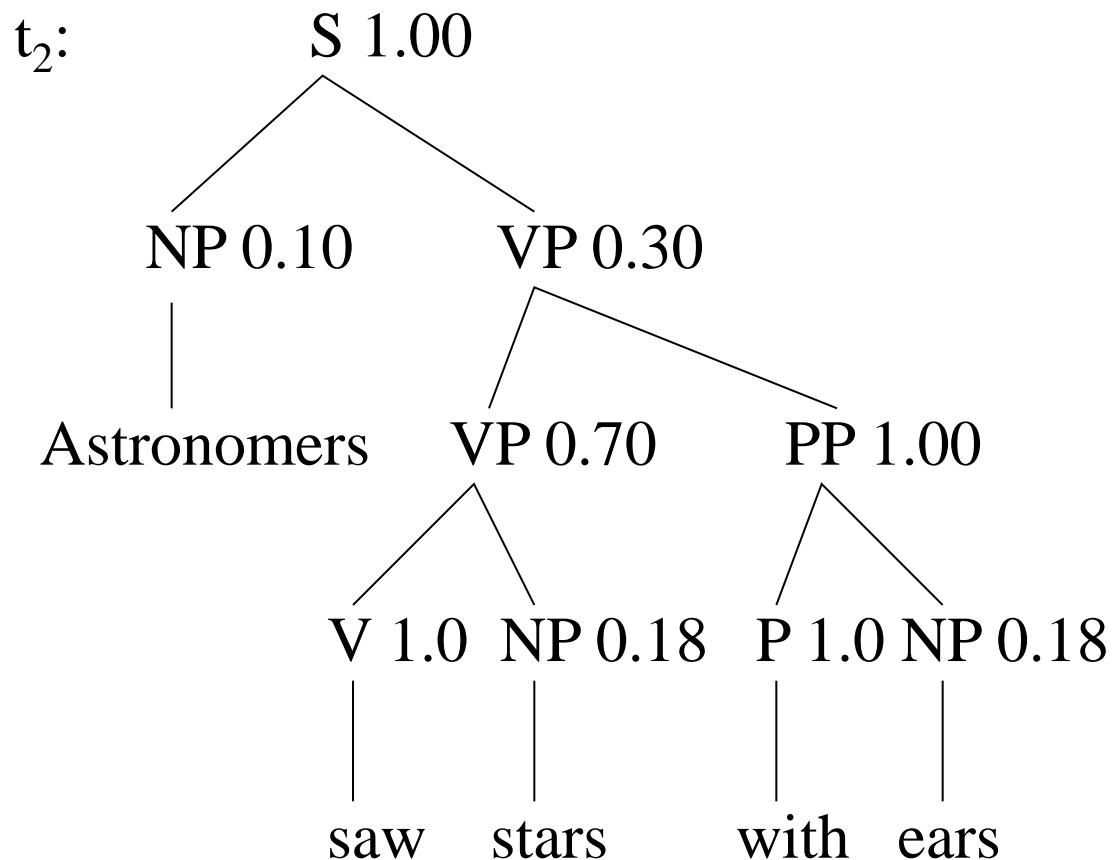
$V \rightarrow \text{saw}, 1.00$

给定句子 S: *Astronomers saw stars with ears.*

3.7 概率上下文无关文法



3.7 概率上下文无关文法



3.7 概率上下文无关文法

□ 计算分析树概率的基本假设

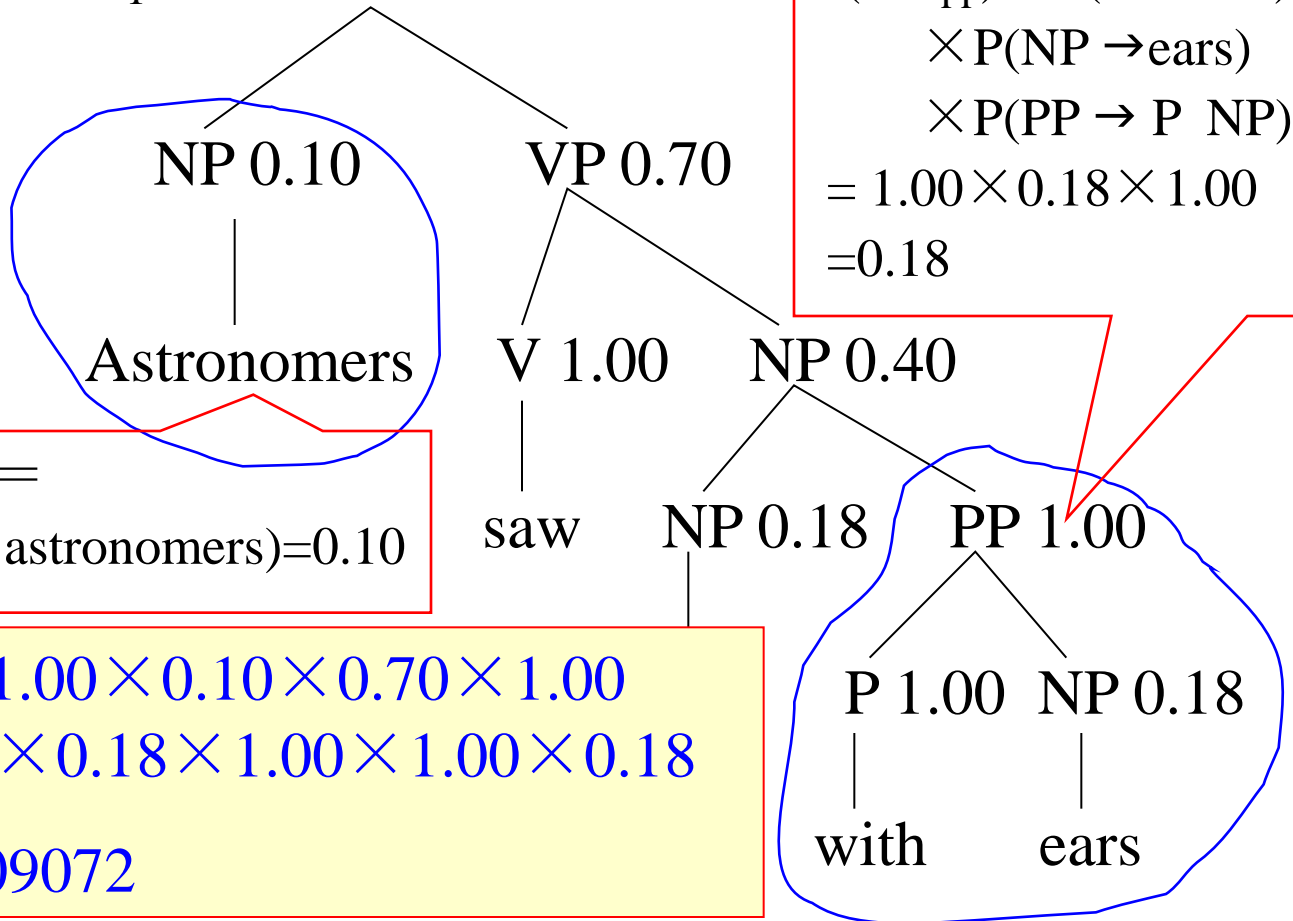
- 位置不变性：子树的概率与其管辖的词在整个句子中所处的位置无关，即对于任意的 k , $P(A_{k(k+C)} \rightarrow w)$ 一样。
- 上下文无关性：子树的概率与子树管辖范围以外的词无关，即 $P(A_{kl} \rightarrow w / \text{任何超出 } k \sim l \text{ 范围的上下文}) = P(A_{kl} \rightarrow w)$ 。

3.7 概率上下文无关文法

- 祖先无关性：子树的概率与推导出该子树的祖先结点无关，即 $P(A_{kl} \rightarrow w \mid \text{任何除 } A \text{ 以外的祖先结点}) = P(A_{kl} \rightarrow w)$ 。

3.7 概率上下文无关文法

t_1 : S 1.00



$$\begin{aligned}
 P(\text{tree}_{pp}) &= P(P \rightarrow \text{with}) \\
 &\quad \times P(\text{NP} \rightarrow \text{ears}) \\
 &\quad \times P(\text{PP} \rightarrow P \text{ NP}) \\
 &= 1.00 \times 0.18 \times 1.00 \\
 &= 0.18
 \end{aligned}$$

$$\begin{aligned}
 P(\text{tree}_{NP}) &= \\
 P(\text{NP} \rightarrow \text{astronomers}) &= 0.10
 \end{aligned}$$

