上海海事大学
SHANGHAI MARITIME UNIVERSITY

自然语言处理

2024–2025 学年第 2 学期

信息工程学院　谢雨波

词汇分析

什么是词?

- **词（Word）**是形式和意义相结合的单位，也是语言中能够独立运用的最小单位
- 掌握一个词汇意味着知道其读音和语义
  - 例如：英文单词“cat”具有的语义是“猫”，读音为“/kæt/”
- 自然语言处理算法中词通常也是基本单元
- 词的处理也是自然语言处理中重要的底层任务，是句法分析、文本分类、语言模型等任务的基础

语言中的词汇

- **词（Word）**通常是由**语素（Morpheme）**构成
  - 语素又称词素，是语言中意义的最小单元
  - 语素与词不同，语素不能够独立运用而词可以
  - 只包含一个语素的词语称为简单词（Simple Word）
  - 包含多个语素的词称为复杂词（Complex Word）
  - 例如：“电灯”，包含“电”和“灯”两个语素

    “incoming”，包含“in-”, “come”, “-ing”三个语素

语言中的词汇

- 根据词在语言中的用途的不同，词还可以被划分为**实义词（Content Word）**和**功能词（Function Word）**
  - 实义词包含事物、行为、属性和观念等概念
  - 例如：椅子 / chair，桌子 / table，书 / book
  - 功能词则是指没有清楚词汇意义或与之有关的明显概念的词
  - 例如：因为 / for，自从 / since，这个 / the

词的形态学

- 由于社会的约定俗成，词的形式具有**服从于某种规则的内在结构**
- 研究单词的内部结构和其构成方式的学科称为**形态学（Morphology）**，又称**构词学**
- 词是由一个或多个语素构成，语素主要分成两类: **词根（Lemma）**和**词缀（Affix）**

词的形态学

- **词根**：也称为原形或字典形，是指能在字典中查的到的语素，通常是一个词最主要的语素
- **词缀**：是其他附着在原形上的语素，帮助在原形基础上衍生出新词，包含前缀（Prefix）、中缀（Infix）、后缀（Suffix）等

例如：

英语单词 unhappy 中，happy 为原形，un- 为前缀

邦托克语单词 fumikas（是强壮的）中，fikas（强壮）为原形，-um- 为中缀

俄语单词 barabanshchik（鼓手）中，baraban（鼓）为原形，-shchik 为后缀

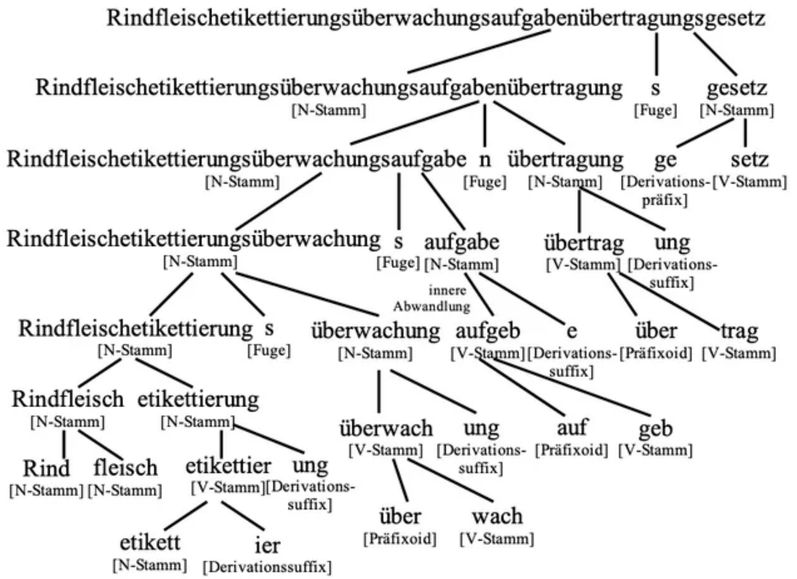
词的形态学

- 有些语言的单词通常只包含一个或者两个语素，但是有一些语言的单词则包含**多达十个以上**的语素

Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz

牛肉标签监管任务委托法

20 个语素！



词的形态学

英语中的常见词形变化

词形变化	说明	举例
屈折 Inflection	通过“词根 + 词缀”的方式构成和原形“同一类型”的词	名词后加 -s 后缀　复数名词（cat+s） 动词后加 -ed 后缀　动词的过去式（walk+ed）
派生 Derivation	通过“词根 + 词缀”的方式构成和原形“不同类型”的词	employ 添加后缀 -ee 变为 employee meaning 添加后缀 -less 变为 meaningless
复合 Compounding	通过组合多个词根构成一个新词，也称组合词	home + work → homework water + proof → waterproof
附着 Cliticization	通过‘词根 + 附着语”的方式“附着”在词根上	I’m 中的’m 代表 am 附着在 I 上 We’re 中’re 代表 are
截搭 Blending	两个词语各自的一部分拼接起来构成新词	smoke（烟）+ fog（雾）→ smog（烟雾） spoon（勺子）+ fork（叉子）→ spork（叉勺）
缩略 Acronym	短语中多个单词首字母组合在一起组合成词	NLP 代表 Natural Language Processing IT 代表 Information Technology
截短 Clipping	将长的单词截为较短的单词	demonstration 简化为 demo refrigerator 简化为 fridge

词的词性

- **词性（Part-of-Speech，POS）**也称词类，是根据词在句子中扮演的语法角色以及与周围词的关系对词的分类

- 例如：表示事物的名字（“钢琴”），地点（“上海”）通常被归为名词（Noun），而表示动作（“踢”），状态（“存在”）的词被归为动词（Verb）

词的词性

- **名词（Noun）**是表示人、物、地点以及抽象概念的一类词

例如：

1）专有名词：Shanghai（上海）New York（纽约）

2）类名词：city（城市）bird（鸟）

3）集体名词：family（家庭）army（军队）

4）物质名词：water（水）light（光）

5）抽象名词：music（音乐）honesty（诚实）

词的词性

- **动词（Verb）**是表示动作或状态的一类词

例如：

1) 及物动词：Boys fly kites.（男孩们放风筝）
2) 不及物动词：Birds fly.（鸟会飞）
3) 连系动词：The rose smells sweet.（玫瑰花香）
4) 助动词：I may have meet him before.（我以前应该见过他）
5) 限定动词：John reads papers every day.（约翰每天都读论文）
6) 不限定动词：I hope to see you this morning.（我希望早上见到你）
7) 短语动词：Tom called up George.（汤姆给乔治打了电话）

词的词性

- **形容词（Adjective）**是用来描写或修饰名词的一类词

例如：

1) 简单形容词：
   a) 由一个单词构成：good（好的）long（长的）
   b) 由现在分词构成：interesting（令人感兴趣的）
   c) 由过去分词构成：learned（博学的）
2) 复合形容词：duty-free（免税的）hand-made（手工制作的）
3) 限制性形容词：an Italian dish （一道意大利菜）
4) 描述性形容词：a delicious Italian dish （一道美味的意大利菜）

词的词性

- **副词（Adverb）**是用来修饰动词、形容词、其他副词以及全句的词

例如：

1) 简单副词：just（刚刚）only（仅仅）
2) 复合副词：somehow（不知怎地）somewhere（在某处）
3) 派生副词：interesting → interestingly（有趣地）
4) 方式副词：quickly（迅速）awkwardly（笨拙地）
5) 方向副词：outside（外面）inside（里面）
6) 时间副词：recently（最近）always（总是）
7) 强调副词：very（很）fairly（相当）

词的词性

- **数词（Numeral）**是表示数目多少或者先后顺序的一类词

例如：

1）基数词：one（1）nineteen（19）

2）序数词：first（第一）fiftieth（第五十）

词的词性

- **代词（Pronoun）**是代替名词以及起名词作用的短语、子句和句子的一类词

例如：

1）人称代词：主格（I）、宾格（me）
2）物主代词：形容词性物主代词（my）、名词性物主代词（mine）
3）自身代词：myself
4）相互代词：each other, one another
5）指示代词：this, that, these, those
6）疑问代词：who, whom, whose, which, what
7）关系代词：who, whom, whose, which, that, as
8）不定代词：some, something, somebody, someone, any, anything, anybody, anyone, no, nothing, nobody, no one

词的词性

- **冠词（Article）**是置于名词之前，说明名词所指的人或事物的一种功能词
- 冠词不能够离开名词而独立存在
- 英语中冠词有三种冠词：
  - 定冠词（Definite Article）：“the”
  - 不定冠词（Indefinite Article）：“a/an”
  - 零冠词（Zero Article）

词的词性

- **介词（Preposition）**又称前置词，是用于表示名词或相当于名词的词语与句中其它词语的关系的一类词

例如：

1) 简单介词： at, in, of, since
2) 复合介词： as for, as to, out of
3) 二重介词： from under, from behind
4) 短语介词： according to, because of
5) 分词介词： including, regarding

词的词性

- **连词（Conjunction）**是连接单词、短语、从句或句子的一类词

例如：

1) 简单连词：and, or, but, if
2) 关联连词：both ... and, not only ... but also
3) 分词连词：supposing, considering
4) 短语连词：as if, as long as, in order that
5) 并列连词：and, or, but, for
6) 从属连词：that, whether, when, because

