词性标注
(Part-Of-Speech Tagging)

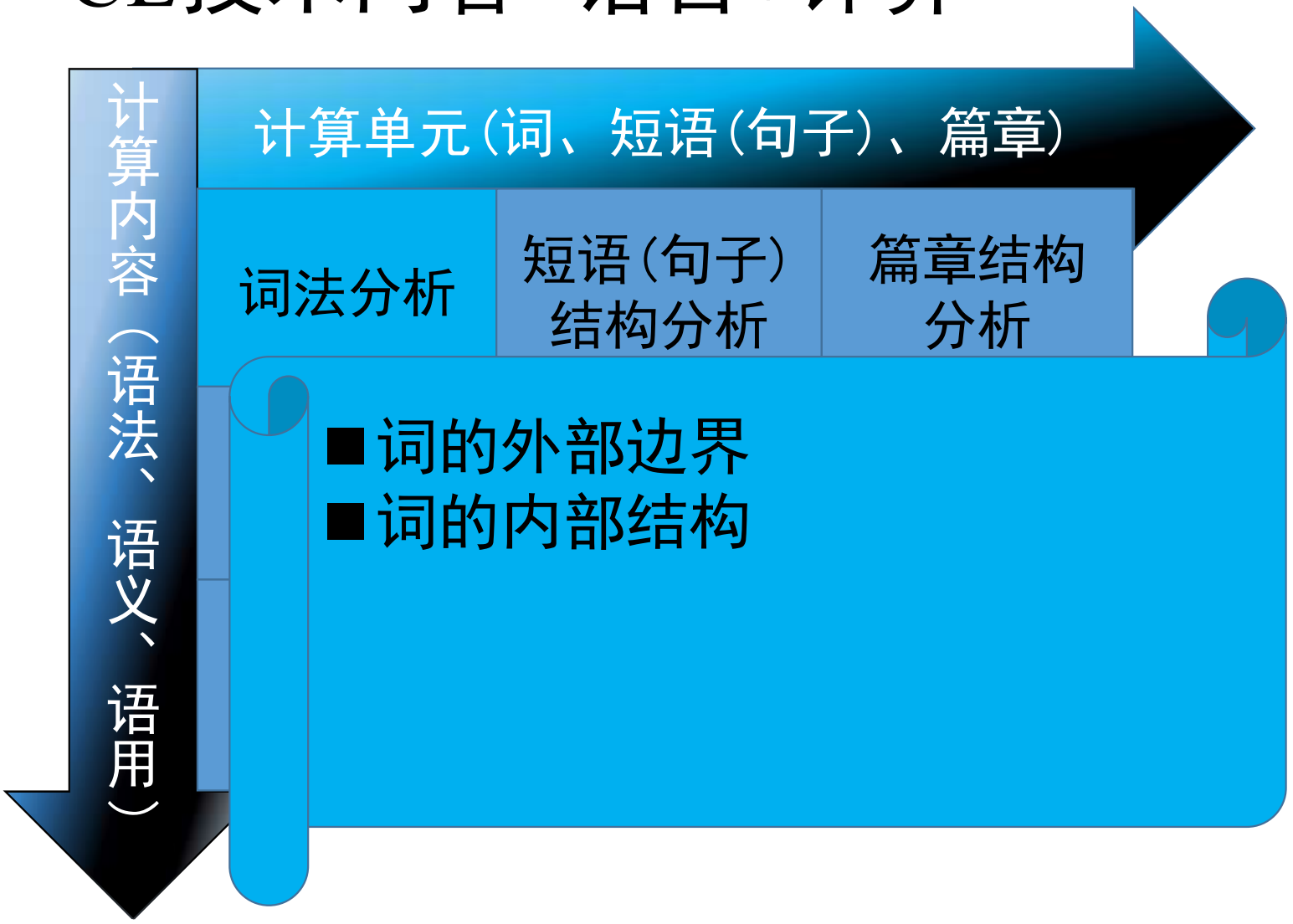
王小捷
智能科学与技术中心
北京邮电大学

大纲

- 引言：短语结构问题
- 词性标注集(POS Tagset)
- 词性标注 (POS Tagging)
- 基于规则的词性标注方法
- 基于统计的词性标注方法
- 总结

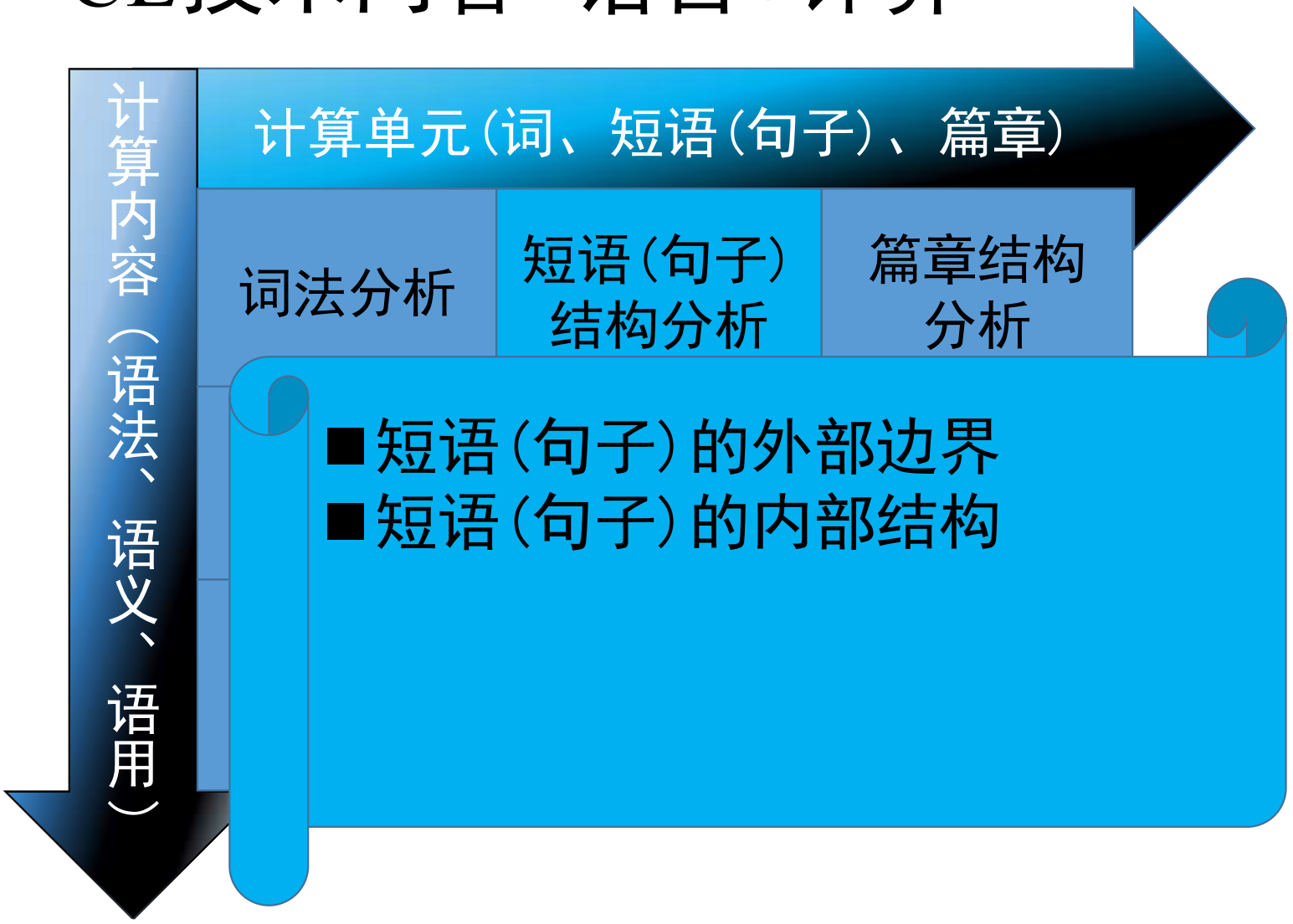
CL技术内容=语言+计算

CL技术内容=语言+计算



CL技术内容=语言+计算

CL技术内容=语言+计算



- 句子边界：连续出现的n个词$w_1,w_2,\ldots w_n$是否构成一个句子？
  - 基于标点符号：大部分情况是平凡问题
  - 标号：。？！；：很可能是句子的分隔符
  - 点号：“”()...-- 也可能是句子的分隔符，但是《》、着重号等等一般不是句子分隔符
  - 符号：*等等一般不是句子分隔符

    - 5月3日，那天一大早，我就听见外面刮着狂风。

- 子句边界：句子中子句边界
  - 我看见你了，你出来吧。(逗号)
- 很多情况下仅基于标点不能实现
  - You know what I mean.
  - 我认为你这样做是不对的。

- 引导词？动词？
- 子句边界实际上又是一个句子内部结构的问题！

- 句子内部结构：
- 粗略地说：句子中的词单元是如何构成句子的，即句子中词的结合关系
- $w_1, w_2, \ldots w_n$是$(w_1, w_2), \ldots w_n)$还是$(w_1, (w_2, \ldots w_n)$等等
  - ((我和他)((一起)((看)(球))))
  - (I (saw ((a boy) (with a telescope))))
  - (I ((saw) (a boy) (with a telescope)))
  - ……

- 短语边界
  - He gave me /a couple of books/.
  - 我们约在/五月的一天/
  - /那个带眼镜的人/在看书
  - 识别短语边界：名词短语、动词短语、…
  - ?