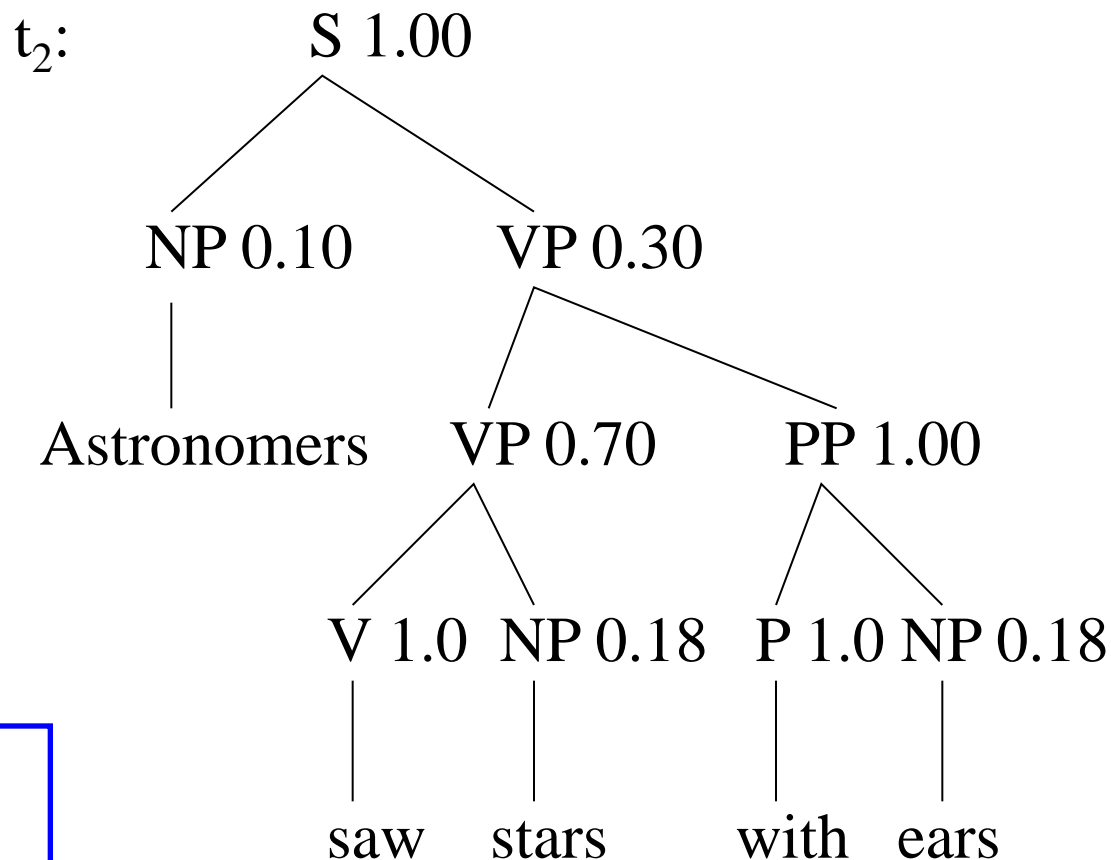
$$\begin{aligned}
 P(t_1) &= 1.00 \times 0.10 \times 0.70 \times 1.00 \\
 &\quad \times 0.40 \times 0.18 \times 1.00 \times 1.00 \times 0.18 \\
 &= 0.0009072
 \end{aligned}$$

3.7 概率上下文无关文法

$$P(t_2) = 1.00 \times 0.10 \times 0.30 \times 0.70 \times 1.00 \times 0.18 \times 1.00 \times 1.00 \times 0.18 = 0.0006804$$

给定的句子 S :

$$P(t_1) > P(t_2)$$





三、词法分析方法

3.8 词法分析概述

词是自然语言中能够独立运用的最小单位，是自然语言处理的基本单位。

自动词法分析就是利用计算机对自然语言的形态 (morphology) 进行分析，判断词的结构和类别等。

词性或称词类 (Part-of-Speech, POS) 是词汇最重要的特性，是连接词汇到句法的桥梁。

3.8 词法分析概述

□ 不同语言的词法分析

曲折语(如, 英语、德语、俄语等): 用词的形态变化表示语法关系, 一个形态成分可以表示若干种不同的语法意义, 词根和词干与语词的附加成分结合紧密。

词法分析: 词的形态分析(形态还原)。

分析语(孤立语)(如: 汉语): 分词。

黏着语(如: 日语等): 分词+形态还原。



3.9 英语的形态分析

- 基本任务
 - ◆ 单词识别
 - ◆ 形态还原

3.9 英语的形态分析

□ 英语单词的识别

例 (1) Mr. Green is a good English teacher.

(2) I'll see prof. Zhang home after the concert.

识别结果:

(1) Mr./ Green/ is/ a/ good/ English/ teacher/.

(2) I/ will/ see/ prof./ Zhang/ home/ after/ the/ concert/.

3.9 英语的形态分析

□ 英语中常见的特殊形式的单词识别

- (1) prof., Mr., Ms. Co., Oct. 等放入词典;
- (2) Let's / let's => let + us
- (3) I'm => I + am
- (4) {it, that, this, there, what, where}'s =>
 {it, that, this, there, what, where} + is
- (5) can't => can + not;
 won't => will + not

3.9 英语的形态分析

(6) {is, was, are, were, has, have, had}n't =>
 {is, was, are, were, has, have, had} + not

(7) X've => X + have;

 X'll=> X + will; X're => X + are

(8) he's => he + is / has => ?

 she's => she + is / has => ?

(9) X'd Y => X + would (如果 Y 为单词原型)
 => X + had (如果 Y 为过去分词)

3.9 英语的形态分析

□ 英语单词的形态还原

1. 有规律变化单词的形态还原

1) -ed 结尾的动词过去时，去掉ed;

*ed → * (e.g., worked → work)

*ed → *e (e.g., believed → believe)

*ied → *y (e.g., studied → study)

3.9 英语的形态分析

2) -ing 结尾的现在分词,

*ing → * (e.g., developing → develop)

*ing → *e (e.g., saving → save)

*ying → *ie (e.g., dying → die)

3) -s 结尾的动词单数第三人称;

*s → * (e.g., works → work)

*es → * (e.g., discusses → discuss)

*ies → *y (e.g., studies → study)

3.9 英语的形态分析

4) -ly 结尾的副词

*ly → * (e.g., hardly → hard)

... ..

5) -er/est 结尾的形容词比较级、最高级

*er → * (e.g., colder → cold)

*ier → *y (e.g., easier → easy)

.....

3.9 英语的形态分析

6) s/ses/xes/ches/shes/oes/ies/ves 结尾的名词复数，
ies/ves 结尾的名词还原时做相应变化：

bodies → body, shelves → shelf,

boxes → box, etc.

7) 名词所有格 X's, Xs'

3.9 英语的形态分析

2. 动词、名词、形容词、副词不规则变化单词的形态还原

一 建立不规则变化词表

例: choose, chose, chosen

axis, axes

bad, worse, worst

3.9 英语的形态分析

3. 对于表示年代、时间、百分数、货币、序数词的数字形态还原

- 1) 1990s → 1990, 标明时间名词;
- 2) 87th → 去掉 th 后, 记录该数字为序数词;
- 3) \$20 → 去掉\$, 记录该数字为名词(20美圆);
- 4) 98.5% → 98.5% 作为一个数词。

3.9 英语的形态分析

4. 合成词的形态还原

1) 基数词和序数词合成的分数词, e.g., one-fourth 等。

2) 名词+名词、形容词+名词、动词+名词等组成的合成名词, e.g., Human-computer, multi-engine, mixed-initiative, large-scale 等。

3.9 英语的形态分析

3) 形容词+名词+ed、形容词+现在分词、副词+现在分词、名词+过去分词、名词+形容词等组成的合形成形容词, e.g., machine-readable, hand-coding, non-adjacent, context-free, rule-based, speaker-independent 等。

