搭配的统计分析

常宝宝
北京大学计算语言学研究所
chbb@pku.edu.cn

什么是搭配(collocation)？

- A COLLOCATION is an expression consisting of two or more words that correspond to some conventional way of saying things.

  --- Manning, C.D. & Schutze, H., *Foundations of Statistical Natural Language Processing*, The MIT press, 1999, 151

- Within the area of corpus linguistics, COLLOCATION is defined as a pair of words (the 'node' and the 'collocate') which co-occur more often than would be expected by chance.

  --- From Wikipedia, the free encyclopedia

搭配举例

- 形容词+名词

strong tea	*powerful computer*
strong computer (×)	*powerful tea* (×)
a *stiff breeze*	a *stiff wind* (×)
a *strong breeze*	a *strong wind*

- 动词+名词

knock at the *door*	*knock* on his *door*
watch the *TV*	*see* the *film*

- ……
- 词及其搭配词可能比邻出现，也可能中间间隔一些其他的词汇。

搭配构成的一般原则

- 有限组合性(non-compositionality)
  - 搭配的意义一般不是其组成词汇意义的简单相加。
    someone has *kicked* the *bucket* --- some one has died
    white wine --- yellow wine
  - 搭配在译成另外一种语言时，通常不能逐词翻译，而应作为一个整体进行翻译。
    blue film --- 黄色电影 (兰色电影)
    black tea --- 红茶 (黑茶)
- 完全不能由组成成分判断整体意义的搭配包括固定搭配(fixed collocation)和成语(idiom)等。

搭配构成的一般原则

- 有限替换性(non-substitutability)
  - 搭配的组成词汇通常不能用意义相近的词汇替换。
    white wine --- yellow wine
    strong tea --- powerful tea
    powerful computer --- strong computer
- 有限修饰性(Non-modifiability)
  - 搭配的组成词汇通常不能再被其他的词汇修饰。
    ... has *kicked the blue bucket* ...(×)

- 搭配的狭义理解和广义理解
  - 广义上的搭配 指 语法上合法的词语序列

常用的搭配提取方法

- 统计方法 与 规则方法
- 常用的统计方法
  - 基于频率的方法(frequency-based approach)
  - 基于方差的方法(variance-based approach)
  - 假设检验法(hypothesis testing)
  - 互信息法 (mutual information)

频率法

- 如果两个词总在一起出现，则这两个词很可能构成一个搭配。
- 因此可以通过统计两个词(bigram)的共现频率的方法来发现并提取搭配。
- 由于虚词的影响，通常最高频的词语组合是虚词的组合。
  (*New York times* newswire, Aug-Nov,1990)
- 可通过词类组合模式进行过滤，剔除高频的虚词组合。

$C(w^1\ w^2)$	w^1	w^2
80871	of	the
58841	in	the
26430	to	the
21842	on	the
21839	for	the
18568	and	the
16121	that	the
15630	at	the
15494	to	be
13899	in	a
13689	of	a
13361	by	the
13183	with	the
12622	from	the
11428	New	York
10007	he	said
9775	as	a
9231	is	a
8753	has	been
8573	for	a

频率法

- Justeson & Katz 用于过滤的词类组合模式

Tag Pattern	Example
A N	*linear function*
N N	*regression coefficients*
A A N	*Gaussian random variable*
A N N	*cumulative distribution function*
N A N	*mean squared error*
N N N	*class probability function*
N P N	*degrees of freedom*

$C(w^1\ w^2)$	w^1	w^2	tag pattern
11487	New	York	A N
7261	United	States	A N
5412	Los	Angeles	N N
3301	last	year	A N
3191	Saudi	Arabia	N N
2699	last	week	A N
2514	vice	president	A N
2378	Persian	Gulf	A N
2161	San	Francisco	N N
2106	President	Bush	N N
2001	Middle	East	A N
1942	Saddam	Hussein	N N
1867	Soviet	Union	A N
1850	White	House	A N
1633	United	Nations	A N
1337	York	City	N N
1328	oil	prices	N N
1210	next	year	A N
1074	chief	executive	A N
1073	real	estate	A N

搭配窗口(collocational window)

