Natural Language Processing I

7. 序列标注

WU Xiaokun 吴晓堃

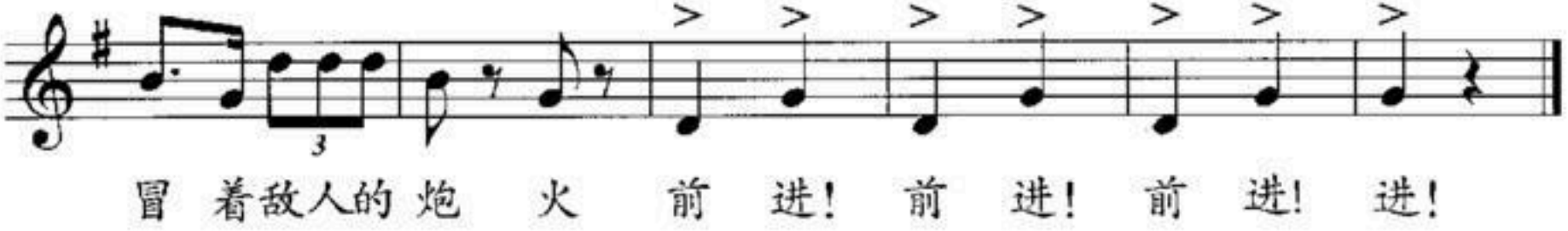
xkun.wu [at] gmail

2022/04/08

语言序列与标注

给每个字一个婉转的音符

To each word a warbling note.
– A Midsummer Night's Dream, Act 5, Scene 1



陌生词问题

词典随历史不断演化：古文中的未知词问题

> *头上戴着**束发嵌宝紫金冠**，齐眉勒着**二龙戏珠金抹额**——《红楼梦》*

陌生词问题

词典随历史不断演化：古文中的未知词问题

> *头上戴着**束发嵌宝紫金冠**，齐眉勒着**二龙戏珠金抹额*——《红楼梦》

但人通常可以**推断**陌生词的大致含义

- “头上戴着”：描述头饰；“冠”：帽子，表身份、地位
- “齐眉”、“抹额”：描述头饰的位置，可能类似导汉巾

先验知识：相似语境，语义通常相近

陌生词问题

词典随历史不断演化：古文中的未知词问题

> *头上戴着**束发嵌宝紫金冠**，齐眉勒着**二龙戏珠金抹额**——《红楼梦》*

但人通常可以**推断**陌生词的大致含义

- “头上戴着”：描述头饰；“冠”：帽子，表身份、地位
- “齐眉”、“抹额”：描述头饰的位置，可能类似导汉巾

先验知识：相似语境，语义通常相近

- 通过形象进行**类比**：辅助推断

> *南海龙王敖钦道：“我有一顶**凤翅紫金冠**哩。”——《西游记》*

从词语到字

词语级别的模型：**低 OOV 召回率**

- 粒度越小：训练集词汇表中越容易出现
- 字级别：更细粒度

从词语到字

词语级别的模型：**低 OOV 召回率**

- 粒度越小：训练集词汇表中越容易出现
- 字级别：更细粒度

字级别的模型：假设能够**识别新词起止范围**，就不局限于词典

- 词典随文本**动态构造**
- 再次印证：分词只是手段，是可选预处理

从词语到字

词语级别的模型：**低 OOV 召回率**

- 粒度越小：训练集词汇表中越容易出现
- 字级别：更细粒度

字级别的模型：假设能够**识别新词起止范围**，就不局限于词典

- 词典随文本**动态构造**
- 再次印证：分词只是手段，是可选预处理

识别新词可以看成**分类问题**

- 每个字对应一个标签；**连续标签**组合成词

> *头上戴着**束发嵌宝紫金冠**，齐眉勒着**二龙戏珠金抹额** —《红楼梦》*

词类：历史

最早的西方语法书（大约公元前100年）：希腊语

- 西方语言学术语的源头：附会文明

词类：历史

最早的西方语法书（大约公元前100年）：希腊语

- 西方语言学术语的源头：附会文明
- **8类词**：名词、动词、代词、介词、副词、连词、分词、冠词
  - 之后2000多年欧洲语法框架的基础

词类：历史

最早的西方语法书（大约公元前100年）：希腊语

- 西方语言学术语的源头：附会文明
- **8类词**：名词、动词、代词、介词、副词、连词、分词、冠词
  - 之后2000多年欧洲语法框架的基础

20世纪50年代：从结构主义到**形式文法**

- Chomsky，《句法结构》，转换-生成语法

词类：历史

最早的西方语法书（大约公元前100年）：希腊语

- 西方语言学术语的源头：附会文明
- **8类词**：名词、动词、代词、介词、副词、连词、分词、冠词
  - 之后2000多年欧洲语法框架的基础

20世纪50年代：从结构主义到**形式文法**

- Chomsky，《句法结构》，转换-生成语法

汉语语法：1898年

- **硬搬西方框架**："言必称希腊"
  - 新文化运动激进派：废除汉字、"直译欧文句法"
    - 鲁迅：我确实说过，而且是推动者

词类：历史

最早的西方语法书（大约公元前100年）：希腊语

- 西方语言学术语的源头：附会文明
- **8类词**：名词、动词、代词、介词、副词、连词、分词、冠词
  - 之后2000多年欧洲语法框架的基础

20世纪50年代：从结构主义到**形式文法**

- Chomsky，《句法结构》，转换-生成语法

汉语语法：1898年

- **硬搬西方框架**："言必称希腊"
  - 新文化运动激进派：废除汉字、"直译欧文句法"
    - 鲁迅：我确实说过，而且是推动者
- 融合新旧思想：**逐步探索**现代中文语法体系

中文西化

具体名词做主语；不常用“的”、“地”

> 西化：他的收入的减少改变了他的生活方式。
> 自然：他因收入减少而改变生活方式。

中文西化

具体名词做主语；不常用“的”、“地”

> 西化：他的收入的减少改变了他的生活方式。
> 自然：他因收入减少而改变生活方式。

不常用“一”；复数不用“们”；不常用连词；不常用“被”（通常表不幸、贬义）

> 西化：你是一个好人，但同学们都知道，我们不适合。请别悲伤和难过。
> 自然：我是好人。诸位同学都知道，我被分手了，但我并不悲伤难过。

中文西化

具体名词做主语；不常用“的”、“地”

> 西化：他的收入的减少改变了他的生活方式。
> 自然：他因收入减少而改变生活方式。

不常用“一”；复数不用“们”；不常用连词；不常用“被”（通常表不幸、贬义）

> 西化：你是一个好人，但同学们都知道，我们不适合。请别悲伤和难过。
> 自然：我是好人。诸位同学都知道，我被分手了，但我并不悲伤难过。

尽量不用介词；不常用“性”；定语短（拆成分句）

> 西化：“在公共场合不准吸烟”是普遍性的和强制性的原则。
> 自然：“公共场合不准吸烟”是普遍原则，通常强制执行。

中文西化

具体名词做主语；不常用“的”、“地”