3.9 英语的形态分析

4) 名词+动词、形容词+动词、副词+动词构成的合成动词, e.g., job-hunt 等。

5) 其他带连字符“-”的合成词, e.g., co-operate, 7-color, bi-directional, inter-lingua, Chinese-to-English, state-of-the-art, part-of-speech, OOV-words, spin-off, top-down, quick-and-dirty, text-to-speech, semi-automatically, *i*-th 等。

3.9 英语的形态分析

□ 形态分析的一般方法

- 1) 查词典，如果词典中有该词，直接确定该词的原形；
- 2) 根据不同情况查找相应规则对单词进行还原处理，如果还原后在词典中找到该词，则得到该词的原形；如果找不到相应变换规则或者变换后词典中仍查不到该词，则作为未登录词处理。
- 3) 进入未登录词处理模块。

3.9 英语的形态分析

下面是英语词法分析的一个基本算法：

repeat

look for word in dictionary,

if not found ,

then modify the word.

until word is found or no further modification possible

其中word是一个变量，其初值就是当前词。

例 用上述算法分析catches ,ladies的过程如下：

catches ladies ； 词典中查不到。

catche ladie ； 修改1，去掉"-s"。

#catch ladi ； 修改2，去掉"-e"。

#lady ； 修改3，变i为y。

上面修改2时就查到了catch，修改3时查到lady。当然更完整的词法分析还应当包括复合词的切分等，这里就不再进一步讨论了。

3.10 汉语自动分词概要

□ 汉语自动分词的重要性

- 自动分词是汉语句子分析的基础
- 词语的分析具有广泛的应用（词频统计，词典编纂，文章风格研究等）
- 文献处理以词语为文本特征
- “以词定字、以词定音”，用于文本校对、同音字识别、多音字辨识、简繁体转换

3.10 汉语自动分词概要

□ 汉语自动分词中的主要问题

◆ 汉语分词规范问题（《信息处理用限定汉语分词规范（GB13715）》）

一 汉语中什么是词？两个不清的界限：

（1）单字词与词素，如：新华社25日讯

（2）词与短语，如：花草，湖边，房顶，鸭蛋，小鸟，担水，一层，翻过？

3.10 汉语自动分词概要

◆ 歧义切分字段处理

1、中国人为了实现自己的梦想 (交集型歧义)

中国/ 人为/ 了/ 实现/ 自己/ 的/ 梦想

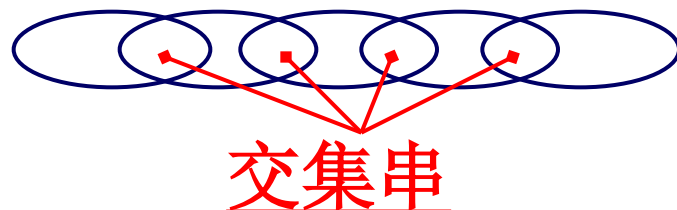
中国人/ 为了/ 实现/ 自己/ 的/ 梦想

中/ 国人/ 为了/ 实现/ 自己/ 的/ 梦想

例如：“大学生”、“研究生物”、“从小学起”、“为人民工作”、“中国产品质量”、“部分居民生活水平”等等

3.10 汉语自动分词概要

- ◆ **定义：链长** 一个交集型切分歧义所拥有的交集串的集合称为交集串链，它的个数称为链长。



例如，结合成分子

“结合”、“合成”、“成分”和“分子”均构成词，交集串的集合为 {合，成，分}，因此，链长为3。



3.10 汉语自动分词概要

类似地,

(1) “为人民工作”

{人, 民, 工}, 歧义字段的链长为3;

(2) “中国产品质量”

{国, 产, 品, 质}, 歧义字段的链长为4;

(3) “部分居民生活水平”

{分, 居, 民, 生, 活, 水}, 歧义字段的链长为6。

3.10 汉语自动分词概要

2、门把手弄坏了。 (组合型歧义)

门/ 把/ 手/ 弄/ 坏/ 了/ 。

门/ 把手/ 弄/ 坏/ 了/ 。

例如，“将来”、“现在”、“才能”、“学生会”等，都是组合型歧义字段。

3.10 汉语自动分词概要

梁南元（1987）曾经对一个含有48,092字的自然科学、社会科学样本进行了统计，结果交集型切分歧义有518个，多义组合型切分歧义有42个。据此推断，中文文本中切分歧义的出现频度约为1.2次/100字，交集型切分歧义与多义组合型切分歧义的出现比例约为12:1。

3.10 汉语自动分词概要

◆ 未登录词的识别

1、人名、地名、组织机构名等，例如：

盛中国，令计划，令狐路线，张建国，蔡国庆，
党政法，蔡英文，水皮，雷地球，彭太发生，
平川三太郎，约翰·斯特朗，詹姆斯·埃尔德

2、新出现的词汇、术语、个别俗语等，例如：

博客，非典，禽流感，恶搞，裸退

3.10 汉语自动分词概要

例如：

- (1) 他还兼任何应钦在福州办的东路军军官学校的政治教官。
- (2) 大不列颠及北爱尔兰联合王国外交和英联邦事务大臣、议会议员杰克·斯特劳阁下在联合国安理会就伊拉克问题发言。
- (3) 坐落于江苏省南京市玄武湖公园内的夏璞墩是晋代著名的文学家、科学家夏璞的衣冠冢(zhong)。

3.10 汉语自动分词概要

□ 汉语自动分词的基本原则

1、语义上无法由组合成分直接相加而得到的字串应该合并为一个分词单位。(合并原则)

如：不管三七二十一（成语），或多或少（副词片语），十三点（定量结构），六月（定名结构），谈谈（重叠结构，表示尝试），辛辛苦苦（重叠结构，加强程度），进出口（合并结构）

3.10 汉语自动分词概要

2、语类无法由组合成分直接得到的字串应该合并为一个分词单位。 (合并原则)

(1)字串的语法功能不符合组合规律，如：好吃，好喝，好听，好看等

(2)字串的内部结构不符合语法规律，如：游水等

3.10 汉语自动分词概要

□ 汉语自动分词的辅助原则

操作性原则，富于弹性，不是绝对的。

1. 有明显分隔符标记的应该切分之 (切分原则)

分隔标记指标点符号或一个词。如：

上、下课 → 上/ 下课

洗了个澡 → 洗/ 了/ 个/ 澡

3.10 汉语自动分词概要

2. 附着性语(词)素和前后词合并为一个分词单位 (合并原则)

如：“吝”是一个附着语素，“不吝”、“吝于”等合并成一个词；

“员”：检查员、邮递员、技术员等；

“化”：现代化、合理化、多变化、民营化等。

3.10 汉语自动分词概要

3. 使用频率高或共现率高的字串尽量合并为一个分词单位 (合并原则)

如：“进出”、“收放”（动词并列）；

“大笑”、“改称”（动词偏正）；

“关门”、“洗衣”、“卸货”（动宾）；

“春夏秋冬”、“轻重缓急”、“男女”（并列）；

“象牙”（名词偏正）；“暂不”、“毫不”、“不再”、“早已”（副词并列）等

3.10 汉语自动分词概要

4. 双音节加单音节的偏正式名词尽量合并为一个分词单位 (合并原则)

如：“线、权、车、点”等所构成的偏正式名词：“国际线、分数线、贫困线”、“领导权、发言权、知情权”、“垃圾车、交通车、午餐车”、“立足点、共同点、着眼点”等。

3.10 汉语自动分词概要

5. 双音节结构的偏正式动词应尽量合并为一个分词单位 (合并原则)

本原则只适合少数偏正式动词，如：“紧追其后”、“组建完成”等，不适合动宾及主谓式复合动词。

3.10 汉语自动分词概要

6. 内部结构复杂、合并起来过于冗长的词尽量切分 (切分原则)

(1) 词组带接尾词

太空/ 计划/ 室、塑料/ 制品/ 业

(2) 动词带双音节结果补语

看/ 清楚、讨论/ 完毕

(3) 复杂结构: 自来水/ 公司、中文/ 分词/ 规范/ 研究/ 计划

(4) 正反问句: 喜欢/ 不/ 喜欢、参加/ 不/ 参加

3.10 汉语自动分词概要

(5) 动宾结构、述补结构的动词带词缀时
写信/ 给、取出/ 给、穿衣/ 去

(6) 词组或句子的专名，多见于书面语，戏剧名、歌曲名等

鲸鱼/ 的/ 生/ 与/ 死、那/ 一/ 年/ 我们/ 都/ 很/ 酷

(7) 专名带普通名词

胡/ 先生、京沪/ 铁路



3.11 汉语自动分词基本算法

- 有词典切分/ 无词典切分
- 基于规则的方法/ 基于统计的方法

3.11 汉语自动分词基本算法

1. 最大匹配法 (Maximum Matching, MM)

