DMWIS

吉林大学
JILIN UNIVERSITY

自然语言处理

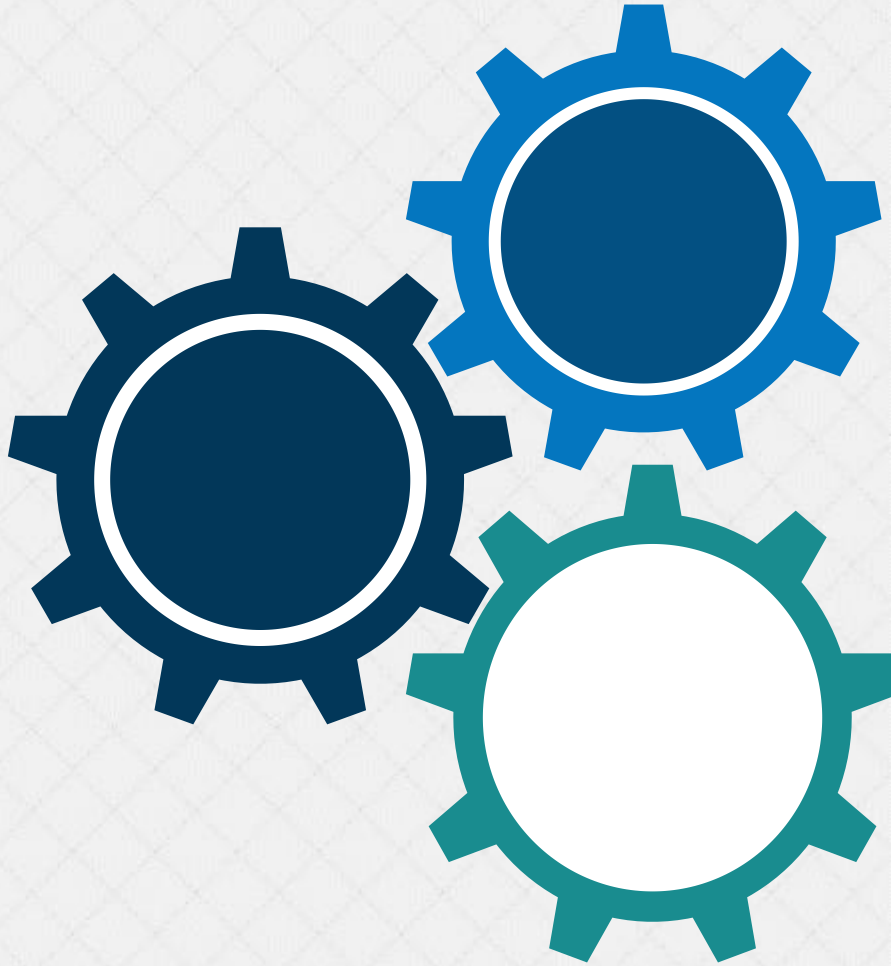
包铁

2023年9月22日

baotie@jlu.edu.cn
Data Mining and Web Information System Group （DMWIS），
College of Computer Science and Technology, Jilin University

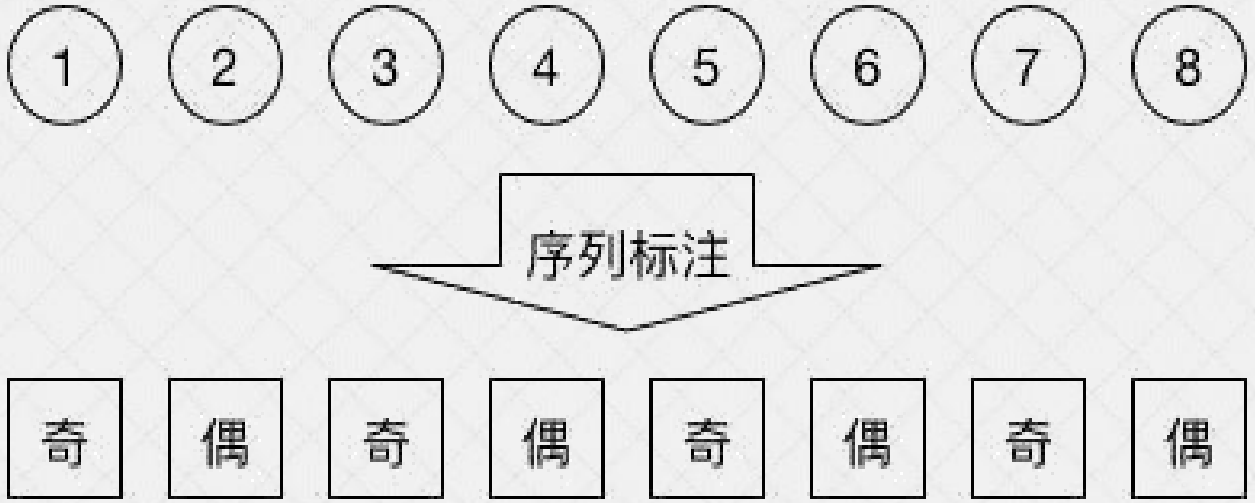
主要内容

1. 基础
2. 隐马尔可夫模型-HMM
3. 感知机
4. 条件随机场
5. 其他任务应用



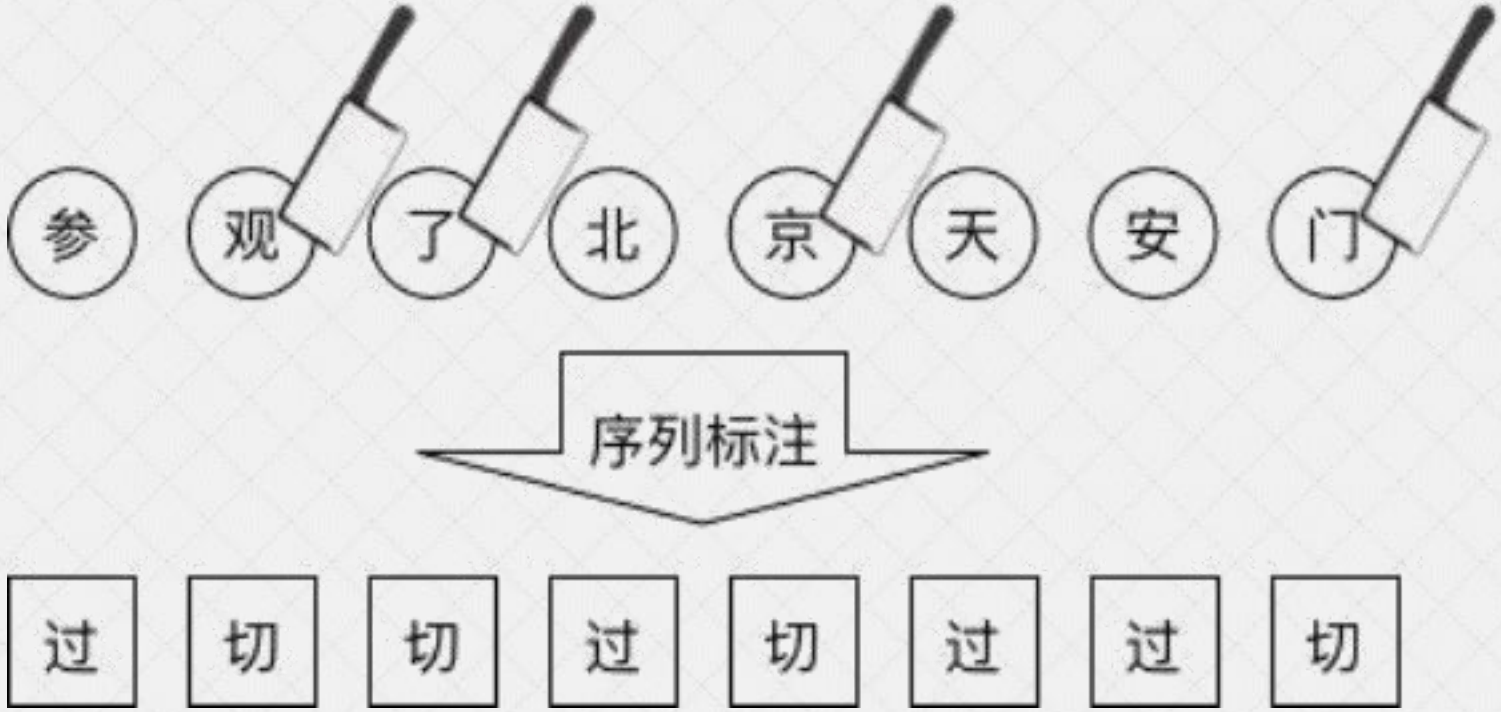
序列标注问题

- **序列标注**（tagging）：给定一个序列$x = x_1 x_2 \ldots x_n$，找出序列中每个元素对应标签$y = y_1 y_2 \ldots y_n$的问题
  - **其中，y是所有可能的取值集合，称为标注集（tagset）**
  - **很多NLP基础任务都可以转化为序列标注问题**