词的词性

- **感叹词（Interjection）**是用来表示喜怒哀乐等情绪或情感的一类词

例如：

Oh, it’s you.　　啊，是你。

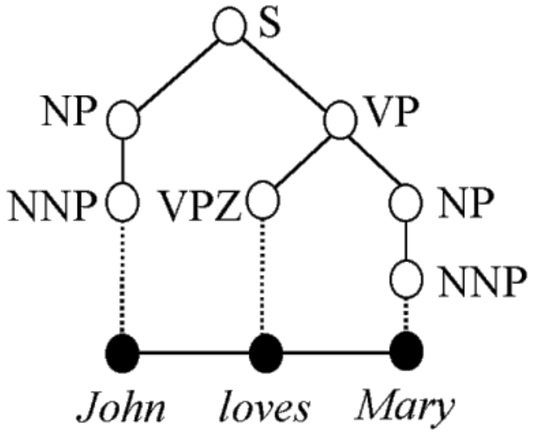
Ah, how pitiful!　呀，多可惜！

词的词性

- 在语言学研究中，对于词性划分的标准、依据甚至目的等都存在大量分歧。到目前为止，还**没有一个被广泛认可的统一划分标准**
- 在不同的语料集中所采用的**划分粒度和标记符号也都不尽相同**
  - 英语宾州树库（Penn Treebank）使用了 36 种不同的词性
  - 汉语宾州树库（Chinese Penn Treebank）中汉语词性被划分为 34 类
  - 布朗语料库（Brown Corpus）中则使用了具有 86 个词性

英语宾州树库（Penn Treebank）

- **树库（Treebank）**：包含大量句子的语言资源，经过了语法分析，以树状结构的形式标注了它们的句法结构

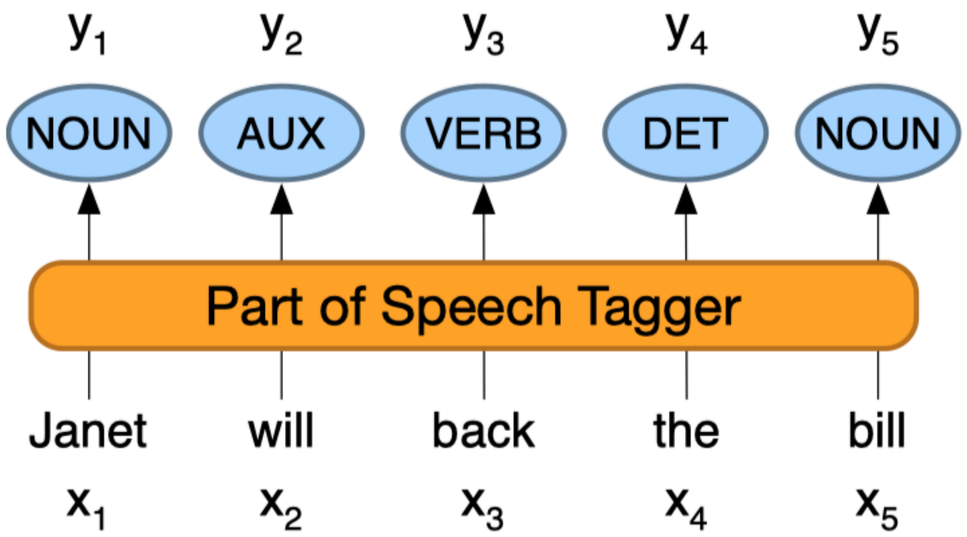


标签	描述	标签	描述
CC	并列连词	CD	数字
DT	限定词	EX	there
FW	外来词	IN	介词或从属连词
JJ	形容词	JJR	形容词比较级
JJS	形容词最高级	LS	列表项标记
MD	情态助动词	NN	名词单数
NNS	名词复数	NNP	专有名词单数
NNPS	专有名词复数	PDT	前限定词
POS	所有格结束词	PRP	人称代名词
PRP$	物主代词	RB	副词
RBR	副词比较级	RBS	副词最高级
RP	小品词	SYM	符号
TO	to	UH	叹词
VB	动词	VBD	动词过去式
VBG	动词现在进行式	VBN	动词过去分词
VBP	动词一般现在式非第三人称单数	VBZ	动词一般现在式第三人称单数
WDT	Wh-限定词	WP	Wh-代词
WP$	所有格 Wh-代词	WRB	Wh-副词

词性标注

词性标注

- **词性标注（Part-of-Speech Tagging）**：为输入文本中的每一个单词都分配一个词性（POS）的过程



词性标注

- 词性标注是一个**去歧义化（Disambiguation）**的过程

例如：*book* that flight
hand me that *book*

Does *that* flight serve dinner?
I thought *that* your flight was earlier.

词性标注

- 词性标注算法的准确率**非常高**！
- Wu and Dredze (2019) 发现：在 Universal Dependency (UD) Treebank 的 15 种语言上达到了超过 97% 的准确率
- 在英语树库上的准确率也达到了 97%（无论使用哪种算法：HMMs, CRFs, BERT 等等）
- 人类的表现：~97%

词性标注

- 词性标注难吗?

	WSJ	Brown
词典中的词:		
非歧义（1 个标签）	44,432 **(86%)**	45,799 **(85%)**
歧义（2+ 个标签）	7,025 **(14%)**	8,050 **(15%)**
文本中的词:		
非歧义（1 个标签）	577,421 **(45%)**	384,349 **(33%)**
歧义（2+ 个标签）	711,780 **(55%)**	786,646 **(67%)**

词性标注

- 大部分的词没有词性歧义，但是实际使用的寻常词汇大都有词性歧义

earnings growth took a *back*/**JJ** seat

a small building in the *back*/**NN**

a clear majority of senators *back*/**VBP** the bill

Dave began to *back*/**VB** toward the door

enable the country to buy *back*/**RP** debt

I was twenty-one *back*/**RB** then

词性标注

- 尽管如此，很多词都很容易去歧义化，因为它们的不同词性的可能性不会相同
  - 例如：a 既可能是冠词也可能是字母，但更有可能是冠词

词性标注的一个简单算法

给定一个有词性歧义的词，直接返回它在训练语料库中出现最频繁的词性。

准确率：92%!

隐马尔可夫模型

- **隐马尔可夫模型：Hidden Markov Model, HMM**
- HMM 是一个**概率的序列模型**
  - 给定一个序列（单词、字母、语素、句子，等等），HMM 计算所有标签序列上的一个概率分布，并选择最好的标签序列

马尔可夫链

- **马尔可夫链（Markov Chain）**：描述一系列随机变量的概率模型，且具有马尔可夫性质
- **马尔可夫性质（Markov Property）**：**系统的下一个状态只依赖于当前状态，而与之前的状态无关**（这意味着，给定当前状态，未来的状态与过去的状态是独立的）

例如：连续七天的天气变化，明天的天气只和今天的天气有关

- 对于一系列状态变量 $s_1, s_2, \cdots, s_i$

$$P(s_i = a \mid s_1, \cdots, s_{i-1}) = P(s_i = a \mid s_{i-1})$$

马尔可夫链

- 马尔可夫链可以用**状态空间**、**转移矩阵**和**初始状态分布**来定义：
  - **状态空间**：$Q=\{q_1,q_2,\cdots,q_N\}$，系统可能处于的所有状态
  - **转移概率矩阵**：$A=\begin{pmatrix}a_{11} & \cdots & a_{1N}\\ \vdots & \ddots & \vdots\\ a_{N1} & \cdots & a_{NN}\end{pmatrix}$，从任一状态转移到另一状态的概率，其中 a_{ij} 表示从状态 q_i 转移到状态 q_j 的概率，$\sum_{j=1}^{N}a_{ij}=1,\forall i$
  - **初始状态分布**：$\pi=(\pi_1,\pi_2,\cdots,\pi_N)$，系统在初始时刻处于每个状态的概率，$\sum_{i=1}^{N}\pi_i=1$

马尔可夫链

状态空间：

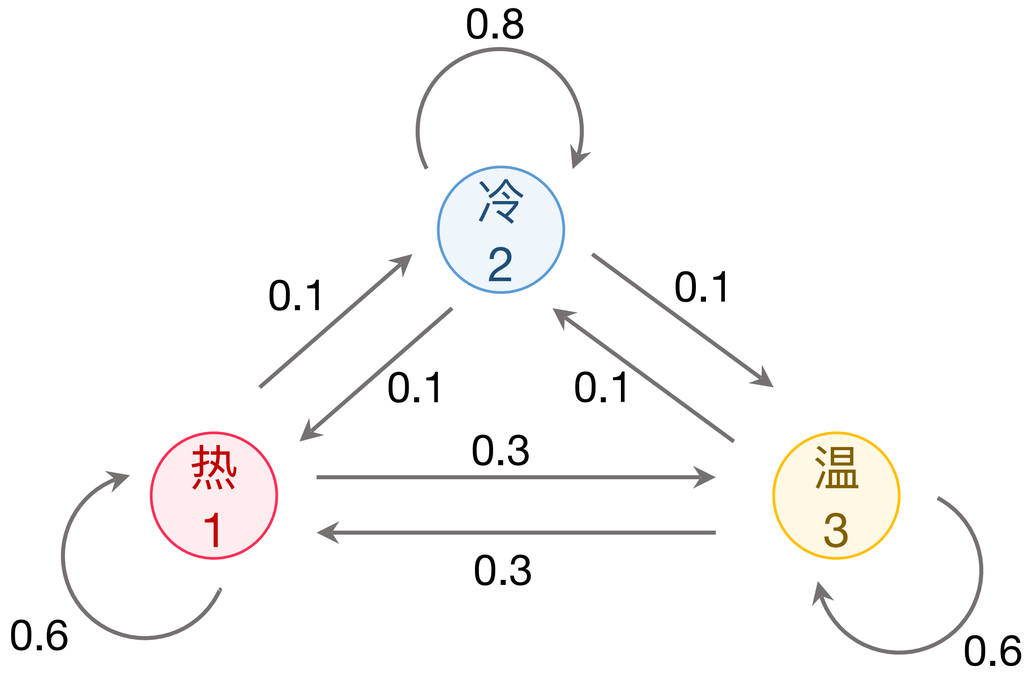
$$Q = \{热, 冷, 温\}$$

转移概率矩阵：

$$A = \begin{pmatrix} 0.6 & 0.1 & 0.3 \\ 0.1 & 0.8 & 0.1 \\ 0.3 & 0.1 & 0.6 \end{pmatrix}$$