—有词典切分，机械切分

- 正向最大匹配算法 (Forward MM, FMM)
- 逆向最大匹配算法 (Backward MM, BMM)
- 双向最大匹配算法 (Bi-directional MM)

假设句子: $S = c_1c_2 \cdots c_n$, 某一词:

$w_i = c_1c_2 \cdots c_m$, m 为词典中最长词的字数。

3.11 汉语自动分词基本算法

◆ FMM 算法描述

- (1) 令 $i=0$ ，当前指针 p_i 指向输入字串的初始位置，执行下面的操作：
- (2) 计算当前指针 p_i 到字串末端的字数（即未被切分字串的长度） n ，如果 $n=1$ ，转(4)，结束算法。否则，令 m =词典中最长单词的字数，如果 $n < m$ ，令 $m=n$ ；

3.11 汉语自动分词基本算法

- (3) 从当前 p_i 起取 m 个汉字作为词 w_i ，判断：
- (a) 如果 w_i 确实是词典中的词，则在 w_i 后添加一个切分标志，转(c)；
 - (b) 如果 w_i 不是词典中的词且 w_i 的长度大于1，将 w_i 从右端去掉一个字，转(a)步；否则 (w_i 的长度等于1)，则在 w_i 后添加一个切分标志，将 w_i 作为单字词添加到词典中，执行 (c) 步；
 - (c) 根据 w_i 的长度修改指针 p_i 的位置，如果 p_i 指向字串末端，转(4)，否则， $i=i+1$ ，返回 (2)；
- (4) 输出切分结果，结束分词程序。

3.11 汉语自动分词基本算法

例：假设词典中最长单词的字数为 7。

输入字符串：他是研究生物化学的。

切分过程：他是研究生物化学的。

$p \uparrow$ |

... ..

他/ 是研究生物化学的。

$p \uparrow$ |

FMM 切分结果：他/ 是/ 研究生/ 物化/ 学/ 的/。

BMM 切分结果：他/ 是/ 研究/ 生物/ 化学/ 的/。

3.11 汉语自动分词基本算法

➤ 优点：

- 程序简单易行，开发周期短；
- 仅需要很少的语言资源（词表），不需要任何词法、句法、语义资源；

➤ 弱点：

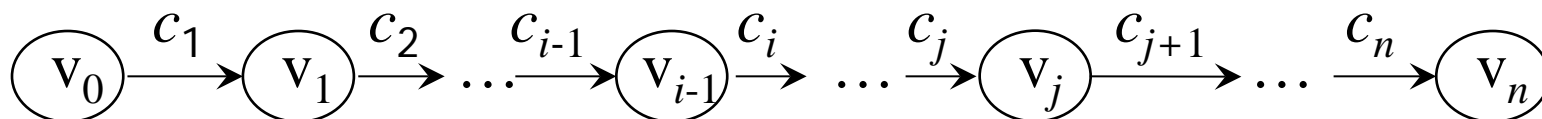
- 歧义消解的能力差；
- 切分正确率不高，一般在95%左右。

3.11 汉语自动分词基本算法

2. 最少分词法（最短路径法）

◆ 基本思想

设待分字串 $S=c_1 c_2 \dots c_n$ ，其中 $c_i (i=1,2,\dots,n)$ 为单个的字， n 为串的长度， $n \geq 1$ 。建立一个节点数为 $n+1$ 的切分有向无环图 G ，各节点编号依次为 $V_0, V_1, V_2, \dots, V_n$ 。

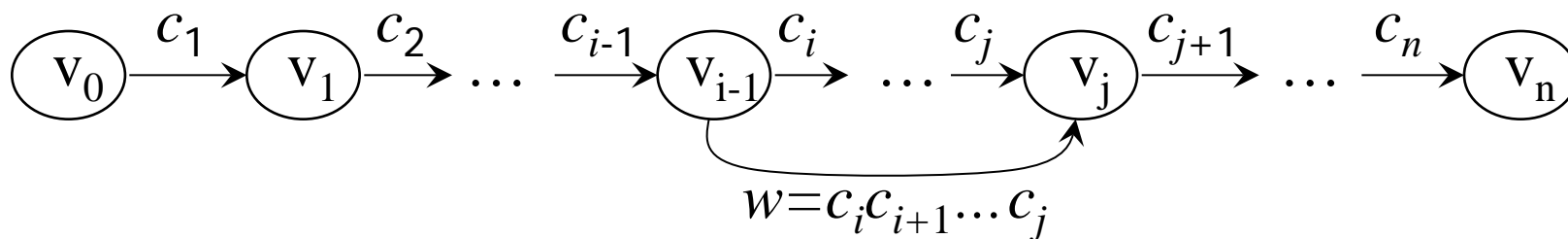


求最短路径：贪心法或简单扩展法。

3.11 汉语自动分词基本算法

◆ 算法描述:

- (1) 相邻节点 v_{k-1}, v_k 之间建立有向边 $\langle v_{k-1}, v_k \rangle$, 边对应的词默认为 c_k ($k=1, 2, \dots, n$)。
- (2) 如果 $w=c_i c_{i+1} \dots c_j$ ($0 < i < j \leq n$) 是一个词, 则节点 v_{i-1}, v_j 之间建立有向边 $\langle v_{i-1}, v_j \rangle$, 边对应的词为 w 。



- (3) 重复步骤(2), 直到没有新路径(词序列)产生。
- (4) 从产生的所有路径中, 选择路径最短的(词数最少的)作为最终分词结果。



3.11 汉语自动分词基本算法



例：(1) 输入字串：他只会诊断一般的疾病。

可能输出：他/ 只会/ 诊断/ 一般/ 的/ 疾病/。(7)

他/ 只/ 会诊/ 断/ 一般/ 的/ 疾病/。(8)

... ..

最终结果：他/ 只会/ 诊断/ 一般/ 的/ 疾病/ 。

(2) 输入字串：他说的确实在理。

可能输出：他/ 说/ 的/ 确实/ 在理/ 。（6）

他/ 说/ 的确/ 实在/ 理/ 。（6）

... ..

3.11 汉语自动分词基本算法

➤ 优点:

- 切分原则符合汉语自身规律
- 需要的语言资源（词表）也不多

➤ 弱点:

- 对许多歧义字段难以区分，最短路径有多条时，选择最终的输出结果缺乏应有的标准。
- 字串长度较大和选取的最短路径数增大时，长度相同的路径数急剧增加，选择最终正确的结果困难越来越大。

3.11 汉语自动分词基本算法

3. 基于统计语言模型的分词方法

◆ 方法描述:

设对于待切分的句子 S , $W = w_1w_2\cdots w_k$
($1 \leq k \leq n$) 是一种可能的切分。

$$\begin{aligned} W^* &= \arg \max_W P(W | S) \\ &= \arg \max_W P(W)P(S | W) \end{aligned}$$

3.11 汉语自动分词基本算法

微软研究院把一个可能的词序列 W 转换成一个可能的词类序列 $C = c_1 c_2 \cdots c_N$ ，即：

- 专有名词的人名PN、地名LN、机构名ON分别作为一类；
- 实体名词中的日期 dat、时间tim、百分数per、货币mon 等作为一类；
- 对词法派生词MW和词表词LW，每个词单独作为一类。

3.11 汉语自动分词基本算法

那么,

$$C^* = \arg \max_C P(C) P(S | C) \quad (7-1)$$

语言模型 \rightarrow $P(C)$ \leftarrow $P(S | C)$ 生成模型

$P(C)$ 可采用3元语法:

$$P(C) = P(c_1)P(c_2 | c_1) \prod_{i=3}^N P(c_i | c_{i-2}c_{i-1}) \quad (7-2)$$

3.11 汉语自动分词基本算法

生成模型在满足独立性假设的条件下，可近似为：

$$P(S | C) \approx \prod_{i=1}^N P(s_i | c_i) \quad (7-3)$$

该公式的含意是，任意一个词类生成汉字串的概率只与自身有关，而与其上下文无关。例如，如果“教授”是词表里的词，那么 $P(s_i=\text{教授} | c_i=\text{LW})=1$ 。