> 西化：他的收入的减少改变了他的生活方式。
> 自然：他因收入减少而改变生活方式。

不常用“一”；复数不用“们”；不常用连词；不常用“被”（通常表不幸、贬义）

> 西化：你是一个好人，但同学们都知道，我们不适合。请别悲伤和难过。
> 自然：我是好人。诸位同学都知道，我被分手了，但我并不悲伤难过。

尽量不用介词；不常用“性”；定语短（拆成分句）

> 西化：“在公共场合不准吸烟”是普遍性的和强制性的原则。
> 自然：“公共场合不准吸烟”是普遍原则，通常强制执行。

> 直译：我不能同意得更多。听起来……的样子。
> 日语：以上。

词类、命名实体

词类 Parts Of Speech (POS)：单词的语法分类，也称**词性**

- 邻接关系：形容词 + 名词，副词 + 动词
- 句法结构：动词依赖于名词

词类、命名实体

词类 Parts Of Speech (POS)：单词的语法分类，也称**词性**

- 邻接关系：形容词 + 名词，副词 + 动词
- 句法结构：动词依赖于名词

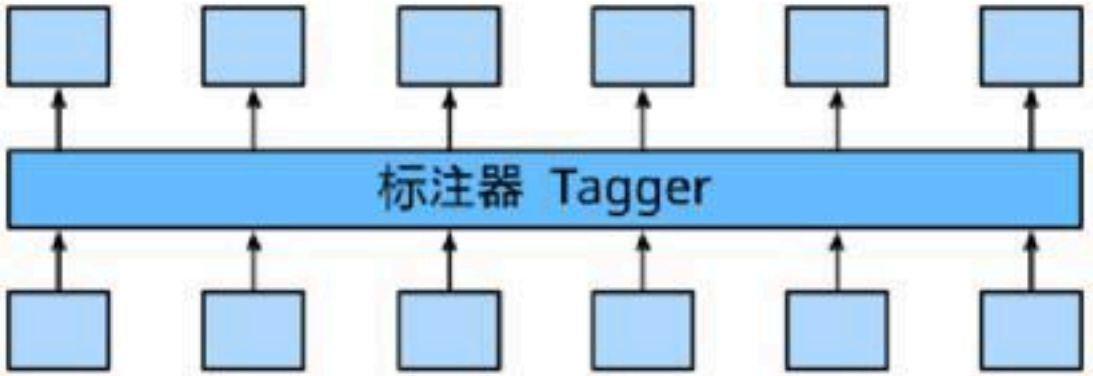
命名实体 Named Entity (NE)：描述实体的词汇

- 人名、地点、组织、日期、价格
- 应用：机器问答、语义关系

序列标注

序列标注 sequence labeling：给每个字一个类别标签

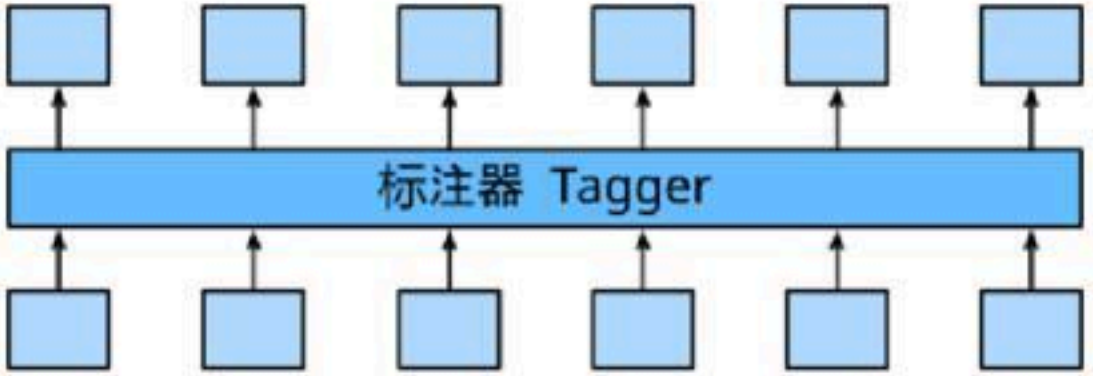
$$f : x_i \mapsto y_i$$



序列标注

序列标注 sequence labeling：给每个字一个类别标签

$$f : x_i \mapsto y_i$$



可以看作**有意义地分词**

- POS标注：字（词） $\mapsto$ 划分 + 词类
- 命名实体识别：字（词） $\mapsto$ 划分 + NE类别

词类：讨论范畴

词类有**语义倾向**

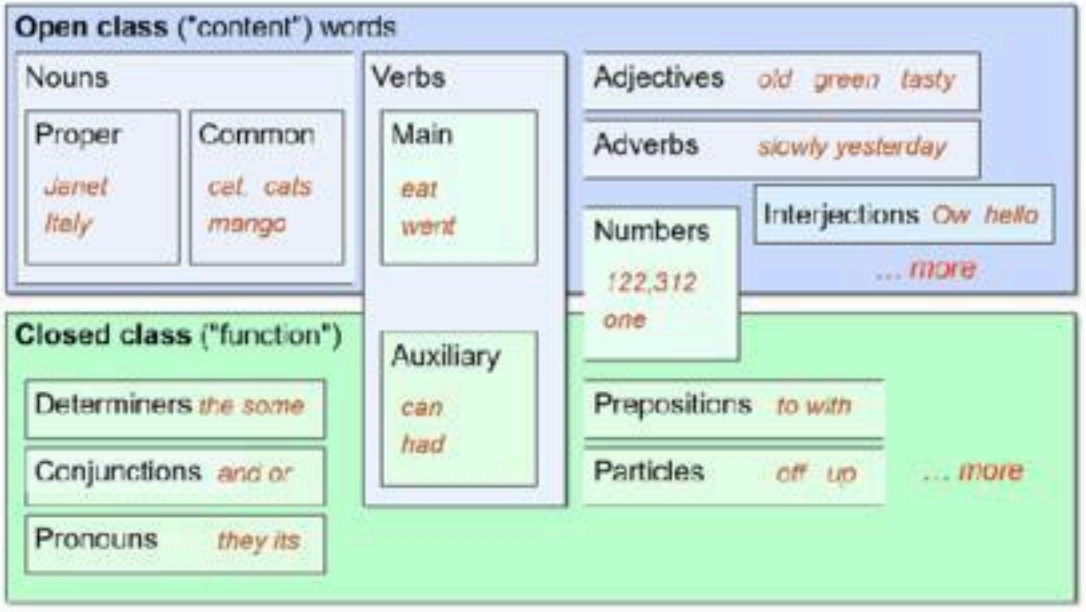
- 形容词：属性；名词：人、物

词类：讨论范畴

词类有**语义倾向**

- 形容词：属性；名词：人、物

NLP中POS主要讨论相邻词之间的**语法关系**



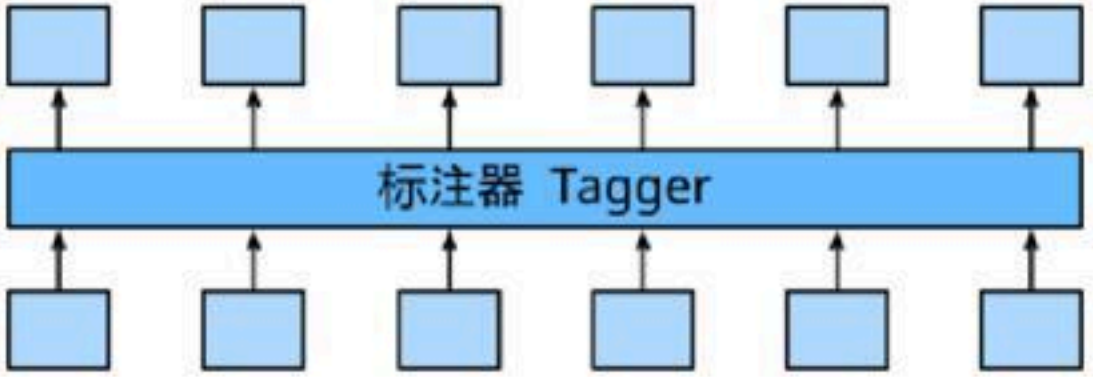
Penn Treebank 标签集

Tag	Description	Example	Tag	Description	Example	Tag	Des
CC	coord. conj.	and, but, or	NNP	proper noun, sing.	IBM	TO	"to"
CD	cardinal number	one, two	NNPS	proper noun, plu.	Carolinas	UH	inte
DT	determiner	a, the	NNS	noun, plural	llamas	VB	verb
EX	existential 'there'	there	PDT	predeterminer	all, both	VBD	verb tens
FW	foreign word	mea culpa	POS	possessive ending	's	VBG	verb
IN	preposition/subordin-conj	of, in, by	PRP	personal pronoun	I, you, he	VBN	verb part
JJ	adjective	yellow	PRP$	possess. pronoun	your, one's	VBP	verb

Tag	Description	Example	Tag	Description	Example	Tag	Des
JJR	comparative adj	bigger	RB	adverb	quickly	VBZ	verb
JJS	superlative adj	wildest	RBR	comparative adv	faster	WDT	wh-
LS	list item marker	1, 2, One	RBS	superlatv. adv	fastest	WP	wh-
MD	modal	can, should	RP	particle	up, off	WP$	wh- pos
NN	sing or mass noun	llama	SYM	symbol	+,%, &	WRB	wh-

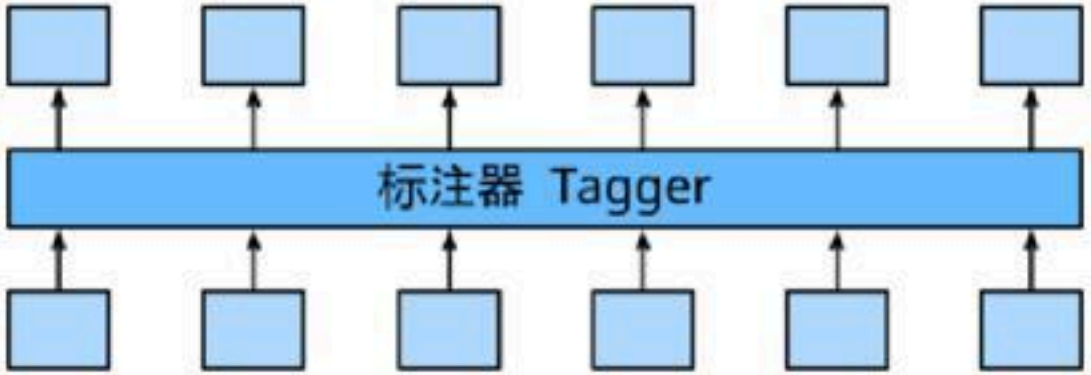
POS标注

词类 Parts Of Speech (POS)：单词的语法分类，也称**词性**



POS标注

词类 Parts Of Speech (POS)：单词的语法分类，也称**词性**



POS标注可以看作**消除歧义**的任务

```
Book that flight
Hand me that book
```

她希望筹建一所希望小学。
她希望成为全村人的希望。

POS标注：有何用途?

其他NLP任务的预处理

- 句法解析：效能依赖于词类的正确判别
- 机器翻译：语序矫正，如形容词（英语到西班牙语）
- 情感分析：形容词提供直接依据
- 语音助手：动词、形容词
  - 重音，如“object”；声调，如“好”

POS标注：有何用途?

其他NLP任务的预处理

- 句法解析：效能依赖于词类的正确判别
- 机器翻译：语序矫正，如形容词（英语到西班牙语）
- 情感分析：形容词提供直接依据
- 语音助手：动词、形容词
  - 重音，如“object”；声调，如“好”

语言学分析、计算任务

- 语义演化、创新词
  - “生而不有”
- 词义相似度量：同类相似原则

POS标注：算法准确度

（英语）大概15%的词目是有歧义的

- 但这些词非常常见
  - 从语言演化角度看：人倾向于**给简单词赋予多重含义**
  - 从信息论编码角度看：码越短，频率越高

POS标注：算法准确度

（英语）大概15%的词目是有歧义的

- 但这些词非常常见
  - 从语言演化角度看：人倾向于**给简单词赋予多重含义**
  - 从信息论编码角度看：码越短，频率越高
- 文本中**大概60%**的词是有歧义的

POS标注：算法准确度

（英语）大概15%的词目是有歧义的

- 但这些词非常常见
  - 从语言演化角度看：人倾向于**给简单词赋予多重含义**
  - 从信息论编码角度看：码越短，频率越高
- 文本中**大概60%**的词是有歧义的

[Wu 2019] Universal Dependency (UD) 树库上的**准确度：97%**

- 英语树库：准确度大致相同，无论什么算法
- [Manning 2011]（英语）人工标注的准确度：97%

POS标注：算法准确度

（英语）大概15%的词目是有歧义的

- 但这些词非常常见
  - 从语言演化角度看：人倾向于**给简单词赋予多重含义**
  - 从信息论编码角度看：码越短，频率越高
- 文本中**大概60%**的词是有歧义的

[Wu 2019] Universal Dependency (UD) 树库上的**准确度：97%**

- 英语树库：准确度大致相同，无论什么算法
- [Manning 2011]（英语）人工标注的准确度：97%

结论：算法与人的表现**差不多**，而且**都很好**

- 是因为POS标注比较简单吗?