初始状态分布：

$$\pi = (0.1,\ 0.7,\ 0.2)$$

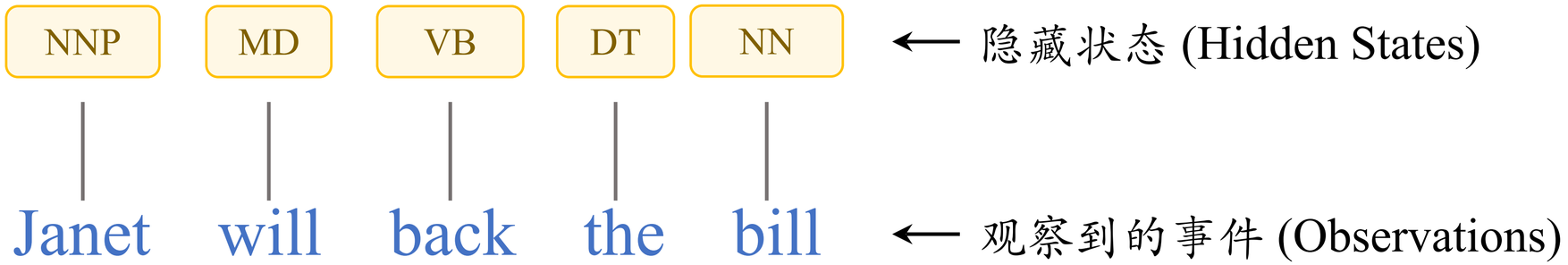


$$P(热热热热) = 0.1 \times 0.6 \times 0.6 \times 0.6$$

$$P(冷热冷热) = 0.7 \times 0.1 \times 0.1 \times 0.1$$

隐马尔可夫模型

- 马尔可夫链可以计算一系列观测事件的概率
- 然而，在某些情况下，我们感兴趣的事件是隐藏的



隐马尔可夫模型

- 隐马尔可夫模型的核心组成部分：
  - **状态空间**： $Q=\{q_1,q_2,\cdots,q_N\}$，系统可能处于的所有隐藏状态
  - **观察空间**： $V=\{v_1,v_2,\cdots,v_M\}$，系统所有可以观察到的结果
  - **转移概率矩阵**： $A=\begin{pmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{N1} & \cdots & a_{NN} \end{pmatrix}$，从任一状态转移到另一状态的概率，其中 a_{ij} 表示从状态 q_i 转移到状态 q_j 的概率，$\sum_{j=1}^{N} a_{ij}=1,\forall i$
  - **观察概率矩阵**： $B=\begin{pmatrix} b_{11} & \cdots & b_{1M} \\ \vdots & \ddots & \vdots \\ b_{N1} & \cdots & b_{NM} \end{pmatrix}$，也称发射概率（Emission Probability），其中 b_{ij} 表示从状态 $q_i \in Q$ 生成观察 $v_j \in V$ 的概率，$\sum_{j=1}^{M} b_{ij}=1,\forall i$
  - **初始状态分布**： $\pi=(\pi_1,\pi_2,\cdots,\pi_N)$，系统在初始时刻处于每个状态的概率，$\sum_{i=1}^{N}\pi_i=1$

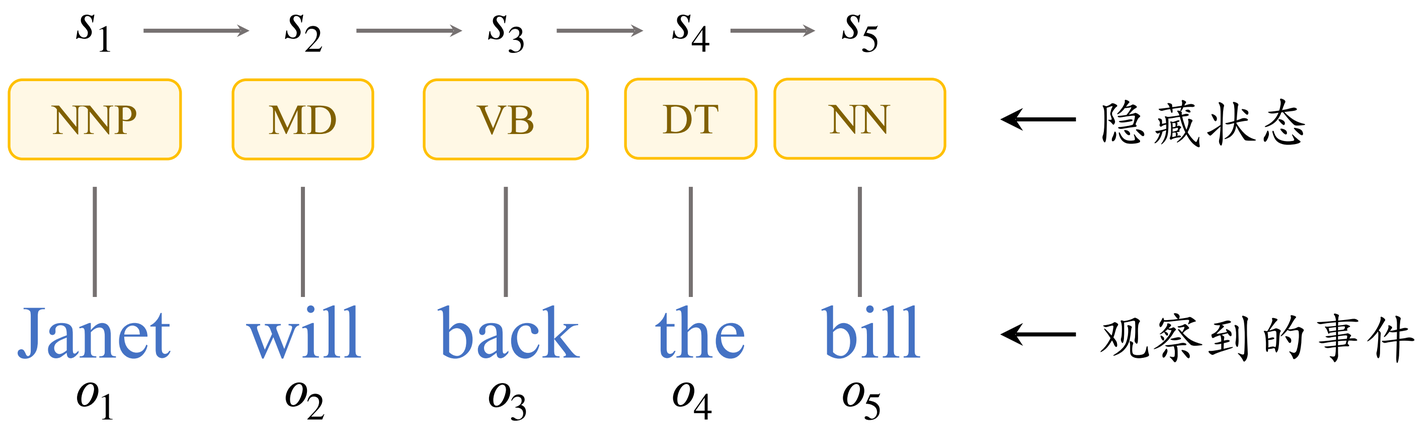
HMM 的性质

- **马尔可夫性质**：某一状态的概率只依赖于前一状态

$$P(s_i = a \mid s_1, \cdots, s_{i-1}) = P(s_i = a \mid s_{i-1})$$

- **观察独立性**：观察 o_i 的概率只依赖于产生它的状态 s_i

$$P(o_i \mid s_1, \cdots, s_i, \cdots, s_T, o_1, \cdots, o_{i-1}, o_{i+1}, \cdots, o_T) = P(o_i \mid s_i)$$



HMM 词性标注

- 转移概率 $P(s_t \mid s_{t-1})$

$$P(s_t \mid s_{t-1}) = \frac{C(s_{t-1}, s_t)}{C(s_{t-1})}$$

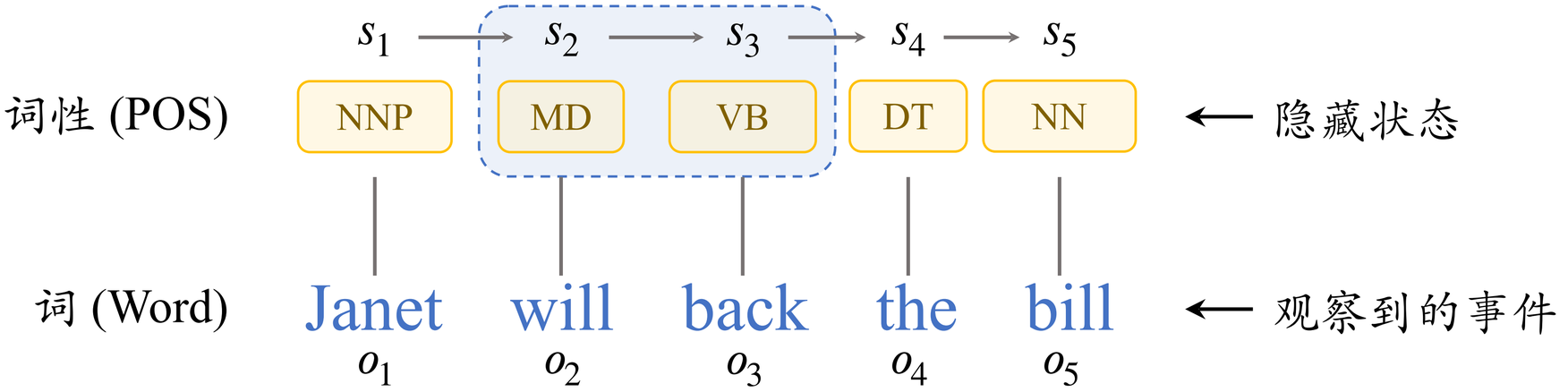
s_{t-1}, s_t 出现的次数

s_{t-1} 出现的次数

（最大似然估计）

WSJ 语料库中：

$$P(\text{VB} \mid \text{MD}) = \frac{C(\text{MD}, \text{VB})}{C(\text{MD})} = \frac{10471}{13124} \approx 0.80$$