3.11 汉语自动分词基本算法

词 类	生成模型 $P(S C)$	语言知识
词表词 (LW)	若 S 是词表词, $P(S LW)=1$, 否则为0;	分词词表
词法派生词 (MW)	若 S 是派生词, $P(S MW)=1$, 否则为0;	派生词词表
人名 (PN)	基于字的二元模型	姓氏表, 中文人名模板
地名 (LN)	基于字的二元模型	地名表、地名关键词表、地名简称表
机构名 (ON)	基于词类的二元模型	机关名关键词表, 机构名简称表
实体名 (FT)	若 S 可用实体名词规则集 G 识别, $P(S G)=1$, 否则为0。	实体名词规则集

3.11 汉语自动分词基本算法

模型的训练由以下三步组成：

- (1) 在词表和派生词表的基础上，用正向最大匹配法切分训练语料，专有名词通过一个专门模块标注，实体名词通过相应的规则和有限状态自动机标注，由此产生一个带词类别标记的初始语料；
- (2) 用带词类别标记的初始语料，采用最大似然估计方法估计语言模型的概率参数；
- (3) 用语言模型（公式(7-1)、(7-2)、(7-3)），对训练语料重新切分和标注，得到新的训练语料；
- (4) 重复(2)(3)步，直到系统的性能不再有明显的变化。



3.11 汉语自动分词基本算法

➤ 优点:

- 减少了很多手工标注的工作
- 在训练语料规模足够大和覆盖领域足够多时，可以获得较高的切分正确率

➤ 弱点:

- 训练语料的规模和覆盖领域不好把握
- 计算量较大

3.11 汉语自动分词基本算法

4. 基于HMM的分词方法

基本思想：

把输入字串(句子) S 作为HMM的输入；
(切分后的)单词串 S_w 为状态的输出，即观察序列 $S_w = w_1 w_2 \cdots w_n$ ($n \geq 1$)；词性序列 S_c 为状态序列，每个词性标记对应HMM中的一个状态 q_i ， $S_c = c_1 c_2 \cdots c_n$ 。

详细解释略，可参见第6章。

3.11 汉语自动分词基本算法

➤ 优点:

- 可以减少很多手工标注的工作
- 在训练语料规模足够大和覆盖领域足够多时，可以获得较高的切分正确率

➤ 弱点:

- 训练语料的规模和覆盖领域不好把握
- 模型实现复杂、计算量较大

3.11 汉语自动分词基本算法

5. 基于统计模型的分词与词性标注一体化方法

基本思想： 设句子 S 由词串组成 $W=w_1w_2\cdots w_n$ ($n\geq 1$), 单词 w_i 的词性标注为 t_i , 即句子 S 相应的词性标注符号序列可表达为 $T=t_1t_2\cdots t_n$ 。那么, 分词与词性标注的任务就是要在 S 所对应的各种切分和标注形式中, 寻找 T 和 W 的联合概率 $P(W, T)$ 为最优的词切分和标注组合。

3.11 汉语自动分词基本算法

如果把词性符号序列作为HMM的中间状态，词序列作为输出，那么， $P(W, T)$ 可以由HMM近似地表示为：

$$P(W, T) = P(W | T)P(T) \approx \prod_{i=1}^n P(w_i | t_i) P(t_i | t_{i-1} t_{i-2}) \quad (7-4)$$

生成模型

基于词性的
语言模型

3.11 汉语自动分词基本算法

反之，如果把词序列作为HMM的中间状态，词性符号作为输出，那么， $P(W, T)$ 的另一种形式为：

$$P(W, T) = P(T | W)P(W) \approx \prod_{i=1}^n P(t_i | w_i) P(w_i | w_{i-1} w_{i-2}) \quad (7-5)$$

生成模型

基于词的
语言模型

3.11 汉语自动分词基本算法

将上述(7-4)和(7-5)综合：

$$P^*(W, T) = \alpha \prod_{i=3}^n P(w_i | t_i) P(t_i | t_{i-1}, t_{i-2}) + \beta \prod_{i=1}^n P(t_i | w_i) P(w_i | w_{i-1}, w_{i-2}) \quad (7-6)$$

显然，这种综合模型的指导思想是希望通过调整参数 α 和 β 的值来确定两个子模型在整个分词与词性标注过程中所发挥作用的比重，从而获得分词与词性标注的整体最优。

3.11 汉语自动分词基本算法

从公式 (7-5) 得到的结果分析可知, $P(t_i | w_i)$ 对分词无帮助, 且在分词确定后对词性标注又会增添偏差。因此, 在实现这一模型时, 可仅取公式 (7-5) 中的语言模型部分, 而舍弃词性标注部分, 并令 $\alpha = 1$, 仅保留加权系统 β , 于是,

$$\begin{aligned} P^{\wedge}(W, T) = & \prod_{i=1}^n P(w_i | t_i) P(t_i | t_{i-1}, t_{i-2}) + \\ & \beta \prod_{i=1}^n P(w_i | w_{i-1}, w_{i-2}) \end{aligned} \quad (7-7)$$

3.11 汉语自动分词基本算法

在确定 β 系数值时，可根据词典中词汇 w 的个数和词性 t 的种类数目，取二者之比，即 $\beta = \text{词典中词 } w \text{ 的个数} / \text{词性 } t \text{ 的种类数}$ 。

3.11 汉语自动分词基本算法

➤ 优点:

- 可以减少很多手工标注的工作
- 在训练语料规模足够大和覆盖领域足够多，各类参数设定适当时，可以获得较高的切分正确率

➤ 弱点:

- 训练语料的规模和覆盖领域不好把握
- β 系数值难以把握

3.11 汉语自动分词基本算法

6. 由字构词的(基于字标注)分词方法 (Character-based tagging)

第一篇由字构词的汉语分词方法的论文[Xue, 2002]发表在2002年的第一届ACL汉语特别兴趣小组SIGHAN (<http://www.sighan.org/>) 组织的研讨会上, 在2005年和2006年的两次Bakeoff 评测中取得好成绩。

3.11 汉语自动分词基本算法

◆ **基本思想**：将分词过程看作是字的分类问题。该方法认为，每个字在构造一个特定的词语时都占据着一个确定的构词位置(即词位)。假定每个字只有4个词位：词首(B)、词中(M)、词尾(E)和单独成词(S)，那么，每个字归属一特定的词位。

这里所说的“字”不仅限于汉字，也可以指标点符号、外文字母、注音符号和阿拉伯数字等任何可能出现在汉语文本中的文字符号，所有这些字符都是由字构词的基本单元。

3.11 汉语自动分词基本算法

例如：

(1) 上海/ 计划/ 到/ 本/ 世纪/ 末/ 实现/ 人均/
国内/ 生产/ 总值/ 五千美元/ 。 /

(2) 上/B 海/E 计/B 划/E 到/S 本/S 世/B 纪/E
末/S 实/B 现/E 人/B 均/E 国/B 内/E 生/B 产/E 总
/B 值/E 五/B 千/M 美/M 元/E 。 /S

3.11 汉语自动分词基本算法

在字标注过程中，对所有的字根据预定义的特征进行词位特征学习，获得一个概率模型，然后在待切分字符串上，根据字与字之间的结合紧密程度，得到一个词位的分类结果，最后根据词位定义直接获得最终的分词结果。

◆ 工具：

- 支持向量机（SVM）
- 条件随机场（CRF）

最常用的两类特征是字本身和词位(状态)的转移概率

3.11 汉语自动分词基本算法

◆ 评价:

该方法的重要优势在于，它能够平衡地看待词表词和未登录词的识别问题，文本中的词表词和未登录词都是用统一的字标注过程来实现的。在学习构架上，既可以不必专门强调词表词信息，也不用专门设计特定的未登录词识别模块，因此，大大地简化了分词系统的设计[黄昌宁，2006]

3.11 汉语自动分词基本算法

□ 其他方法

- ◆ 全切分方法
- ◆ 串频统计和词形匹配相结合的分词方法
- ◆ 规则方法与统计方法相结合
- ◆ 多重扫描法

.....