POS标注：并不简单

尽管文本中大概60%的词是有歧义的

- 歧义对应的标签：**频率不平衡**
  - 例如：“a”作为冠词的可能性要高得多

POS标注：并不简单

尽管文本中大概60%的词是有歧义的

- 歧义对应的标签：**频率不平衡**
  - 例如："a"作为冠词的可能性要高得多

基准算法

对每个有歧义的词：标注为训练集中**频率最高**的词类

POS标注：并不简单

尽管文本中大概60%的词是有歧义的

- 歧义对应的标签：**频率不平衡**
  - 例如："a"作为冠词的可能性要高得多

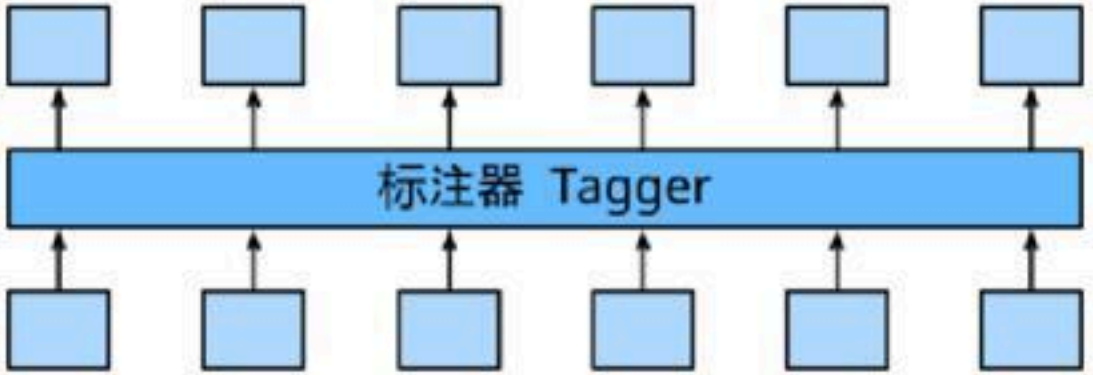
基准算法

对每个有歧义的词：标注为训练集中**频率最高**的词类

- 基准算法的准确度：92%
  - 算法、人工**只能改进5%**的效能：97%
  - 相对比例：$(97-92)/92 = 5.435\%$

NE标注

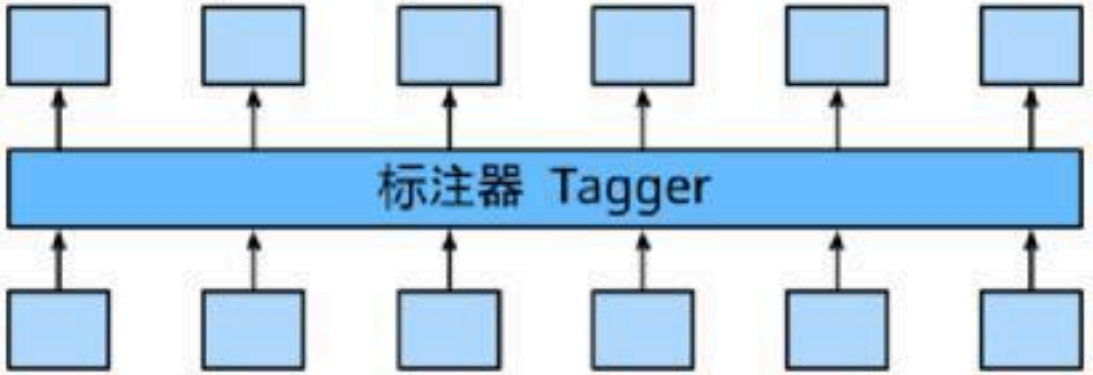
命名实体 Named Entity (NE)：描述实体的词汇



- 人名 PER、组织 ORG、地点 LOC、地域 GPE

NE标注

命名实体 Named Entity (NE)：描述实体的词汇



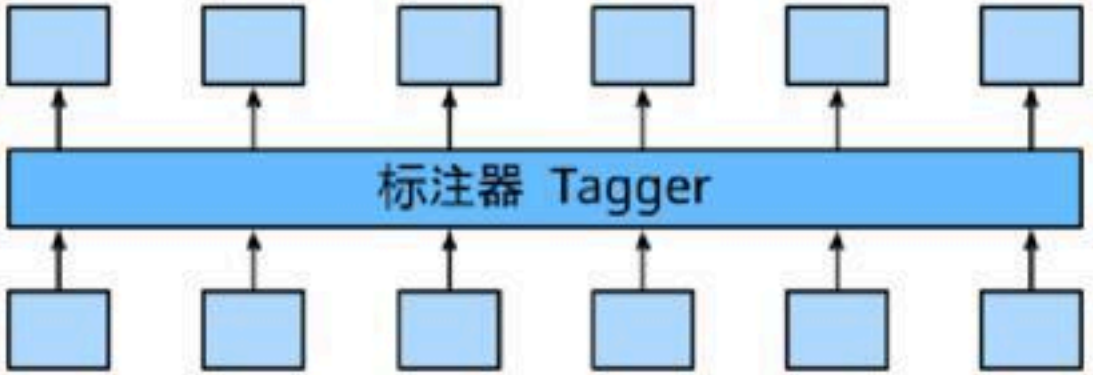
- 人名 PER、组织 ORG、地点 LOC、地域 GPE

NE标注可以看作**专有名词的分词**

- 构成专有名词的起止范围 + 标签类型

NE标注

命名实体 Named Entity (NE)：描述实体的词汇



- 人名 PER、组织 ORG、地点 LOC、地域 GPE

NE标注可以看作**专有名词的分词**

- 构成专有名词的起止范围 + 标签类型

难点：分割，并识别，即分词的高级任务

- 也称**命名实体识别 Named Entity Recognition (NER)**

NE标注：BIO

[Ramshaw 1995] 转换成**逐词标注**的任务

- B：一个子序列的开始，可简化
- I：子序列的内部
- O：不在任何子序列里

北	京	大	学	位	于	北	京	市
B-ORG	I-ORG	I-ORG	I-ORG	O	O	B-GPE	I-GPE	I-GPE
I-ORG	I-ORG	I-ORG	I-ORG	O	O	I-GPE	I-GPE	I-GPE

NE标注：BIO

[Ramshaw 1995] 转换成**逐词标注**的任务

- B：一个子序列的开始，可简化
- I：子序列的内部
- O：不在任何子序列里

北	京	大	学	位	于	北	京	市
B-ORG	I-ORG	I-ORG	I-ORG	O	O	B-GPE	I-GPE	I-GPE
I-ORG	I-ORG	I-ORG	I-ORG	O	O	I-GPE	I-GPE	I-GPE

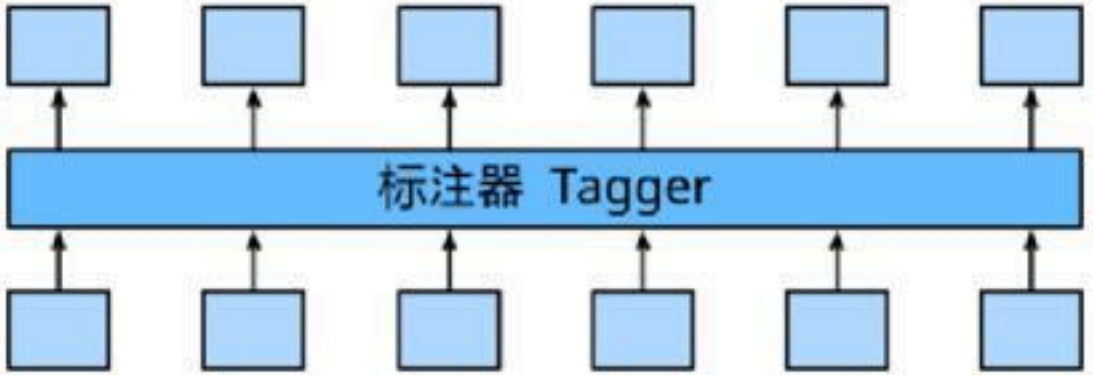
扩展标签种类：总共$2n+1$，或$n+1$

隱式Markov模型

回顾序列标注

序列标注：给每个字一个类别标签

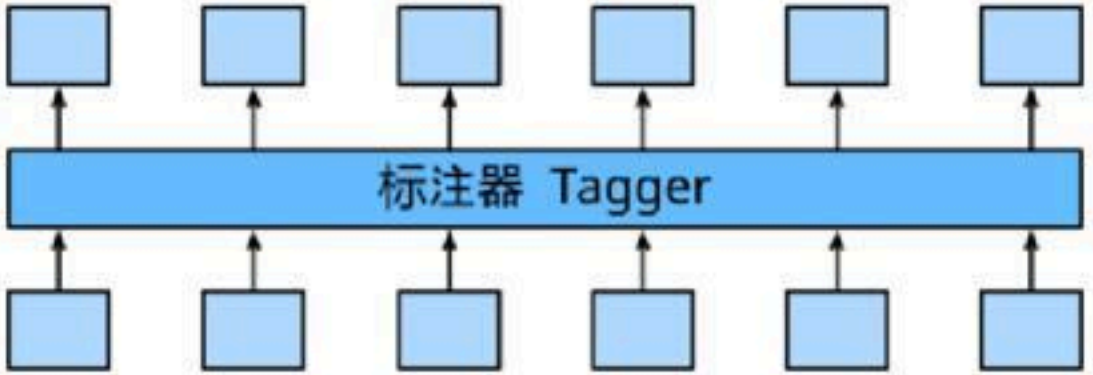
- 将字（词）映射到类别
- 将文字序列映射到**同长度**的标签序列



回顾序列标注

序列标注：给每个字一个类别标签

- 将字（词）映射到类别
- 将文字序列映射到**同长度**的标签序列



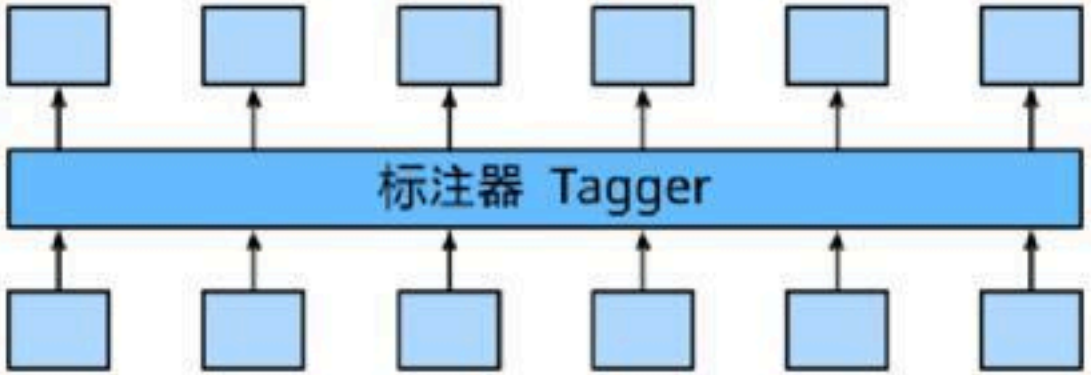
概率建模：计算T^N个**定长标签序列**概率，选出最可能的一个

- 既然是对序列建模：需要表达**链式节点组织关系**

回顾序列标注

序列标注：给每个字一个类别标签

- 将字（词）映射到类别
- 将文字序列映射到**同长度**的标签序列



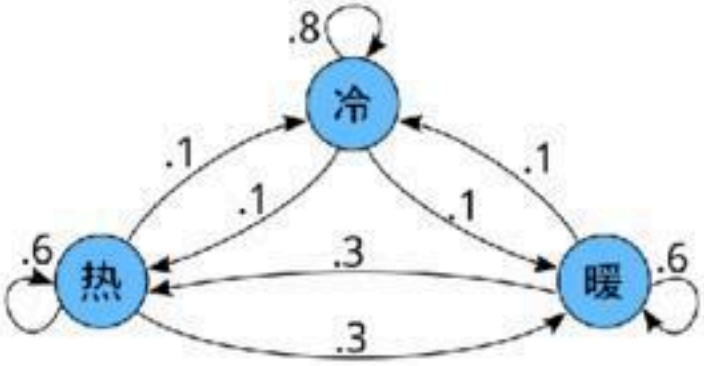
概率建模：计算T^N个**定长标签序列**概率，选出最可能的一个

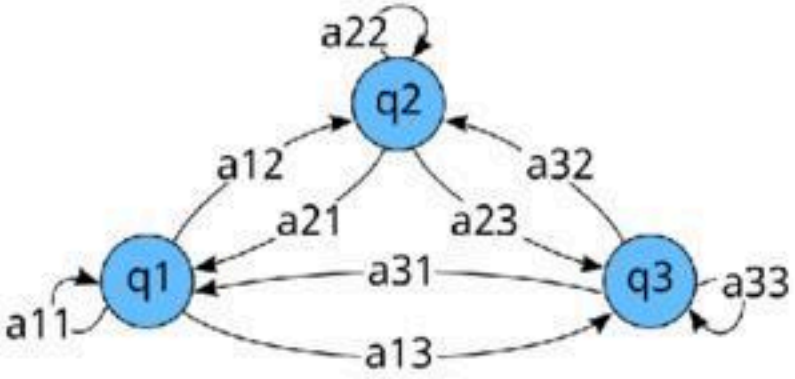
- 既然是对序列建模：需要表达**链式节点组织关系**
- 整条链有T^N种序列选择：等价于每个节点有T个**状态选择**