HMM 词性标注

- 观察概率 $P(o_t \mid s_t)$

$$P(o_t \mid s_t) = \frac{C(s_t, o_t)}{C(s_t)}$$

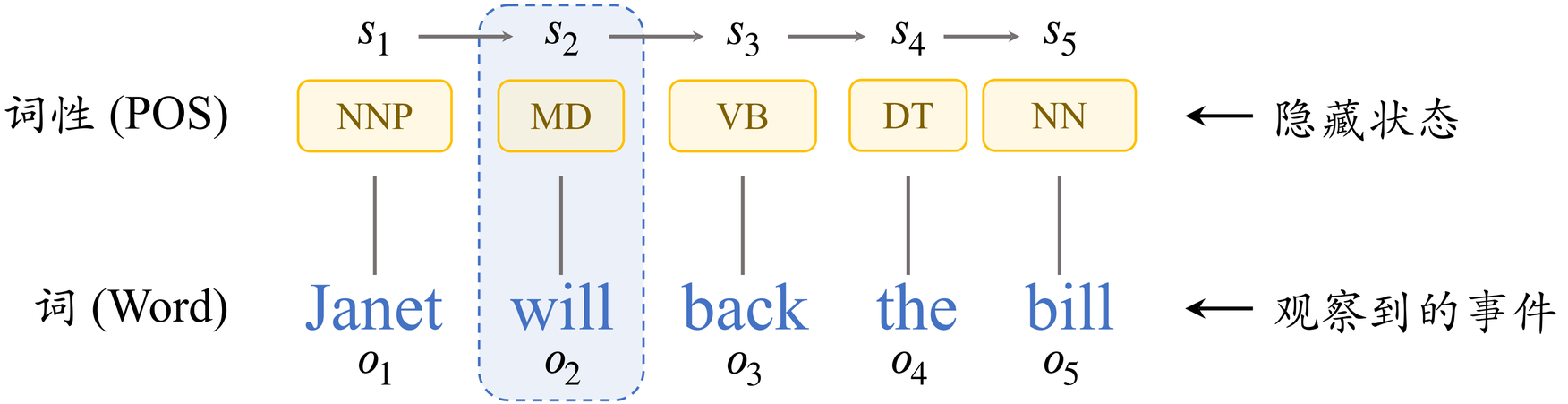
（最大似然估计）

$C(s_t, o_t)$：s_t, o_t 出现的次数

$C(s_t)$：s_t 出现的次数

WSJ 语料库中：

$$P(\text{will} \mid \text{MD}) = \frac{C(\text{MD}, \text{will})}{C(\text{MD})} = \frac{4046}{13124} \approx 0.31$$



HMM 词性标注

- **解码（Decoding）：**

给定 HMM $\lambda = (A, B)$ 和观察序列 $O = o_1, o_2, \cdots, o_T$，找到概率最大的隐藏状态序列 $S = s_1, s_2, \cdots, s_T$

$$\hat{s}_{1:T} = \arg\max_{s_1 \cdots s_T} P(s_1 \cdots s_T | o_1 \cdots o_T)$$

HMM 词性标注

$$\hat{s}_{1:T} = \arg\max_{s_1 \cdots s_T} P(s_1 \cdots s_T | o_1 \cdots o_T)$$

$$= \arg\max_{s_1 \cdots s_T} \frac{P(o_1 \cdots o_T | s_1 \cdots s_T) P(s_1 \cdots s_T)}{P(o_1 \cdots o_T)}$$

贝叶斯公式

$$= \arg\max_{s_1 \cdots s_T} P(o_1 \cdots o_T | s_1 \cdots s_T) \, P(s_1 \cdots s_T)$$

去掉分母

观察独立性

马尔可夫性质

$$\approx \arg\max_{s_1 \cdots s_T} \prod_{t=1}^{T} P(o_t | s_t) \prod_{t=1}^{T} P(s_t | s_{t-1})$$

HMM 词性标注

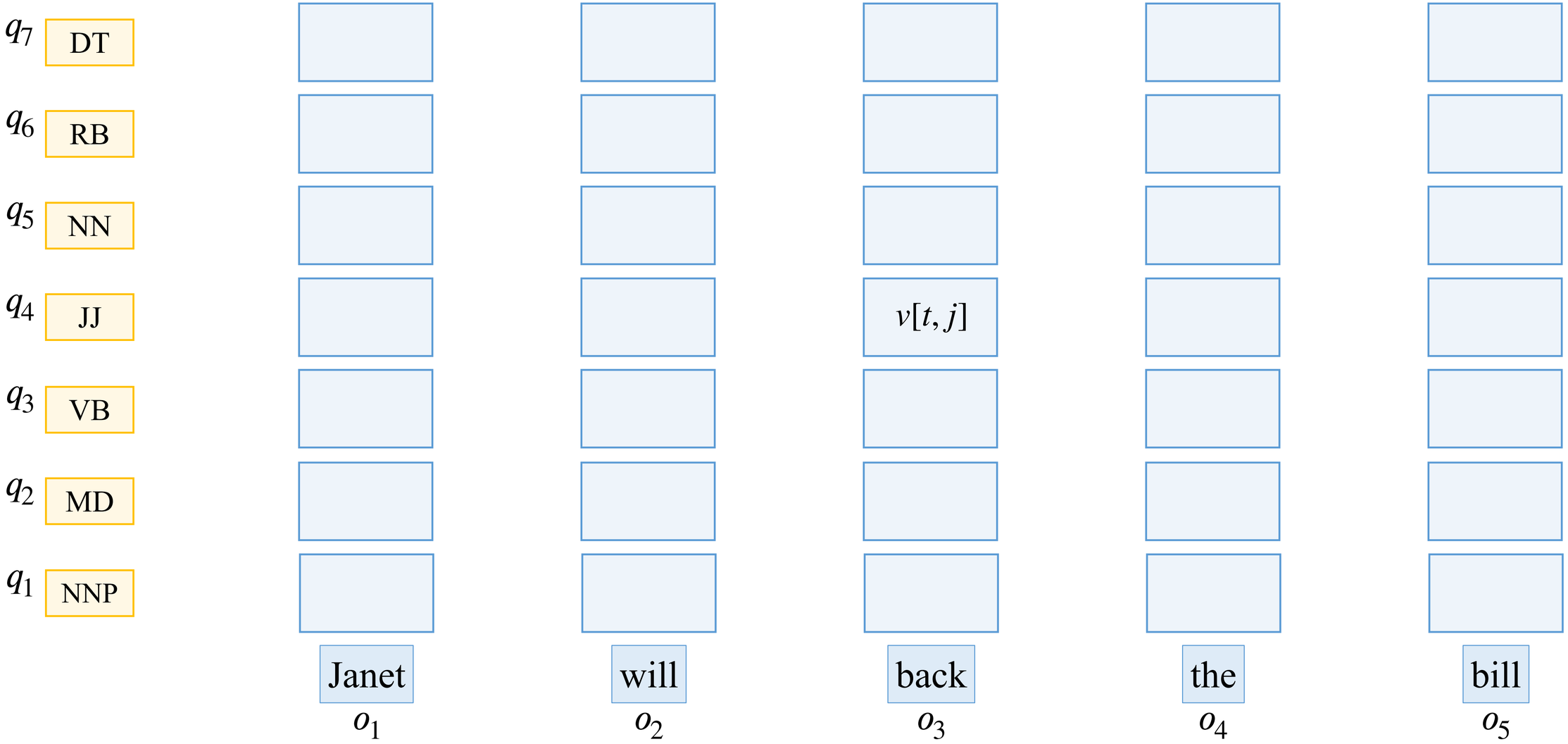
$$\hat{s}_{1:T} = \arg\max_{s_1 \cdots s_T} P(s_1 \cdots s_T | o_1 \cdots o_T)$$

$$\approx \arg\max_{s_1 \cdots s_T} \prod_{t=1}^{T} P(o_t | s_t) \prod_{t=1}^{T} P(s_t | s_{t-1})$$

$$= \arg\max_{s_1 \cdots s_T} \prod_{t=1}^{T} P(o_t | s_t) P(s_t | s_{t-1})$$

观察概率　　转移概率

HMM 解码：Viterbi 算法



HMM 解码：Viterbi 算法

$$v[t,j] = \max_{s_1 \cdots s_{t-1} \in Q} P(s_1, \cdots, s_{t-1}, o_1, \cdots, o_t, s_t = q_j)$$

在观察到 $o_1, o_2, \cdots, o_t$ 并且经过概率最大的前 $t-1$ 个状态 $s_1, \cdots, s_{t-1}$ 之后，系统处于状态 q_j 的概率

$$= \max_{i=1:N,\, s_1 \cdots s_{t-2} \in Q} P(s_1, \cdots, s_{t-2}, s_{t-1} = q_i, o_1, \cdots, o_t, s_t = q_j)$$

$$= \max_{i=1:N,\, s_1 \cdots s_{t-2} \in Q} P(s_1, \cdots, s_{t-2}, o_1, \cdots, o_{t-1}, s_{t-1} = q_i) \cdot P(o_t, s_t = q_j | \cancel{s_1, \cdots, s_{t-2}, o_1, \cdots, o_{t-1}}, s_{t-1} = q_i)$$

$$= \max_{i=1:N,\, s_1 \cdots s_{t-2} \in Q} P(s_1, \cdots, s_{t-2}, o_1, \cdots, o_{t-1}, s_{t-1} = q_i) \cdot P(o_t, s_t = q_j | s_{t-1} = q_i)$$

HMM 解码：Viterbi 算法

$$v[t,j] = \max_{i=1:N,\, s_1 \cdots s_{t-2} \in Q} P(s_1, \cdots, s_{t-2}, o_1, \cdots, o_{t-1}, s_{t-1} = q_i) \cdot P(o_t, s_t = q_j \mid s_{t-1} = q_i)$$

$$= \max_{i=1:N,\, s_1 \cdots s_{t-2} \in Q} P(s_1, \cdots, s_{t-2}, o_1, \cdots, o_{t-1}, s_{t-1} = q_i) \cdot P(s_t = q_j \mid s_{t-1} = q_i) \cdot P(o_t \mid s_t = q_j, \cancel{s_{t-1} = q_i})$$

$$= \max_{i=1:N} \max_{s_1 \cdots s_{t-2} \in Q} P(s_1, \cdots, s_{t-2}, o_1, \cdots, o_{t-1}, s_{t-1} = q_i) \cdot P(s_t = q_j \mid s_{t-1} = q_i) \cdot P(o_t \mid s_t = q_j)$$

$$= \max_{i=1:N} v[t-1,i]\, a_{ij}\, b_j(o_t)$$

转移概率

观察概率

$$v[t,j] = \max_{i=1:N} v[t-1,i] a_{ij} b_j(o_t)$$

动态规划

HMM 解码：Viterbi 算法

function Viterbi($o_1, \cdots, o_T, A, B, \pi$):