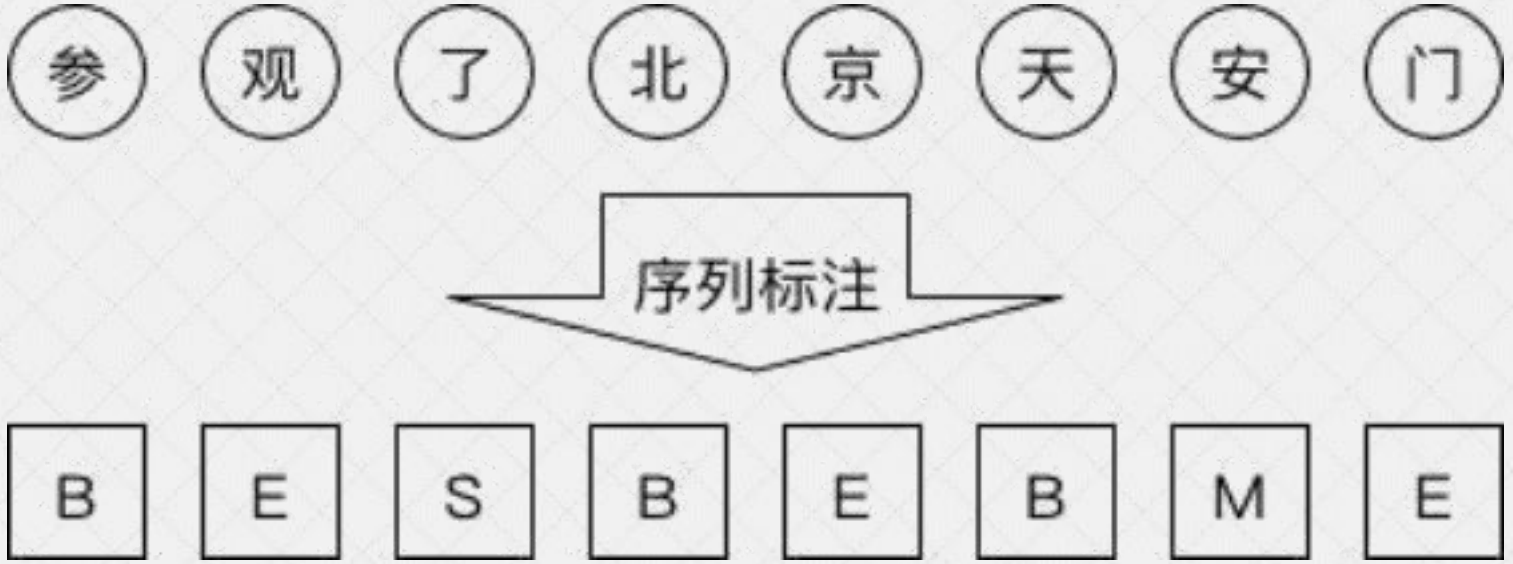
序列标注与中文分词

- 中文分词转化为标注集为{ 切,过 }的序列标注问题-标注简单



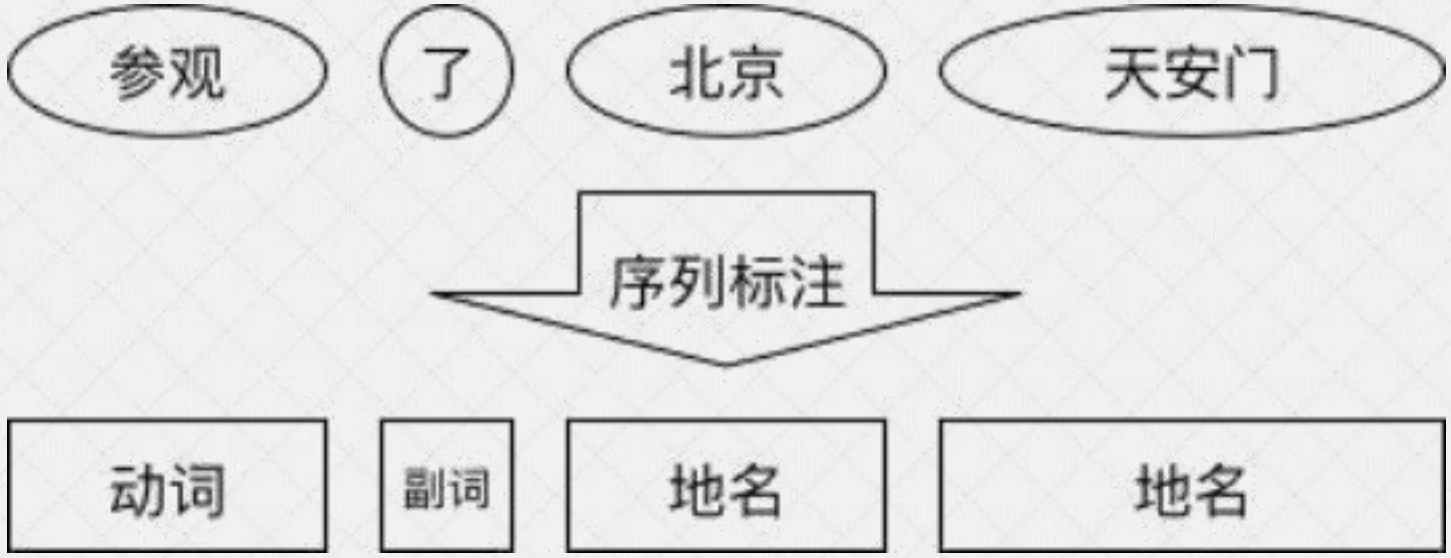
序列标注与中文分词

- **{ B,M,E,S }标注集**-汉字分别作为词语首尾（Begin、End）、词中（Middle）以及单字成词（Single）



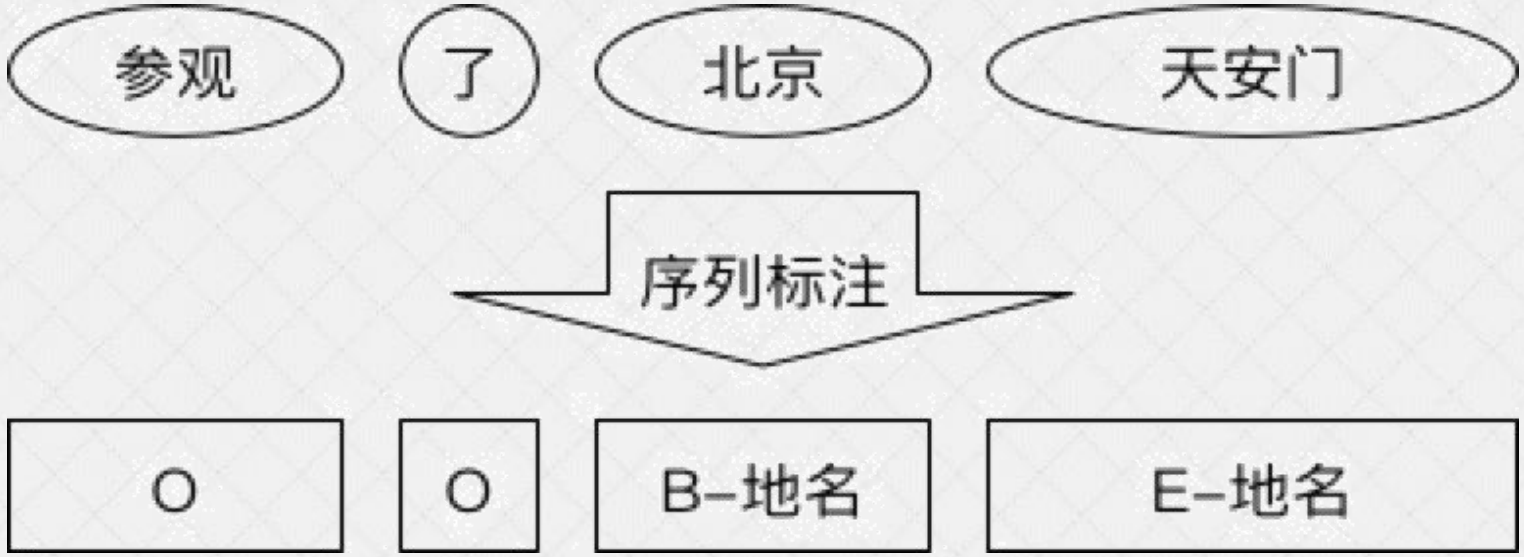
序列标注与词性标注

- **词性标注集：**863标注集（词性数量少一些、颗粒度大一些）、北大标注集



序列标注与命名实体识别

- 人名、地名、机构名常常由多个单词组成（称为复合词）
  - 可以复用BMES标注集，附加命名实体类标签（O表示Outside）



隐马尔可夫模型

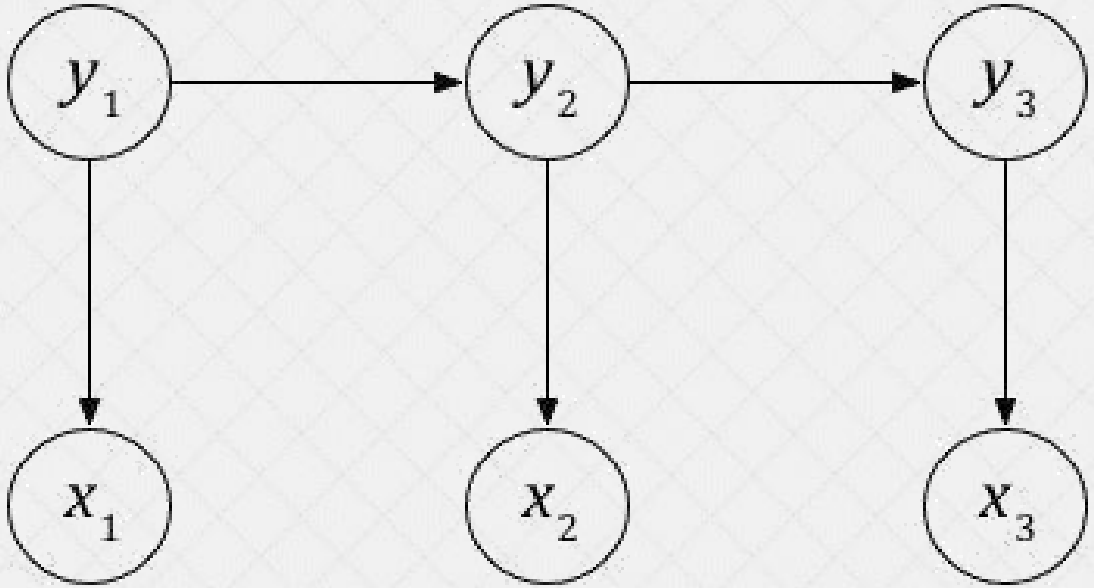
- **隐马尔可夫模型**（Hidden Markov Model，HMM）是描述两个时序序列联合分布$p(\boldsymbol{x}, \boldsymbol{y})$的概率模型，满足马尔可夫假设
  - $\boldsymbol{x}$序列外界可见（外界指观测者），称为**观测序列**（observation sequence），也称为**显状态**（visible state）
    - 观测x为单词
  - $\boldsymbol{y}$序列外界不可见，称为**状态序列**（state sequence），也称为**隐状态**（hidden state）
    - 状态y为词性

从马尔可夫假设到隐马尔可夫模型

- **马尔可夫假设**：每个事件的发生概率只取决于前一个事件
  - 将满足该假设的连续多个事件串联在一起，就构成了马尔可夫链
- **隐马尔可夫模型**的理解：将马尔可夫假设作用于状态序列
  - **假设①**：当前状态y_t仅仅依赖于前一个状态y_{t-1}，连续多个状态构成**隐马尔可夫链**y
  - **假设②**：任意时刻的观测x_t只依赖于该时刻的状态y_t，与其他时刻的状态或观测独立无关

从马尔可夫假设到隐马尔可夫模型

- **隐马尔可夫模型状态序列与观测序列的依赖关系**
  - 箭头表示事件的依赖关系（箭头终点是结果，依赖于起点的原因）

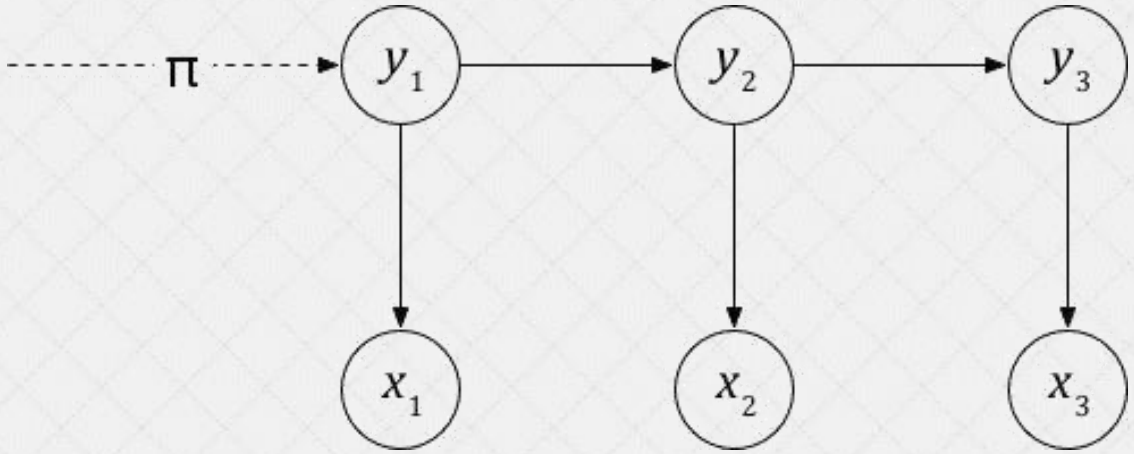


从马尔可夫假设到隐马尔可夫模型

- **隐马尔可夫模型**利用三个要素来模拟时序序列的发生过程
  - 初始状态概率向量
  - 状态转移概率矩阵
  - 发射概率矩阵（也称作观测概率矩阵）

初始状态概率向量

- 系统启动时进入的第一个状态y_1称为**初始状态**
  - 假设y有N种可能的取值，即$y \in \{s_1, \cdots, s_N\}$，那么$y_1$就是一个独立的离散型随机变量，由$p(y_1|\boldsymbol{\pi})$描述
  - 其中$\boldsymbol{\pi} = (\pi_1, \cdots, \pi_N)^{\mathrm{T}}, 0 \leq \pi_i \leq 1, \sum_{i=1}^{N} \pi_i = 1$ 是概率分布的参数向量，称为**初始状态概率向量**



