(概率)上下文无关文法

上下文无关文法(CFG)

上下文无关文法，Context Free Grammar，CFG是一个四元组：

G =（N，T，P，S），其中

- N是非终结符（Nonterminals）的有限集合；
- T是终结符（Terminals）的有限集合，且N∩T=Φ；
- P是产生式（Productions）的有限集合，形如：
  A→α，其中A∈N（左部），α∈(N∪T)*（右部），
  若α=ε，则称A→ε为空产生式（也可以记为A →）；
- S是非终结符，称为文法的开始符号（Start symbol）。

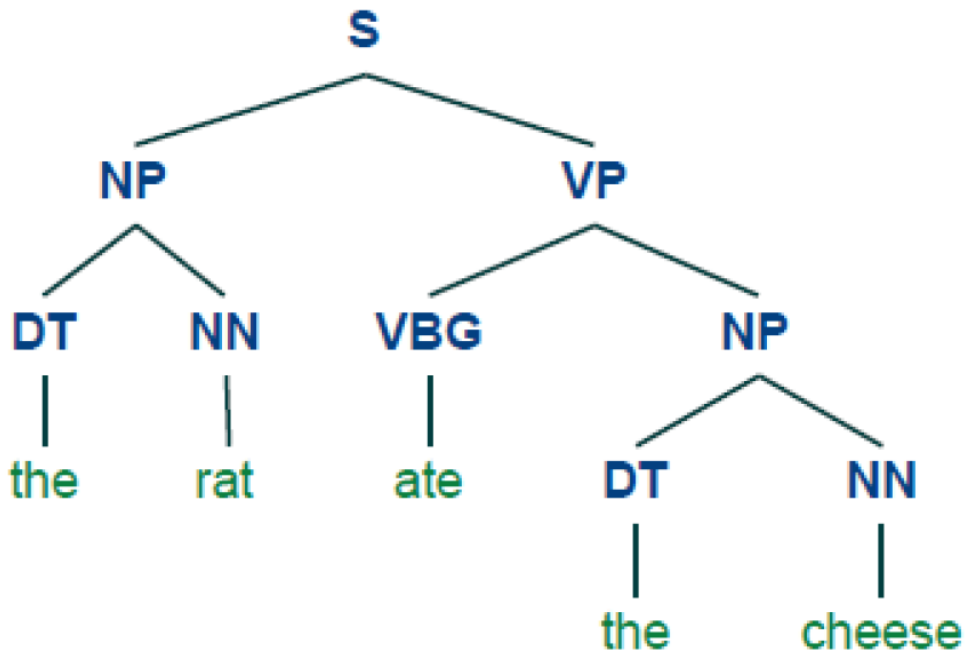
- 终结符集: rat, the, ate, cheese
- 非终结行集: S, NP, VP, DT, VBD, NN
- 产生式集合:
  - S → NP VP
  - NP → DT NN
  - VP → VBD NP
  - DT → the
  - NN → rat
  - NN → cheese
  - VBD → ate
- 开始符号: S

S → NP VP
NP → DT NN
VP → VBD NP
DT → the
NN → rat
NN → cheese
VBD → ate

- 求句子the rat ate the cheese的分析树

S → NP VP
NP → DT NN
VP → VBD NP
DT → the
NN → rat
NN → cheese
VBD → ate

- 求句子the rat ate the cheese的分析树



句法分析 (Syntactic Parsing)