创建数组 $v[T, N]$

for i **from** 1 **to** N **do**　　// 初始化

　$v[1,i] \leftarrow \pi_i \cdot b_i(o_1)$

　backtrack$[1,i] \leftarrow 0$

for t **from** 2 **to** T **do**　　// 迭代

　for j **from** 1 **to** N **do**

　　$v[t,j] \leftarrow \max_{i=1:N} v[t-1,i] \cdot a_{ij} \cdot b_j(o_t)$

　　backtrack$[t,j] \leftarrow \arg\max_{i=1:N} v[t-1,i] \cdot a_{ij} \cdot b_j(o_t)$

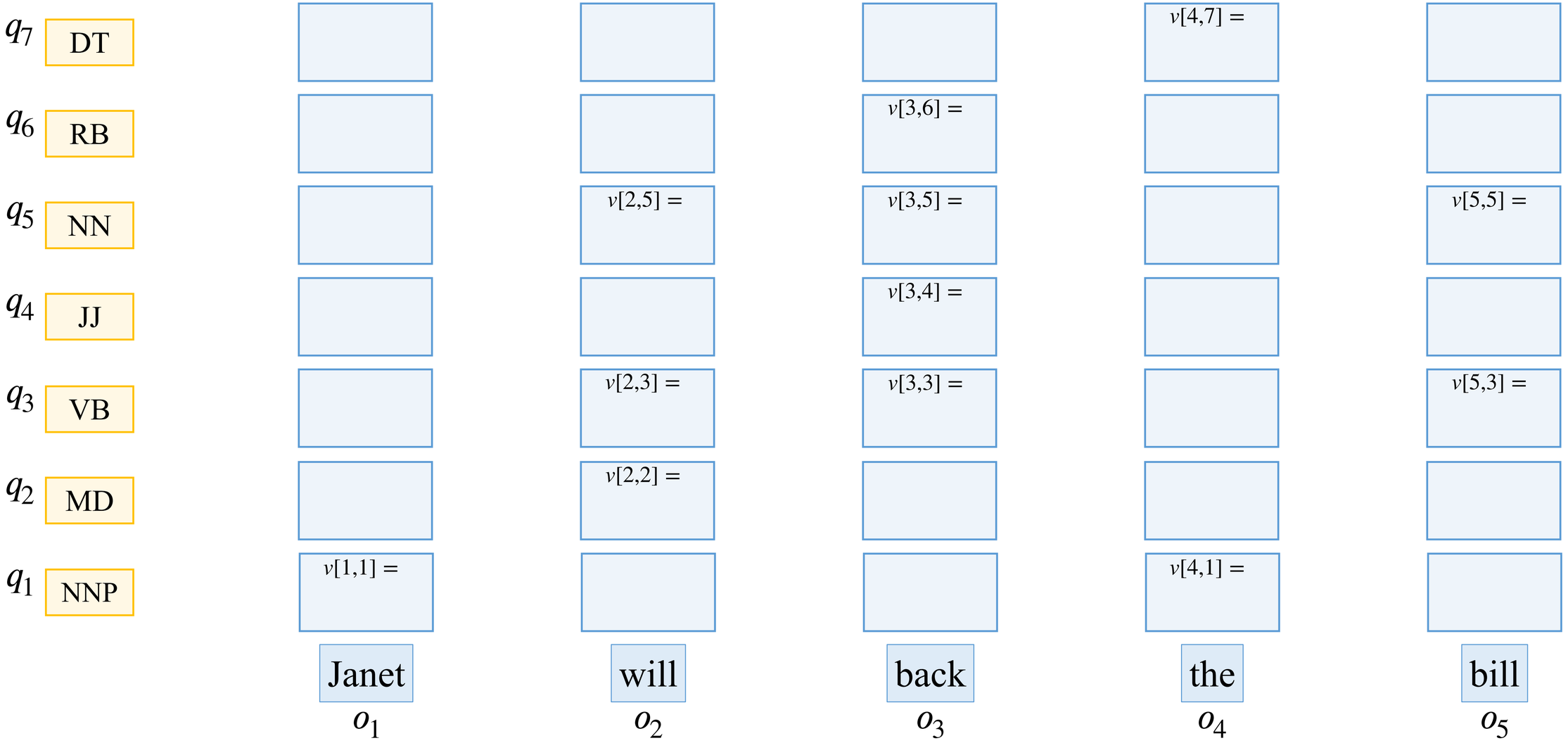
bestprob $\leftarrow \max_{i=1:N} v[T,i]$

bestpointer $\leftarrow \arg\max_{i=1:N} v[T,i]$

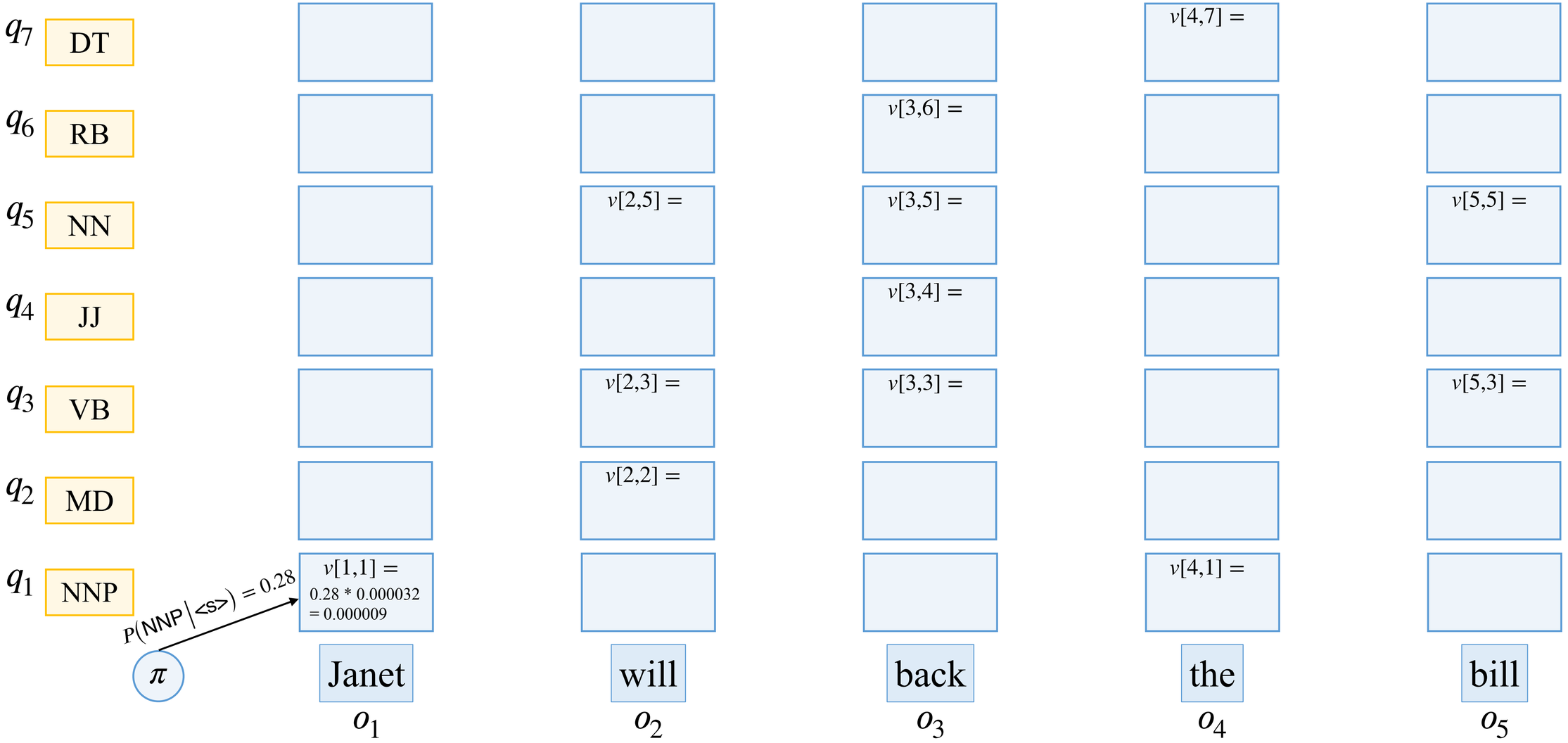
bestpath ← 从 bestpointer 开始，根据 backtrack 往前回溯