- 词语及其搭配词未必比邻出现。例如：
  - she *knocked* on his *door* (3)
  - they *knocked* at the *door* (3)
  - 100 women *knocked* on Donaldson's *door* (5)
  - a man *knocked* on the metal front *door* (5)
- 前述频率法不能直接应用，此时可以通过定义搭配窗口的方法进行解决，统计词语和窗口范围内的其他所有词的共现频率。
  大小为[-5,+5]的搭配窗口

  w_{-5} w_{-4} w_{-3} w_{-2} w_{-1} $\boldsymbol{w}$ w_{+1} w_{+2} w_{+3} w_{+4} w_{+5}

方差法

- 若w_1和w_2出现的位置相对固定，则二者有可能构成一个搭配。
- 计算w_1和w_2两个词在语料库中位置偏移的均值 μ。
- 计算位置偏移的方差 σ^2

$$\sigma^2 = \frac{\sum_{i=1}^{n}(d_i - \mu)^2}{n-1}$$

这里 n 是w_1和w_2共现的次数，d_i 是第 i 次共现二者的位置偏移，μ 是位置偏移的均值。

- 均值和方差刻画了两个词之间距离的分布情况。
- 如果两个词的距离的方差较小，则有可能二者构成一个搭配。较小的方差意味着两个词之间的距离相对固定。

方差计算举例

- 对于*knock*和*door*
- 均值是

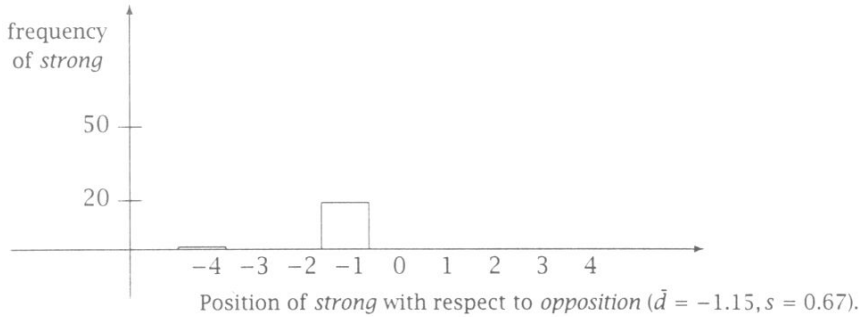
$$\frac{1}{4}(3+3+5+5)=4.0$$

- 样本标准差是

$$\sigma=\sqrt{\frac{1}{3}((3-4.0)^2+(3-4.0)^2+(5-4.0)^2+(5-4.0)^2)}\approx 1.15$$

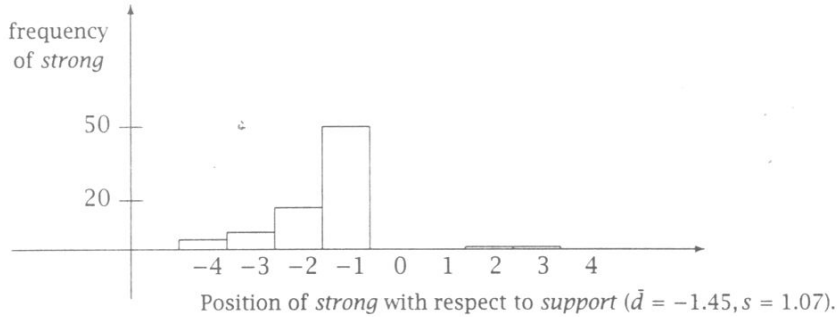
搭配词间的距离分布

strong & opposition



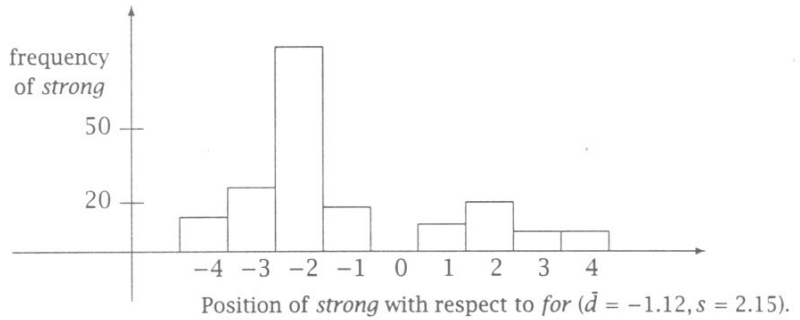
Position of *strong* with respect to *opposition* ($\bar{d} = -1.15, s = 0.67$).

strong & support



Position of *strong* with respect to *support* ($\bar{d} = -1.45, s = 1.07$).

strong & for



Position of *strong* with respect to *for* ($\bar{d} = -1.12, s = 2.15$).