3.11 汉语自动分词基本算法

□ 方法比较

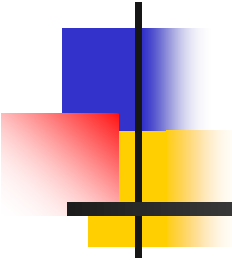
- (1) 最大匹配分词算法是一种简单的基于词表的分词方法，有着非常广泛的应用。这种方法只需要最少的语言资源（仅需要一个词表，不需要任何词法、句法、语义知识），程序实现简单，开发周期短，是一个简单实用的方法，但对歧义字段的处理能力不够强大。

3.11 汉语自动分词基本算法

- (2) 全切分方法首先切分出与词表匹配的所有可能的词，然后运用统计语言模型和决策算法决定最优的切分结果。这种切分方法的优点是发现所有的切分歧义，但解决歧义的方法很大程度上取决于统计语言模型的精度和决策算法，需要大量的标注语料，并且分词速度也因搜索空间的增大而有所缓慢。

3.11 汉语自动分词基本算法

- (3) 最短路径分词方法的切分原则是使切分出来的词数最少。这种切分原则多数情况下符合汉语的语言规律，但无法处理例外的情况，而且如果最短路径不止一条时，系统往往不能确定最优解。
- (4) 统计方法具有较强的歧义区分能力，但需要大规模标注 (或预处理) 语料库的支持，需要的系统开销也较大。



3.12 未登录词识别

3.12 未登录词识别

□ 命名实体(Named Entity, NE) (专有名词)

人名（中国人名和外国译名）、地名、组织机构名、数字、日期、货币数量

□ 其他新词

专业术语、新的普通词汇等。

3.12 未登录词识别

□ 关于中文姓名

- 台湾出版的《中国姓氏集》收集姓氏 5544 个，其中，单姓 3410 个，复姓 1990 个，3字姓 144 个
- 中国目前仍使用的姓氏共 737 个，其中，单姓 729 个，复姓 8 个
- 根据我们收集的 300 万个人名统计，姓氏有974 个，其中，单姓 952个，复姓 23 个，300万人名中出现汉字4064个。（曹文洁，2002a, 2002b）

3.12 未登录词识别

□ 中文姓名识别的难点

- 名字用字范围广，分布松散，规律不很明显。
- 姓氏和名字都可以单独使用用于特指某一人。
- 许多姓氏用字和名字用字（词）可以作为普通用字或词被使用，例如，姓氏：于（介词），张（量词），江（名词）等；名字：建国，国庆，胜利，文革等，全名本身也是普通词汇，如：万里，温馨，高山，高升，高飞，周密，江山等。

3.12 未登录词识别

➤ 缺乏可利用的启发标记。

例如: (1) 祝贺老总百战百胜。

(2) 林徽因此时已经离开了那里。

(3) 赵微笑着走了。

(4) 南京市长江大桥。

3.12 未登录词识别

□ 中文姓名识别方法

- ◆ 姓名库匹配，以姓氏作为触发信息，寻找潜在的名字
- ◆ 计算潜在姓名的概率估值及相应姓氏的姓名阈值(threshold value)，根据姓名概率评价函数和修饰规则对潜在的姓名进行筛选。

3.12 未登录词识别

□ 计算概率估计值

设姓名 $Cname = Xm_1m_2$ ，其中 X 表示姓， m_1m_2 分别表示名字首字和名字尾字。分别用下列公式计算姓氏和名字的使用频率：

$$F(X) = \frac{X \text{ 用作姓氏}}{X \text{ 出现的总次数}}$$

$$F(m_1) = \frac{m_1 \text{ 作为名字首字出现的次数}}{m_1 \text{ 出现的总次数}}$$

$$F(m_2) = \frac{m_2 \text{ 作为名字尾字出现的次数}}{m_2 \text{ 出现的总次数}}$$

3.12 未登录词识别

字符串 $Cname$ 可能为姓名的概率估值:

$$P(Cname) = \begin{cases} F(X) \times F(m_1) \times F(m_2) & \text{复名情况} \\ F(X) \times F(m_2) & \text{单名情况} \end{cases}$$

□ 确定阈值

姓氏 X 构成姓名的最小阈值:

$$T_{\min}(X) = \begin{cases} F(X) \times \text{Min}(F(m_1) \times F(m_2)) & \text{复名情况} \\ F(X) \times \text{Min}(F(m_2)) & \text{单名情况} \end{cases}$$

3.12 未登录词识别

□ 设计评估函数

姓名的评价函数：

$$f = \ln P(Cname)$$

对于特定的姓氏 X 通过训练语料得到一
阈值 β_X ，当 f 大于 β_X 时，该识别的汉字串确
定为中文姓名。

3.12 未登录词识别

□ 使用修饰规则：

如果姓名前是一个数字，或者与“.”字符的距离小于 2 个字节，则否定此姓名。

◆ 确定潜在的姓名边界

➤ 左界规则：若潜在姓名前面是一称谓，或一标点符号，或者潜在姓名在句首，或者潜在的姓名的姓氏使用频率为100%，则姓名的左界确定。

3.12 未登录词识别

➤ 右界规则：若姓名后面是一称谓，或者是一指界动词(如，说，是，指出，认为等)或标点符号，或者潜在的姓名在句尾，或者潜在姓名的尾字使用频率为100%，则姓名的右界确定。

◆ 校正潜在的姓名

依据：含重合部分的潜在姓名不可能同时成立。利用各种规则消除冲突的潜在姓名。

3.12 未登录词识别

□ 中文地名识别方法

◆ 困难

- 地名数量大，缺乏明确、规范的定义。《中华人民共和国地名录》(1994)收集88026个，不包括相当一部分街道、胡同、村庄等小地方的名称。
- 真实语料中地名出现情况复杂。如地名简称、地名用词与其他普通词冲突、地名是其他专用名词的一部分，地名长度不一等。

3.12 未登录词识别

◆ 基本资源

- 建立地名资源知识库
 - 地名库、地名用字库、地名用词库
- 建立识别规则库
 - 筛选规则、确认规则、否定规则



3.12 未登录词识别

◆ 基本方法

- 统计模型
- 通过训练语料选取阈值
- 地名初筛选
- 寻找可以利用的上下文信息
- 利用规则进一步确定地名

3.12 未登录词识别

□ 中文机构名称的识别

◆ 中文机构名称的构成

- 词法角度: 偏正式(修饰格式)的复合词
{名词|形容词|数量词|动词} + 名词
- 句法角度: “定语 + 名词性中心语”型的名词短语(定名型短语)
- 中心语: 机构称呼词, 如: 大学, 学院, 研究所, 学会, 公司等。

3.12 未登录词识别

◆ 中文机构名称的类型

- 地名，如：北京大学，武汉大学
- 人名，如：中山大学，哈佛大学
- 学科、专业 and 部门系统，如：公安部，教育委员会
- 研究、生产或经营等活动的对象，如：软件研究所，卫星制造厂
- 上述情况的综合，如：白求恩医科大学

3.12 未登录词识别

- 大机构、团体、组织和职业的名称，如：中国人民解放军洛阳外国语学院，中国发明家学会等
- 专造的机构名，如：复旦大学，四通公司，微软研究院
- 创办、工作的方式，如：某某股份公司，中央电视大学

3.12 未登录词识别

◆ 机构名称识别方法

- 找到一机构称呼词
- 根据相应规则往前逐个检查名词作为修饰名词的合法性，直到发现非法词
- 如果所接受的修饰词同机构称呼词构成一个合法的机构名称，则记录该机构名称
- 统计模型



3.13 词性标注面临的问题

3.13 词性标注面临的问题

□ 概要

词性(part-of-speech, POS)标注(tagging)的主要任务是消除词性兼类歧义。

例如，在英语中：

1) **Time flies like an arrow.**

2) **I want you to web our annual report.**

对 **Brown** 语料库的统计，**55%**词次兼类。汉语中常用词兼类现象严重，《现代汉语八百词》兼类占 **22.5%**。

3.13 词性标注面临的问题

◆ 汉语中的词性兼类现象

(1) 形同音不同，如：“好（hao3，形容词）、好（hao4，动词）”

这个人什么都好，就是好酗酒。

(2) 同形、同音，但意义毫不相干，如：“会（会议，名词）、会（能够、动词）”