返回 bestpath, bestprob

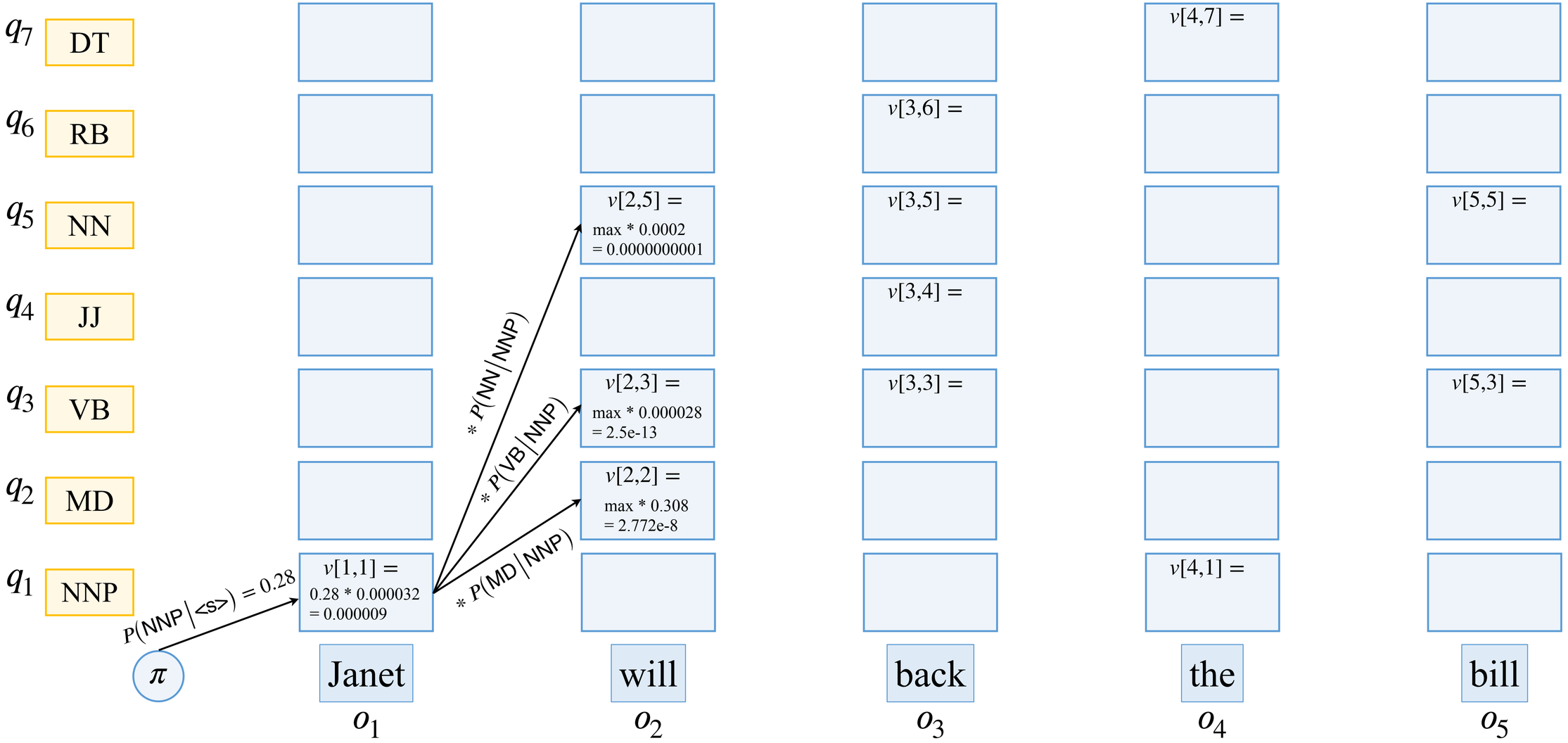
HMM 解码：Viterbi 算法



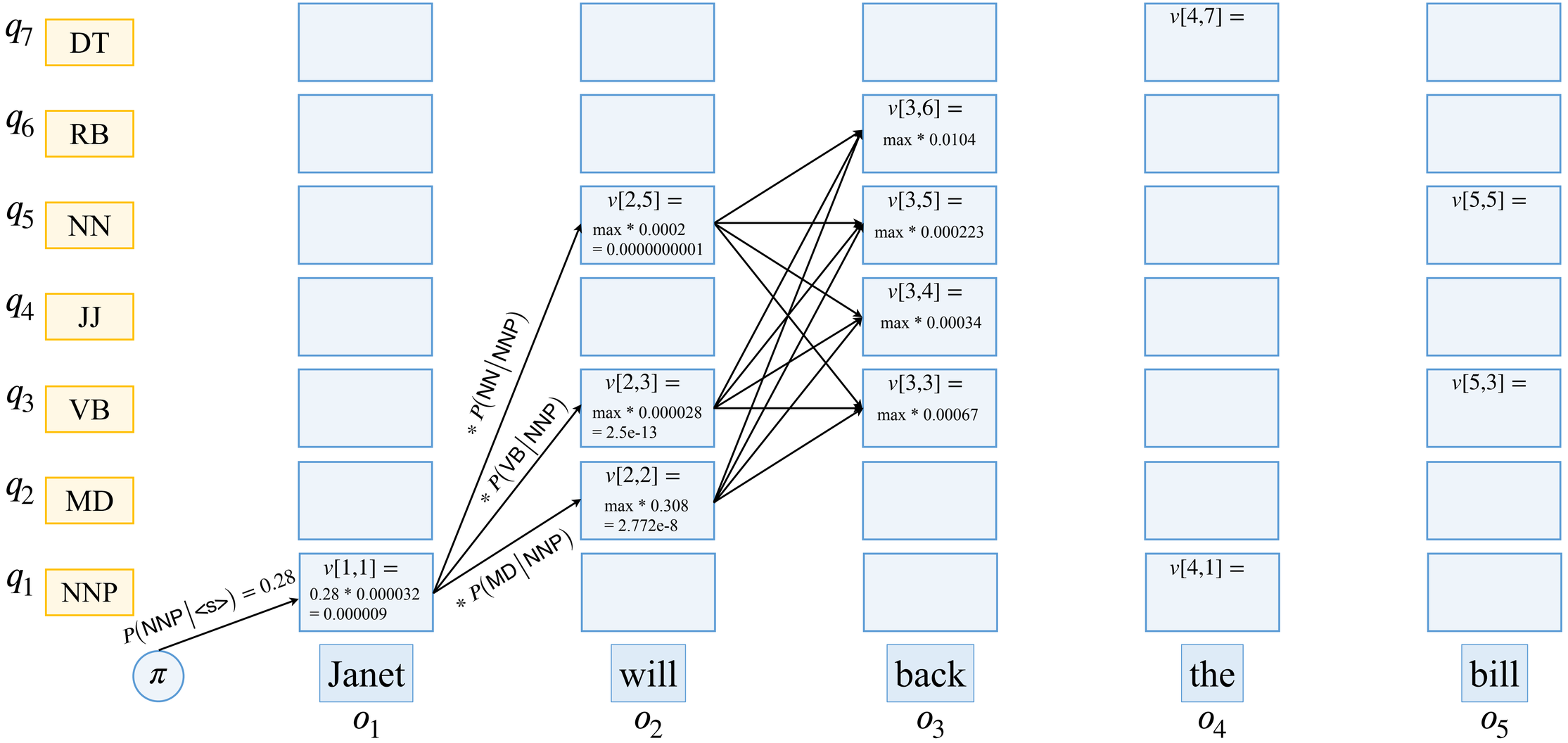
HMM 解码：Viterbi 算法



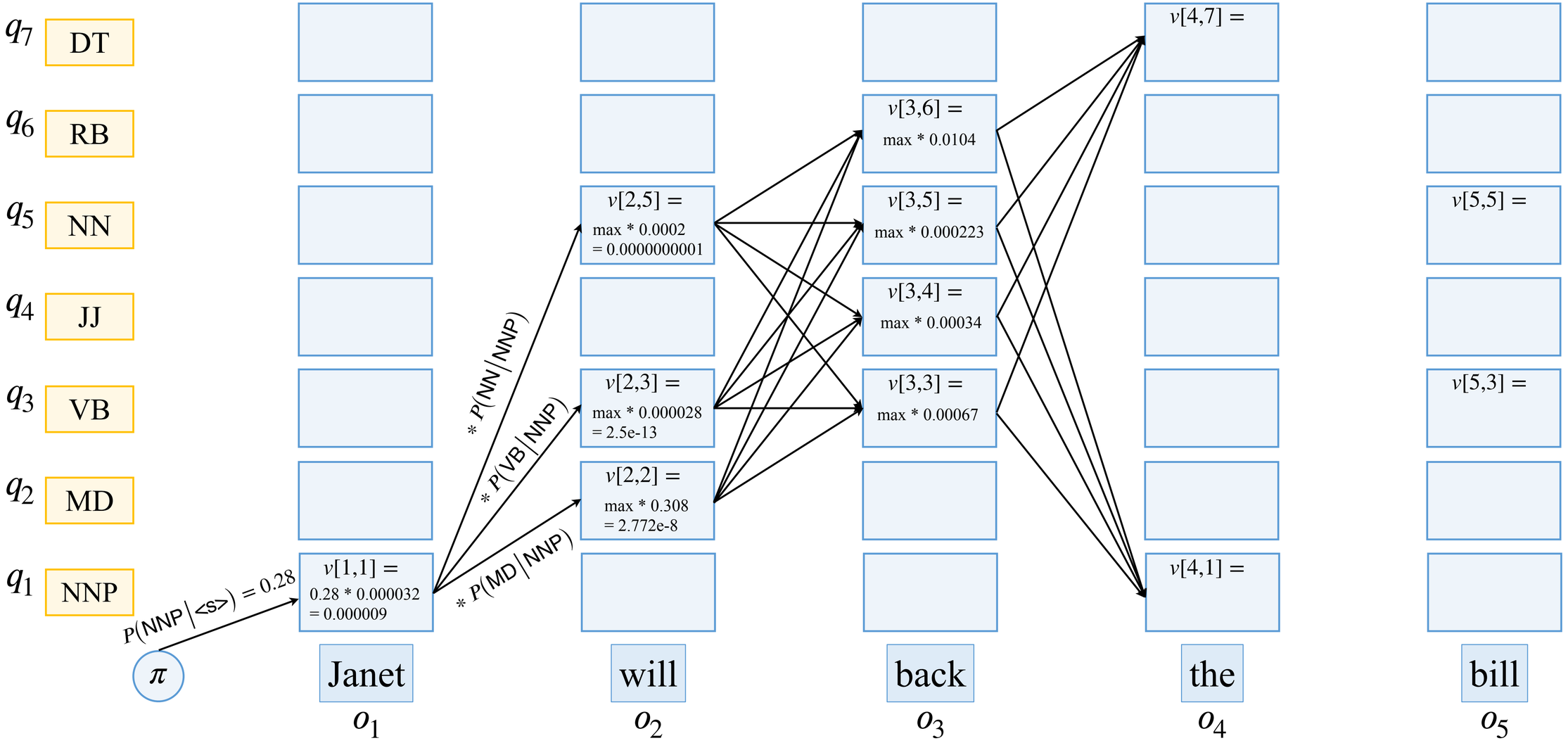