Markov链

Markov链：计算**随机变量序列**的概率

- 状态：随机变量，如气温体感
- 状态集：如{冷、热、暖}



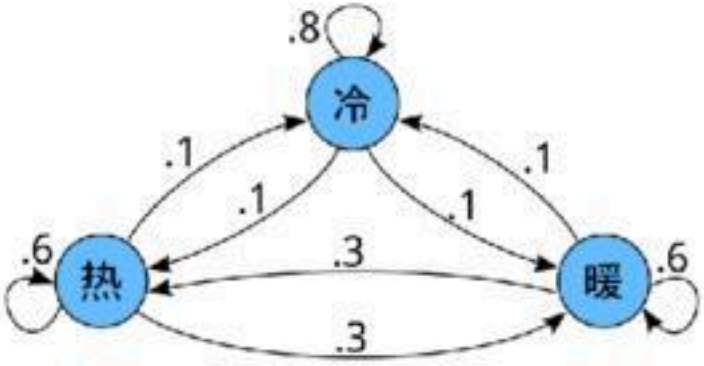


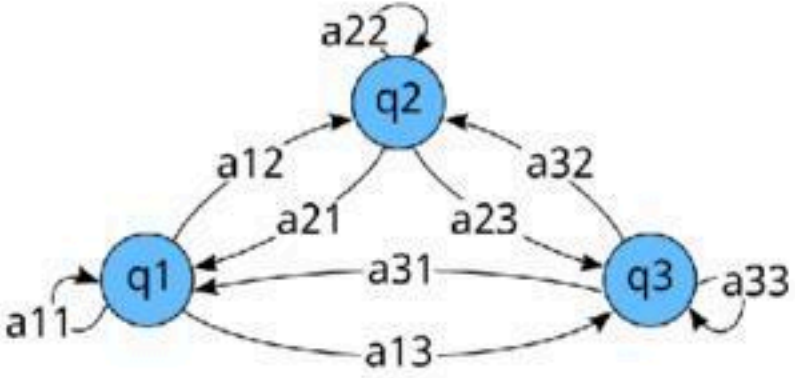
Markov假设

一阶**Markov假设**：未来的预测与过去无关，只看当前状态

$$P(q_i|q_1..q_{i-1}) = P(q_i|q_{i-1})$$

- $q_1..q_i$是状态序列



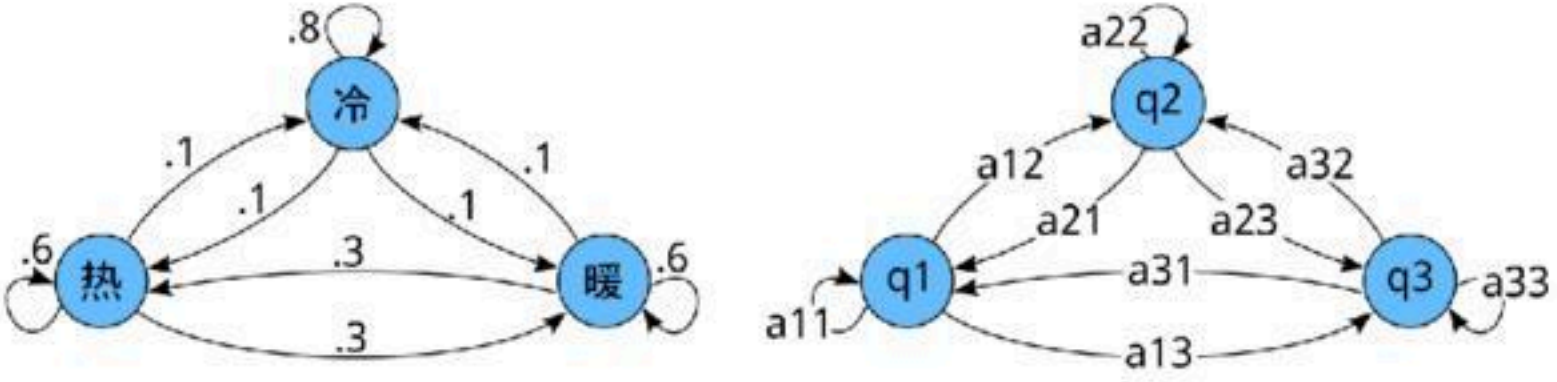


Markov假设

一阶**Markov假设**：未来的预测与过去无关，只看当前状态

$$P(q_i|q_1..q_{i-1}) = P(q_i|q_{i-1})$$

- $q_1..q_i$是状态序列

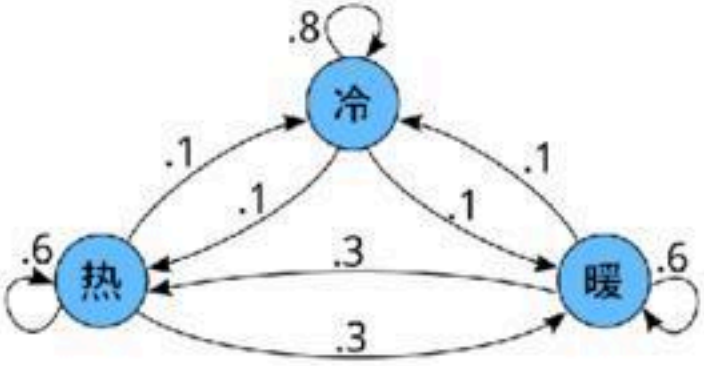


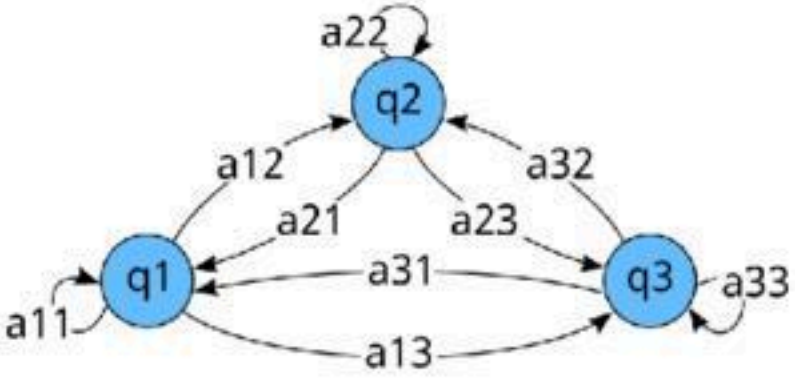
- 一阶Markov模型：等价于**二元语法**模型

Markov链：概率

假设**初始概率分布**：$\pi = [0.1, 0.7, 0.2]$

- 初始最可能的状态："冷"；概率：.7

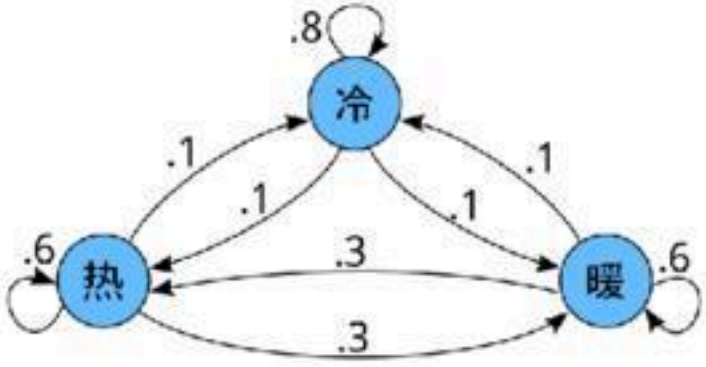


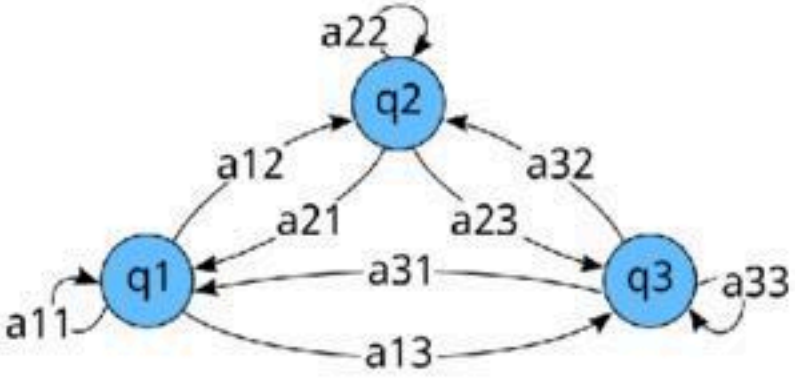


Markov链：概率

假设**初始概率分布**：$\pi = [0.1, 0.7, 0.2]$

- 初始最可能的状态："冷"；概率：.7





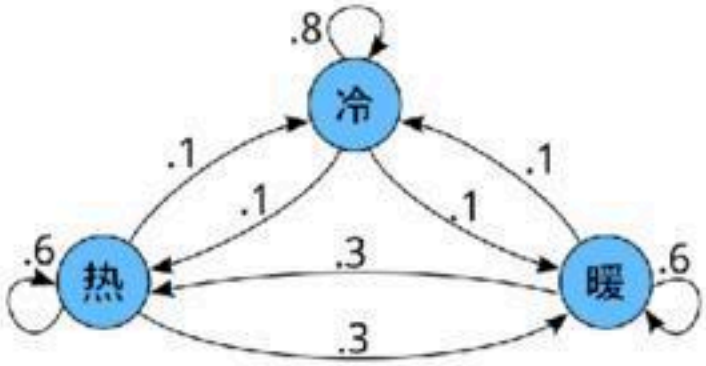
转移概率矩阵：状态变化的概率

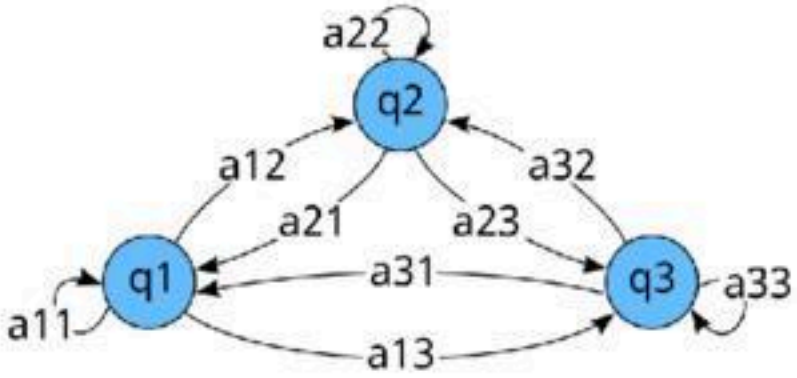
- 每行都是可能性集合

	热	冷	暖
热	.6	.1	.3
冷	.8	.1	.1
暖	.6	.3	.1

Markov链：序列概率计算举例

假设初始概率分布：$\pi = [0.1, 0.7, 0.2]$





观测序列：计算连续4天的气温

- 热 热 热 热: $.1 \times .6 \times .6 \times .6 = .0216$
- 冷 热 冷 热: $.7 \times .1 \times .1 \times .1 = .0007$

Markov链：严格定义

$Q = q_1 q_2 .. q_N$	N **states**
$A = a_{11} a_{12} .. a_{N1} .. a_{NN}$	**transition probability matrix** A, each a_{ij} representing the probability of moving from state q_i to state q_j, s.t. $\sum_{j=1}^{n} a_{ij} = 1, \forall i$
$\pi = \pi_1, \pi_2, ..., \pi_N$	**initial probability distribution** over states. π_i is the probability that the Markov chain will start in state q_i. Some states q_j may have $\pi_j = 0$, meaning that they cannot be initial states. Also, $\sum_{i=1}^{n} \pi_i = 1$

隐变量

Markov链：状态、事件必须**可观测**

- 根据观测结果计算初始、转移概率

隐变量

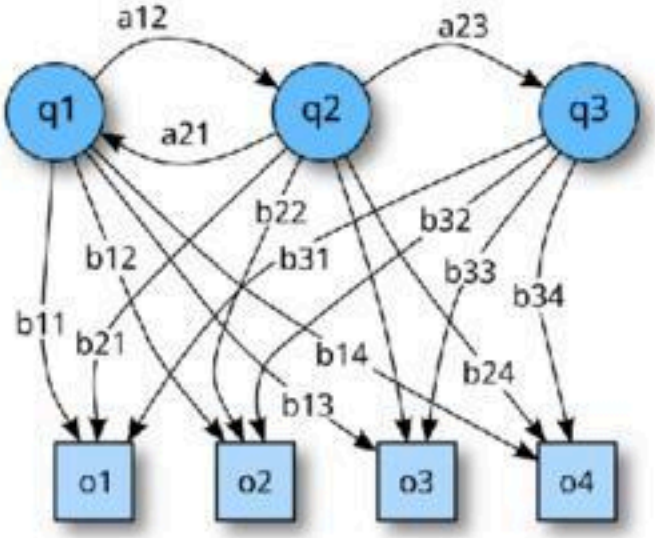
Markov链：状态、事件必须**可观测**

- 根据观测结果计算初始、转移概率

实际情况：状态、事件**不能被观测**

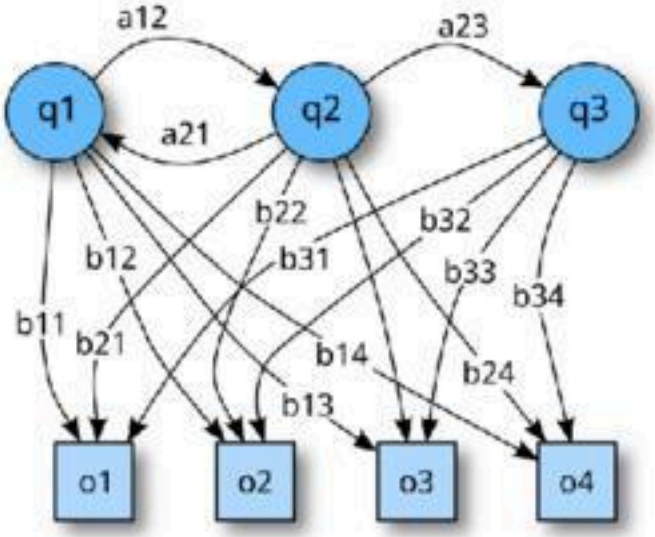
- 例如：推断词类标签的概率
- 此类变量称为：**隐 hidden 变量**

隐式Markov模型 (HMM)



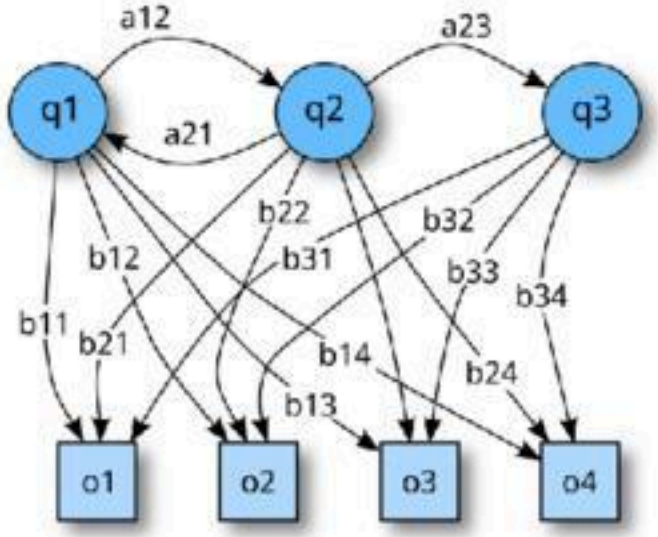
- 隐状态：Markov链，相互转化的**内因**
  - 《扁鹊见蔡桓公》：“**疾**在腠理”，“**病**在骨髓”，常人看不出“健康、疾、病”

隐式Markov模型 (HMM)



- 隐状态：Markov链，相互转化的**内因**
  - 《扁鹊见蔡桓公》：“**疾**在腠理”，“**病**在骨髓”，常人看不出“健康、疾、病”
- 输出变量：由隐状态直接生成的**表象**，例如发热

隐式Markov模型 (HMM)



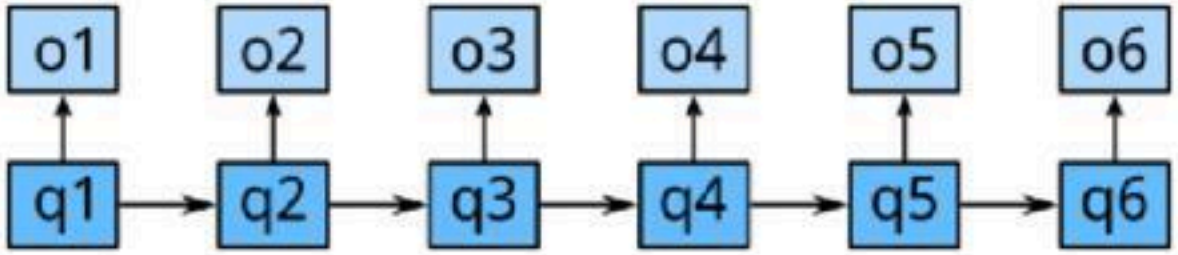
- 隐状态：Markov链，相互转化的**内因**
  - 《扁鹊见蔡桓公》：“**疾**在腠理”，“**病**在骨髓”，常人看不出“健康、疾、病”
- 输出变量：由隐状态直接生成的**表象**，例如发热

问题：**依赖链长**，计算复杂。例如：$(q_1 \to q_2)_{\times n} \to q_3 \to o_4$

- “病入骨髓”需要从健康状态开始计算吗?