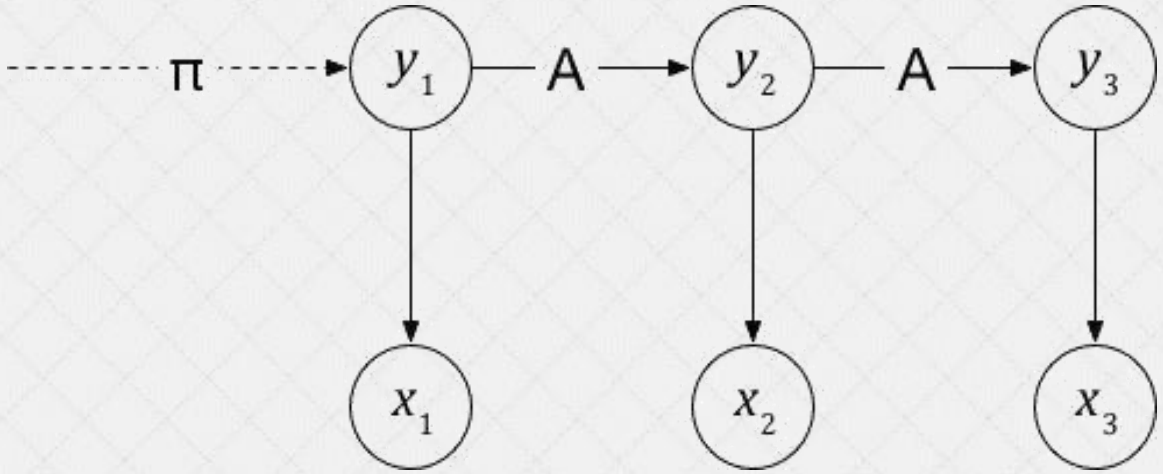
初始状态概率向量

- $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_N)^{\mathrm{T}}, 0 \leq \pi_i \leq 1, \sum_{i=1}^{N} \pi_i = 1$ 是概率分布的参数向量，称为**初始状态概率向量**
  - 比如在中文分词中，采用BMES标注集，可能的隐马尔可夫模型的初始状态概率向量为π=[0.7,0,0,0.3]

$$p(y_1 = \mathrm{B}) = 0.7$$
$$p(y_1 = \mathrm{M}) = 0$$
$$p(y_1 = \mathrm{E}) = 0$$
$$p(y_1 = \mathrm{S}) = 0.3$$

状态转移概率矩阵

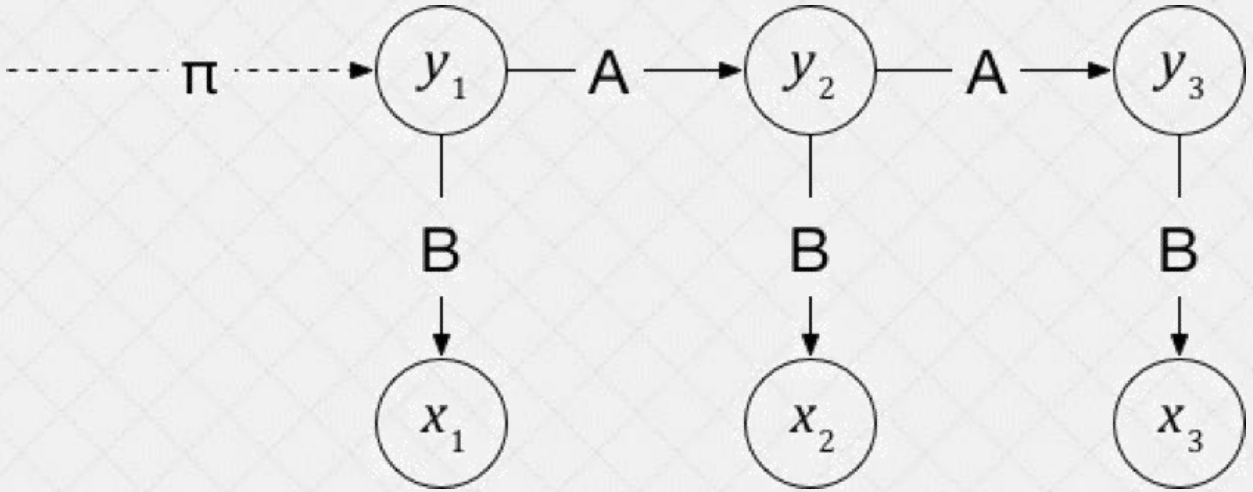
- 从状态s_i到状态s_j的概率就构成了一个$N \times N$的方阵，称为**状态转移概率矩阵**$\boldsymbol{A}$：$\boldsymbol{A} = [p(y_{t+1} = s_j | y_t = s_i)]_{N \times N}$
  - 其中下标i、j分别表示状态的第i、j种取值，比如约定1表示BMES标注集中的B，依序类推。
  - **概率**模拟实际**语言现象**



发射概率矩阵

- 给定y，x都是独立的离散型随机变量，其参数对应一个向量
  - 这些参数向量构成了$N \times M$的矩阵，称为发射概率矩阵$\boldsymbol{B}$。

$$\boldsymbol{B} = [p(x_t = o_i | y_t = s_j)]_{N \times M}$$



模型的应用

1 **样本生成问题**：给定模型$\lambda = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$，生成满足模型约束的样本，即一系列观测序列及其对应的状态序列$\{(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})\}$。

2 **模型训练问题**：给定训练集$\{(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})\}$，估计模型参数$\lambda = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$。

3 **序列预测问题**：已知模型参数$\lambda = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B})$，给定观测序列$\boldsymbol{x}$，求最可能的状态序列$\boldsymbol{y}$。

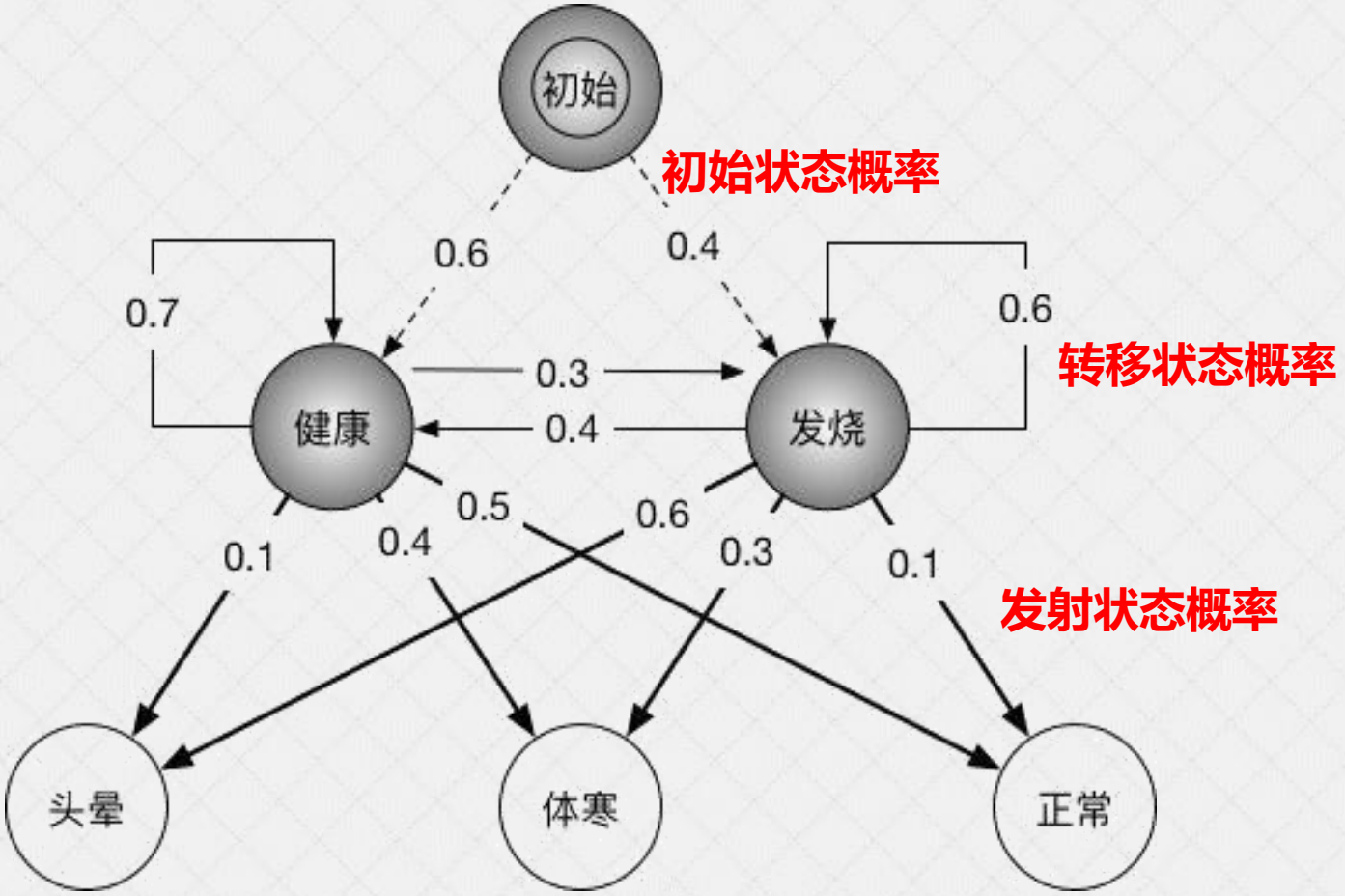
模型的应用-样本生成

- 某医院开发“智能”医疗诊断系统，用来辅助感冒诊断。已知
  - ①来诊者只有两种状态：健康、发烧。
  - ②来诊者不确定自己的状态，只能回答感觉头晕、体寒或正常。
  - ③感冒这种病，只跟病人前一天的状态有关。
  - ④当天的病情决定当天的身体感觉。
  - 系统功能-来诊者的病历卡上完整地记录了最近T天的身体感受（头晕、体寒或正常），预测这T天的身体状态（健康或发烧）

模型的应用-样本生成

- 病情与体感规律-医生经验（生成样本作为测试数据）

案例满足HMM条件，
基于已知参数的HMM模型，
随机采样可以生成样本。



模型的应用-训练

实例	词性
1. Bill will pay the bill.	1. P M V D N
2. Will Mike pay the bill?	2. M P V D N
3. Mike will drink milk.	3. P M V N
4. Jane will drink water.	4. P M V N
5. Give Mike some drink.	5. V P D N

词性标注含义

P：proper noun-专有名词

M：modal-情态动词

V：verb-动词

N：noun-名词

D：determiner-限定词

待词性标注语句：Bill will pay the drink

模型的应用-训练

- 统计y_1的所有取值的频次记作向量c，然后用类似的方法归一化，**初始状态概率**可以估计为：

$$\hat{\pi}_i = \frac{c_i}{\sum_{i=1}^{N} c_i}, \quad i = 1,2,\cdots,N$$

模型的应用-训练

1. P M V D N
2. M P V D N
3. P M V N
4. P M V N
5. V P D N

	P	N	M	V	D	<E>
<S>	3	0	1	1	0	
P	0	0	3	1	1	
N	0	0	0	0	0	5
M	1	0	0	3	0	
V	1	2	0	0	2	
D	0	3	0	0	0	

模型的应用-训练

- 记样本序列在时刻t处于状态s_i，时刻$t+1$转移到状态s_j
- 统计这样的转移频次计入矩阵元素$A_{i,j}$
- 根据极大似然估计，从s_i到s_j的**转移概率**$a_{i,j}$可估计为矩阵第i行频次的归一化：

$$\hat{a}_{i,j} = \frac{A_{i,j}}{\sum_{j=1}^{N} A_{i,j}}, \quad i,j = 1,2,\cdots,N$$

模型的应用-训练

1. P M V D N
2. M P V D N
3. P M V N
4. P M V N
5. V P D N

	P	N	M	V	D	<E>
<S>	3/5	0	1/5	1/5	0	
P	0	0	3/5	1/5	1/5	
N	0	0	0	0	0	5/5
M	1/4	0	0	3/4	0	
V	1/5	2/5	0	0	2/5	
D	0	3/3	0	0	0	

模型的应用-训练

- 统计样本中状态为s_i且观测为o_j的频次，计入矩阵元素$\boldsymbol{B}_{i,j}$，则状态s_i**发射**观测o_j的**概率**估计为：

$$\hat{b}_{ij} = \frac{\boldsymbol{B}_{i,j}}{\sum_{j=1}^{M} \boldsymbol{B}_{i,j}}, \quad i = 1,2,\cdots,N; j = 1,2,\cdots,M$$

模型的应用-训练

实例

1. Bill will pay the bill.
2. Will Mike pay the bill?
3. Mike will drink milk.
4. Jane will drink water.
5. Give Mike some drink.