HMM 解码：Viterbi 算法



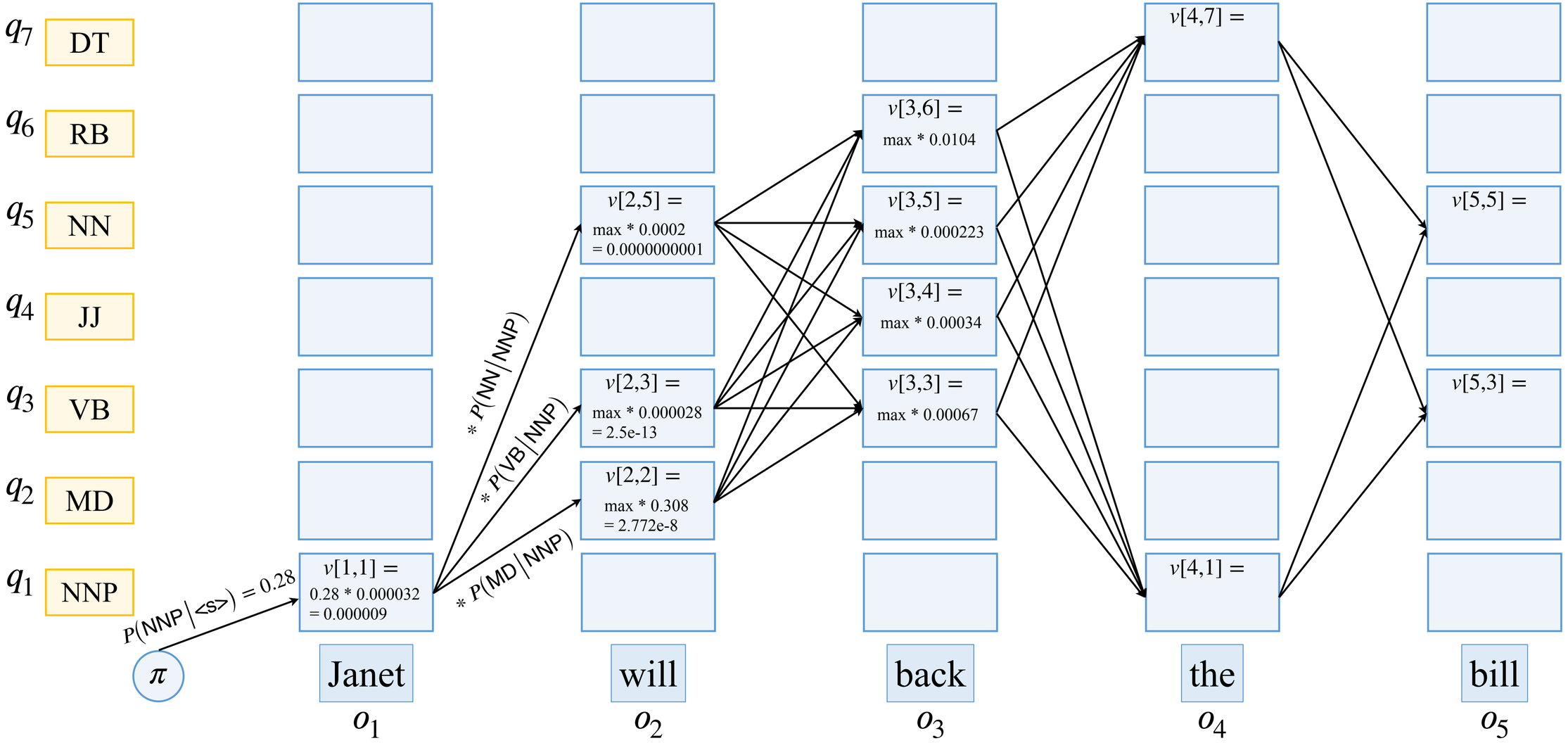
HMM 解码：Viterbi 算法



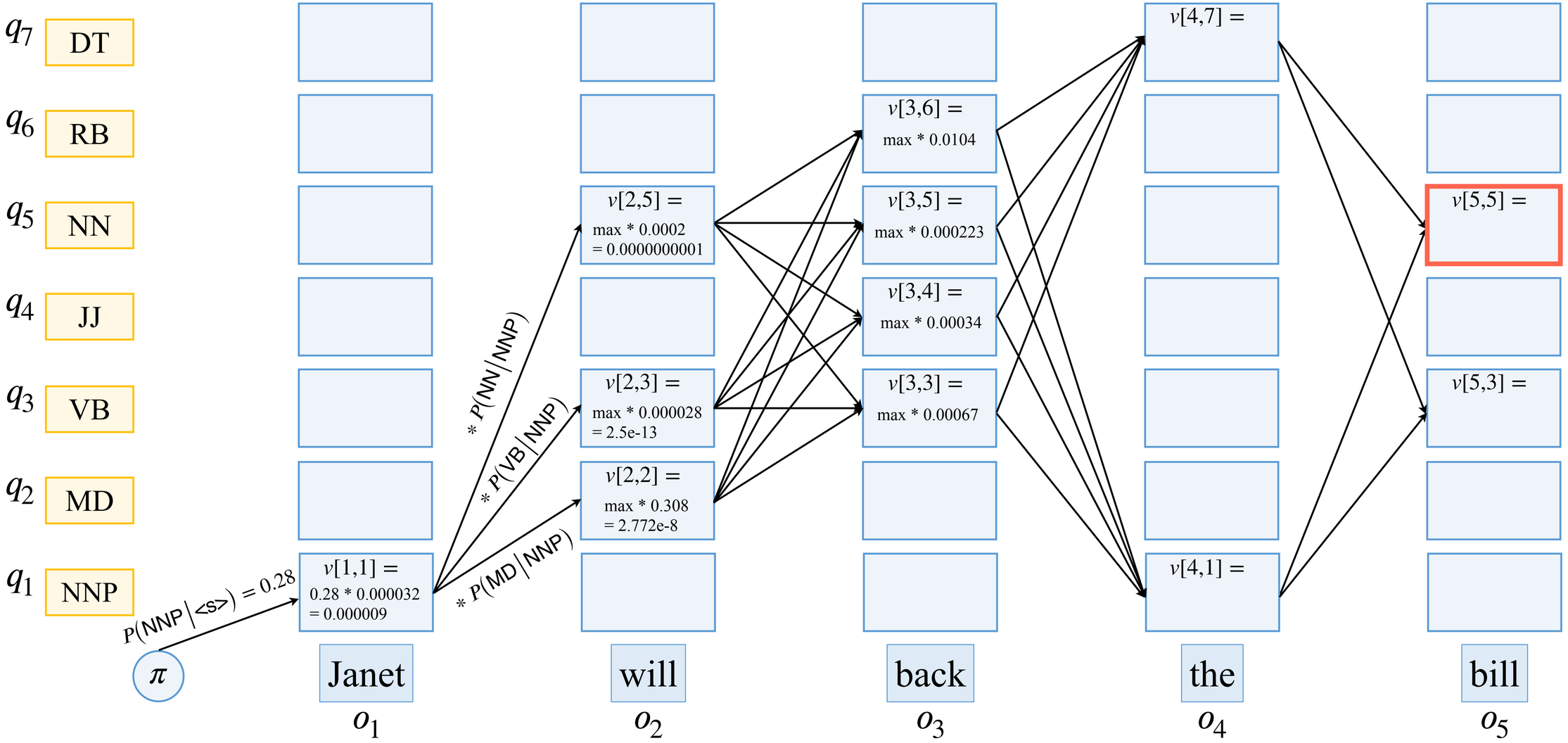
HMM 解码：Viterbi 算法



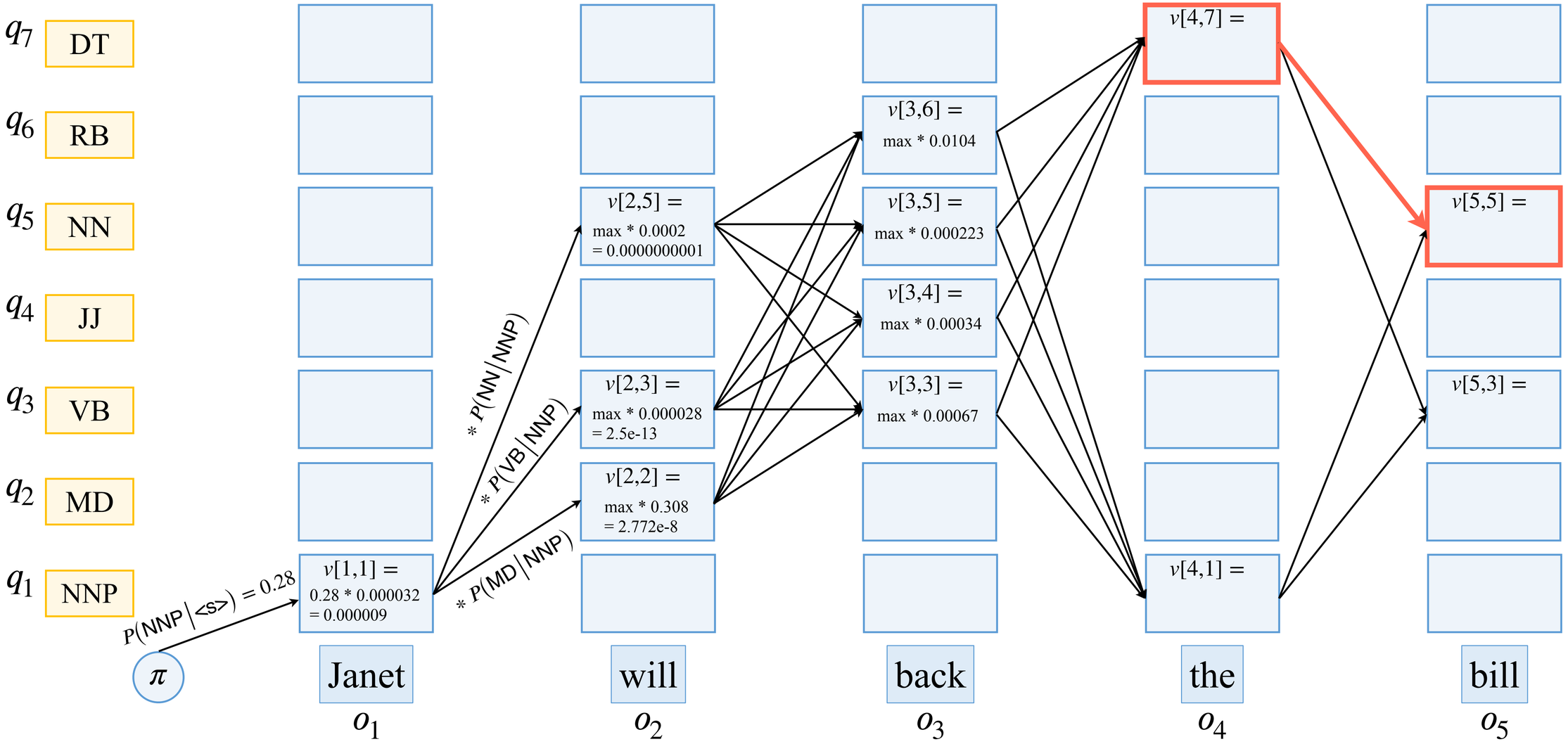
HMM 解码：Viterbi 算法