HMM：一阶模型

一阶HMM的两个假设

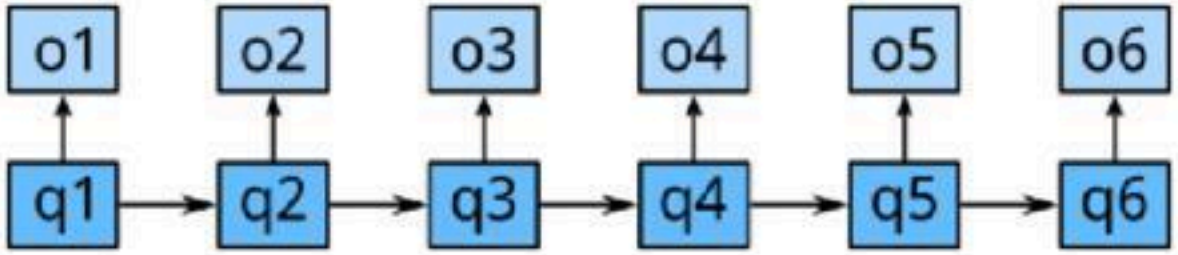


- 一阶Markov假设：（隐）状态概率只取决于**前一个**（隐）状态
  - 称为**状态转移概率**

$$P(q_i|q_1,..,q_{i-1}) \approx P(q_i|q_{i-1})$$

HMM：一阶模型

一阶HMM的两个假设



- 一阶Markov假设：（隐）状态概率只取决于**前一个**（隐）状态
  - 称为**状态转移概率**

$$P(q_i|q_1,..,q_{i-1}) \approx P(q_i|q_{i-1})$$

- 独立性假设：输出变量的概率只取决于**直接关联**的隐状态
  - 称为**观测似然，或"发射概率"**

$$P(o_i|q_1,..,q_T,o_1,..,o_T) \approx P(o_i|q_i)$$

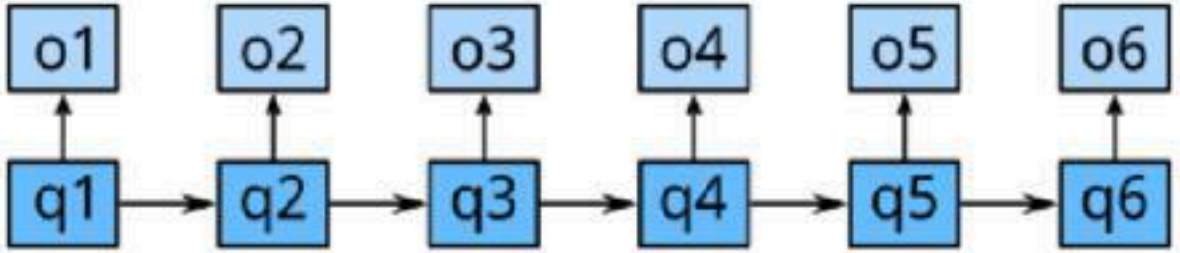
HMM：严格定义

$Q = q_1 q_2 .. q_N$	N **states**
$A = a_{11} a_{12} .. a_{N1} .. a_{NN}$	**transition probability matrix** A, each a_{ij} representing the probability of moving from state q_i to state q_j, s.t. $\sum_{j=1}^{n} a_{ij} = 1, \forall i$
$\pi = \pi_1, \pi_2, ..., \pi_N$	**initial probability distribution** over states. π_i is the probability that the Markov chain will start in state q_i. Some states q_j may have $\pi_j = 0$, meaning that they cannot be initial states. Also, $\sum_{i=1}^{n} \pi_i = 1$
$O = o_1 o_2 ... o_T$	T **observations**, each one drawn from a vocabulary $V = v_1, v_2, ..., v_V$
$B = b_i(o_t)$	**observation likelihoods**, or **emission probabilities**, each expressing the probability of an observation o_t being generated from a state q_i

HMM标注器：状态转移概率

状态转移概率$P(t_i|t_{i-1})$：标签出现的条件概率

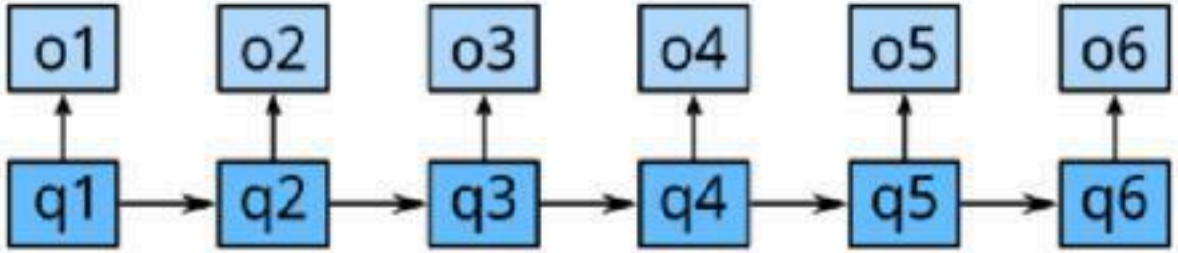
- 例如：P(名词|形容词)，P(动词|副词)



HMM标注器：状态转移概率

状态转移概率$P(t_i|t_{i-1})$：标签出现的条件概率

- 例如：P(名词|形容词)，P(动词|副词)



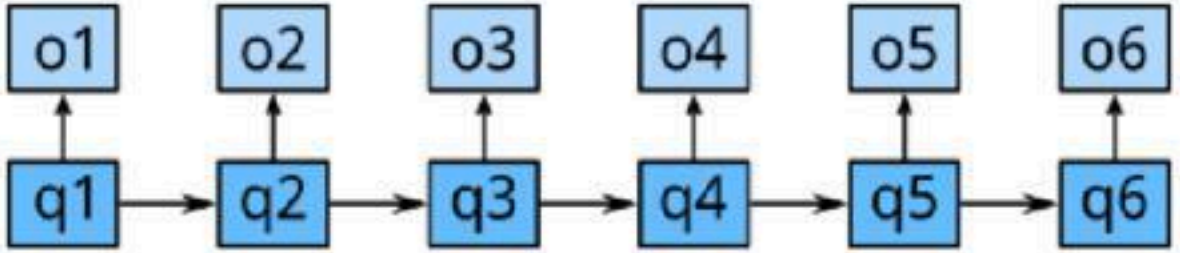
- 最大似然估计 MLE：标签t_{i-1}的邻居中，出现**标签对**(t_{i-1}, t_i)**的可能性**是多少?

$$P(t_i|t_{i-1}) = \frac{\Gamma(t_{i-1}, t_i)}{\Gamma(t_{i-1})}$$

HMM标注器：观测似然

观测似然$P(w_i|t_i)$：由标签生成词的概率

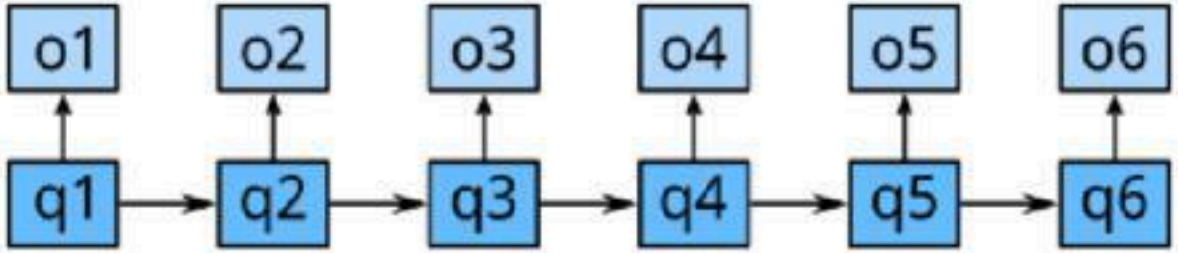
- 例如：P(好|形容词)，P(快|副词)



HMM标注器：观测似然

观测似然$P(w_i|t_i)$：由标签生成词的概率

- 例如：P(好|形容词)，P(快|副词)



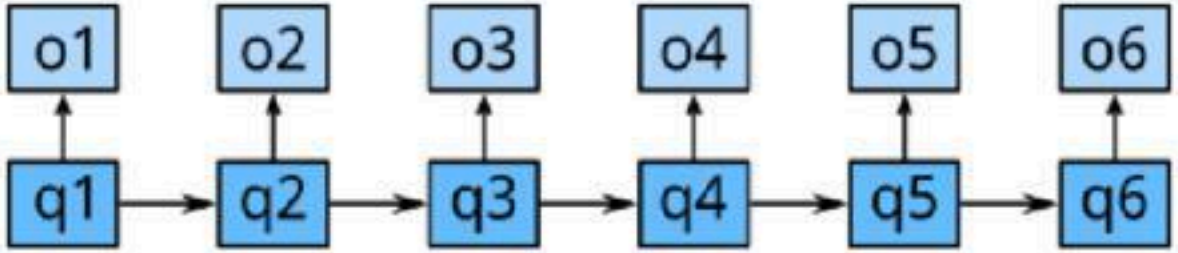
- MLE：假设要生成标签t_i，**词**w_i**的可能性**是多少?

$$P(w_i|t_i) = \frac{\Gamma(t_i, w_i)}{\Gamma(t_i)}$$

HMM标注器：观测似然

观测似然$P(w_i|t_i)$：由标签生成词的概率

- 例如：P(好|形容词)，P(快|副词)



- MLE：假设要生成标签t_i，**词w_i的可能性**是多少?

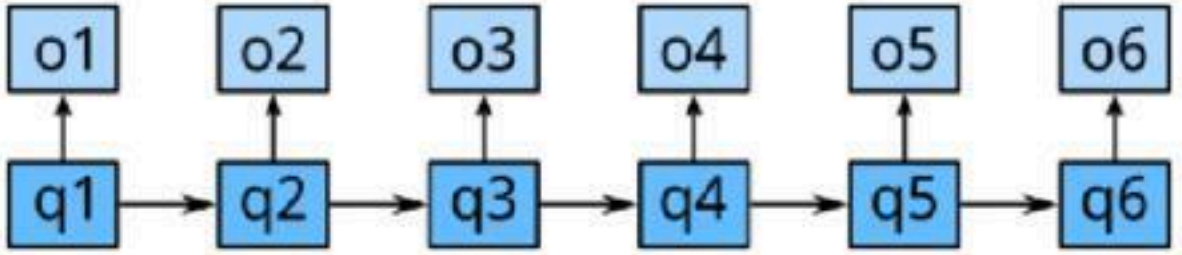
$$P(w_i|t_i) = \frac{\Gamma(t_i, w_i)}{\Gamma(t_i)}$$

- 对比后验$P(t_i|w_i)$：**词w_i最可能的标签**是什么?
  - 即标注任务的根本目标

HMM标注器：解码观点

解码：给定观测序列，确定隐变量序列；等价于解读未知密码

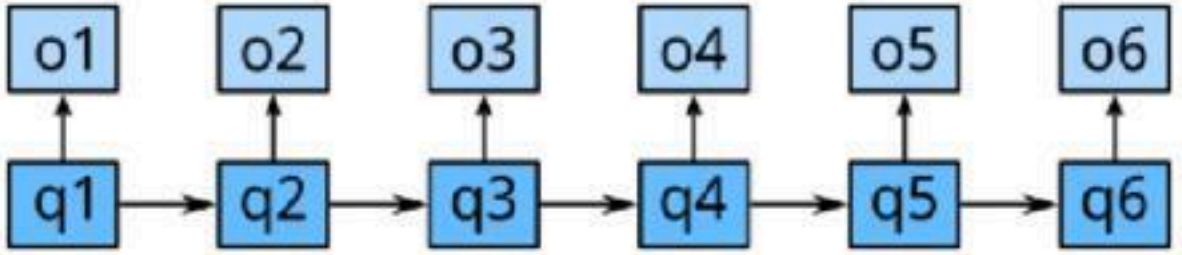
$$
\begin{aligned}
\hat{t}_{1:n} &= \arg\max_{t_1..t_n} P(t_1..t_n|w_1..w_n) \\
&= \arg\max_{t_1..t_n} P(w_1..w_n|t_1..t_n)P(t_1..t_n) \\
&\approx \arg\max_{t_1..t_n} \prod_{i=1}^{n} \overbrace{P(w_i|t_i)}^{\text{emission}} \overbrace{P(t_i|t_{i-1})}^{\text{transition}}
\end{aligned}
$$



HMM标注器：解码观点

解码：给定观测序列，确定隐变量序列；等价于解读未知密码

$$
\begin{aligned}
\hat{t}_{1:n} &= \arg\max_{t_1..t_n} P(t_1..t_n|w_1..w_n) \\
&= \arg\max_{t_1..t_n} P(w_1..w_n|t_1..t_n)P(t_1..t_n) \\
&\approx \arg\max_{t_1..t_n} \prod_{i=1}^{n} \overbrace{P(w_i|t_i)}^{\text{emission}} \overbrace{P(t_i|t_{i-1})}^{\text{transition}}
\end{aligned}
$$