方差法

s	$\bar{d}$	Count	Word 1	Word 2
0.43	0.97	11657	New	York
0.48	1.83	24	previous	games
0.15	2.98	46	minus	points
0.49	3.87	131	hundreds	dollars
4.03	0.44	36	editorial	Atlanta
4.03	0.00	78	ring	New
3.96	0.19	119	point	hundredth
3.96	0.29	106	subscribers	by
1.07	1.45	80	strong	support
1.13	2.57	7	powerful	organizations
1.01	2.00	112	Richard	Nixon
1.05	0.00	10	Garrison	said

假设检验

- 高频共现以及低方差都有可能是机会导致，未必一定意味着搭配现象。
  - 若 w_1高频出现、 w_2高频出现，则 w_1w_2通常也会高频共现，但未必构成搭配。
- 采用假设检验的方法
  - 首先假设w_1w_2是在语料库中是机会共现(co-occur by chance)，该假设通常称为原假设(null hypothesis)
  - 基于原假设，利用样本数据进行检验，若不能推翻原假设，则w_1w_2不构成搭配，若原假设不成立，则w_1w_2构成搭配，即认为备择假设成立。(alternative hypothesis)

假设检验

- 若w_1w_2为机会共现，则w_1、w_2相互独立，即
  $p(w_1w_2) = p(w_1)\,p(w_2)$
- 语料库可被视为N个*bigram*所组成的序列，定义随机变量X，若w_1w_2出现，则X=1，否则X=0。则随机序列服从二项分布。
- $p(w_1)$、$p(w_2)$可作如下估计

$$p(w_1) = \frac{c(w_1)}{N} \qquad p(w_2) = \frac{c(w_2)}{N}$$

t-检验

- *t*-检验的基本原则是假定样本数据来自均值为 μ的分布，然后通过对比样本均值和预期的均值μ之间的差异，判断样本是否来自于所假设的分布，从而推断出原假设是否成立。
- *t*-检验的计算公式(二项分布)

$$t = \frac{\bar{x} - \mu}{SE} = \frac{np_1 - np_2}{\sqrt{n}SD} = \frac{np_1 - np_2}{\sqrt{n}\sqrt{p_1(1-p_1)}} = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n}}}$$

这里 $\bar{x}$ 是样本均值，*SD* 是样本标准差，*n* 是样本大小， μ 是假设的分布的均值。

t-检验: 例子

- 假定在语料库中，*new* 出现了15,828次，*companies*出现了4,675次，语料库中共含14,307,668词次。*new companies* 出现了8 次。
- 原假设是 :

$$p(\textit{new companies}) = p(\textit{new})p(\textit{companies})$$

$$= \frac{15828}{14307668} \times \frac{4675}{14307668} \approx 3.615 \times 10^{-7}$$

- 在假设的分布中，$p_2 = 3.615 \times 10^{-7}$

t-检验: 例子

- 在样本中，$p_1 = p(new\ companies) = 8/14307668 = 5.591 \times 10^{-7}$
- 样本方差是$p_1(1-p_1)$，但由于p_1通常很小，故$p_1\,(1-p_1) \approx p_1$，即5.591×10^{-7}
- *new companies*的*t*值可计算如下

$$t = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n}}} \approx \frac{p_1 - p_2}{\sqrt{\frac{p_1}{n}}} = \frac{5.591 \times 10^{-7} - 3.615 \times 10^{-7}}{\sqrt{\frac{5.591 \times 10^{-7}}{14307668}}} = 0.999932$$

t-检验: 例子

- 由于t-值 0.999932小于2.576（置信水平为α=0.005），所以不能推翻原假设，故*new companies*不构成搭配。

	p	0.05	0.025	0.01	**0.005**	0.001	0.0005
	C	90%	95%	98%	**99%**	99.8%	99.9%
d.f.	1	6.314	12.71	31.82	63.66	318.3	636.6
	10	1.812	2.228	2.764	3.169	4.144	4.587
	20	1.725	2.086	2.528	2.845	3.552	3.850
(Z)	∞	1.645	1.960	2.326	**2.576**	3.091	3.291