- 短语结构：
- 短语中词的关系
- w_i和w_j的关系，是以w_i为主还是w_j为主，还是别的？
  - 偏正结构：旧世界、美丽心灵、……
  - 述谓结构：开动脑筋、写论文……
  - 并列结构：学习学习、老人与海、.....
  - ……

- 短语或句子内部结构(构成方式)
- 语言学的角度
  - 引入一个中间概念：词性，给短语中的每一个词赋予一个词性，通过词性序列来帮助揭示语言中的结构信息：
  - 例如：
  - 影响 团结 (动词 名词)：述谓结构
  - 重大 影响(形容词 名词)：偏正
  - ……

- 词性(POS\词类\词汇范畴\)
  - POS (Part-Of-Speech)
  - 词的聚合关系：具有相似语法性质的词构成一类
    - I\you\she\he\we\...；apple\table\room\street\...

	语法	语义
聚合	词性(POS\词类\词汇范畴\)	同义
组合	句法结构	语义结构

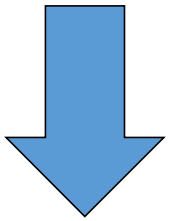
- 为文本中的词标注其词性：词性标注(POS tagging)
- 两个子任务
  - 选择词性标注集(POS Tagset)
    - 对一个语言(任务)而言有多少POS：确定POS tagset
    - 对每个词而言有多少个不同的POS可能
      - 或不为每个词分别指定一个POS子集，而直接认为每个词都可以取所有可能的POS
  - 词性标注(POS Tagging)
    - 在上下文中为词指派POS

Part-Of-Speech(POS)

- 通常的
  - 名词：人民、学校
  - 动词：打动、袭击
  - 形容词：美丽的、善良的
  - …
- 更多的
  - 名词细分：专有名词、集合名词. .
  - ……

词性标注(POS tagging)

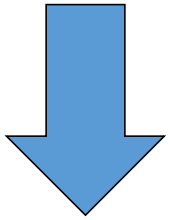
- 任务
  - Secretariat is expected to race tomorrow.



  - Secretariat/N is/V expected/V to/Prep race/V tomorrow/N

词性标注(POS tagging)

- 任务
  - 受 暖湿 气流 影响，……



  - 受/V 暖湿/ADJ 气流/N 影响/V，/w …….

词性标注的一些作用

- 词性是短语结构和句子结构分析的基础
  - 后面句法分析可以进一步看到
- 对很多NLP应用的性能提高有帮助
  - 机器翻译
    - 不同词性意义不同：例如：“hide”：兽皮(N)，隐藏(V)
  - 信息检索
    - 尤其是语言学习时：例如：找出play的常见搭配
      - 使用 Play + N 检索可以避免找出play a, play the
  - …

大纲

- 引言：短语结构问题
- 词性标注集(POS Tagset)
- 词性标注 (POS Tagging)
- 基于规则的词性标注方法
- 基于统计的词性标注方法
- 总结

POS标注集（POS tagset)

- 例子：一个简单POS tagset{Noun，Verb，Adj，Other}
  - 人民： Noun，Verb，Adj，Other
  - 创造： Noun，Verb，Adj，Other
  - 。： Noun，Verb，Adj，Other
  - …
- 或通过使用额外的语言知识获得，即使如此也有词有多个可能的POS
  - 人民：Noun
  - 创造：Noun，Verb
  - 。：other
  - …

POS标注集 (POS tagset)

- 多样性
  - 不同的语言学家有不同的
    - 动词/名词
    - …
  - 不同的任务可以设计不同的
    - 名词/人名/时间词
    - 动词/及物动词/不及物动词
    - …

一些有代表性的标注集

英语	
语料库	标注集大小
Brown语料	87
Penn 树库(Treebank)	45
Lancaster UCREL C5(BNC)	61
Lancaster C7	146
汉语	
Penn 汉语树库(Treebank)	33
北京大学	39

Penn 树库标注集

Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	*and, but, or*	SYM	Symbol	+,%, &
CD	Cardinal number	*one, two, three*	TO	"to"	*to*
DT	Determiner	*a, the*	UH	Interjection	*ah, oops*
EX	Existential 'there'	*there*	VB	Verb, base form	*eat*
FW	Foreign word	*mea culpa*	VBD	Verb, past tense	*ate*
IN	Preposition/sub-conj	*of, in, by*	VBG	Verb, gerund	*eating*
JJ	Adjective	*yellow*	VBN	Verb, past participle	*eaten*
JJR	Adj., comparative	*bigger*	VBP	Verb, non-3sg pres	*eat*
JJS	Adj., superlative	*wildest*	VBZ	Verb, 3sg pres	*eats*
LS	List item marker	*1, 2, One*	WDT	Wh-determiner	*which, that*
MD	Modal	*can, should*	WP	Wh-pronoun	*what, who*
NN	Noun, sing. or mass	*llama*	WP$	Possessive wh-	*whose*
NNS	Noun, plural	*llamas*	WRB	Wh-adverb	*how, where*
NNP	Proper noun, singular	*IBM*	$	Dollar sign	*$*
NNPS	Proper noun, plural	*Carolinas*	#	Pound sign	#
PDT	Predeterminer	*all, both*	"	Left quote	(' or ")
POS	Possessive ending	*'s*	"	Right quote	(' or ")
PP	Personal pronoun	*I, you, he*	(	Left parenthesis	([, (, {, <)
PP$	Possessive pronoun	*your, one's*	)	Right parenthesis	( ],), }, >)
RB	Adverb	*quickly, never*	,	Comma	,
RBR	Adverb, comparative	*faster*	.	Sentence-final punc	(. ! ?)
RBS	Adverb, superlative	*fastest*	:	Mid-sentence punc	(: ; ... – -)
RP	Particle	*up, off*			

Penn 树库 POS示例

NN	noun	**JJ**	adjective
NNP	proper noun	**CC**	coord conj
DT	determiner	**CD**	cardinal number
IN	preposition	**PRP**	personal pronoun
VB	verb	**RB**	adverb

-R	comparative
-S	superlative or plural
-$	possessive

Penn 树库 POS示例：动词

VBP	base present	*take*
VB	infinitive	*take*
VBD	past	*took*
VBG	present participle	*taking*
VBN	past participle	*taken*
VBZ	present 3sg	*takes*
MD	modal	*can, would*

UCREL C5

Tag	Description	Example
PNX	reflexive pronoun	*itself, ourselves*
POS	possessive *'s* or *'*	
PRF	the preposition *of*	
PRP	preposition (except *of*)	*for, above, to*
PUL	punctuation – left bracket	(or [
PUN	punctuation – general mark	. ! , : ; - ? ...
PUQ	punctuation – quotation mark	' ' "
PUR	punctuation – right bracket	) or]
TO0	infinitive marker *to*	
UNC	unclassified items (not English)	
VBB	base forms of *be* (except infinitive)	*am, are*
VBD	past form of *be*	*was, were*
VBG	-ing form of *be*	*being*
VBI	infinitive of *be*	
VBN	past participle of *be*	*been*
VBZ	-s form of *be*	*is, 's*
VDB	base form of *do*(except infinitive)	*does*
VDD	past form of *do*	*did*
VDG	-ing form of *do*	*doing*
VDI	infinitive of *do*	*to do*
VDN	past participle of *do*	*done*
VDZ	-s form of *do*	*does*
VHB	base form of *have* (except infinitive)	*have*
VHD	past tense form of *have*	*had, 'd*
VHG	-ing form of *have*	*having*
VHI	infinitive of *have*	
VHN	past participle of *have*	*had*
VHZ	-s form of *have*	*has, 's*
VM0	modal auxiliary verb	*can, could, will, 'll*
VVB	base form of lexical verb (except infin.)	*take, live*
VVD	past tense form of lexical verb	*took, lived*
VVG	-ing form of lexical verb	*taking, living*
VVI	infinitive of lexical verb	*take, live*
VVN	past participle form of lex. verb	*taken, lived*
VVZ	-s form of lexical verb	*takes, lives*
XX0	the negative *not* or *n't*	
ZZ0	alphabetical symbol	*A, B, c, d*

Brown语料中的POS标签

Television/NN has/HVZ yet/RB to/TO work/VB out/RP a/AT living/RBG arrangement/NN with/IN jazz/NN ,/, which/VDT comes/VBZ to/IN the/AT medium/NN more/QL as/CS an/AT uneasy/JJ guest/NN than/CS as/CS a/AT relaxed/VBN member/NN of/IN the/AT family/NN ./.