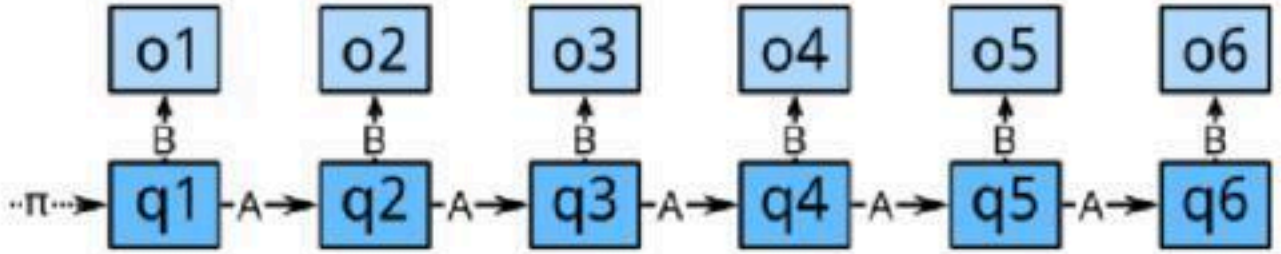
反过来即可**生成观测**：例如写作前先构思框架

HMM：应用

HMM的基本用法

HMM 的**参数三元组**：$\lambda = (\pi, A, B)$

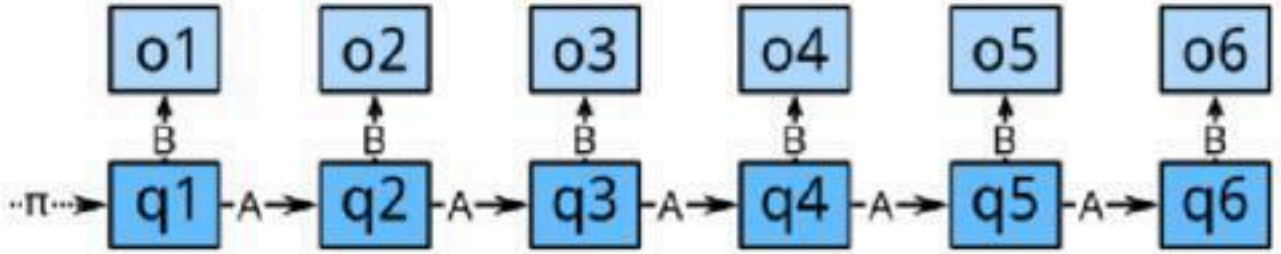
- 初始概率分布、状态转移矩阵、观测似然矩阵
- 用于**描述隐状态**，例如标签



HMM的基本用法

HMM 的**参数三元组**：$\lambda = (\pi, A, B)$

- 初始概率分布、状态转移矩阵、观测似然矩阵
- 用于**描述隐状态**，例如标签

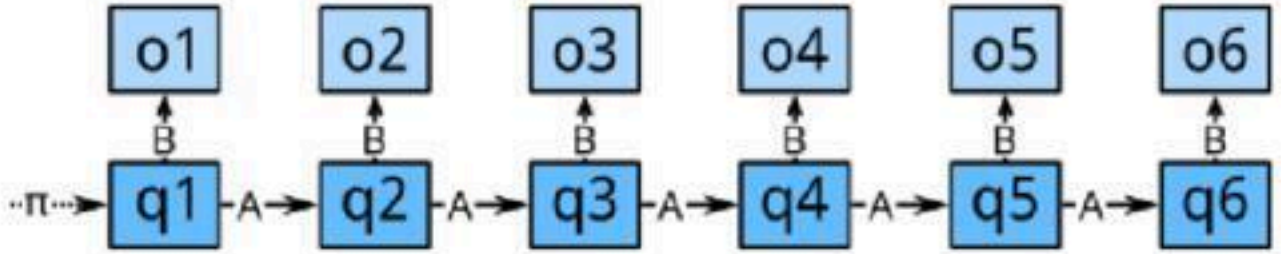


- **生成样本**：$s = \lambda(t; \pi, A, B)$
  - 连续采样生成文本序列$S = s_1..s_n$

HMM的基本用法

HMM 的**参数三元组**：$\lambda = (\pi, A, B)$

- 初始概率分布、状态转移矩阵、观测似然矩阵
- 用于**描述隐状态**，例如标签

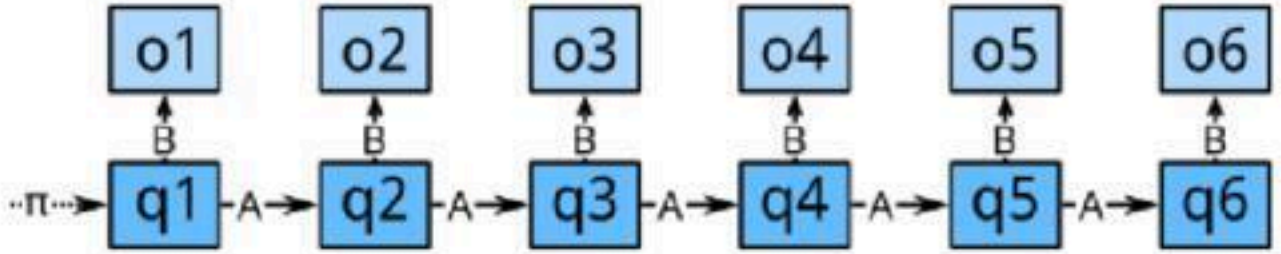


- **生成样本**：$s = \lambda(t; \pi, A, B)$
  - 连续采样生成文本序列$S = s_1..s_n$
- **训练模型**：给定训练集$\{(s^{(i)}, t^{(i)})\}$，估计模型参数$\lambda = (\pi, A, B)$

HMM的基本用法

HMM 的**参数三元组**：$\lambda = (\pi, A, B)$

- 初始概率分布、状态转移矩阵、观测似然矩阵
- 用于**描述隐状态**，例如标签

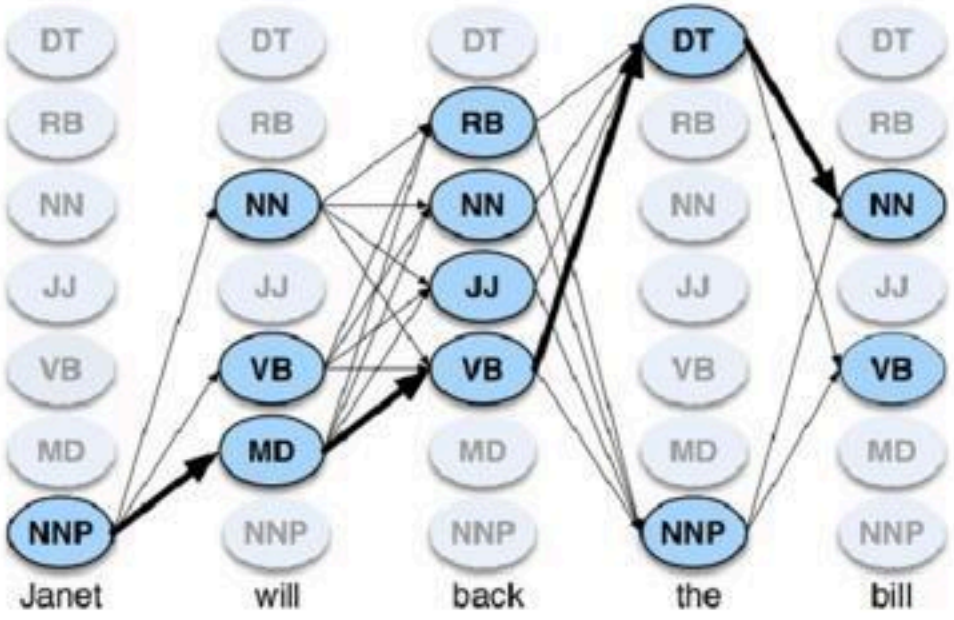


- **生成样本**：$s = \lambda(t; \pi, A, B)$
  - 连续采样生成文本序列$S = s_1..s_n$
- **训练模型**：给定训练集$\{(s^{(i)}, t^{(i)})\}$，估计模型参数$\lambda = (\pi, A, B)$
- **预测标签**：$t = \lambda^{-1}(s) = \arg\max_t \lambda(t; \pi, A, B)$
  - 等价于采样$(t, s) \sim \lambda$，不区分顺序

Viterbi 算法

首先计算状态-观测概率矩阵

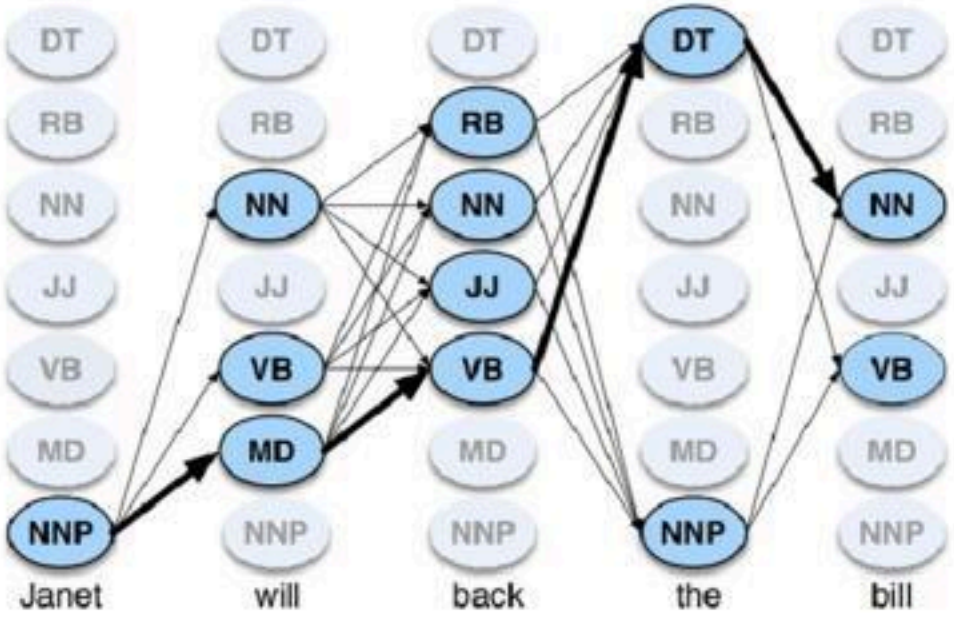
- 每一列：所有可能的标签集合
  - **标签集合**：从词典查出
- 行：单个隐状态，即标签



Viterbi 算法

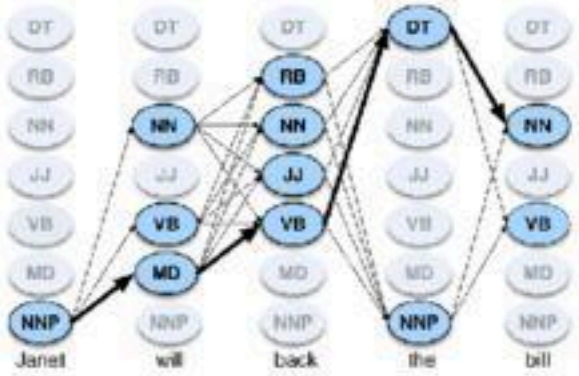
首先计算**状态-观测概率矩阵**

- 每一列：所有可能的标签集合
  - **标签集合**：从词典查出
- 行：单个隐状态，即标签



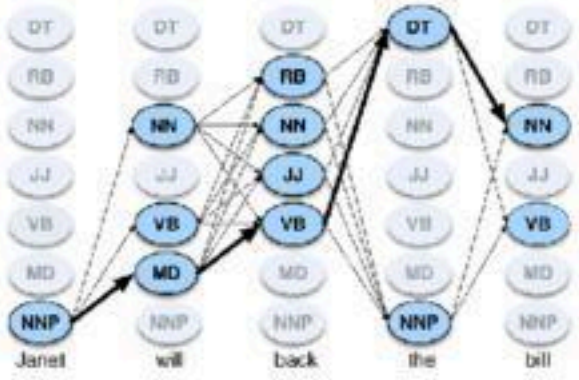
动态规划：寻找一条**最可能的路径**

Viterbi 算法：观测似然



	Janet	will	back	the	bill
NNP	0.000032	0	0	0.000048	0
MD	0	0.308431	0	0	0
VB	0	0.000028	0.000672	0	0.000028
JJ	0	0	0.000340	0	0
NN	0	0.000200	0.000223	0	0.002337
RB	0	0	0.010446	0	0
DT	0	0	0	0.506099	0