- *t*-检验和其他检验常用来给搭配排序，置信水平在这里用处不大，即*t*-值越大，w_1w_2越可能是一个搭配。

t-检验

通过了*t*-检验 ($t > 2.756$) 所以: 可以拒绝原假设，因此所考察的*bigram*形成搭配

t	$C(w_1)$	$C(w_2)$	$C(w_1\ w_2)$	w_1	w_2
4.4721	42	20	20	Ayatollah	Ruhollah
4.4721	41	27	20	Bette	Midler
1.2176	14093	14776	20	like	people
0.8036	15019	15629	20	time	last

未通过*t*-检验，($t > 2.756$) 所以: 不能拒绝原假设，因此所考察的bigram不形成搭配

频率法无法做出判别，因为这些bigram共现频率相同

χ^2 检验

- 在搭配研究中，另外一种常用的检验方法是χ^2 检验。
- *t*-检验假设概率分布为正态分布，这通常并不成立，χ^2 检验没有这个要求。
- χ^2 检验的主要思想是对比预期频率以及观察频率，若二者差别较大，则拒绝原假设。
- 最简单的χ^2 检验使用2×2的联列表，分别列出不同的频率

	w_1	$\neg w_1$
w_2	a	b
$\neg w_2$	c	d

χ^2 检验

- $\chi 2$ 用下述公式进行计算

$$X^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

这里 *i* 代表联列表的行， *j*代表联列表的列, O_{ij} 代表单元格 *(i, j)*中的观察值，E_{ij} 代表相应单元格的预期值

	$w_1 = new$	$w_1 \neq new$
$w_2 = companies$	8 (*new companies*)	4667 (e.g., *old companies*)
$w_2 \neq companies$	15820 (e.g., *new machines*)	14287181 (e.g., *old machines*)

χ^2检验

- 计算预期频率 E_{ij}

Expected	w_1 = new	$w_1 \neq$ new
W_2 = companies	5.17 c(new) x c(companies) / N 15828 x 4675 / 14307676	4669.83 c(companies) x c(~new) / N 4675 x 14291848 / 14307676
$w_2 \neq$ companies	15 822.83 c(new) x c(~companies) / N 15828 x 14303001 /14307676	14 287 178.17 c(~new) x c(~companies) / N 14291848 x 14303001 / 14307676

- 原假设假定w_1和w_2相互独立，所以

$$E_{ij} = p_i(*)p_j(*)N$$

- E_{11}的值应为

$$\frac{8+4667}{N} \times \frac{8+15820}{N} \times N = 5.17$$

χ^2检验

- 计算χ^2值

$$\chi^2 = \frac{(8-5.17)^2}{5.17} + \frac{(4667-4669.83)^2}{4669} + \frac{(15\,820-15\,822.83)^2}{15\,823} + \frac{(14\,287\,181-14\,287\,178.17)^2}{14\,287\,186} \approx 1.55$$

	p	0.99	0.95	0.10	**0.05**	0.01	0.005	0.001
d.f.	1	0.00016	0.0039	2.71	**3.84**	6.63	7.88	10.83
	2	0.020	0.10	4.60	5.99	9.21	10.60	13.82
	3	0.115	0.35	6.25	7.81	11.34	12.84	16.27
	4	0.297	0.71	7.78	9.49	13.28	14.86	18.47
	100	70.06	77.93	118.5	124.3	135.8	140.2	149.4

- 查自由度为1的χ^2值表，置信度为α=0.05的χ^2值应为3.84
- 由于1.55 < 3.84，所以不能拒绝原假设，new companies不形成搭配

互信息

- 点间互信息的概念来自于信息论

$$I(x,y) = log_2 \frac{p(x,y)}{p(x)p(y)}$$

- 事件x和y间的互信息描述了:
  - 一个事件中所蕴含的关于另外一个事件的信息量
  - 两个事件之间的关联度
    - 若两个事件独立，则有$I(x,y)=0$
    - 若两个事件高度依赖，一个出现必然意味着另外一个事件出现，则有

$$I(x,y) = log_2 \frac{p(x,y)}{p(x)p(y)} = log_2 \frac{p(x)}{p(x)p(y)} = log_2 \frac{1}{p(y)}$$