	P	N	M	V	D
bill	1	2	0	0	0
will	0	0	4	0	0
pay	0	0	0	2	0
the	0	0	0	0	2
mike	3	0	0	0	0
drink	0	1	0	2	0
milk	0	1	0	0	0
jane	1	0	0	0	0
give	0	0	0	1	0
some	0	0	0	0	1
water	0	1	0	0	0

模型的应用-训练

实例

1. Bill will pay the bill.
2. Will Mike pay the bill?
3. Mike will drink milk.
4. Jane will drink water.
5. Give Mike some drink.

	P	N	M	V	D
bill	1/5	2/5	0	0	0
will	0	0	4/4	0	0
pay	0	0	0	2/5	0
the	0	0	0	0	2/3
mike	3/5	0	0	0	0
drink	0	1/5	0	2/5	0
milk	0	1/5	0	0	0
jane	1/5	0	0	0	0
give	0	0	0	1/5	0
some	0	0	0	0	1/3
water	0	1/5	0	0	0

模型的应用-预测

- 沿隐马尔可夫链，首先$t = 1$时初始状态没有前驱状态，发生概率由π决定

$$p(y_1 = s_i) = \pi_i$$

- 接着对$t \geq 2$，状态y_t由前驱状态y_{t-1}转移而来，其转移概率由矩阵A决定

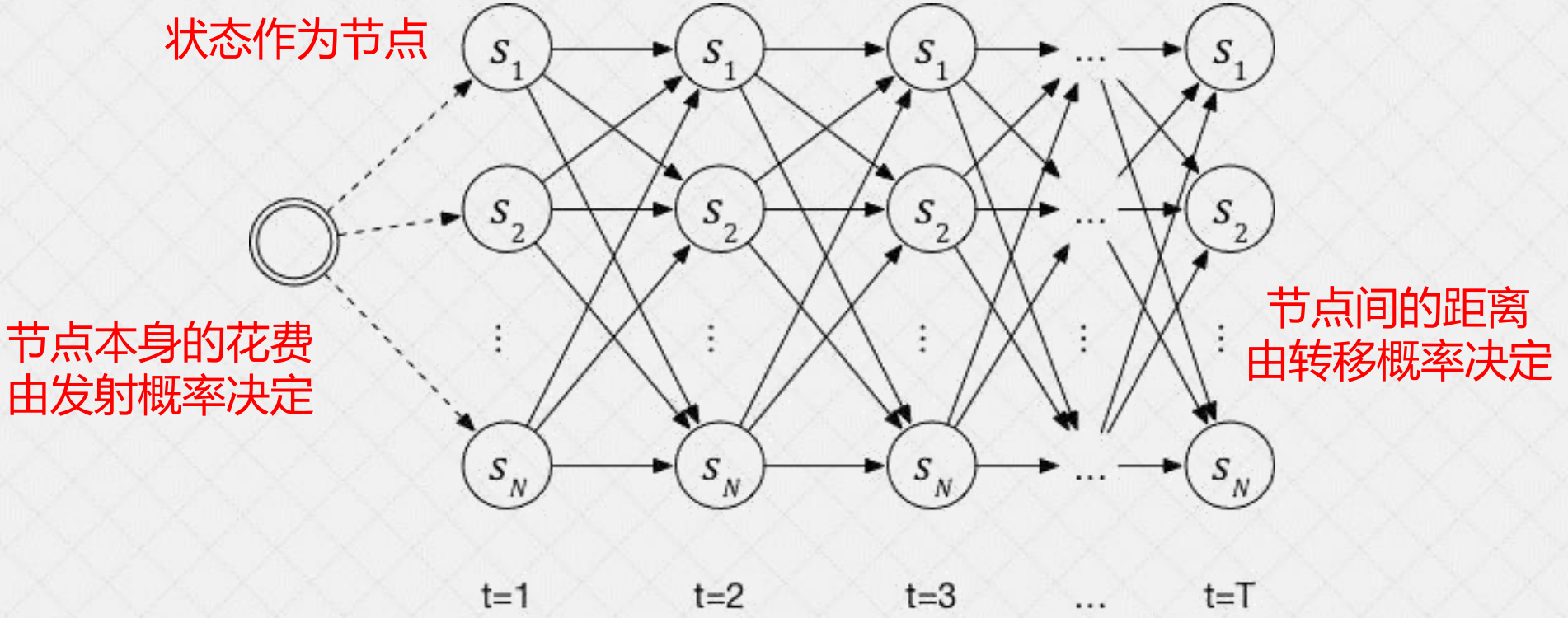
$$p(y_t = s_j | y_{t-1} = s_i) = A_{i,j}$$