Viterbi 算法：转移概率



	NNP	MD	VB	JJ	NN	RB	DT
<s>	0.2767	0.0006	0.0031	0.0453	0.0449	0.0510	0.2026
NNP	0.3777	0.0110	0.0009	0.0084	0.0584	0.0090	0.0025
MD	0.0008	0.0002	0.7968	0.0005	0.0008	0.1698	0.0041
VB	0.0322	0.0005	0.0050	0.0837	0.0615	0.0514	0.2231
JJ	0.0366	0.0004	0.0001	0.0733	0.4509	0.0036	0.0036
NN	0.0096	0.0176	0.0014	0.0086	0.1216	0.0177	0.0068
RB	0.0068	0.0102	0.1011	0.1012	0.0120	0.0728	0.0479

DT	0.1147	0.0021	0.0002	0.2157	0.4744	0.0102	0.0017

实验：隐式Markov模型

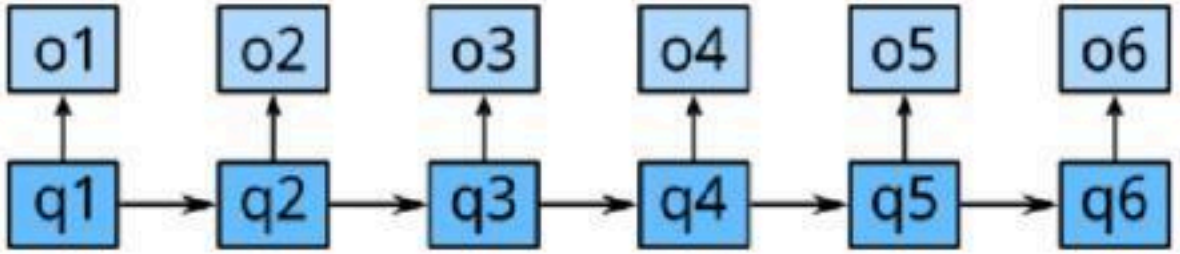
条件随机场

HMM：回顾

HMM计算：本质是**Bayes生成模型**

- 词序列：$S = s_1..s_n$；标签序列：$T = t_1..t_n$

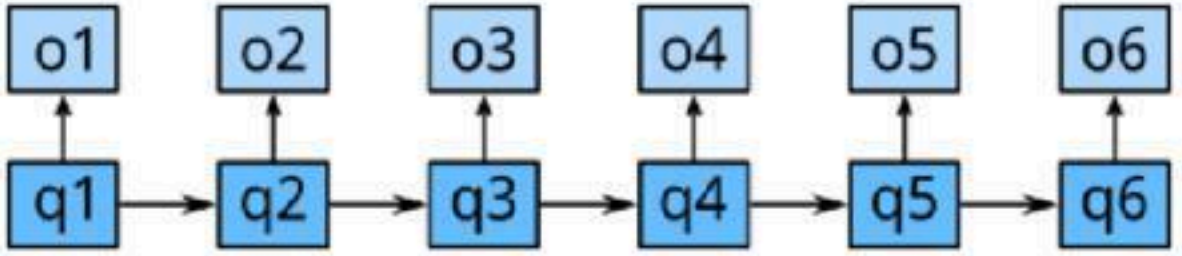
$$\hat{T} = \arg\max_T P(S|T)P(T) = \arg\max_T P(T,S) = \arg\max_T P(T|S)$$
$$= \arg\max_T \prod_i P(s_i|t_i) \prod_i P(t_i|t_{i-1})$$



HMM：局限性

HMM 的问题来源于假设的限制，需要足够的**扩充**才能解决实际 NLP 任务

$$\hat{T} = \arg\max_{T} \prod_{i} P(s_i|t_i) \prod_{i} P(t_i|t_{i-1})$$

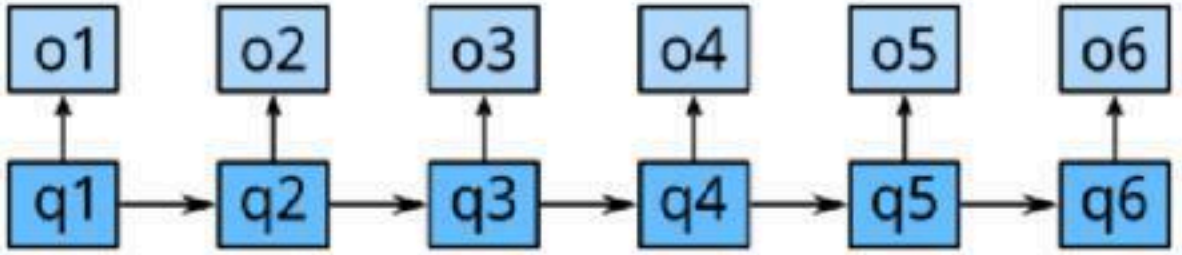


- 独立性假设太强：字之间**并非完全独立**
  - 例如：“蝴”之后几乎一定是“蝶”

HMM：局限性

HMM 的问题来源于假设的限制，需要足够的**扩充**才能解决实际 NLP 任务

$$\hat{T} = \arg\max_{T} \prod_{i} P(s_i|t_i) \prod_{i} P(t_i|t_{i-1})$$



- 独立性假设太强：字之间**并非完全独立**
  - 例如："蝴"之后几乎一定是"蝶"
- 未知词：创新词的出现远比词典更新频繁
  - 观测与任何隐状态都没有连接：需要**扩充规则集合**
  - 模型必须重建：**似然估计**依赖词典的统计数据

从生成模型到判别模型

生成模型通常难以直接添加（输出）特征：等价于更新整个模型

$$\hat{c} = \arg\max_{c \in C} \overbrace{P(d|c)}^{\text{likelihood}} \overbrace{P(c)}^{\text{prior}}$$

思考：有没有可以灵活添加特征的模型?

从生成模型到判别模型

生成模型通常难以直接添加（输出）特征：等价于更新整个模型

$$\hat{c} = \arg\max_{c \in C} \overbrace{P(d|c)}^{\text{likelihood}} \overbrace{P(c)}^{\text{prior}}$$

思考：有没有可以灵活添加特征的模型?

有：**判别模型**。例如逻辑回归，将特征看作输入参数

- 可以任意添加、组合特征；但**不能输出序列**

$$\hat{c} = \arg\max_{c \in C} \overbrace{P(c|d)}^{\text{posterior}}$$

从生成模型到判别模型

生成模型通常难以直接添加（输出）特征：等价于更新整个模型

$$\hat{c} = \arg\max_{c \in C} \overbrace{P(d|c)}^{\text{likelihood}} \overbrace{P(c)}^{\text{prior}}$$

思考：有没有可以灵活添加特征的模型?

有：**判别模型**。例如逻辑回归，将特征看作输入参数

- 可以任意添加、组合特征；但**不能输出序列**

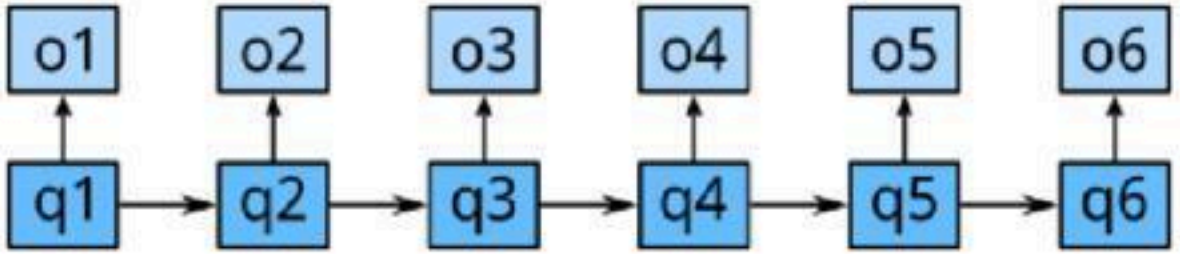
$$\hat{c} = \arg\max_{c \in C} \overbrace{P(c|d)}^{\text{posterior}}$$

思考：只要构造出**判别式序列模型**即可解决 HMM 的扩充问题

概率图模型

概率图模型 (PGM)：用图表示、推断随机变量的**联合概率分布**

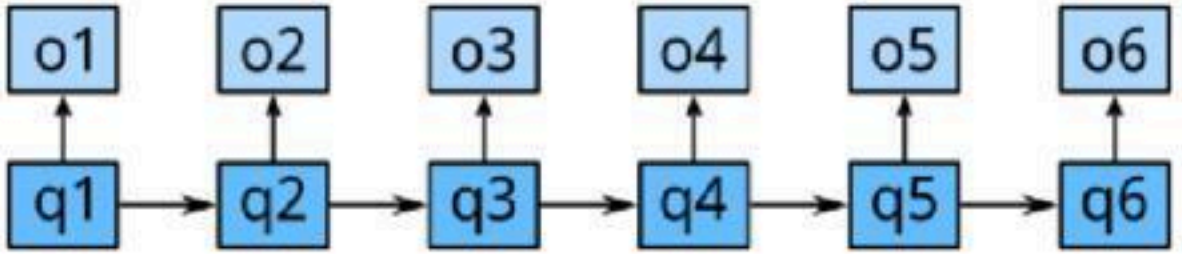
- 子图分解：不相关的变量**互不影响计算**
  - 特殊情况：确定性事件阻断传递链



概率图模型

概率图模型 (PGM)：用图表示、推断随机变量的**联合概率分布**

- 子图分解：不相关的变量**互不影响计算**
  - 特殊情况：确定性事件阻断传递链



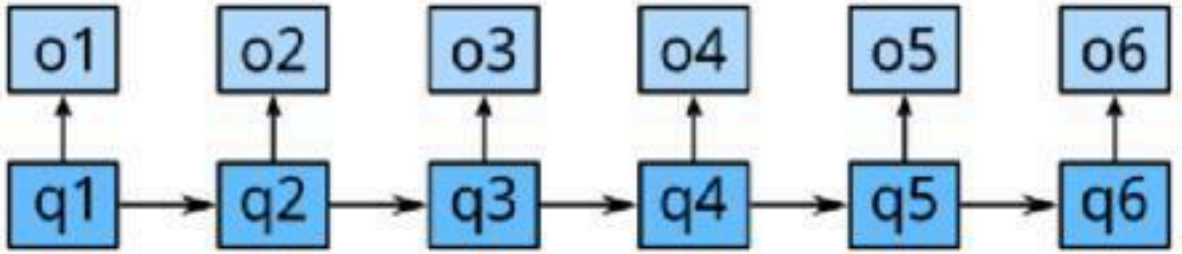
有向图模型 (DGM)：按事件的先后顺序连接

- 通常用生成模型实现，例如 HMM

概率图模型

概率图模型 (PGM)：用图表示、推断随机变量的**联合概率分布**

- 子图分解：不相关的变量**互不影响计算**
  - 特殊情况：确定性事件阻断传递链

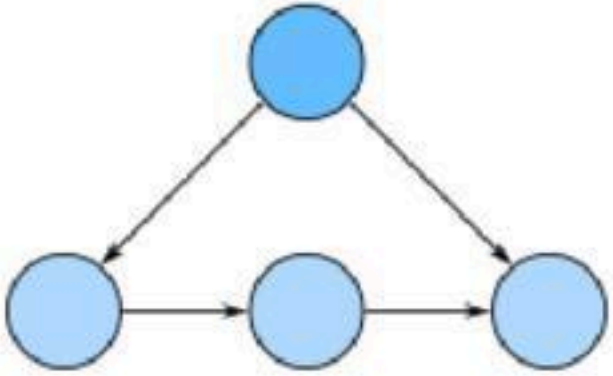


有向图模型 (DGM)：按事件的先后顺序连接

- 通常用生成模型实现，例如 HMM
- 分解为条件概率之积：$P(x,y) = \prod_v P(v|\pi(v))$
  - 可以表示**因果推断**

DGM：因果推断举例

1854年宽街霍乱爆发事件

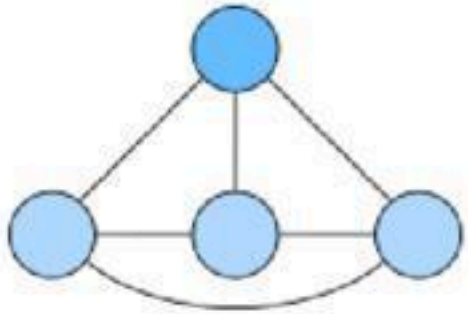


- “瘴气”理论
- 贫穷、水源、水泵、霍乱

无向图模型

无向图模型：不探究因果关系、不涉及条件概率

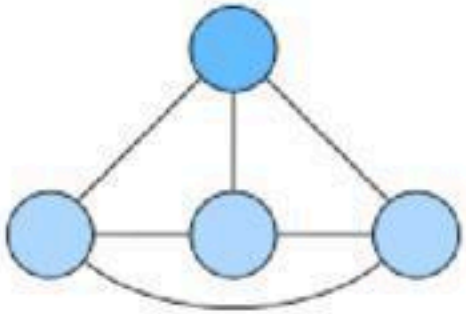
- **最大团 clique**：相互关联的随机变量集合，无法继续细分
  - 所有变量两两互联：计算必须同时考虑**所有两两组合**