互信息

- 假定:

$$c(\text{Ayatollah}) = 42$$
$$c(\text{Ruhollah}) = 20$$
$$c(\text{Ayatollah, Ruhollah}) = 20$$
$$N = 143\ 076\ 668$$

- 则:

$$I(\text{Ayatollah, Ruhollah}) = log_2 \left(\frac{\frac{20}{14\,307\,668}}{\frac{42}{14\,307\,668} \times \frac{20}{14\,307\,668}} \right) \approx 18.38$$

- 互信息也给出了$w_1 w_2$是否可能成为搭配的一种排序。

互信息

- 互信息

$I(w_1,w_2)$	$C(w_1)$	$C(w_2)$	$C(w_1\ w_2)$	w_1	w_2
18.38	42	20	20	Ayatollah	Ruhollah
17.98	41	27	20	Bette	Midler
0.46	14093	14776	20	like	people
0.29	15019	15629	20	time	last

- t-检验

t	$C(w_1)$	$C(w_2)$	$C(w_1\ w_2)$	w_1	w_2
4.4721	42	20	20	Ayatollah	Ruhollah
4.4721	41	27	20	Bette	Midler
1.2176	14093	14776	20	like	people
0.8036	15019	15629	20	time	last

- 对于上述个例，互信息和t-检验排序结果相同

互信息

- 互信息对于两个事件是否独立可以给出较好的判别。
  - 互信息值接近 0 → 两个事件相互独立
- 但对于两个事件互相依赖，仅依靠互信息值有缺陷。
  - 互信息值与事件的频率有关
  - 低频率事件有可能获得较高的互信息值，因而对于稀疏数据，互信息结果未必可靠
  - 改进措施 $c(w_1 w_2)\, I(w_1, w_2)$

部分分析技术简介

- 前面介绍的基于上下文无关文法的句法分析均属于完全句法分析。完全句法分析技术的目标是得到待分析句子完全详尽的句法结构。
- 完全句法分析技术目前还是一个远远没有得到解决的问题，这表现在目前所有的完全分析技术都还不能真正走向实用。或者只能在领域非常受限的应用系统中发挥作用。

完全句法分析技术的困难

- 健壮性问题
  - 完全句法分析技术大都假定待分析的句子是完全合乎句法的句子(well-formed sentence)。但在实际语料中，常常包含一些不那么合法的句子(ill-formed sentence)。在遇到这些句子时，完全句法分析技术大多只能以失败告终，完全句法分析技术在无法面对真实文本或不受限的应用领域。
- 完备性问题
  - 完全句法分析技术要成功工作，需要依赖完备的句法规则和完备的词典，这实际上很难做到。要使句法规则覆盖足够多的真实文本，往往意味着要撰写更多的规则，规则之间的一致性、规则的管理和维护不易保证。
- 效率问题
  - 完全分析技术大多效率不高。尤其是句子较长或句子中包含较多歧义的时候更是如此，在许多应用场合无法使用。

部分句法分析

- 由于完全句法分析技术存在困难，但又面临着越来越多的处理真实文本的压力，而且在有些应用场合，并不一定需要完全的句法结构。因此人们提出了部分句法分析(partial parsing)的概念。
- 部分分析、浅层分析、组块分析
- 部分分析技术试图通过牺牲分析深度而换取分析效率和分析结果的可靠性。
- 与完全句法分析技术不同，部分分析技术多基于正则语法或有限状态自动机，处理的可靠性和性能都比较高，但往往得不到句子详尽的句法结构。

部分句法分析

- 部分分析技术并不排斥完全分析技术的存在，相反部分分析技术的引入可以降低完全分析技术的复杂度。
- 因为部分分析技术可以用做完全分析技术的前处理，对已经经过部分分析的句子再进行完全分析，问题复杂度会有所降低。
- 部分分析技术一般力图抽出一个句子的某些结构，而不是全部结构。对英语而言，部分分析技术一般比较感兴趣的结构有：
  - 名词短语的核心部分，即不含有名词短语的名词短语(nonrecursive noun phrase)，也称为基本名词短语(BaseNP)。
  - 其他一些主要类别的短语的核心部分，如非递归的动词短语
  - 不包含其他子句的非递归简单子句