- 最后，对每个$y_t = s_i$，都会“发射”一个$x_t = o_j$，其发射概率由矩阵B决定

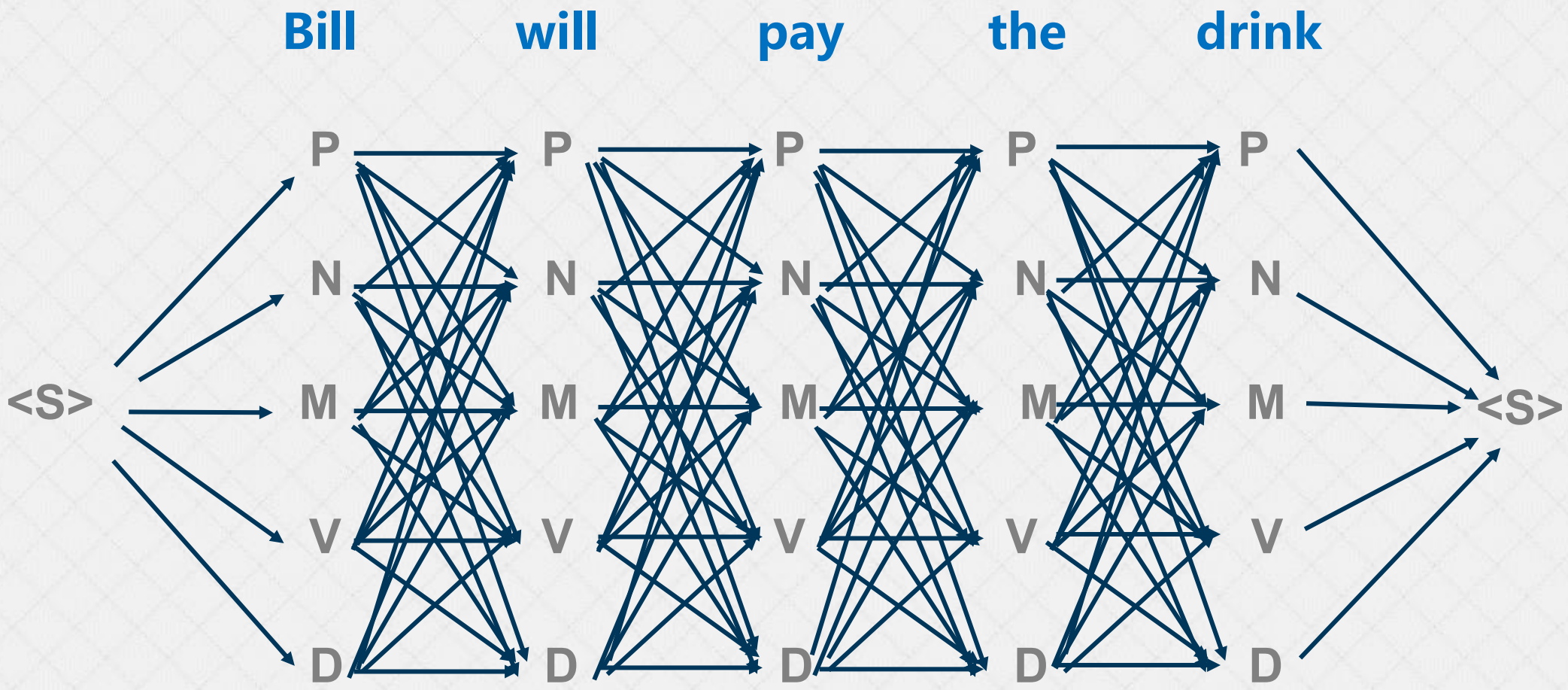
$$p(x_t = o_j | y_t = s_i) = B_{i,j}$$

模型的应用-预测

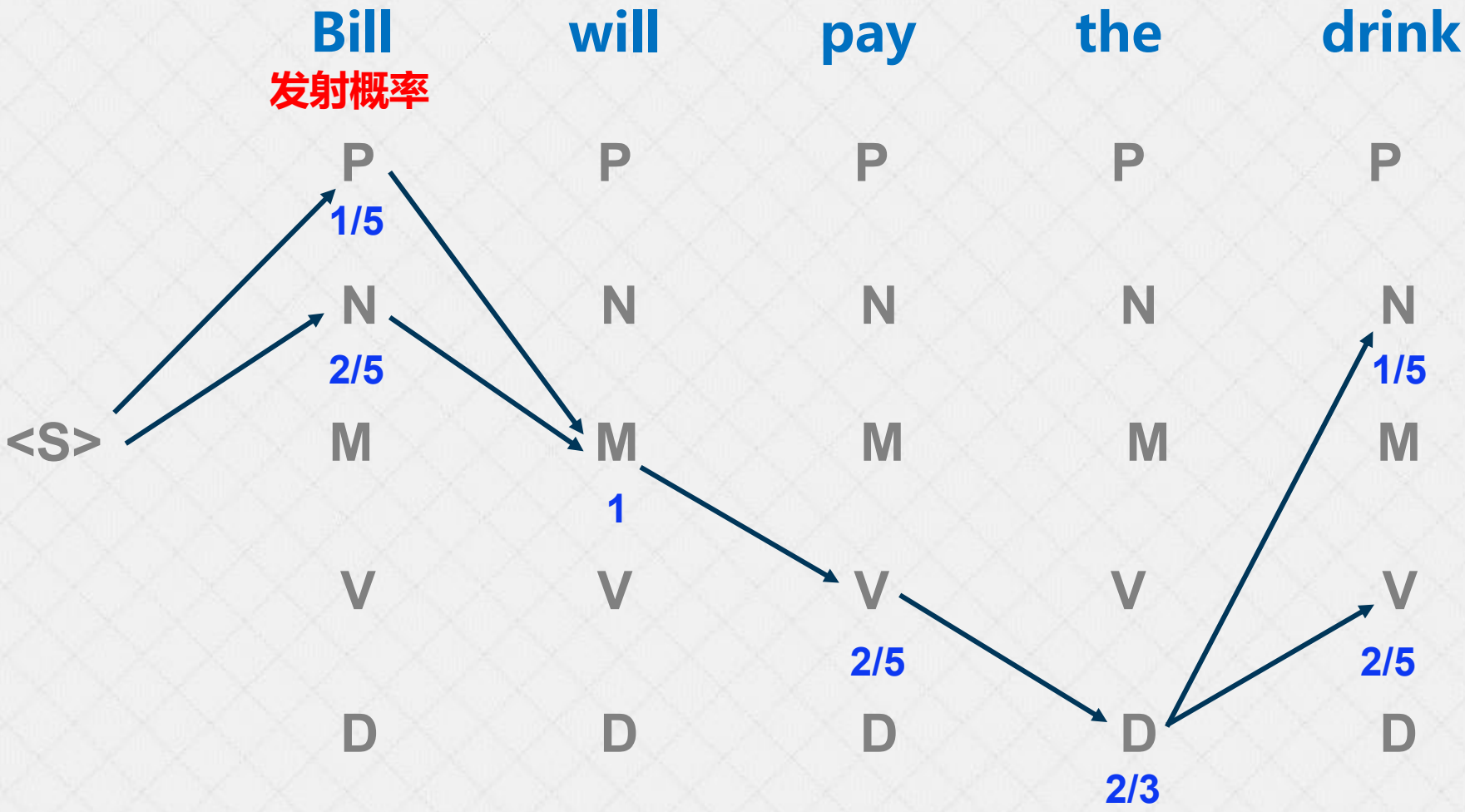
- 搜索状态序列的**维特比算法**：备选状态构成一幅有向无环图，待求的概率最大的状态序列就是图中的最长路径



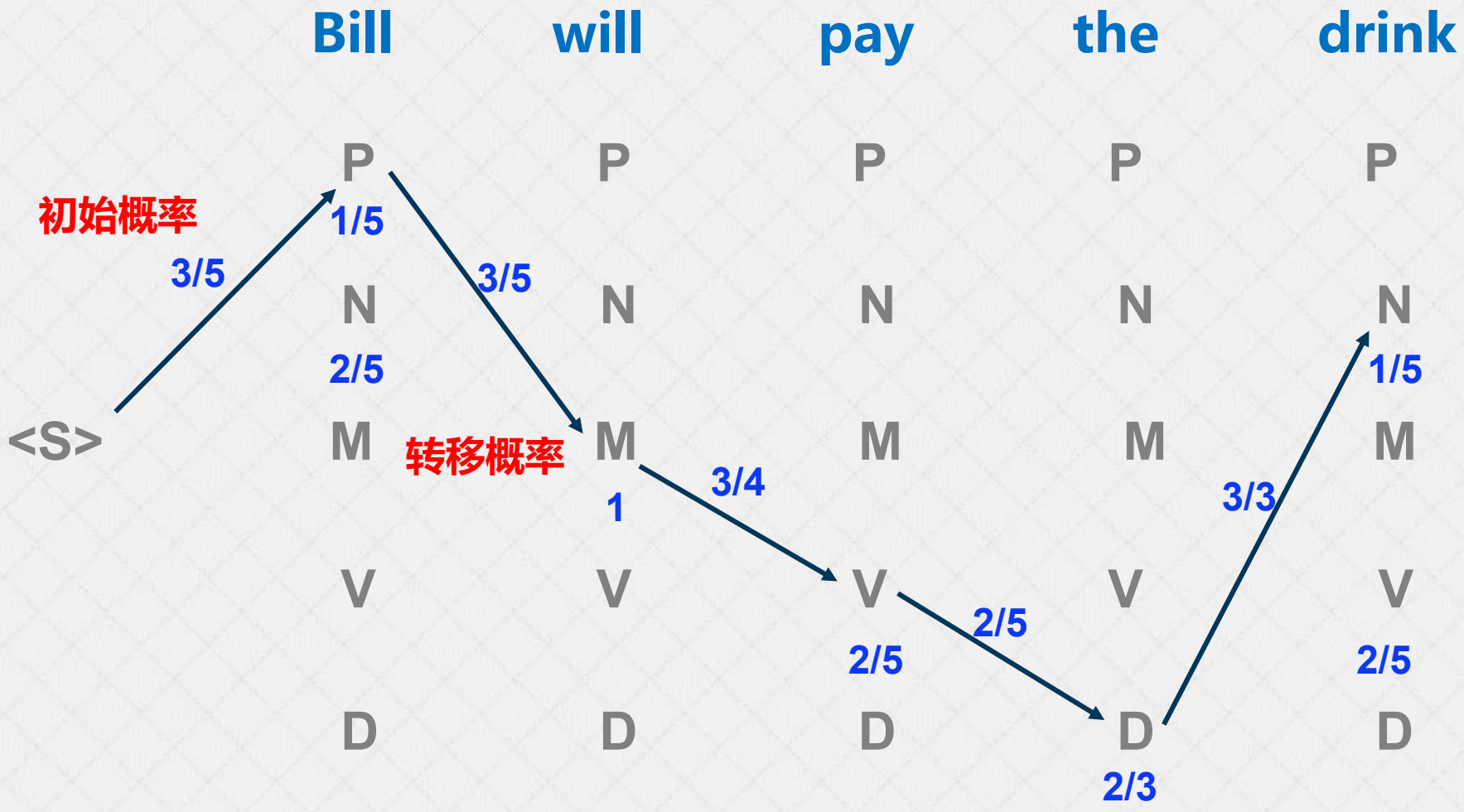
模型的应用-预测



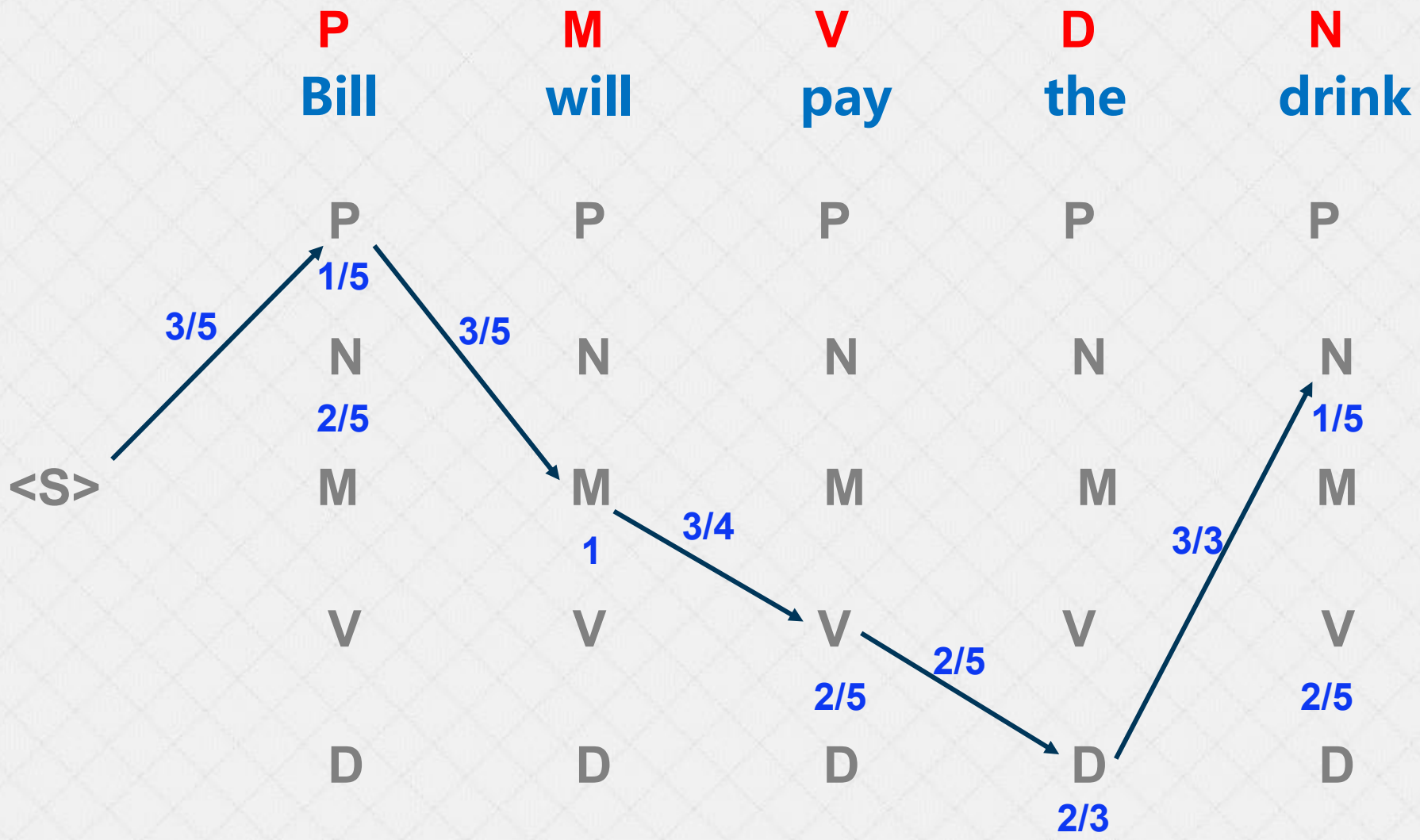
模型的应用-预测



模型的应用-预测


Bill
will
pay
the
drink
P
N
M
V
D
<S>
初始概率
3/5
1/5
3/5
2/5
转移概率
3/4
1
2/5
2/5
3/3
1/5
2/5
2/3

模型的应用-预测


P
M
V
D
N
Bill
will
pay
the
drink
P
N
M
V
D
<S>
3/5
1/5
3/5
2/5
1
3/4
2/5
2/5
3/3
2/3
1/5
2/5

感知机的引入

- HMM实现了基于序列标注的中文分词器，但是效果不理想
  - HMM假设语句取决于隐藏的BMES序列-简单
  - 特征捕捉少-包括前一个标签、当前字符
- 线性模型可以引入更多特征-感知机
  - 一系列提取特征的特征函数ϕ-如：特征模板提取人名特征
  - 相应的权重w

分类问题

- **分类**（classification）指的是预测样本所属类别的一类问题
  - 分类问题的目标就是给定输入样本$\boldsymbol{x}$，将其分配给K种类别$\mathcal{C}_k$中的一种，其中$k = 1, \cdots, K$
  - 如果$K = 2$，则称为**二分类**（binary classification）
  - 否则称为**多分类**（multiclass classification）
- 二分类也可以解决任意类别数的多分类问题
  - one-vs-one-类别两两组合交给分类器投票
  - one-vs-rest-分类器判断的正类作为结果（如有多个比较置信度）