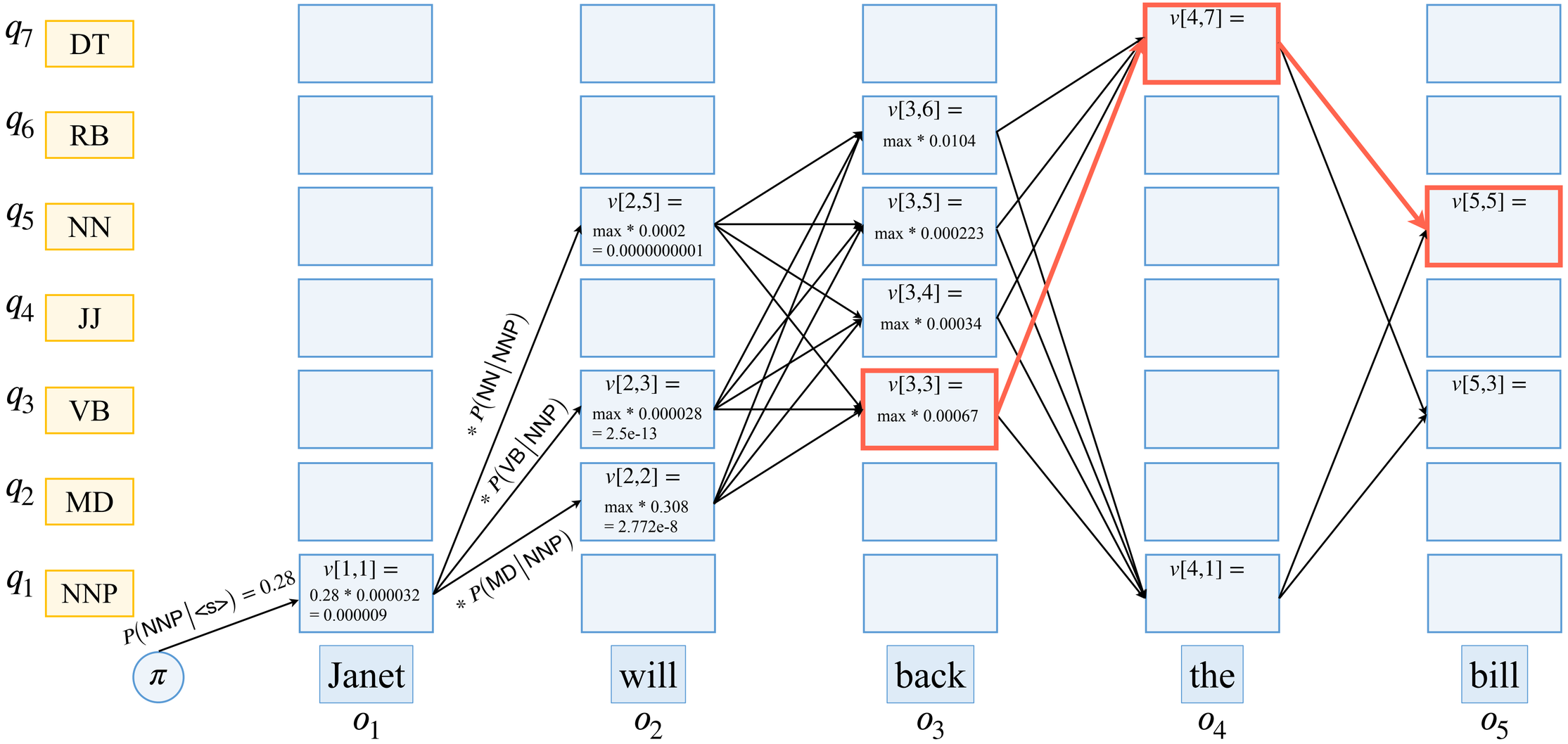
HMM 解码：Viterbi 算法



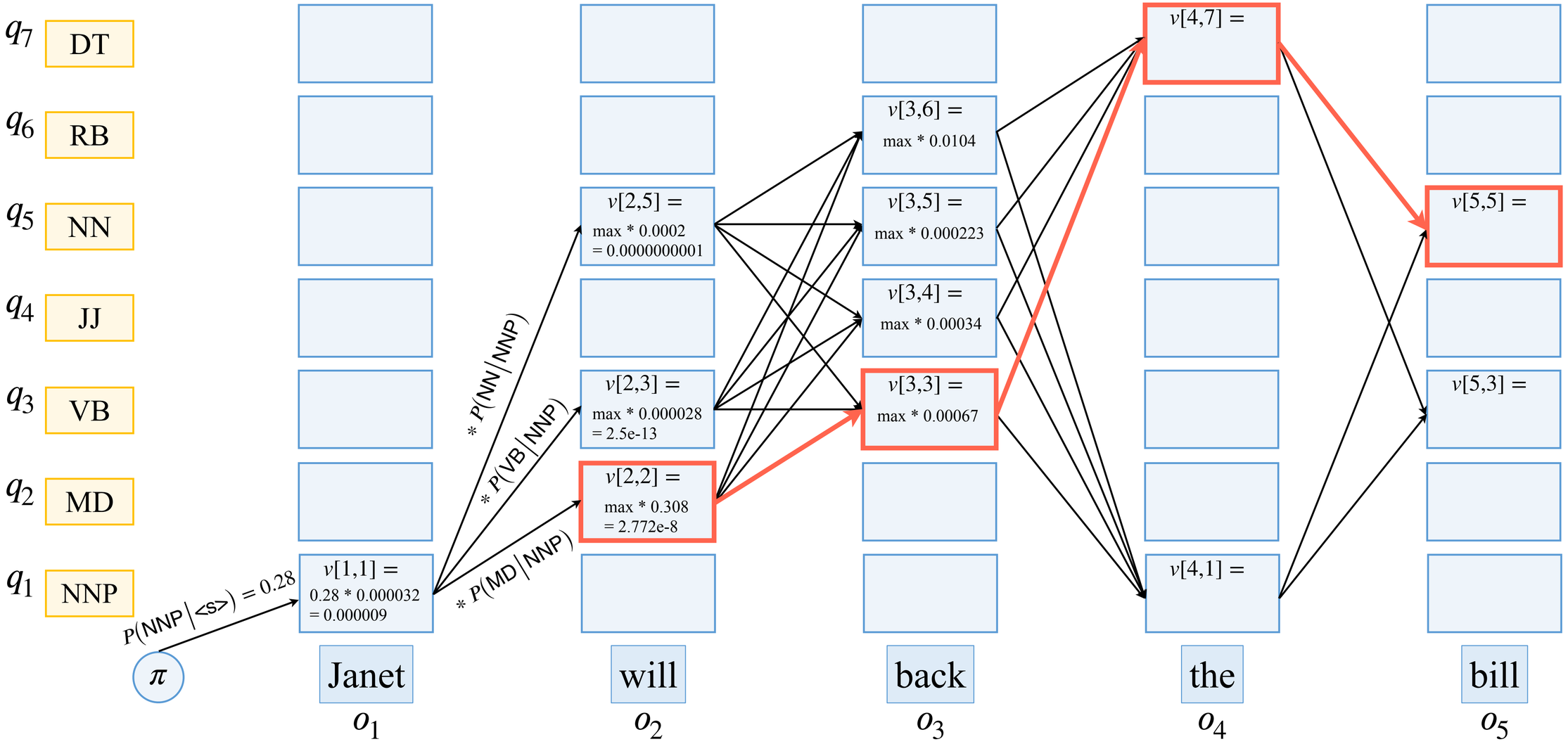
HMM 解码：Viterbi 算法



HMM 解码：Viterbi 算法



HMM 解码：Viterbi 算法



HMM 解码：Viterbi 算法