BNC基于SGML的POS标记


```
<div1 complete=y org=seq>
 <head>
<s n=00040> <w NN2>TROUSERS <w VVB>SUIT
</head>
<caption>
<s n=00041> <w EX0>There <w VBZ>is <w PNI>nothing <w AJ0>masculine
<w PRP>about <w DT0>these <w AJ0>new <w NN1>trouser <w NN2-
VVZ>suits <w PRP>in <w NN1>summer<w POS>'s <w AJ0>soft <w
NN2>pastels<c PUN>.
<s n=00042> <w NP0>Smart <w CJC>and <w AJ0>acceptable <w PRP>for
<w NN1>city <w NN1-VVB>wear <w CJC>but <w AJ0>soft <w AV0>enough
<w PRP>for <w AJ0>relaxed <w NN2>days
</caption>
```

北京大学标注集

■Ag: 形语素	a:形容词	ad:副形词	an:名形词
■Bg	b:区别词	c:连词	Dg:副语素
■d:副词	e:叹词	f:方位词	g:语素
■h:前接成分	i:成语	j:简称略语	k:后接成分
■l:习用语	Mg:数语素	m:数词	Ng:名语素
■n:名词	nr:人名	ns:地名	nt:机构团体
■Nx	nz:其他专名	o:拟声词	p:介词
■Qg	q:量词	Rg:代语素	r:代词
■s:处所词	Tg:时语素	t:时间词	Ug:助语素
■u:助词	Vg:动语素	v:动词	vd:副动词
■Vn:名动词	w:标点符号	x:非语素字	Yg:语气语素
■y:语气词	z:状态词		

汉语语素

- 语素是最小的音义结合体
- 依所含音节数来划分，可分为：
  - 单音节语素
    - 大多数汉字是一个语素：我、飞、过
  - 双音节语素
    - 琵琶、蜻蜓、葡萄、蹊跷
  - 多音节语素
    - 主要是拟声词、专用名词、音译外来词：噼里啪啦、葡萄牙、喜马拉雅、中华人民共和国
  - 非音节语素
    - 儿化音节:花儿huor

- 依构词能力来划分，可分为：
  - 自由语素
    - 能独立成词：你，好
  - 半自由语素
    - 不能单独成词，可以在前后加上别的语素组成一个词语。如民：人民、民众
  - 不自由语素
    - 不能单独成词，可以在固定位置加上别的语素组成一个词语。如阿：阿爸、阿妈

- 语素与字的区别
  - 大多数汉字都是语素：我、飞、过
  - 但并非每个汉字都是语素：蜻、蜓
- 语素与词的区别
  - 词：能独立使用的最小音义结合体
  - 词一定是语素构成的
    - 由一个语素构成的词成为单纯词：好
    - 由多个语素构成的词成为复合词：正确
  - 语素不一定是词
    - 浩、鸿…

北京大学语料的POS标记 示例

■19980105-04-007-004/m 慕/nr 凌飞/nr 先生/n 一生/n 热衷/v 于/p 社会/n 公益/n 事业/n ，/w 在/p [毛/nr 主席/n 纪念堂/n]ns 、/w [天安门/ns 城楼/n]ns 、/w [周/nr 恩来/nr 邓/nr 颖超/nr 纪念馆/n]ns 、/w [平津战役/nz 纪念馆/n]ns 等/u 处/n ，/w 均/d 留/v 有/v 他/r 的/u 鸿/Ag 幅/Ng 巨制/n 。/w 赈灾/v 义卖/v 、/w 支/v 教/Ng 助残/v ，/w 甚至/d 对/p 劳教/vn 人员/n 帮教/v ，/w 慕/nr 凌飞/nr 都/d 以/p 饱满/a 的/u 热情/an ，/w 慈善/a 的/u 心地/n ，/w 予以/v 真诚/a 无私/b 的/u 贡献/n 。/w 据/p 不/d 完全/a 统计/v ，/w 慕/nr 凌飞/nr 以/p 各种/r 名义/n 捐赠/v 书画/n １５００/m 多/m 件/q ，/w 被/p 誉为/v 德/n 艺/Ng 双/m 馨/Ng 的/u 画家/n 。/w

POS标注集的影响

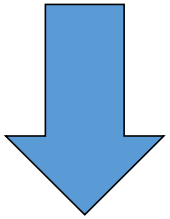
- 标注集的选择极大影响标注任务的难度
  - 45 v 146
  - 哪个难？为何？
- 标注集选择时需要平衡：
  - 信息丰富程度
  - 不会太难标

大纲

- 引言：短语结构问题
- 词性标注集(POS Tagset)
- 词性标注 (POS Tagging)
- 基于规则的词性标注方法
- 基于统计的词性标注方法
- 总结

标注(Tagging)

- 标注中的问题
  - 影响 暖湿 气流 走向，……



  - 影响/V 暖湿/ADJ 气流/N 走向/N，/w …….
  - 影响V 与 影响V &N

标注(Tagging)

- 标注中的问题
  - 歧义(Ambiguity)
    - 当一个词有多个可能的POS时，如何为之指派合适的POS
    - 影响/？　暖湿　气流　走向
    - 暖湿　气流　给 该 地区 带来 巨大　影响/？

POS标注中的歧义

- 广泛存在
- 大多数高频词都有多个可能的POS

- Brown语料中：
  - 11.5% 的word type　是有多个POS的
  - 40%　的word tokens是有多个POS的

■汉语

■吕叔湘 “汉语800词”: 22.5% 有2个以上词性.

■汉语高频词

词	可能的POS
的	助词、名次、形容词
了	动词、副词
我	代词(主格、宾格)
是	动词、名词、代词、形容词、叹词
一	数词、副词、名词
在	动词、副词、名词(姓氏)
不	副词
他	代词(主格、宾格)
人	名词

词性标注的另一个问题

- OOV(未登录词)
  - 词典未见，未规定词性
  - 所有可能的词性中的一个
    - 名词多：命名实体

标注方法

- 确定一个词的词性需要什么知识？
  - 该词可能的词性
  - 不同词性之间的搭配约束(上下文)

标注方法

- 基于规则的方法
  - Engtwol
- 基于统计的方法：POS tagging是典型的序标问题
  - HMM(Hidden Markov Model)
  - MEMM
  - CRF
  - RNN/LSTM
- 结合规则与统计的方法
  - Transformation-based Learning

大纲

- 引言：短语结构问题
- 词性标注集(POS Tagset)
- 词性标注 (POS Tagging)
- 基于规则的词性标注方法
- 基于统计的词性标注方法
- 总结

- 基于规则的方法：两步
  - 为每个词打上其所有可能的POS标记
  - 对于有多个POS的词，基于一些规则从中选择一个POS
    - 规则预先手工制定

- 规则示例
  - 合力：V，N
  - if POS(x_{-1})=Verb　then　POS(合力)=N
  - Engtwol(now ENGCG-2)：1100 规则
    - http://archive.is/Vakmz
    - 早期在Brown corpus上取得 70%的准确率

大纲

- 引言：短语结构问题
- 词性标注集(POS Tagset)
- 词性标注 (POS Tagging)
- 基于规则的词性标注方法
- 基于统计的词性标注方法
  - HMM (Hidden Markov Model)
- 总结

HMM：五元组

■状态集： $q_t \in S = \{s_1, s_2 \ldots s_N\}$

■观测集： $o_t \in V = \{v_1, v_2, \ldots\ldots v_M\}$

■状态转移概率：

$$A = (a_{ij})_{N\times N} \qquad a_{ij} = P(q_{t+1} = s_j \mid q_t = s_i)$$

(隐含假设：当前状态只依赖前一时刻状态)

■状态-观测输出概率(发射概率)

$$B = (b_i(o_t))_{N\times M} \qquad b_i(o_t) = P(o_t = v_k \mid q_t = s_i)$$

(隐含假设：当前观测只依赖当前状态)