部分分析技术

- 部分分析结果示例

```
[S
    [NP The resulting formations]
    [VP are found]
    [PP along [NP an escarpment]]
][RC
    [WhNP that]
    [VP is known]
    [PP as[NP the Fischer anomly]]
]
```

- 常用的部分分析技术包括
  (1)基于HMM的部分分析技术、(2)基于转换的部分分析技术和(3)基于分层有限状态自动机的部分分析技术

基于HMM的部分分析技术

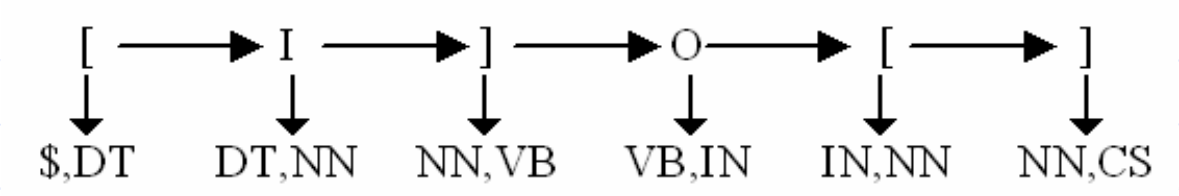
- 和词类标注方法类似，可以采用HMM来进行部分分析。
- 通过向词性序列中插入括号的方式实现组块的识别。
- 以名词组块为例，名词组块的词性排列有一定规律。

	The	prosecuter	said	in	closing	that
$	DT	NN	VBD	IN	NN	CS
	[	I]	O	[	]	

- HMM输出 词类标记对，如：*$,DT DT,NN NN,VB…*
- HMM状态 组块边界标记，如：[,],][, I, O

基于HMM的部分分析技术

- HMM状态的含义
  (1) [名词组块的开始；
  (2)] 名词组块的结尾；
  (3)][两个名词组块相邻；
  (4) I 处在组块的内部；
  (5) O 处在组块的外部。