- 句法解析就是指在给定词串的情况下，例如the rat ate the cheese，来识别其可能的句法结构。

咬 死 了 猎人 的 狗

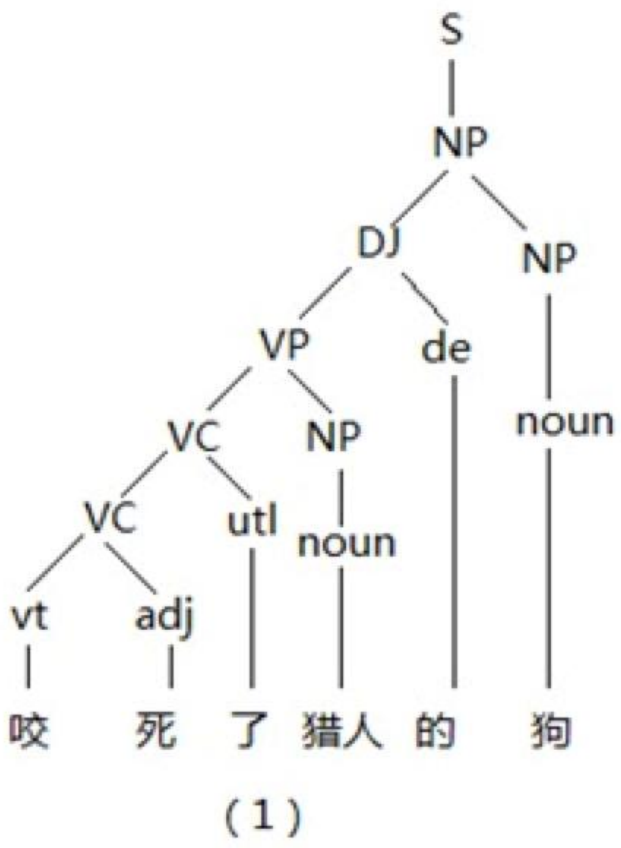
S → NP
S → VP
VP → VC NP
NP → DJ NP
NP → noun
DJ → VP de
DJ → NP de
VC → vt adj
VC → VC utl

vt → 咬
adj → 死
utl → 了
noun → 猎人
noun → 狗
de → 的

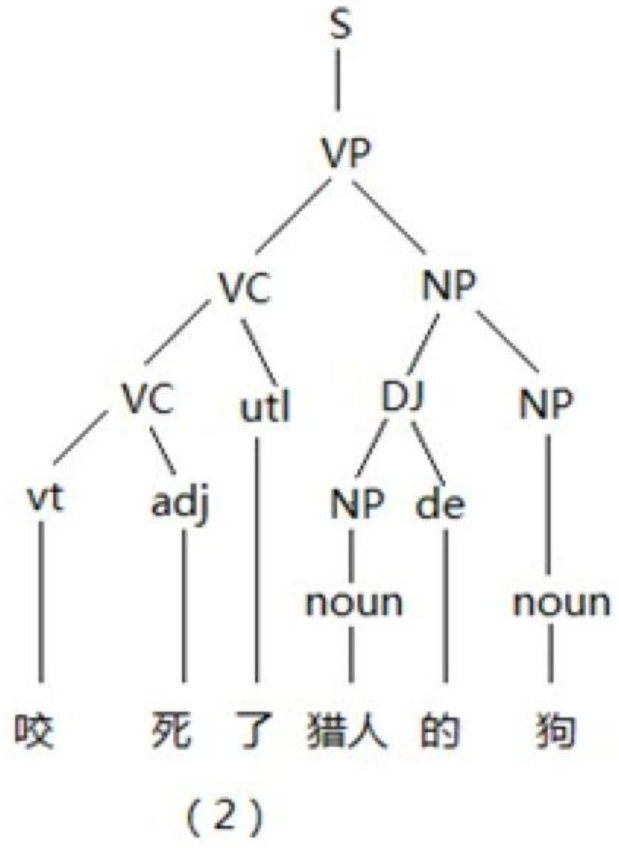
S → NP
S → VP
VP → VC NP
NP → DJ NP
NP → noun
DJ → VP de
DJ → NP de
VC → vt adj
VC → VC utl

vt → 咬
adj → 死
utl → 了
noun → 猎人
noun → 狗
de → 的

咬 死 了 猎人 的 狗



(1)



(2)

怎么判断哪个是合理的??

概率上下文无关文法(CFG)

概率上下文无关文法，Probabilistic Context Free Grammar，
CFG是一个四元组：

G =（N，T，P，S），其中

- N是非终结符（Nonterminals）的有限集合；
- T是终结符（Terminals）的有限集合，且N∩T=Φ；
- P是产生式（Productions）的有限集合，形如：
  A→α [p]，其中A∈N（左部），α∈(N∪T)*（右部），
  若α=ε，则称A→ε为空产生式（也可以记为A →），
  p 是0~1 之间的值，表示该产生式的概率P(α|A)；
- S是非终结符，称为文法的开始符号（Start symbol）。

- A→α [p]
- 产生式左边 LHS (left hand side)
- 产生式右边RHS (right hand side)