■初始状态： q_0 ， 终止状态： q_{end}

$$a_{0i} = P(q_1 = i \mid q_0)$$

$$q_{i,end} = P(q_{end} \mid q_T = i)$$

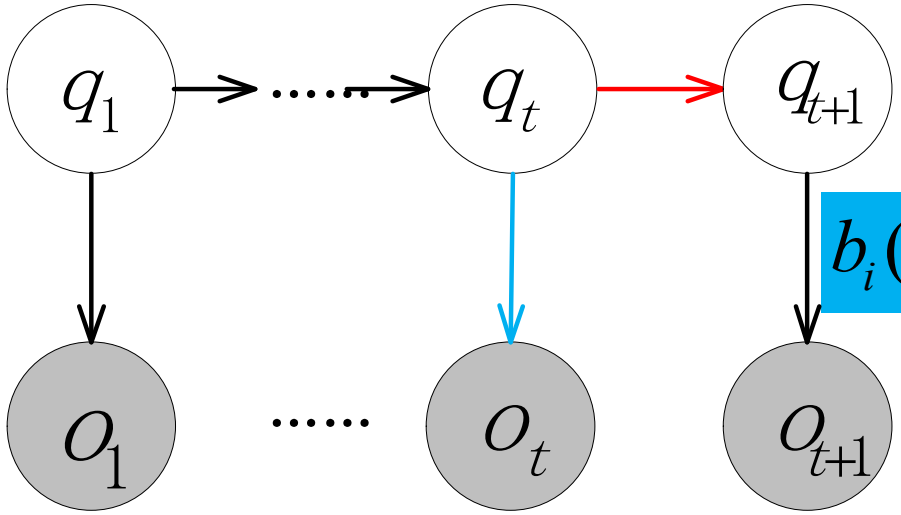
HMM模型参数：$\lambda = (A, B)$

HMM: 图表示

$$a_{ij} = P(q_{t+1} = s_j \mid q_t = s_i)$$

■状态序列

■观测序列

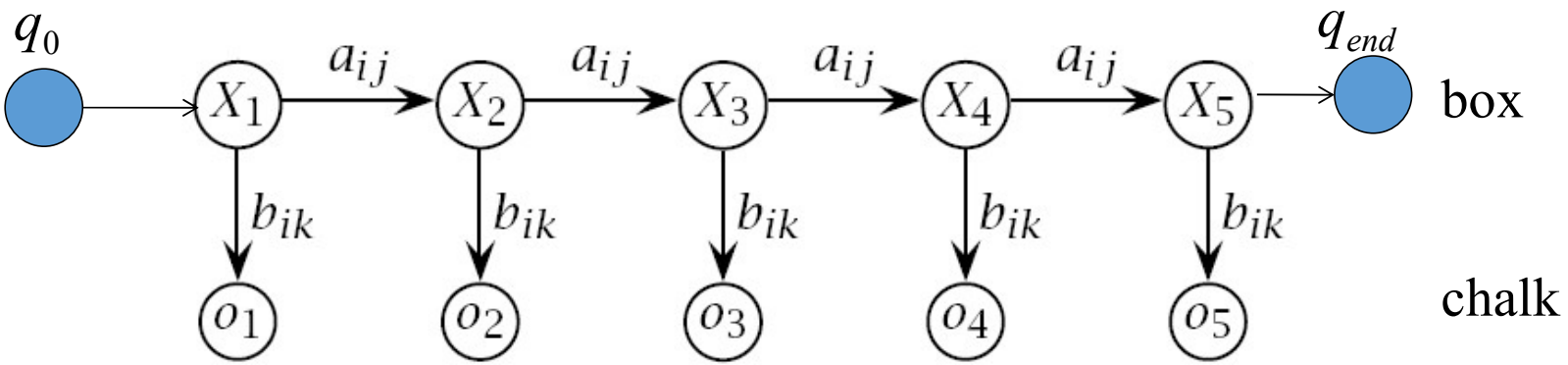


$$b_i(o_t) = P(o_t = v_k \mid q_t = s_i)$$

■(隐)状态序列具有马氏性

■盒中取粉笔问题：

- ■粉笔颜色：观测
- ■盒编号：状态
- ■选盒：状态转移A
- ■取笔：发射概率B

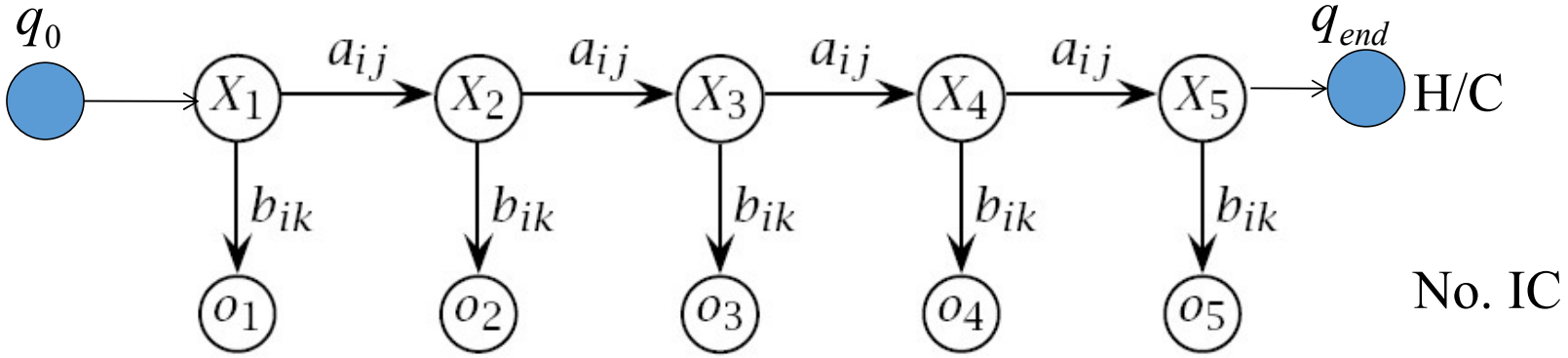


$X_1, X_2, \ldots X_5 \in \text{boxset} = \{1,2,3\}$

$o_1, o_2, \ldots o_5 \in \text{chalkset} = \{\text{red,blue,green}\}$

■天气考古问题

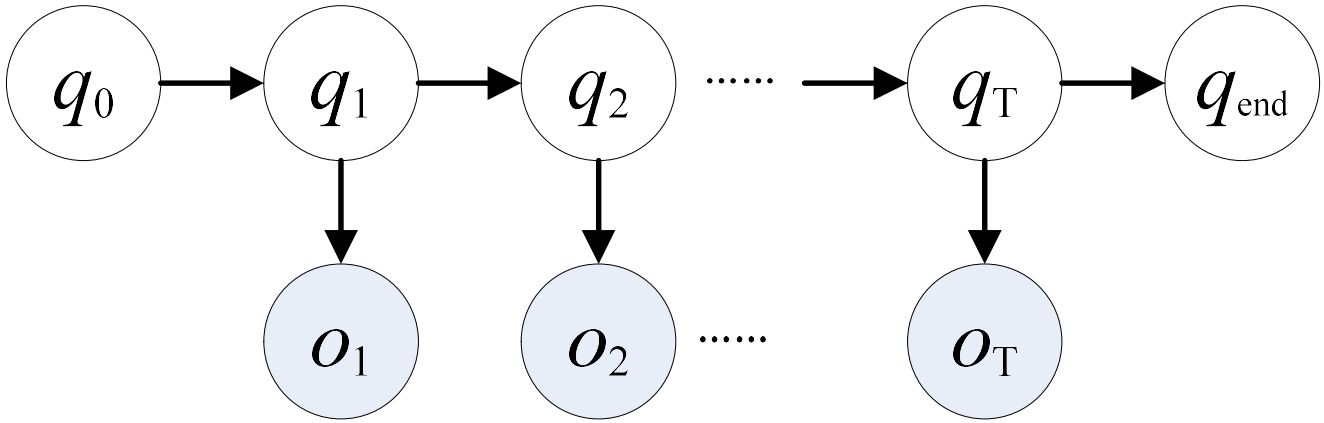
- 冰淇淋数量：观测
- 天气(冷/热)：状态
- 天气变化：状态转移
- 什么天气吃多少冰淇淋：发射概率



$X_1, X_2, \ldots X_5 \in$ Weatherset ={H,C}

$o_1, o_2, \ldots o_5 \in$ No. of Ice Cream={1,…,100}

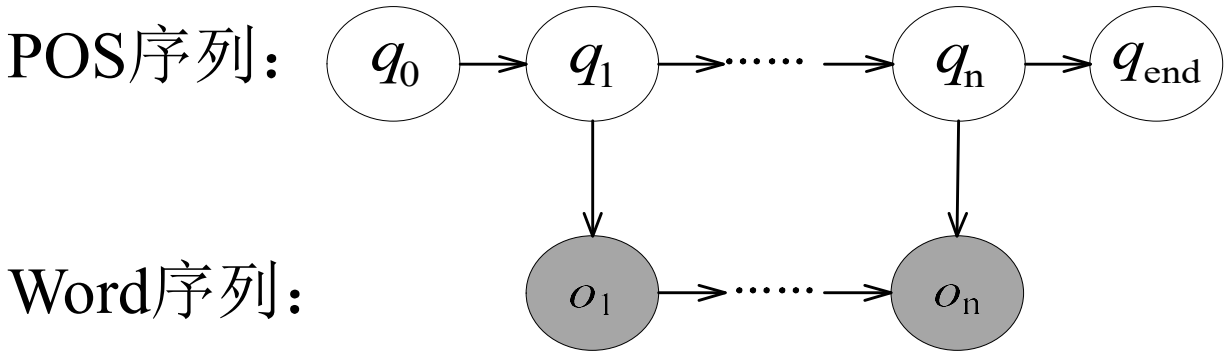
- 共性
  - 观测构成的序列
  - 状态构成的序列(隐变量)
  - 观测由状态决定
  - 状态间存在马氏性