基于HMM的部分分析技术

- 组块不能嵌套（组块内部不能再有组块）
  [[，]]，]][等不是合法HMM状态
- 不存在空组块 []不是合法HMM状态

- 训练语料　对语料进行组块标注
- 利用训练语料，统计状态转移概率以及状态输出概率
  [→[、] →]、O →]、] →I、] →][等转移概率应为 0

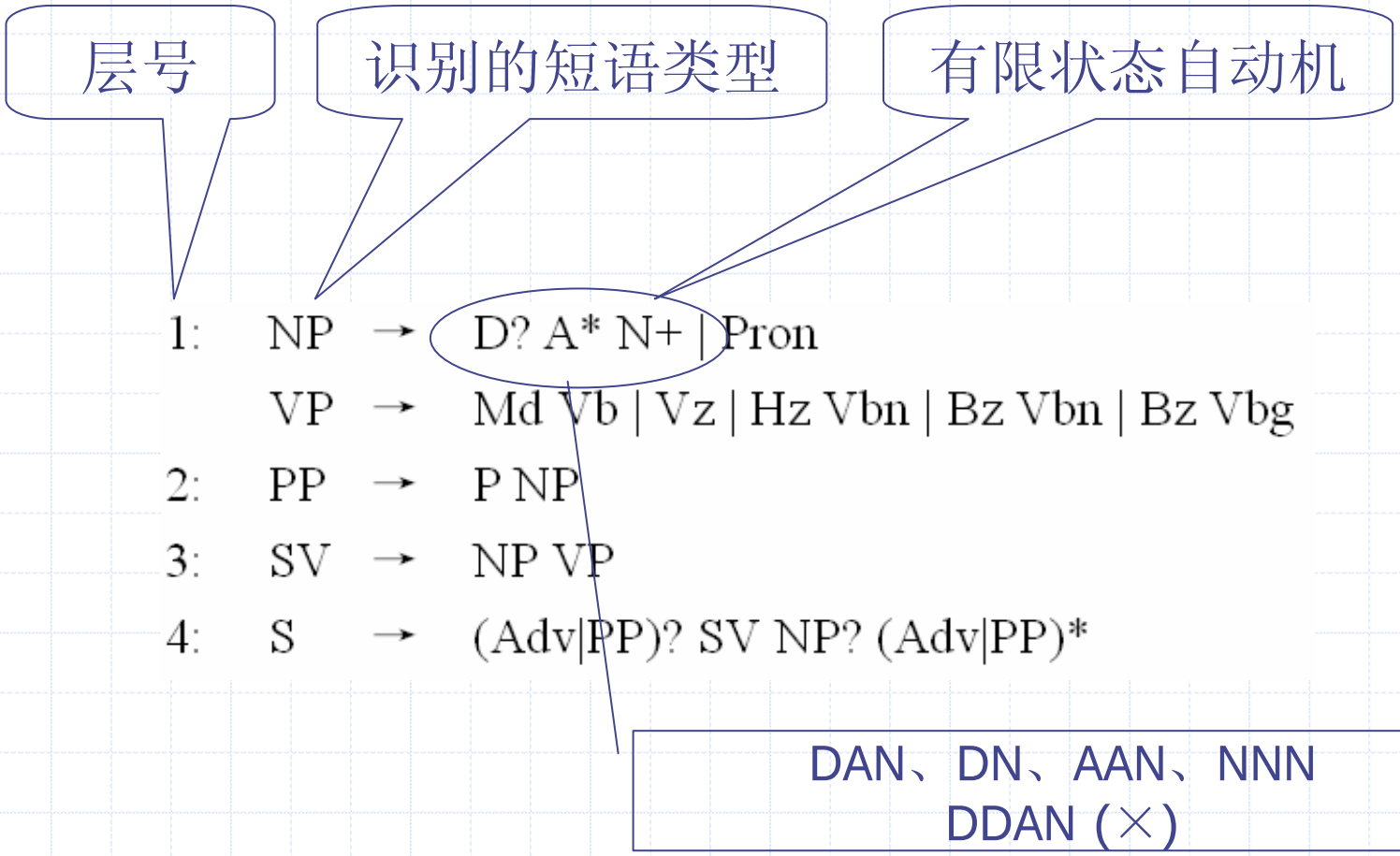
- 利用韦特比算法求解最佳状态转移序列作为组块分析结果。

分层有限状态自动机和部分分析

- Abney①于1996年提出用有限状态自动机进行部分分析。
- 有限状态自动机 同 三型文法 等价。
  描述能力有限，无法胜任描述自然语言结构的任务。
  对于 完全句法分析而言，不使用有限状态自动机。
  部分分析 分析深度较浅 可以使用有限状态机。
- 使用有限状态机进行部分分析通常效率优于完全句法分析。
- 为了识别不同类型的短语，使用分层的有限状态机
  每层识别一种或几种类型的短语。(Cascading Finite State Automata)

① Abney, S., Partial Parsing via Finite-State Cascades,
In *Proceedings of the ESSLLI'96 Robust Parsing Workshop*, 1996.

分层的有限状态机和部分分析



分层的有限状态机和部分分析

- 单层分析过程
  ① 用自动机识别符号串前缀子串
  ② 若 识别成功 输出短语类别 删除前缀子串
  　否则 输出并删除符号串最前面的符号
  ③ 转到 ① 继续处理

- 不同层次的自动机处理不同层次的符号串
  第*n*层的自动机的输入是第*n*-1层自动机的输出

  先识别NP VP，再识别PP、SV、S等。

分层的有限状态机和部分分析

> He said he read a book by a famous writer yesterday

> Pron Vz Pron Vz D N P D A N Adv

第一层识别结果

>NP VP NP VP NP P NP Adv

第二层识别结果

>NP VP NP VP NP PP Adv

第三层识别结果

>SV SV NP PP Adv

第四层识别结果

>S S

分层有限状态机和部分分析

[S
[SV [NP He] [VP said]]
]
[S
[SV [NP he] [VP read]]
[NP a book]
[PP by [NP a famous writer]]
yesterday
]

- 部分分析结果 有时缺乏 语言学解释 权益之策