每次他都会在会上制造点新闻。

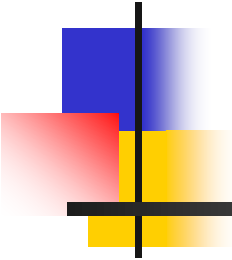
3.13 词性标注面临的问题

(3) 具有典型意义的兼类词，如：“典型(名词或形容词)”、“教育(名词或动词)”

让孩子接受那样的教育简直是对教育事业的侮辱。

(4) 上述情况的组合，如：“行(xing2, 动词/形容词; hang2, 名词/量词)”

每当他走过那行白杨树时，他都感觉好像每一棵树都在向他行注目礼。



3.14 词性标注集

3.14 词性标注集

□ 标注集的确定原则：

不同语言中，词性划分基本上已经约定俗成。
自然语言处理中对词性标记要求相对细致。

◆ 一般原则：

- 标准性：普遍使用和认可的分类标准和符号集；
- 兼容性：与已有资源标记尽量一致，或可转换；
- 可扩展性：扩充或修改。

3.14 词性标注集

◆ UPenn Treebank 的词性标注集确定原则：

- 可恢复性(recoverability)：从标注语料能恢复原词汇或借助于句法信息能区分不同词类；
- 一致性(consistency)：功能相同的词应该属于同一类；
- 不明确性(indeterminacy)：为了避免标注者在不明确的情况下任意决定标注类型，允许标注者给出多个标记（限于一些特殊情况）。

—[Marcus et al., 1993]

3.14 词性标注集

◆ UPenn Treebank 的词性标注集

□ 33 类

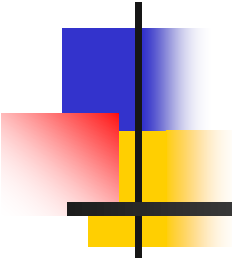
□ **NN** 名词、**NR** 专业名词、**NT** 时间名词、**VA** 可做谓语的形容词、**VC** “是”、**VE** “有”作为主要动词、**VV** 其他动词、**AD** 副词、**M** 量词等。

3.14 词性标注集

◆ 北大计算语言研究所的词性标注集

□ 26个基本词类代码，74个扩充代码，标记集中共有106个代码。

名词(n)、时间词(t)、处所词(s)、方位词(f)、数词(m)、量词(q)、区别词(b)、代词(r)、动词(v)、形容词(a)、状态词(z)、副词(d)、介词(p)、连词(c)、助词(u)、语气词(y)、叹词(e)、拟声词(o)、成语(i)、习用语(l)、简称(j)、前接成分(h)、后接成分(k)、语素(g)、非语素字(x)、标点符号(w)。



3.15 词性标注方法



3.15 词性标注方法

- 基于规则的词性标注方法
- 基于统计模型的词性标注方法
- 规则和统计方法相结合的词性标注方法
- 基于有限状态变换机的词性标注方法
- 基于神经网络的词性标注方法



3.15 词性标注方法

□ 基于规则的词性标注方法

◆ TAGGIT 词性标注系统(Bwon University)

- 86 种词性, 3300 规则
- 手工编写词性歧义消除规则
- 机器自动学习规则

3.15 词性标注方法

□ 山西大学的词性标注系统 [刘开瑛, 2000]

◆ 手工编写消歧规则

➤ 建立非兼类词典

➤ 建立兼类词典

- 词性可能出现的概率高低排列

➤ 构造兼类词识别规则

3.15 词性标注方法

(1) 并列鉴别规则

如：体现了人民的要求(N/V ?)和愿望(N, 非兼类)。

(2) 同境鉴别规则

如：一个优秀的企业必须具备一流的产品(名词, 非兼类)、一流的管理(N/V ?)和一流的服务(N/V ?)。

3.15 词性标注方法

(3) 区别词鉴别规则(区别词只能直接修饰名词)

如：他们搞的这次大型(鉴别词，非兼类) 调查(V/N ?)历时半年。

(4) 唯名形容词鉴别规则(有些形容词只能直接修饰名词)

如：重大（唯名形容词）损失（N/V ?）

巨大（唯名形容词）影响（N/V ?）

3.15 词性标注方法

➤ 根据词语的结构建立词性标注规则

(1) 词缀（前缀、后缀）规则

- 形容词：蓝茵茵，绿油油，金灿灿，...
- 数量词：一片片，一次次，一回回，...
- 人名简称：李总，张工，刘老，...
- 其他：年轻化，知识化，...{化}
 篮球赛，足球赛，...{赛}



3.15 词性标注方法

(2) 重叠词规则

一 看看，瞧瞧，高高兴兴，热热闹闹，...

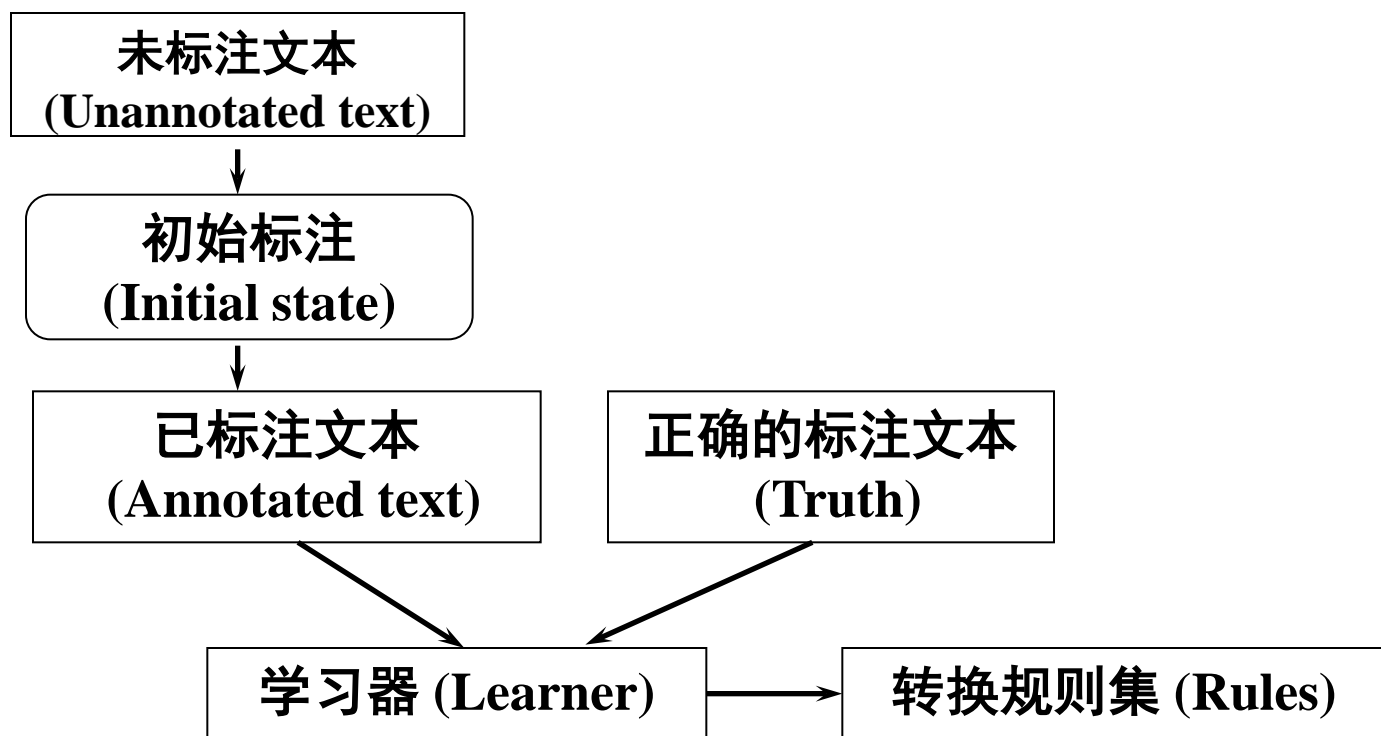
3.15 词性标注方法

□ 基于错误驱动的机器学习方法

- 初始词性赋值
- 对比正确标注的句子，自动学习结构转换规则
- 利用转换规则调整初始赋值

— [E. Brill, 1992]

3.15 词性标注方法



基于转换规则的错误驱动的机器学习方法

3.15 词性标注方法

□ 基于统计模型的词性标注方法

◆ 基于 n -gram 的语言模型

应用系统: (1) 1983年 Mashall 提出的 LOB 语料库的标注系统: CLAWS (Constituent-Likelihood Automatic Word-tagging System)
(2) DeRose 对 CLAWS 改进后 VOLSUNGA 系统 (bi-gram)。