如何用HMM建模POS tagging

- 语言生成的假设
  - N-gram： $o_t \rightarrow o_{t+1}$ (以bigram为例)
- 语言生成的另一种假设：以q记POS标签
  - $q_t \rightarrow o_t$
  - $q_i \rightarrow q_{i+1}$
- 与HMM比较：
  - 词：观测
  - POS：状态
  - POS间存在马氏转移：状态转移
  - 词由当前POS概率决定：发射概率

■HMM模型描述POS和词的关系:



■$q_1,q_2,\ldots q_n$ 取值于 POS集S，存在转移：$a_{ij} = p(q_{t+1} = s_j \mid q_t = s_i)$

■$o_1,o_2,\ldots o_n$ 取值于单词表V，存在发射概率：$b_i(o_t) = P(o_t = v_k \mid q_t = s_i)$

■POS tagging问题：

■已知$o_1,o_2,\ldots o_n$，求解出最优的$q_1,q_2,\ldots q_n$？

从Bayes推断开始给出数学描述：

■词序列: $o_1 . . . o_n$

■POS序列: $q_1 . . . q_n$

■POS tagging任务的概率模型:

$$(\hat{q}_1, \ldots, \hat{q}_n) = \underset{\forall q_t \in S, t=1,\ldots,n}{argmax} P(q_1 . . q_n \mid o_1 . . o_n)$$

■使用Bayes法则

$$(\hat{q}_1 . . \hat{q}_n) = argmax \frac{p(o_1 . . o_n \mid q_1 . . q_n) p(q_1 . . q_n)}{p(o_1 . . o_n)}$$

■进一步

$$(\hat{q}_1 . . \hat{q}_n) = argmax\ p(o_1 . . o_n \mid q_1 . . q_n) p(q_1 . . q_n)$$

$$(\hat{q}_1 . . \hat{q}_n) = argmax\ p(o_1 . . o_n \mid q_1 . . q_n) p(q_1 . . q_n)$$

似然(Likelihood):

$$p(o_1 . . o_n \mid q_1 . . q_n) = p(o_{1,n} \mid q_{1,n}) = p(o_1^n \mid q_1^n)$$

两种不同缩写记号

先验(prior):

$$p(q_1 . . q_n) = p(q_{1,n}) = p(q_1^n)$$

如何计算？难！

在HMM下：运用其中的假设

- 假设1(马尔科夫假设): 当前POS只依赖于前N个POS (一般N=1)
- 假设2(独立性假设)：当前词只依赖于其POS
- 推出：
- 假设3(条件独立性假设)：词之间是条件独立的
- 基于真实语言的简化

$$(\hat{q}_1 . . \hat{q}_n) = argmax\ p(o_1 . . o_n \mid q_1 . . q_n) p(q_1 . . q_n)$$

$$p(o_1, \ldots, o_n \mid q_1, \ldots, q_n) \overset{\text{假设3}}{====} \prod_{t=1}^{n} p(o_t \mid q_1, \ldots, q_n)$$

$$\overset{\text{假设2}}{====} \prod_{t=1}^{n} p(o_t \mid q_t)$$

$$p(q_1, \ldots, q_n) = p(q_n \mid q_1, \ldots, q_{n-1}) p(q_{n-1} \mid q_1, \ldots, q_{n-2}) . . p(q_2 \mid q_1) p(q_1 \mid q_0)$$

$$\overset{\text{假设1}}{=====} p(q_n \mid q_{n-1}) p(q_{n-1} \mid q_{n-2}) . . p(q_2 \mid q_1) p(q_1 \mid q_0)$$

$$= \prod_{t=1}^{n} p(q_t \mid q_{t-1})$$

$$\hat{q}_{1,n} = \arg\max \prod_{t=1}^{n} p(o_t \mid q_t) p(q_t \mid q_{t-1})$$

简化后如何计算？

$$\hat{q}_{1,n} = \arg\max \prod_{t=1}^{n} p(o_t \mid q_t) p(q_t \mid q_{t-1})$$

如果有训练数据(?)，则可用ML估计:

$$P(o_t = w_i \mid q_t = s_j) = \frac{C(w_i, s_j)}{C(s_j)} \qquad P(q_t = s_i \mid q_{t-1} = s_j) = \frac{C(s_j, s_i)}{C(s_j)}$$

之后，对每一种可能的POS序列 $(q_1, \ldots, q_t, \ldots, q_n)$

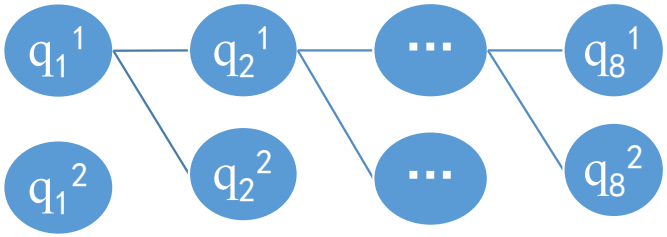
计算一个：$\prod_{t=1}^{n} p(o_t \mid q_t) p(q_t \mid q_{t-1})$

使该值最大的POS序列就是所求。

计算时的问题

$$\hat{q}_{1,n} = \arg\max \prod_{t=1}^{n} p(o_t \mid q_t) p(q_t \mid q_{t-1})$$

■对于8个词的句子，如每个词有2个可能的POS，则可能的POS序列是: 2**8



$$q_{1,n}^{1} = p(o_1 \mid q_1^1) p(q_2^1 \mid q_1^1) p(o_2 \mid q_2^1) p(q_3^1 \mid q_2^1) \ldots$$

$$q_{1,n}^{2} = p(o_1 \mid q_1^2) p(q_2^1 \mid q_1^1) p(o_2 \mid q_2^1) p(q_3^1 \mid q_2^1) \ldots$$

.....

■总共需要计算 2^8 个乘积项
■其中包含大量重复计算

如何优化？➔
HMM下的Viterbi解码算法

Viterbi算法之前先区分一下一种局部最有：

- 另一种可能的标准：
  - 给定观测序列O的最优状态序列Q：Q是由每一个时刻t时最有可能处于的状态q_t所构成的。
  - 在这种标准下的q_t为:

$$\hat{q}_t = \arg\max P(q_t \mid o_1, \ldots o_t) \quad for\ t = 1 \ldots T$$

  - 一个一个时刻分别找q的最优
- 对比：

$$(\hat{q}_1, \ldots, \hat{q}_n) = \underset{\forall q_t \in S, t=1,\ldots,n}{argmax} P(q_1 \ldots q_n \mid o_1 \ldots o_n)$$

- 问题
  - 可能得到某些不可能出现的状态序列为最优序列：某些状态之间的转移概率为0。
  - 问题示例：下页

■例：Box1、box2

Box1	
颜色	数量
白	3
红	7

Box2	
颜色	数量
白	7
红	3

转移概率	box1	Box2
box1	1	0
box2	0	1

■观测粉笔序列：红-白

■t=1时观测为红，p(红/box1)>p(红/box2)，则 q1=box1

■t=2时观测为白，p(白/box1)<p(白/box2)，则 q2=box2，

■则最优解(状态序列)：box1-box2

■不可能: p(box2/box1)=0

- 原因
  - 该最优标准仅考虑在每一个孤立时刻位于某一状态的可能，而没有考虑整个内部状态序列是否存在出现的可能性
- 而

$$\hat{Q} = \arg\max P(Q \mid \lambda, O) \quad \text{其中} Q = (q_1, \ldots, q_T),\ O = (o_1, \ldots, o_T)$$

是状态序列在观察序列***O***的条件下，最可能的内部状态序列（一个完整的最优路经），是HMM建模下Viterbi算法要求解的

■前面已经进行转化得到

$$(\hat{q}_1,...,\hat{q}_T) = \text{argmax}\, P(q_1..q_T \mid o_1..o_T;\lambda)$$

$$\underline{Bayes\text{公式}}\ \text{argmax}\, P(o_1..o_T \mid q_1..q_T;\lambda)P(q_1..q_T)$$

$$\underline{\text{独立性假设}}\ \text{argmax}\prod_i P(o_i \mid q_i)P(q_i \mid q_{i-1})$$