分类的应用

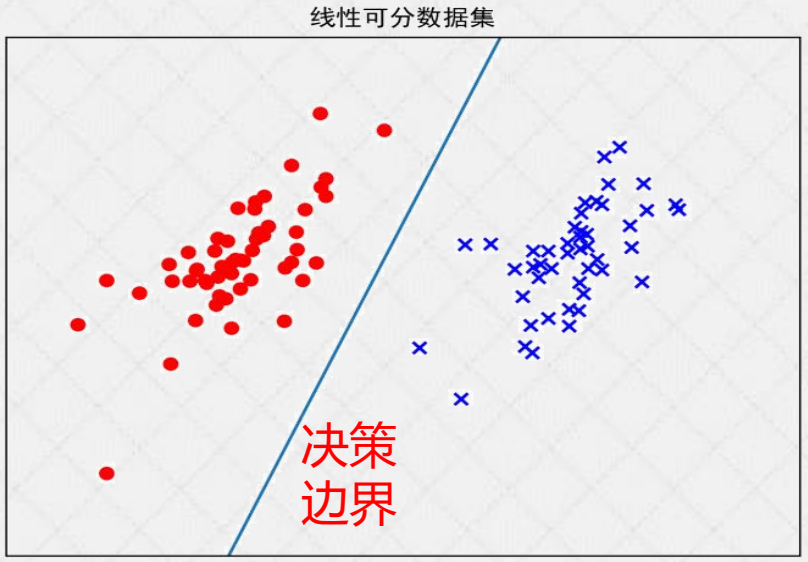
- 文本分类本身就是一个分类问题
- 关键词提取时，对文章中的每个单词判断是否属于关键词，于是转化为二分类问题
- 在指代消解问题中，对每个代词和每个实体判断是否存在指代关系，又是一个二分类问题
- 预测天气阴晴雨雪、照片对应哪种事物、声波是否由某个人发出

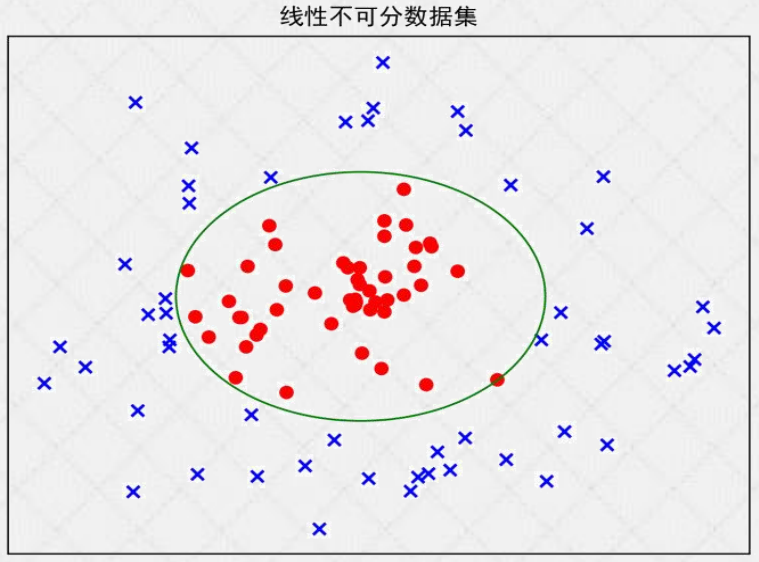
线性分类模型

- **线性模型**（linear model）是传统机器学习方法中最简单最常用的分类模型，用一条线性的直线或高维平面将数据一分为二
- 线性模型由特征函数ϕ，以及相应的权重向量w组成

特征向量

样本空间





线性不可分

增加特征

特殊函数

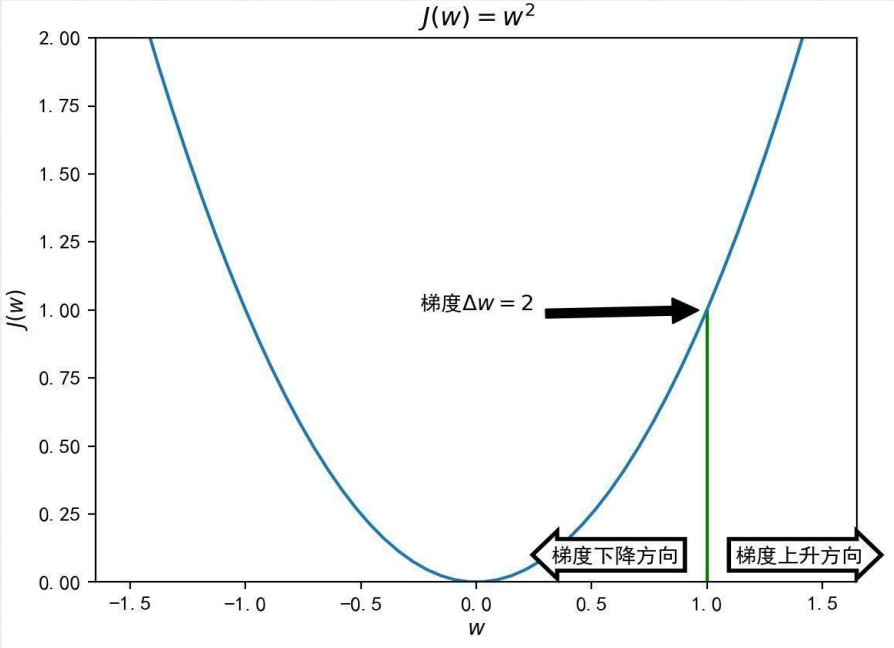
感知机算法-训练算法

- **感知机算法** - 迭代找出分离超平面

  1) 读入训练样本$(\boldsymbol{x}^{(i)}, y^{(i)})$，执行预测$\hat{y} = \text{sign}(\boldsymbol{w} \cdot \boldsymbol{x}^{(i)})$。

  2) 如果$\hat{y} \neq y^{(i)}$，则更新参数$\boldsymbol{w} \leftarrow \boldsymbol{w} + y^{(i)}\boldsymbol{x}^{(i)}$。

- **投票感知机**-要求储存多个模型及加权
- **平均感知机**-取多个模型的权重的平均



结构化预测问题

- **NLP问题大致可分为两类**
  - 一种是分类问题，另一种就是结构化预测问题
- **结构化预测**（structured prediction）是预测对象结构的一类监督学习问题
  - 相应的模型训练过程称作结构化学习
  - NLP中有很多结构化预测任务：序列标注预测的结构是一个序列，句法分析预测的结构是一棵句法树，机器翻译预测的结构是一段完整的译文

结构化感知机

- 定义新的特征函数$\boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{y})$，把结构$\boldsymbol{y}$也作为一种特征，输出新的“结构化特征向量” $\boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{y}) \in \mathbb{R}^{D\times 1}$
- 新特征向量与权重向量做点积后，得到一个标量，将其作为分数：

$$\mathbf{score}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{w}\cdot\boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{y})$$

- 取分值最大的结构作为预测结果，得到结构化预测函数：

$$\hat{\boldsymbol{y}} = \underset{\mathbf{y}\in Y}{\operatorname{argmax}}(\boldsymbol{w}\cdot\boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{y}))$$

结构化感知机算法

- 读入样本$(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})$，执行结构化预测$\hat{\boldsymbol{y}} = \text{argmax}_{\boldsymbol{y}\in Y}\left(\boldsymbol{w}\cdot\boldsymbol{\phi}(\boldsymbol{x}^{(i)}, \boldsymbol{y})\right)$。

- 与正确答案对比，若$\hat{\boldsymbol{y}} \neq \boldsymbol{y}^{(i)}$，则更新参数：奖励正确答案触发的特征函数的权重$\boldsymbol{w} \leftarrow \boldsymbol{w} + \boldsymbol{\phi}(\boldsymbol{x}^{(i)}, \boldsymbol{y})$，惩罚错误结果触发的特征函数的权重$\boldsymbol{w} \leftarrow \boldsymbol{w} - \boldsymbol{\phi}(\boldsymbol{x}^{(i)}, \boldsymbol{y})$。奖惩可以合并到一个式子里：$\boldsymbol{w} \leftarrow \boldsymbol{w} + \boldsymbol{\phi}(\boldsymbol{x}^{(i)}, \boldsymbol{y}) - \boldsymbol{\phi}(\boldsymbol{x}^{(i)}, \hat{\boldsymbol{y}})$，还可以调整学习率：$\boldsymbol{w} \leftarrow \boldsymbol{w} + \alpha\left(\boldsymbol{\phi}(\boldsymbol{x}^{(i)}, \boldsymbol{y}) - \boldsymbol{\phi}(\boldsymbol{x}^{(i)}, \hat{\boldsymbol{y}})\right)$。

结构化感知与序列标注

- 序列标注中最重要的结构特点是标签间的依赖性，这种依赖性在HMM中体现为初始概率向量和转移概率矩阵，对序列中的连续标签提取如下**转移特征**：

$$\phi_k(y_{t-1}, y_t) = \begin{cases} 1 & \text{if } y_{t-1} = s_i \text{ and } y_t = s_j; \\ 0 & \text{otherwise.} \end{cases} \qquad i = 0, \cdots, N; j = 1, \cdots, N$$

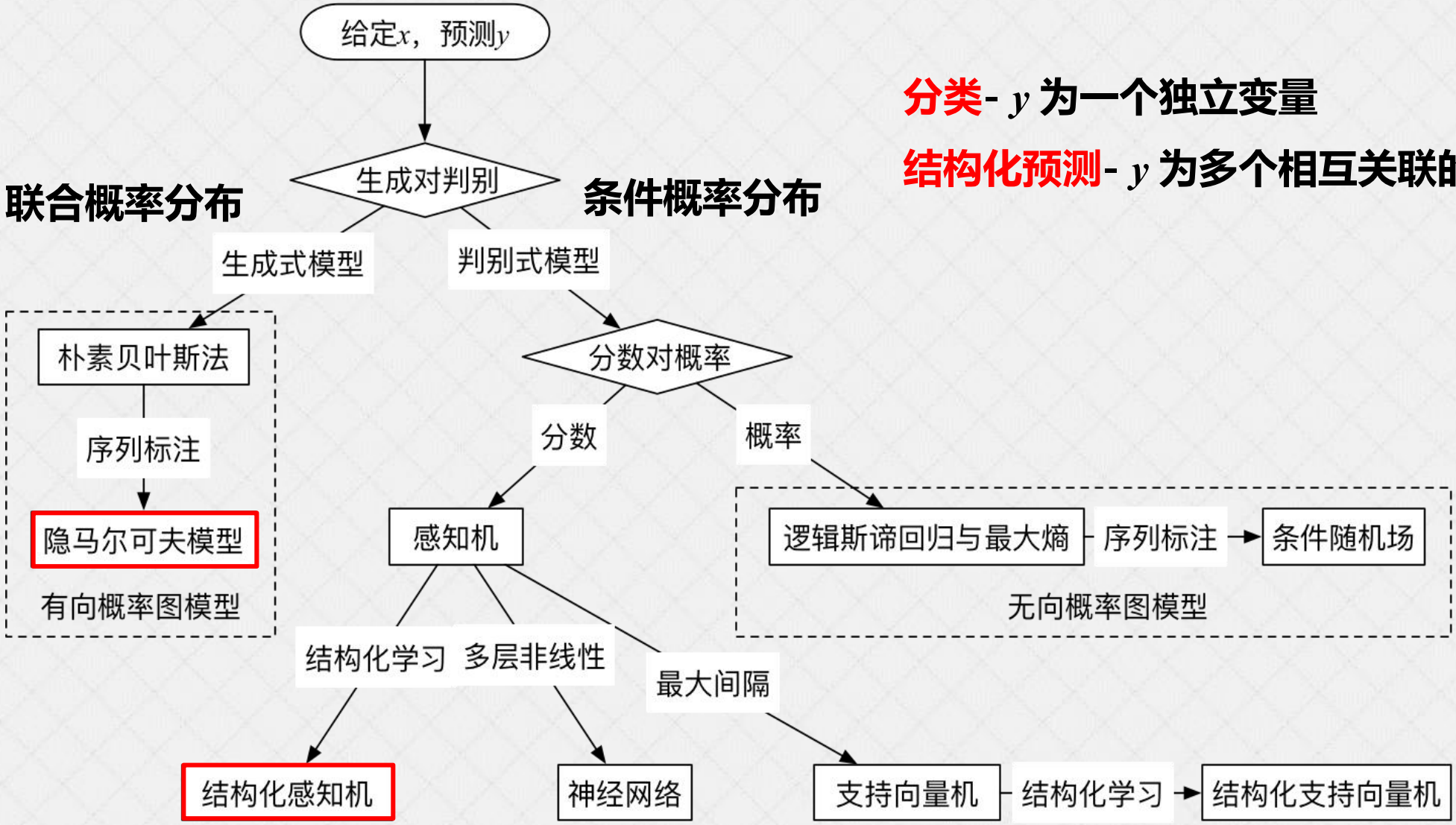
- 定义每个时刻的**状态特征**为：$\phi_l(\mathbf{x}_t, y_t) = \begin{cases} 1 \\ 0 \end{cases}$