3.15 词性标注方法

◆ 基于 HMM 的词性标注方法

- 状态集 (词性序列, 状态数: 词类符号数)
- 输出符号 (单词序列, 词汇量)
- 初始状态概率
- 状态转移概率
- 符号输出概率

— [Manning, 2001] pp. 357-359:

. **Jelink's Method**

. **Kupier's Method**

3.15 词性标注方法

□ 规则和统计相结合的词性标注方法

- ◆ 规则消歧，统计概率引导
- ◆ 或者统计方法赋初值，规则消歧

—[周强，1995；张民，1998]



3.16 分词与词性标注结果评测



3.16 分词与词性标注结果评测

□ 两种测试

- 封闭测试 / 开放测试
- 专项测试 / 总体测试

3.16 分词与词性标注结果评测

□ 评测指标

◆ **正确率**(Correct ratio/Precision, C): 测试结果中正确结果的个数占系统所有输出结果的比例。假设系统输出 N 个, 其中, 正确的结果为 n 个, 那么,

$$C = \frac{n}{N} \times 100\%$$

3.16 分词与词性标注结果评测

◆ 召回率(找回率) (Recall ratio, R): 测试结果中正确结果的个数占标准答案总数的比例。假设系统输出 N 个结果, 其中, 正确的结果为 n 个, 而标准答案的个数为 M 个, 那么,

$$R = \frac{n}{M} \times 100\%$$

3.16 分词与词性标注结果评测

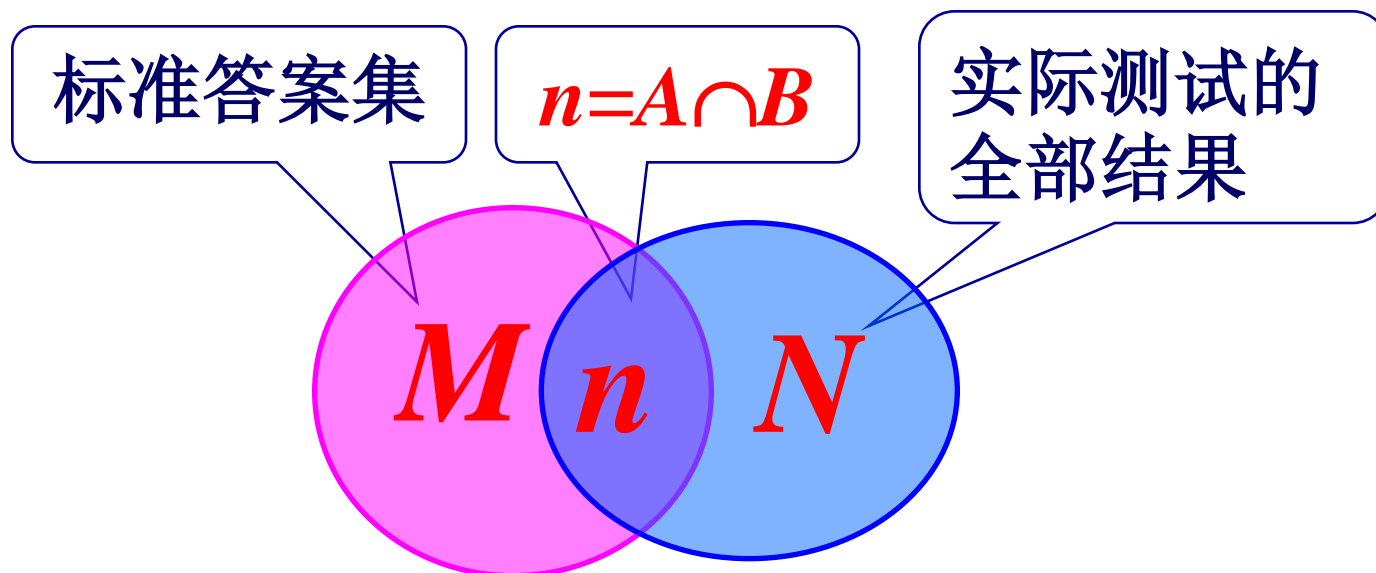
◆ F-测度值(F-Measure): 正确率与找回率的综合值。计算公式为:

$$F - measure = \frac{(\beta^2 + 1) \times C \times R}{\beta^2 \times C + R} \times 100\%$$

一般地, 取 $\beta = 1$, 即

$$F1 = \frac{2 \times C \times R}{C + R} \times 100\%$$

3.16 分词与词性标注结果评测



$$C = \frac{n}{N} \times 100\%$$

$$R = \frac{n}{M} \times 100\%$$

3.16 分词与词性标注结果评测

□ 2003年国家863评测部分结果

◆ 分词

● 最好成绩: $C=93.44\%$, $R=93.69\%$,
 $F1=93.46\%$

● 最差成绩: $C=91.42\%$, $R=89.27\%$,
 $F1=90.33\%$

3.16 分词与词性标注结果评测

◆ 词性标注

- 最好成绩: $C=87.47\%$, $R=87.52\%$,
 $F1=87.50\%$
- 最差成绩: $C=68.65\%$, $R=68.99\%$,
 $F1=68.82\%$

3.16 分词与词性标注结果评测

◆ 人名识别

- 最好成绩: $C=72.35\%$, $R=78.07\%$,
 $F1=68.33\%$
- 最差成绩: $C=27.27\%$, $R=43.29\%$,
 $F1=33.46\%$

3.16 分词与词性标注结果评测

◆ 机构名识别

- 最好成绩: $C=81.51\%$, $R=77.38\%$,
 $F1=68.56\%$
- 最差成绩: $C=4.65\%$, $R=10.60\%$,
 $F1=6.52\%$

3.16 分词与词性标注结果评测

□ 2005年SIGHAN 汉语分词评测结果(使用MSR语料)

评测方式	系统排名	性能指标				
		召回率	精确率	F-值	R_{ooV}	R_{iV}
封闭测试	最好	0.962	0.966	0.964	0.717	0.968
	最差	0.898	0.896	0.897	0.327	0.914
开放测试	最好	0.980	0.965	0.972	0.59	0.99
	最差	0.788	0.818	0.803	0.37	0.8

R_{ooV} 表示集外词的召回率, R_{iV} 表示集内词的召回率。

3.16 分词与词性标注结果评测

◆ 说明：

如果汉语自动分词与词性标注一体化进行，对于词性标注来说，可以用“召回率”衡量词性标注系统的性能，但是，如果不是分词与词性标注一体化进行，而是词性标注系统对已经切分好的汉语词汇进行词性标注，那么，一般不采用“召回率”指标衡量词性标注系统的性能。



本章小结

- 词法分析的任务（英语汉语有所不同）
- 英语形态分析
 - ◆ 单词识别
 - ◆ 形态还原
- 汉语自动分词
 - ◆ 汉语分词中的主要问题
 - ◆ 基本原则和辅助原则
 - ◆ 几种基本方法

(MM、最少分词法、统计法等)



本章小结

□ 未登录词识别

- ◆ 人名、地名、机构名等

□ 词性标注

- 问题(兼类、标注集、规范)
- 方法(规则方法、统计方法、综合方法)

□ 分词与词性标注结果评测

- 正确率、找回率、F-测度值

习题

1. 设计并实现算法用于还原英语动词。
2. 设计一个有限状态自动机用于识别缩写 {he, she}'s 是 he / she has 还是 he / she is，并编写程序实现该自动机。
3. 编写程序实现汉语逆向最大分词算法（可采用有限词表），并利用该程序对一段中文文本进行分词实验，校对切分结果，计算该程序分词的正确率、召回率及F-测度。



习题

4. 设计并实现一个汉语未登录词的识别算法(可限定条件)，并通过实验分析该算法的优缺点。
5. 了解目前常见的几种汉语词性标注集，比较它们的差异，并阐述你个人的观点。
6. 掌握各种词性标注方法的要点，了解目前汉语词性标注的几种主要方法。



习题

7. 试参考前人的工作，提出消除汉语自动分词中组合歧义的几点设想。
8. 阅读《信息处理用现代汉语分词规范》(中华人民共和国国家标准 GB13715)，了解规范的基本内容。



Thanks

谢谢!