- 但存在计算中的效率问题：重复计算！
- 解决方案： Viterbi算法 (动态规划)
- 定义状态以保留中间结果，相同的只计算一次
- 基于状态变化的递推关系推进计算

- 定义Viterbi变量 (状态的值)

$$v_t(i) = \max_{1 \le k \le N} P(o_1, o_2 .. o_t, q_1, q_2, .., q_{t-1} = s_k, q_t = s_i \mid \lambda)$$

  - 即: 所有可能的前一个状态下最优的当前状态
    - 当前最优 + 考虑历史

  - 状态间有递推关系:

$$v_{t+1}(j) = \max_{1 \le i \le N} v_t(i) a_{ij} b_j(o_{t+1})$$

- 基于Viterbi变量的Viterbi算法
  - 1前向递推计算Viterbi变量
  - 2反向回溯获得最优状态序列

- 例子：
  - HMM的参数
    - 观测空间：词表(5个词)
    - 状态空间：POS表(5个POS)
  - 转移概率矩阵：A
  - 发射概率矩阵：B
  - 状态初始和结束转移概率：

$P(.|q_0)$：
(0.2,0.2,0.3,0.1,0)

$P(q_{end}|.)$：
(0.3,0.2,0.4,0.1,0)

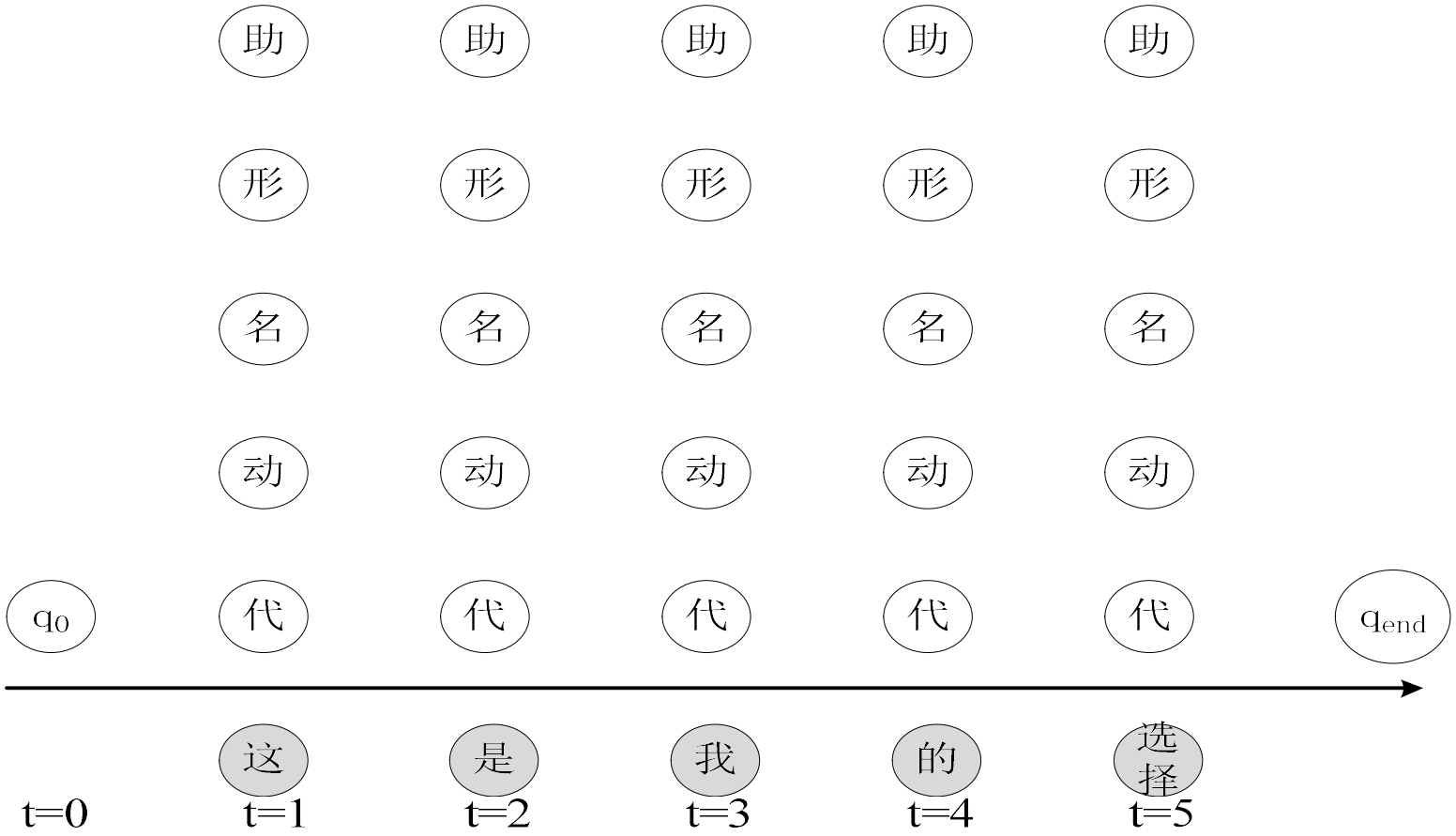
A	代	动	名	形	助	…
代	0.05	0.2	0.1	0.05	0.2	…
动	0.2	0.1	0.3	0.2	0.1	…
名	0.05	0.3	0.2	0.1	0.25	…
形	0.2	0	0.5	0	0.1	…
助	0.1	0	0.3	0.15	0	…

B	这	是	我	的	选择	…
代	0.025	0.002	0.1	0.01	0	…
动	0	0.03	0	0	0.1	…
名	0	0.02	0	0.005	0.07	…
形	0	0.001	0.003	0	0	…
助	0	0	0	0.2	0	…