- $\mathrm{NP} \rightarrow \mathrm{DT\ NN} \qquad [\mathrm{p} = 0.45]$
- $\mathrm{NN} \rightarrow \mathrm{leprechaun} \qquad [\mathrm{p} = 0.0001]$

对于一个给定的LHS要满足加和等于1。

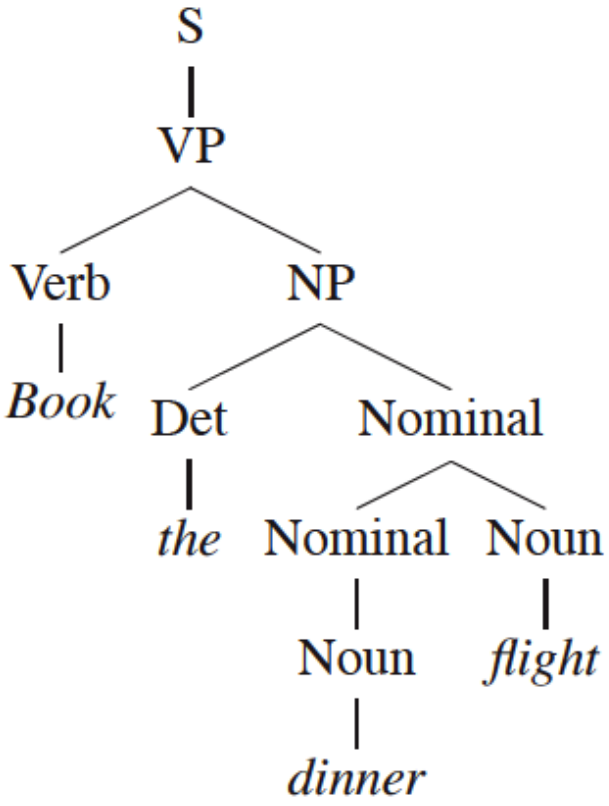
$S \rightarrow NP\ VP$	[.80]	$Det \rightarrow that$ [.10] | a [.30] | the [.60]
$S \rightarrow Aux\ NP\ VP$	[.15]	$Noun \rightarrow book$ [.10] | $flight$ [.30]
$S \rightarrow VP$	[.05]	| $meal$ [.15] | $money$ [.05]
$NP \rightarrow Pronoun$	[.35]	| $flights$ [.40] | $dinner$ [.10]
$NP \rightarrow Proper\text{-}Noun$	[.30]	$Verb \rightarrow book$ [.30] | $include$ [.30]
$NP \rightarrow Det\ Nominal$	[.20]	| $prefer$; [.40]
$NP \rightarrow Nominal$	[.15]	$Pronoun \rightarrow I$ [.40] | she [.05]
$Nominal \rightarrow Noun$	[.75]	| me [.15] | you [.40]
$Nominal \rightarrow Nominal\ Noun$	[.20]	$Proper\text{-}Noun \rightarrow Houston$ [.60]
$Nominal \rightarrow Nominal\ PP$	[.05]	| TWA [.40]
$VP \rightarrow Verb$	[.35]	$Aux \rightarrow does$ [.60] | can [40]
$VP \rightarrow Verb\ NP$	[.20]	$Preposition \rightarrow from$ [.30] | to [.30]
$VP \rightarrow Verb\ NP\ PP$	[.10]	| on [.20] | $near$ [.15]
$VP \rightarrow Verb\ PP$	[.15]	| $through$ [.05]
$VP \rightarrow Verb\ NP\ NP$	[.05]	
$VP \rightarrow VP\ PP$	[.15]	
$PP \rightarrow Preposition\ NP$	[1.0]	