- 于是，结构化感知机的特征函数就是转移特征和状态特征的合集：

$$\boldsymbol{\phi} = [\phi_k; \phi_l] \qquad k = 1, \cdots, N^2 + N; l = N^2 + N + 1, \cdots$$

条件随机场

机器学习的模型谱系



分类- y 为一个独立变量

结构化预测- y 为多个相互关联的变量

生成式模型与判别式模型

- **生成式模型**模拟数据的生成过程，两类随机变量存在因果先后关系：先有因素$\boldsymbol{y}$，后有结果$\boldsymbol{x}$。这种因果关系由联合分布模拟：

$$p(\boldsymbol{x},\boldsymbol{y}) = p(\boldsymbol{y})p(\boldsymbol{x}|\boldsymbol{y})$$

- 生成式模型其实间接建模了$p(\boldsymbol{x})$

$$p(\boldsymbol{x}) = \sum_{\boldsymbol{y}\in\mathcal{Y}} p\,(\boldsymbol{x},\boldsymbol{y})$$

- $p(\boldsymbol{x})$很难准确估计，特征之间一般并非相互独立

生成式模型与判别式模型

- **判别式模型**跳过了$p(\boldsymbol{x})$，直接对条件概率$p(\boldsymbol{y}|\boldsymbol{x})$建模，$\boldsymbol{x}$内部的关联不影响判断，可以利用有关联的特征
- 一些判别式模型并不介意输出的到底是$[0,1]$区间内的概率$p(\boldsymbol{y}|\boldsymbol{x})$，还是一个分值$\text{score}(\boldsymbol{x},\boldsymbol{y})$

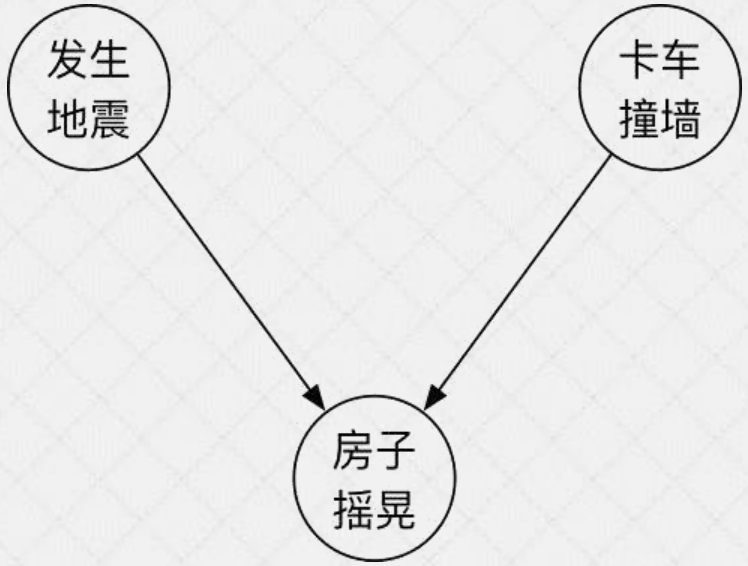
$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{\exp(\text{score}(\boldsymbol{x},\boldsymbol{y}))}{\sum_{\boldsymbol{x},\boldsymbol{y}} \exp\left(\text{score}(\boldsymbol{x},\boldsymbol{y})\right)}$$

有向与无向概率图模型

- **概率图模型**（Probabilistic Graphical Model，PGM）是用来表示与推断多维随机变量联合分布$p(\boldsymbol{x}, \boldsymbol{y})$的强大框架
  - 利用节点V来表示随机变量
  - 用边E连接有关联的随机变量
  - 将多维随机变量分布表示为图$G = (V, E)$
    - 整个图可以分解为子图再进行分析
    - 子图中的随机变量更少，建模更加简单

有向与无向概率图模型

- **有向图模型**（Directed Graphical Model，DGM）按事件的先后因果顺序将节点连接为有向图。如果事件A导致事件B，则用箭头连接两个事件$A \rightarrow B$
  - 因果关系可能比较复杂



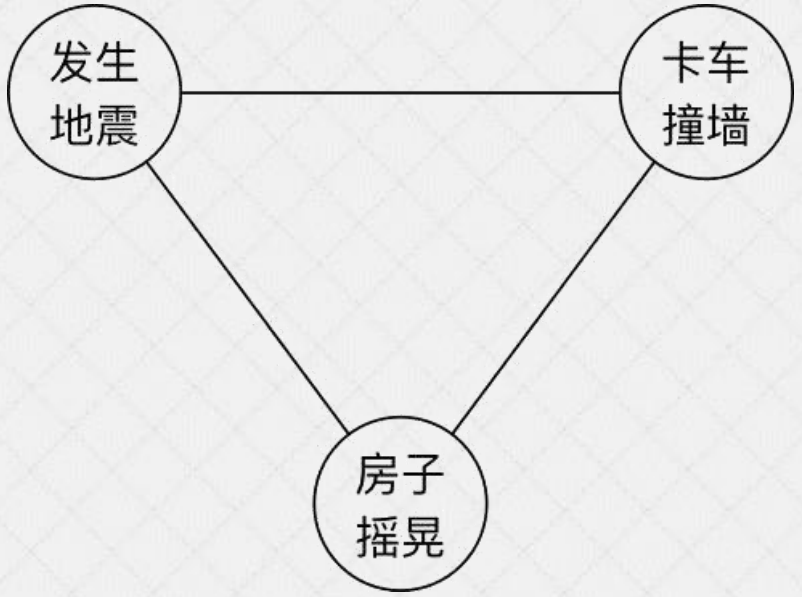
有向与无向概率图模型

- 有向图模型将概率有向图分解为一系列条件概率之积。定义$\boldsymbol{\pi}(v)$表示节点v的所有前驱节点，则多维随机变量的分布分解为：

$$p(\boldsymbol{x}, \boldsymbol{y}) = \prod_{v \in V} p\left(v | \boldsymbol{\pi}(v)\right)$$

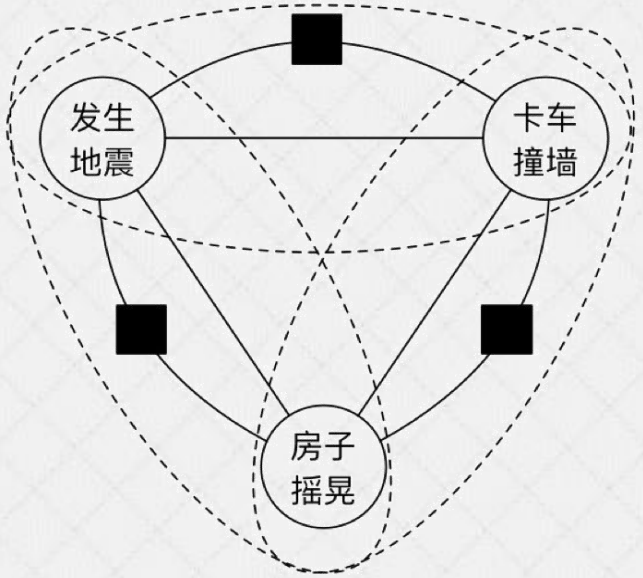
有向与无向概率图模型

- **无向图模型**则不探究每个事件的因果关系，即不涉及条件概率分解，无向图模型的边没有方向，仅仅代表两个事件有关联，不表示谁是因谁是果



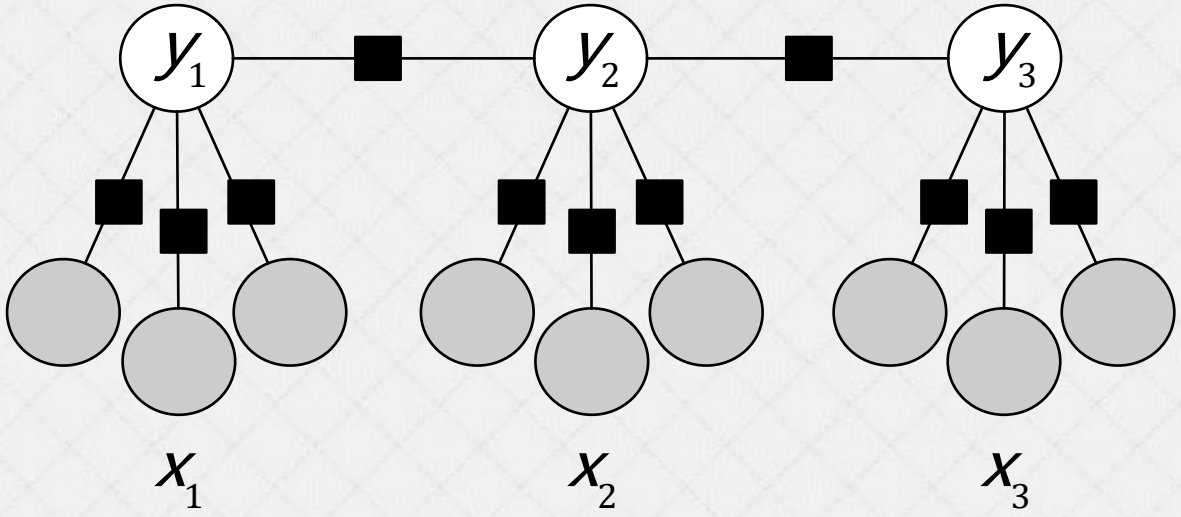
有向与无向概率图模型

- 无向图模型将概率分解为所有最大团上的某种函数之积
  - **最大团**（maximal clique）指的是满足所有节点相互连接的最大子图
  - 无向图模型定义了一些虚拟的**因子节点**（factor），每个因子节点只连接部分节点，组成更小的最大团



条件随机场

- **条件随机场**（Conditional Random Field，CRF）是一种给定输入随机变量$\boldsymbol{x}$，求解条件概率$p(\boldsymbol{y}|\boldsymbol{x})$的概率无向图模型
- 用于序列标注时，特例化为线性链（linear-chain）条件随机场



条件随机场

- 线性链条件随机场的定义如下：

$$p(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{t=1}^{T} \exp\left\{ \sum_{k=1}^{K} w_k f_k(y_{t-1}, y_t, \mathbf{x}_t) \right\}$$

- 其中，$Z(\boldsymbol{x})$为归一化函数：

$$Z(\boldsymbol{x}) = \sum_{\boldsymbol{y}} \prod_{t=1}^{T} \exp\left\{ \sum_{k=1}^{K} w_k f_k(y_{t-1}, y_t, \boldsymbol{x}_t) \right\}$$