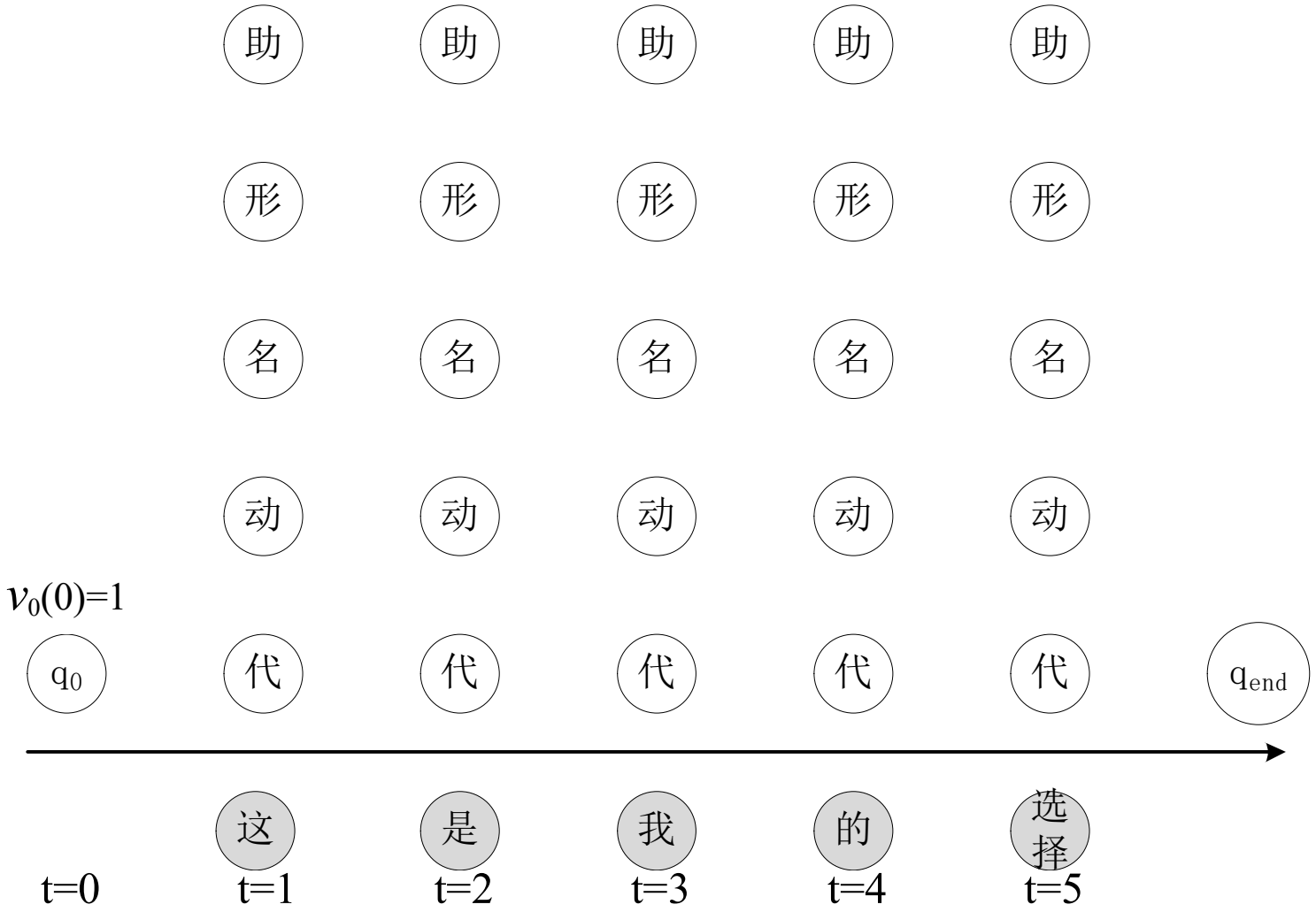
任务：为句子进行POS tagging：这 是 我 的 选择

■构建结构

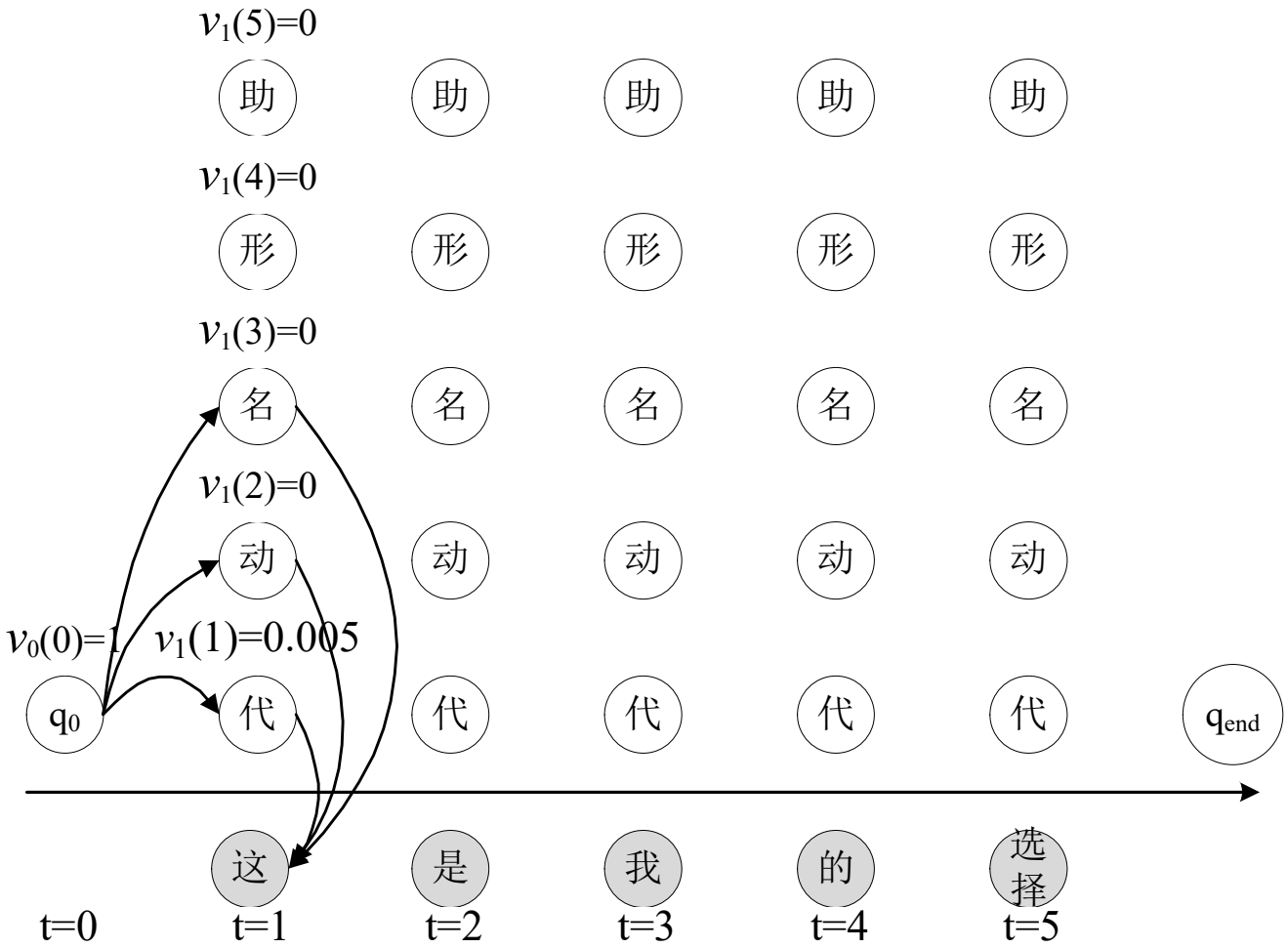
■状态：每个时刻的每个词性，为每一个词性计算其Viterbi变量，并记录产生该Viterbi变量的前一时刻状态