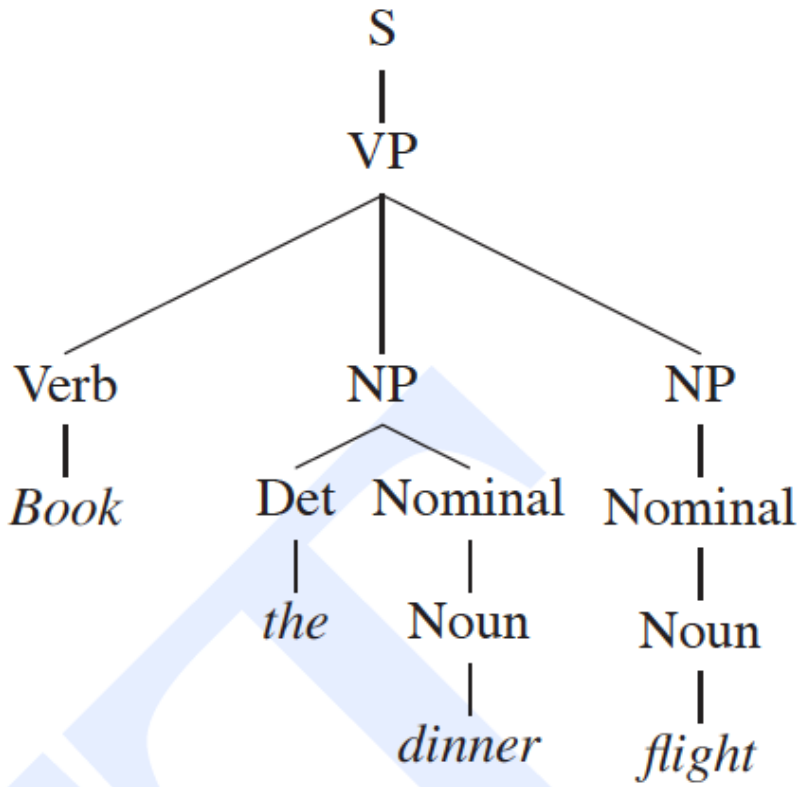
$$P(T) = \prod_{i=1}^{n} P(\mathrm{RHS}_i | \mathrm{LHS}_i)$$

句法树T的概率

一般来说，概率值大的更可能是正确的句法树 。

$$P(T) = P(S \rightarrow VP) \times$$
$$P(VP \rightarrow Verb\ NP) \times$$
$$P(Verb \rightarrow \textit{Book}) \times$$
$$P(NP \rightarrow Det\ Nominal) \times$$
$$P(Det \rightarrow \textit{the}) \times$$
$$P(Nominal \rightarrow Nominal\ Noun) \times$$
$$P(Noun \rightarrow \textit{dinner}) \times$$
$$P(Noun \rightarrow \textit{flight}) = 2.2 \times 10^{-6}$$





$$P = 6.1 \times 10^{-7}$$

- 给定一个PCFG (概率上下文无关文法)和一个句子, 如何构造该句子概率最大的句法树?

- CNF (Chomsky Normal Form)
- CYK算法

CNF (Chomsky Normal Form)

- 如果一个上下文无关文法的每个产生式的形式为:

  A → BC 或 A → a

  即规则的右部或者是两个非终结符或者是一个终结符.

- 任何CFG都可以转变成一个弱等价的Chomsky范式语法。

CYK算法 (CKY)

- 给定一个句子$s = w_1 w_2, \ldots, w_n$, 和一个上下文无关文法PCFG，$G=(T, N, S, R)$;

- 定义一个跨越单词 i到j的概率最大的语法成分π:

$$\pi(i,j,X) \quad (i , j \in 1\ldots n , X \in N),$$

- 目标是找到一个属于 $\pi [1 , n , S]$的所有树中概率最大的那棵。

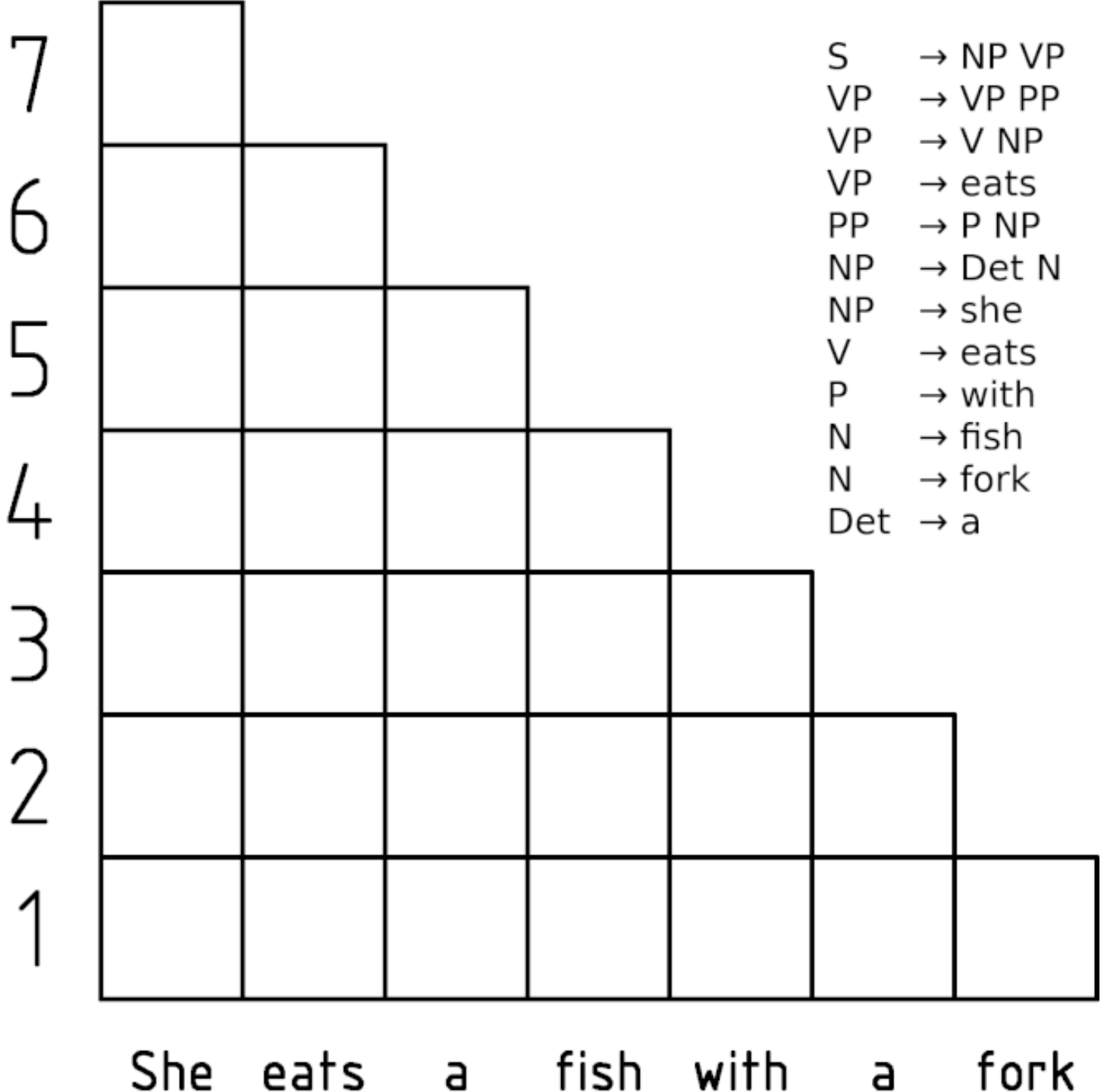
CYK算法用于PCFG下的句法分析：

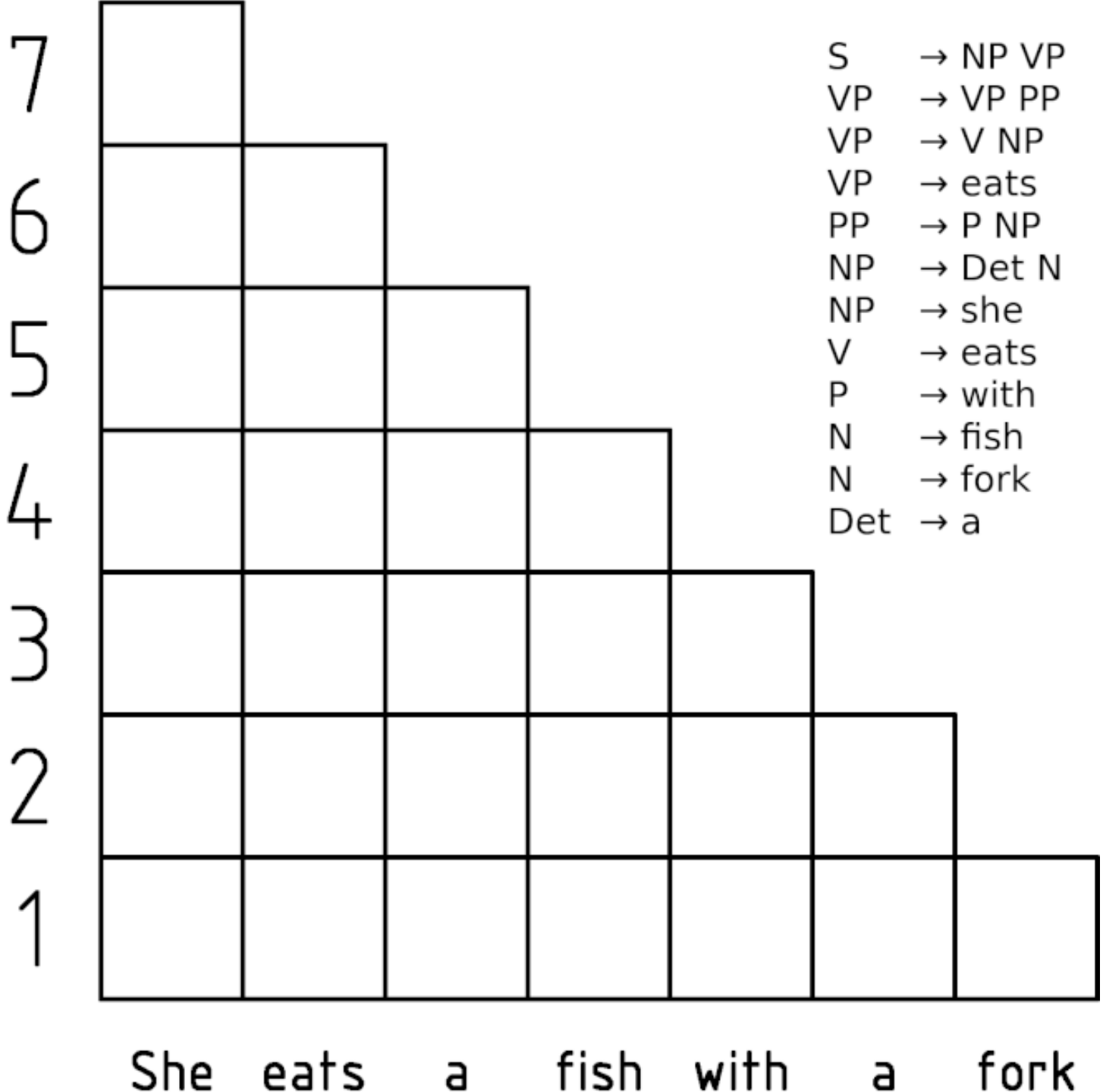
- 基本定义：for all i=1,…,n, X ∈N

  $\pi(i,i,X) = q(X \rightarrow wi)$ (if X → wi 没有出现在语法中，则定义q(X → wi)=0)

- 递归定义：for all i=1,…,n, j=(i+1),…,n, X ∈N

  $\pi(i,j,X) = max(q(X \rightarrow YZ) \times \pi(i,k,Y) \times \pi(k+1,j,Z))$ (i≤k≤j−1)


7
6
5
4
3
2
1
S → NP VP
VP → VP PP
VP → V NP
VP → eats
PP → P NP
NP → Det N
NP → she
V → eats
P → with
N → fish
N → fork
Det → a
She
eats
a
fish
with
a
fork


7
6
5
4
3
2
1
S → NP VP
VP → VP PP
VP → V NP
VP → eats
PP → P NP
NP → Det N
NP → she
V → eats
P → with
N → fish
N → fork
Det → a
She
eats
a
fish
with
a
fork

S → NP VP	0.9
S → VP	0.1
VP → V NP	0.5
VP → V	0.1
VP → V @VP_V	0.3
VP → V PP	0.1
@VP_V → NP PP	1.0
NP → NP NP	0.1
NP → NP PP	0.2
NP → N	0.7
PP → P NP	1.0

N → *people*	0.5
N → *fish*	0.2
N → *tanks*	0.2
N → *rods*	0.1
V → *people*	0.1
V → *fish*	0.6
V → *tanks*	0.3
P → *with*	1.0

求 fish people fish tanks的最优分析树.