与感知机的联系

- 条件随机场与结构化感知机的联系：
  - 条件随机场和结构化感知机的特征函数完全一致；
  - 结构化感知机对某预测打分越高，条件随机场给予该预测的概率也越大。
- 由于条件随机场的对数似然函数为凸函数（concave function），所以可以利用许多凸优化算法。
  - 与感知机采用相同的特征函数、权重向量、打分函数、预测算法，同为结构化学习，但采用更好的训练算法

词性标注

- **词性**（Part-Of-Speech，POS）指的是单词的语法分类，也称为词类。同一个类别的词语具有相似的语法性质
  - 所有词性的集合称为**词性标注集**。
  - 当下游应用遇到OOV时，可以通过OOV的词性猜测用法
  - 词性也可以直接用于抽取一些信息，比如抽取所有描述特定商品的形容词等

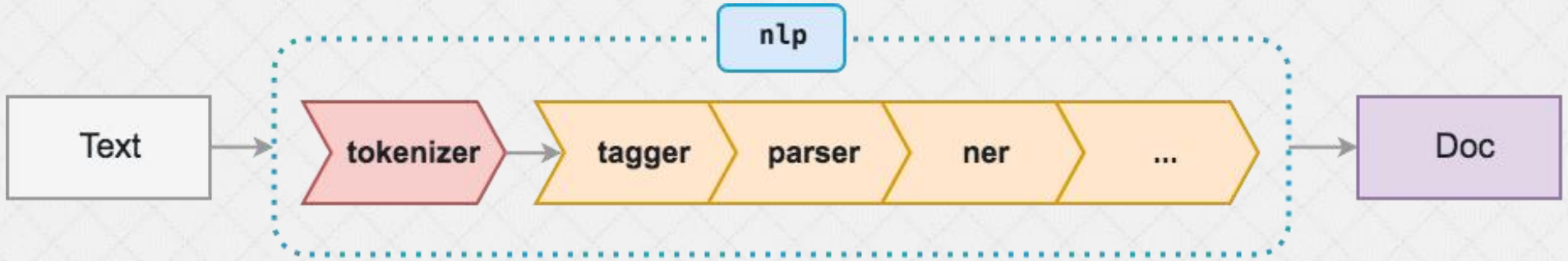
词性标注的难题

- **词性标注**指的是为句子中每个单词预测一个词性标签的任务
  - 汉语中一个单词多个词性的现象很常见（称作兼类词）
  - OOV是任何自然语言处理任务的难题
  - 统计方法中的**序列标注模型**为以上两个难题提供了解决方案
- 同时进行多个任务的模型称为**联合模型**（joint model）

商	品	和	服	务
B-名词	E-名词	S-连词	B-名词	E-名词

词性标注模型

- **实际采用流水线模式**
  - 中文分词语料库远远多于词性标注语料库
  - 实际工程上通常在大型分词语料库上训练分词器
  - 然后与小型词性标注语料库上的词性标注模型灵活组合为一个异源的流水线式词法分析器



词性标注语料库与标注集

- 目前还没有一个被广泛接受的汉语词性划分标准
- 《人民日报》语料库与PKU标注集：
  - １９９７年/t １２月/t ３１日/t 午夜/t，/w 聚集/v 在/p 日本/ns 东京/ns 增上寺/ns 的/u 善男信女/i 放飞/v 气球/n，/w 祈祷/v 新年/t 好运/n 。
- 国家语委语料库与863标注集
  - 其词类体系分为20个一级类、29个二级类
- 更具时效性的一些标注新闻领域材料（网络文本）的语料库

其他任务应用

命名实体识别

- 文本中有一些描述实体的词汇，比如人名、地名、组织机构名、股票基金、医学术语等，称为**命名实体**（named entity）
  - 数量无穷-宇宙中的恒星、蛋白质、病菌的名称
  - 构词灵活-简称、嵌套，如：中国工商银行
  - 类别模糊-地名与机构名，如：国家博物馆
  - 基于不同的领域和关注点可能变化

命名实体识别

- 识别出句子中命名实体的**边界**与**类别**的任务称为**命名实体识别**（Named Entity Recognition，NER）
  - 对于规则性较强的命名实体，完全可以通过正则表达式处理-网址、Email
  - 对于较短的命名实体，比如人名，完全可以通过分词确定边界，通过词性标注模块确定类别-微软系列的语料库，命名实体标注为一个词，粒度较大
  - 在一些语料库中（如PKU等），机构名这样的复合词是拆开的，此时就需要一个专门的命名实体识别模块了-可转化为序列标注问题，附着BMES

基于规则的命名实体识别

- 一段待识别的文本中，若音译字符连续出现，则很有可能来自一个音译人名（需建立常见音译字词典）
  - 莎士比亚、亚历山大、华莱士、福克斯、马尔蒂尼
- 数词英文识别
  - 牛奶三〇〇克壹佰块、牛奶300克100块、牛奶300g100rmb
- 效果有限，适用于一些语料匮乏的专门领域

基于统计方法的命名实体识别

- 基于序列标注的命名实体识别
  - 可以通过B-nt等附加类别的标签来确定边界和类别
  - 可以基于HMM、感知机、条件随机场
- 命名实体语料针对性较强-标注关注的命名实体
  - 1998年《人民日报》语料库
  - 微软命名实体识别语料库

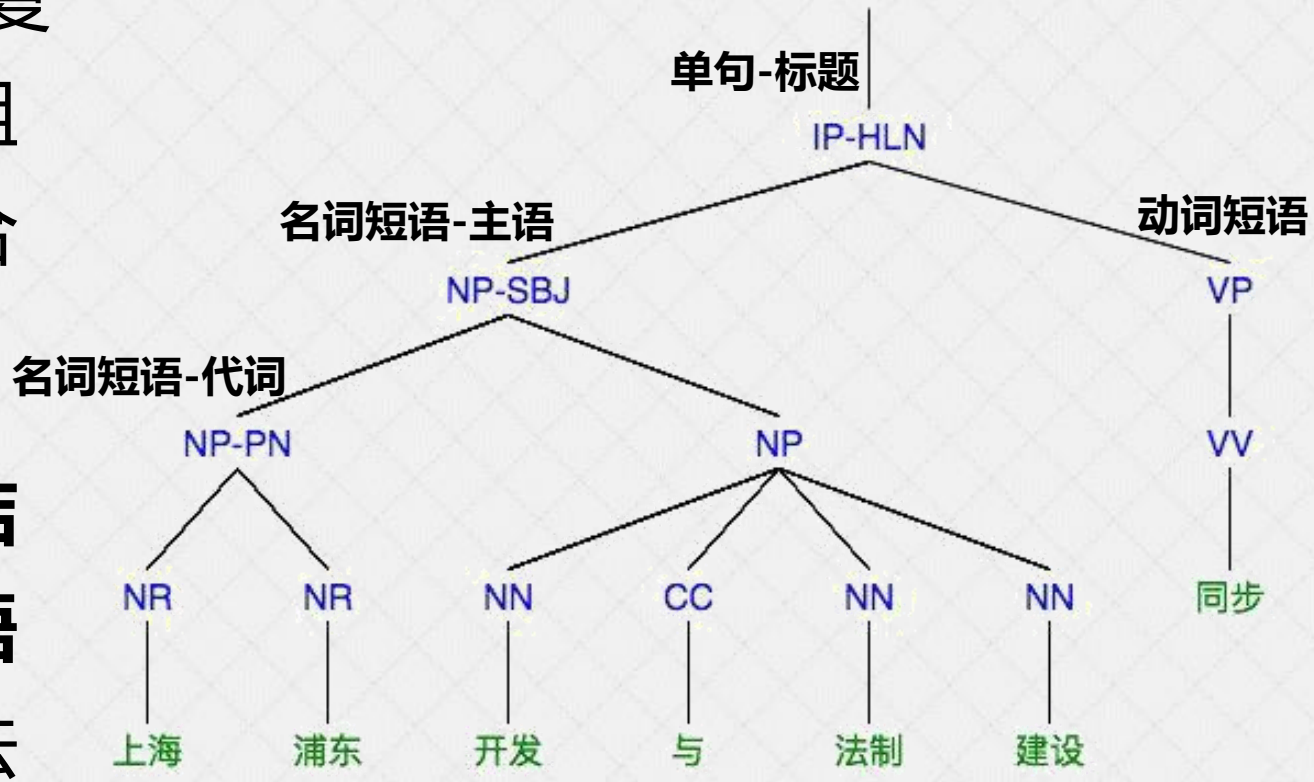
依存句法分析

- **语法分析**的目标是分析句子的语法结构，并将其表示为容易理解的结构（通常是树形结构）
  - NLP中的重要任务，也是较为高级和复杂的任务
  - 语法形式 - 短语结构树、依存句法树

其他任务应用

短语结构树

- 语言满足**复合性原理**-一个复杂表达式的意义是由其各组成部分的意义以及用以结合它们的规则来决定的
- 这样的树形结构称为**短语结构树**，相应的语法称为**短语结构语法**或上下文无关文法



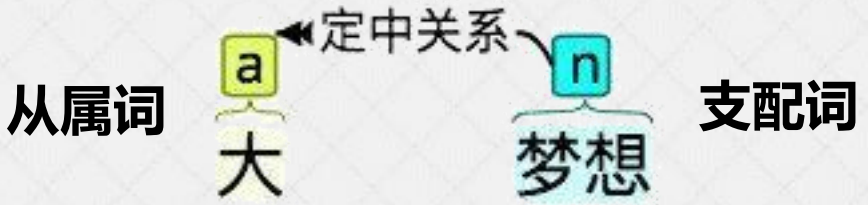
宾州树库和中文树库

- 语言学家制定短语结构语法规范，将大量句子人工分解为树形结构，形成了一种语料库，称为**树库**（treebank）
- 常见的英文树库有宾州树库，相应地，中文领域有CTB

标记	释义
IP-HLN	单句-标题
NP-SBJ	名词短语-主语
NP-PN	名词短语-代词
NP	名词短语
VP	动词短语

依存句法树

- 关注句子中词语之间的语法联系，并且将其约束为树形结构
  - 依存语法理论认为词与词之间存在主从关系，是一种二元不等价的关系
  - 如果一个词修饰另一个词，则称修饰词为从属词，被修饰的词语称为支配词，两者之间的语法关系称为依存关系



依存句法树

- 将一个句子中所有词语的依存关系以有向边的形式表示出来，就会得到一棵树，称为依存句法树（UD依存句法树库）

依存句法理论

- 现代依存语法中，语言学家Robinson对依存句法树提了4个约束性的公理。

1. 有且只有一个词语（ROOT，虚拟根节点，简称虚根）不依存于其他词语。根节点唯一性
2. 除此之外所有单词必须依存于其他单词。连通
3. 每个单词不能依存于多个单词。无环
4. 如果单词A依存于B，那么位置处于A和B之间的单词C只能依存于A、B或AB之间的单词。投射性（projective）

依存句法分析

- 分析句子的依存语法的一种中高级NLP任务，其输入通常是词语和词性，输出则是一棵依存句法树。
  - **基于图**-依存句法树其实是完全图（每对顶点都相连的图）的一个子图，为完全图中的每条边属于句法树与否的可能性打分，利用Prim之类的算法找出最大生成树（MST）作为依存句法树
  - **基于转移**-将一棵依存句法树的构建过程表示为动作，计算机就能够根据这些动作拼装出正确的依存句法树了，这种拼装动作称为转移（transition），而这类算法统称为基于转移的依存句法分析

课后实践

- 用Python语言实现HMM，并将其运用到中文分词
- 对比基于HMM的中文分词与其他分词方法的效果