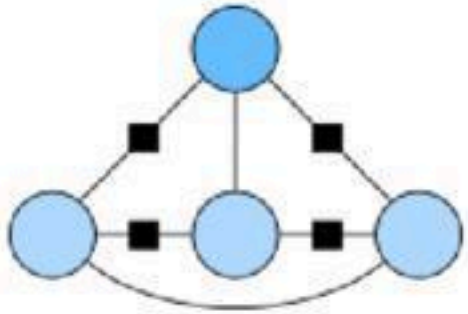


因子图模型

无向图模型：不探究因果关系、不涉及条件概率

- **最大团 clique**：相互关联的随机变量集合，无法继续细分
  - 所有变量两两互联：计算必须同时考虑**所有两两组合**





因子图模型：因子节点只连接需要计算的变量，描述**实际关联关系**

- 最大团的因子之积：$P(x,y) = \frac{1}{Z}\prod_a \Psi_a(x_a, y_a)$
  - 归一化因子：$Z = \sum_{x,y}\prod_a \Psi_a(x_a, y_a)$

条件随机场 CRF

条件随机场 conditional random field (CRF)：直接计算后验，判别标签序列

$$\hat{T} = \arg\max_{T \in \mathcal{T}} P(T|S)$$

- 对比 HMM：$\hat{T} = \arg\max_T P(S|T)P(T)$

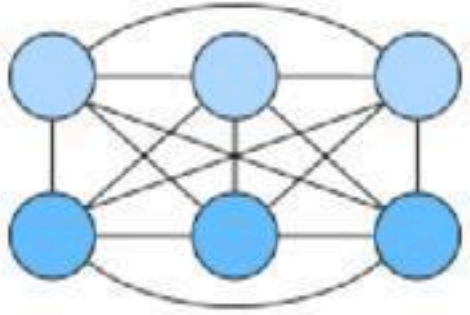
条件随机场 CRF

条件随机场 conditional random field (CRF)：直接计算后验，判别标签序列

$$\hat{T} = \arg\max_{T \in \mathcal{T}} P(T|S)$$

- 对比 HMM：$\hat{T} = \arg\max_T P(S|T)P(T)$

问题：每个时间步都计算**整个标签序列**T的概率



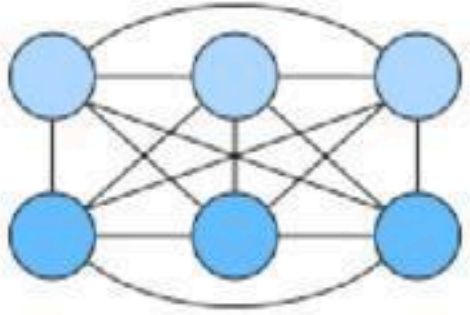
条件随机场 CRF

条件随机场 conditional random field (CRF)：直接计算后验，判别标签序列

$$\hat{T} = \arg\max_{T \in \mathcal{T}} P(T|S)$$

- 对比 HMM：$\hat{T} = \arg\max_T P(S|T)P(T)$

问题：每个时间步都计算**整个标签序列**T的概率



- 参考独立性假设：拆解成相关的**局部特征**，然后聚集并归一化

CRF：严格计算公式

- 首先引入**特征**$F_k(S,T)$：描述序列组合（的可能性）

$$P(T|S) \propto \frac{F_k(S,T)}{\sum_{T'} F_k(S,T')}$$

CRF：严格计算公式

- 首先引入**特征**$F_k(S,T)$：描述序列组合（的可能性）

$$P(T|S) \propto \frac{F_k(S,T)}{\sum_{T'} F_k(S,T')}$$

- 其次引入**权重**w_k：区分不同特征的重要性，例如红色判定红灯

$$P(T|S) \propto \frac{\sum_k w_k F_k(S,T)}{\sum_{k,T'} w_k F_k(S,T')}$$

CRF：严格计算公式

- 首先引入**特征**$F_k(S,T)$：描述序列组合（的可能性）

$$P(T|S) \propto \frac{F_k(S,T)}{\sum_{T'} F_k(S,T')}$$

- 其次引入**权重**w_k：区分不同特征的重要性，例如红色判定红灯

$$P(T|S) \propto \frac{\sum_k w_k F_k(S,T)}{\sum_{k,T'} w_k F_k(S,T')}$$

- 转换成**概率**：类似于sigmoid函数

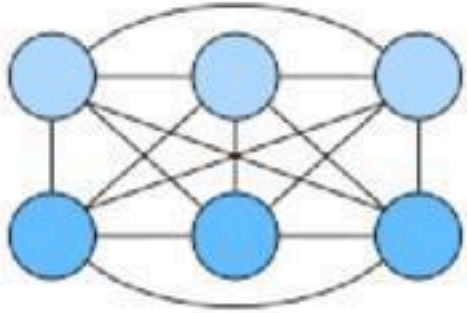
$$P(T|S) = \frac{\exp\left(\sum_{k=1}^{K} w_k F_k(S,T)\right)}{\sum_{T' \in \mathcal{T}} \exp\left(\sum_{k=1}^{K} w_k F_k(S,T')\right)}$$

特征拆解

全局特征$F_k(S,T)$：每个都是整个输入序列S和输出序列T的特征

$$P(T|S)=\frac{1}{Z(S)}\exp\left(\sum_{k=1}^{K}w_kF_k(S,T)\right)$$

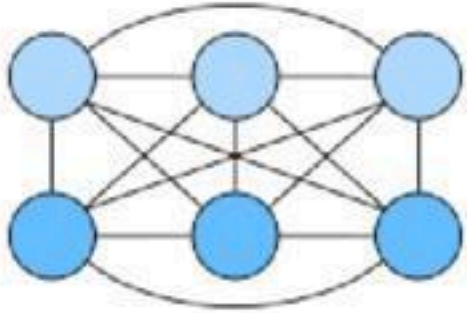
- $Z(S)=\sum_{T'\in\mathcal{T}}\exp\left(\sum_{k=1}^{K}w_kF_k(S,T')\right)$



特征拆解

全局特征$F_k(S,T)$：每个都是整个输入序列S和输出序列T的特征

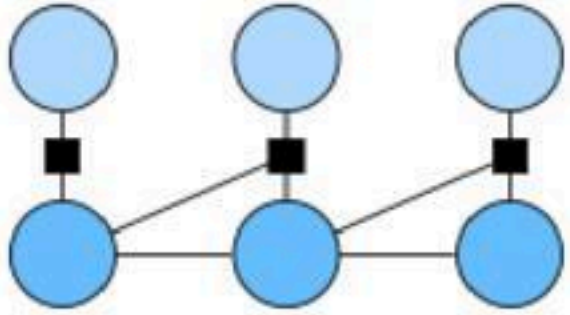
$$P(T|S) = \frac{1}{Z(S)} \exp\left(\sum_{k=1}^{K} w_k F_k(S,T)\right)$$



- $Z(S) = \sum_{T' \in \mathcal{T}} \exp\left(\sum_{k=1}^{K} w_k F_k(S,T')\right)$

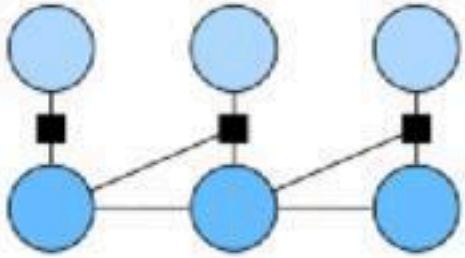
简化计算：拆解成T每个位置上的**局部特征**之和

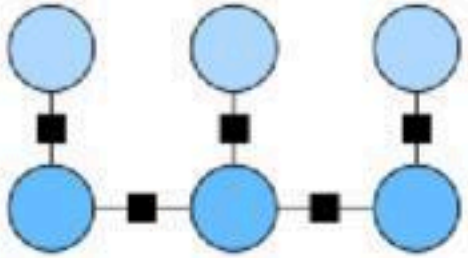
$$F_k(S,T) = \sum_{i=1}^{n} f_k(t_{i-1}, t_i, s_i, i)$$

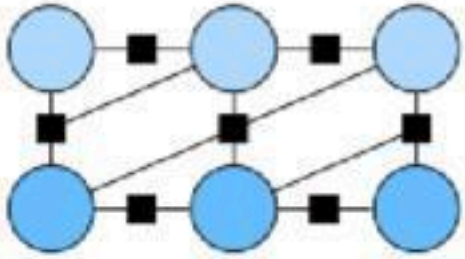


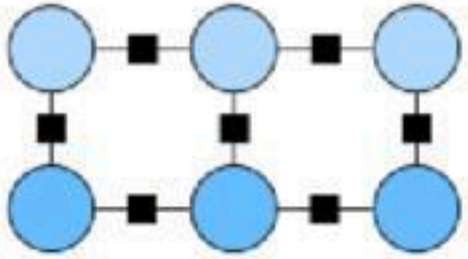
- 称为**线性链式CRF**：特征计算只依赖于局部输出t_{i-1}, t_i

CRF：因子图的灵活性









CRF：推断与训练

$$
\begin{aligned}
\hat{T} &= \arg\max_{T \in \mathcal{T}} P(T|S) \\
&= \arg\max_{T \in \mathcal{T}} \frac{1}{Z(S)} \exp\left(\sum_{k=1}^{K} w_k F_k(S,T)\right) \\
&= \arg\max_{T \in \mathcal{T}} \exp\left(\sum_{k=1}^{K} w_k \sum_{i=1}^{n} f_k(t_{i-1}, t_i, s_i, i)\right) \\
&= \arg\max_{T \in \mathcal{T}} \sum_{i=1}^{n} \sum_{k=1}^{K} w_k f_k(t_{i-1}, t_i, s_i, i)
\end{aligned}
$$

CRF：推断与训练

$$
\begin{aligned}
\hat{T} &= \arg\max_{T \in \mathcal{T}} P(T|S) \\
&= \arg\max_{T \in \mathcal{T}} \frac{1}{Z(S)} \exp\left(\sum_{k=1}^{K} w_k F_k(S,T)\right) \\
&= \arg\max_{T \in \mathcal{T}} \exp\left(\sum_{k=1}^{K} w_k \sum_{i=1}^{n} f_k(t_{i-1}, t_i, s_i, i)\right) \\
&= \arg\max_{T \in \mathcal{T}} \sum_{i=1}^{n} \sum_{k=1}^{K} w_k f_k(t_{i-1}, t_i, s_i, i)
\end{aligned}
$$

- 推断最优标签$\hat{T}$：Viterbi 算法

CRF：推断与训练

$$
\begin{aligned}
\hat{T} &= \arg\max_{T \in \mathcal{T}} P(T|S) \\
&= \arg\max_{T \in \mathcal{T}} \frac{1}{Z(S)} \exp\left(\sum_{k=1}^{K} w_k F_k(S,T)\right) \\
&= \arg\max_{T \in \mathcal{T}} \exp\left(\sum_{k=1}^{K} w_k \sum_{i=1}^{n} f_k(t_{i-1}, t_i, s_i, i)\right) \\
&= \arg\max_{T \in \mathcal{T}} \sum_{i=1}^{n} \sum_{k=1}^{K} w_k f_k(t_{i-1}, t_i, s_i, i)
\end{aligned}
$$

- 推断最优标签$\hat{T}$：Viterbi 算法
- 训练参数w_k：前向积累、反向传递

Review

本章内容

语言序列与标注。隐式Markov模型 (HMM)。条件随机场 (CRF)。

重点：词类、词类标注；命名实体、命名实体识别；一阶HMM的两个基本假设、架构、计算方法；HMM的三种应用方法：采样、训练、预测；线性链式CRF的架构、计算方法。

难点：条件随机场 (CRF)的动机；有向图模型与因果推断；因子图模型与最大团。

学习目标

- 理解序列标注的两大基本问题：词类标注、命名实体识别
- 理解隐式Markov模型 (HMM)的架构
- 理解一阶HMM的两个基本假设：一阶Markov假设、独立性假设
- 理解一阶HMM的计算方法
- 理解HMM的三种应用方法：采样、训练、预测
- 了解条件随机场 (CRF)的动机：解决HMM模型架构更新难的问题
- 了解有向图模型与因果推断的联系
- 了解因子图模型的动机：描述实际关联关系（而非最大团的全连接）
- 理解线性链式CRF的架构、计算方法

问题

简述序列标注任务，及其两大基本问题。

绘图并简述隐式Markov模型 (HMM)的架构。

简述一阶HMM的两个基本假设、架构、计算方法。

简述HMM的三种应用方法。

(*) 简述条件随机场 (CRF)的动机（解决了HMM的什么问题）。

(*) 简述有向图模型与因果推断的联系。

(*) 简述因子图模型的动机（解决了最大团的什么问题）。

简述线性链式CRF的架构、计算方法。