■初始化



■前向计算



■1时刻之前的0时刻只有一个状态：

$v_1(1)=v_0(0)*a_{01}b_1(这)$

$=1*0.2.*0.025=0.005$

$bt_1(1)=0$

$v_1(2)=v_0(0)*a_{02}b_2(这)$

$=1*0.2*0=0$

$bt_1(2)=0$

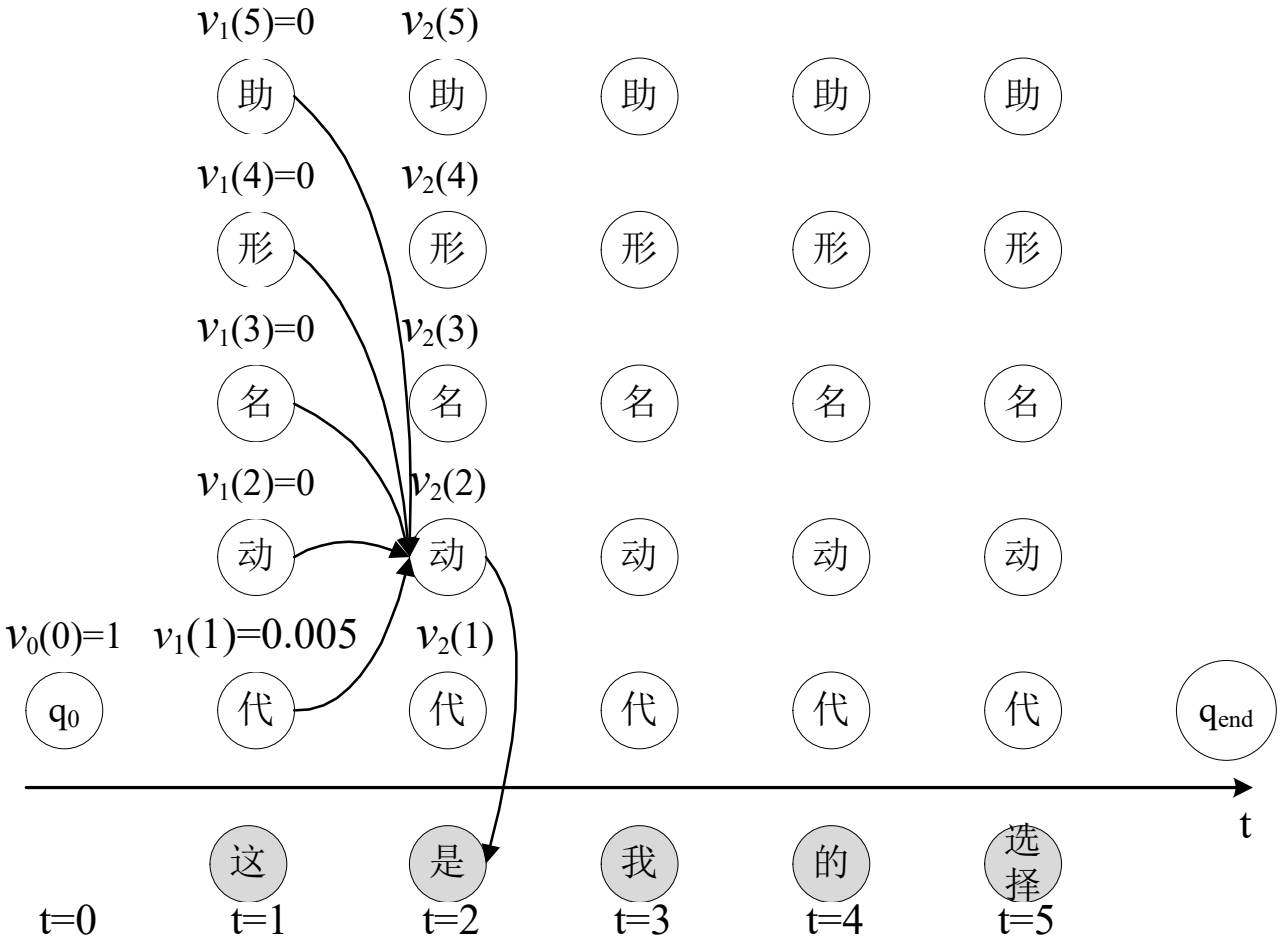
$v_1(3)=v_0(0)*a_{03}b_3(这)$

$=1*0.3*0=0$

$bt_1(3)=0$

……

■前向计算



■2时刻之前的1时刻有5个状态，以下以2时刻的第二个状态为例：

$v_2(2)=\max\{$

$v_1(1)*a_{12}b_2(\text{动})=0.005*0.2*0.03$

$=3*10^{-5}$,

$v_1(2)*a_{22}b_2(\text{动})=0$,

$v_1(3)*a_{32}b_2(\text{动})=0$,

$v_1(4)*a_{42}b_2(\text{动})=0$,

$v_1(5)*a_{52}b_2(\text{动})=0$

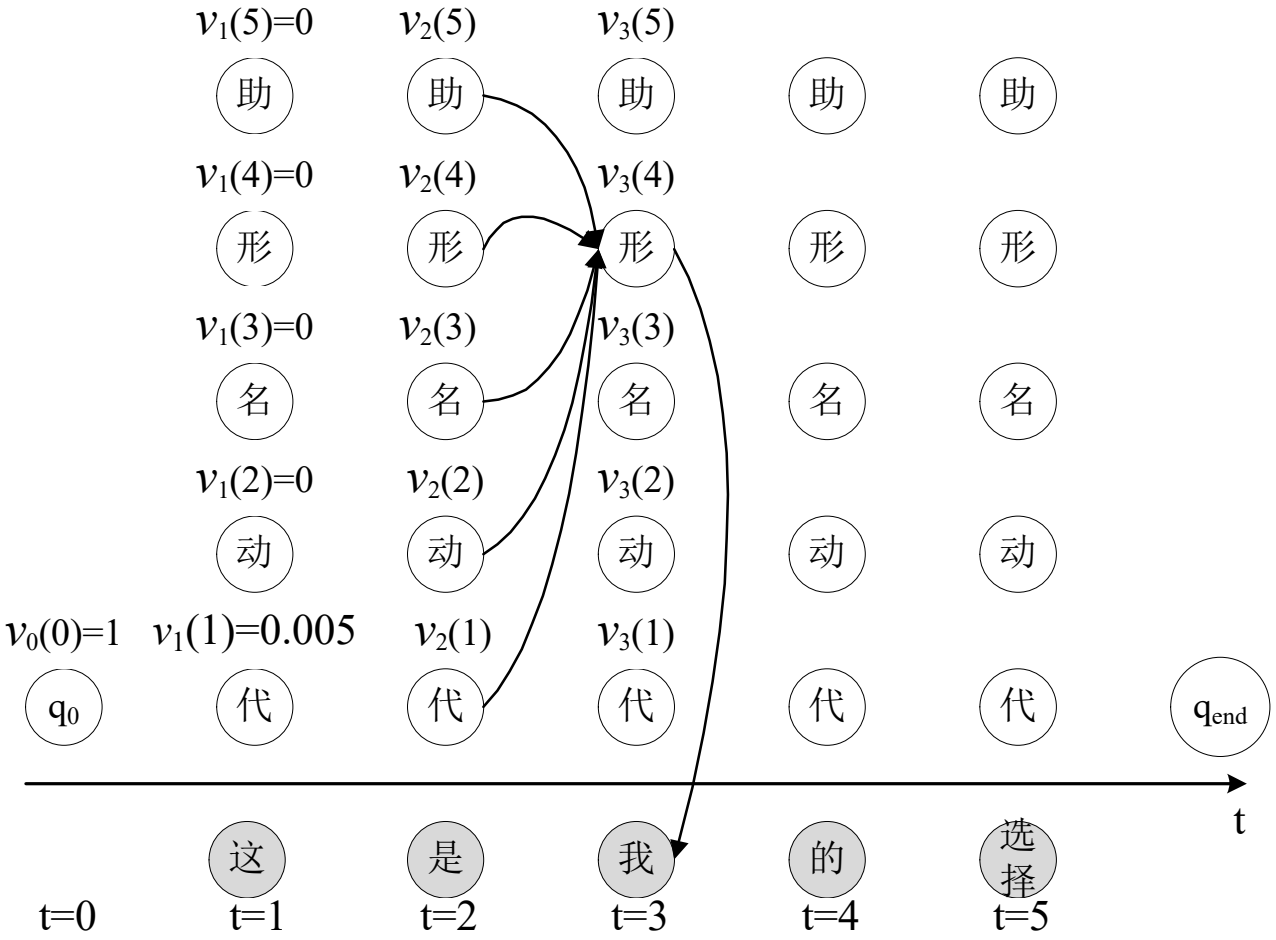
$\}$

$= 3*10^{-5}$

$bt_2(2)=1$

……

■前向计算



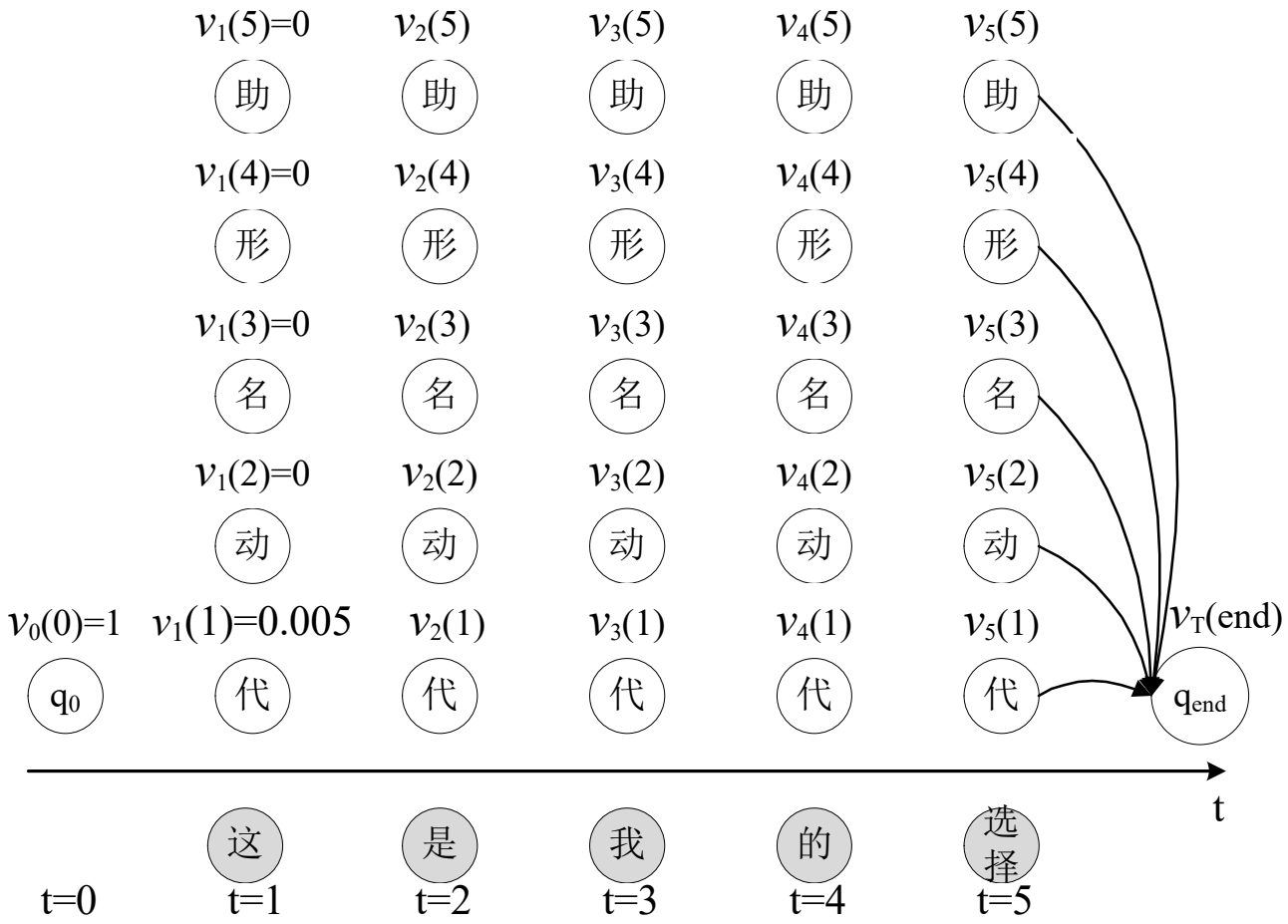
■3时刻之前的2时刻有5个状态，以下以3时刻的第二个状态为例：

$v_3(4)=\max\{$
$v_2(1)*a_{14}b_3(形),$
$v_2(2)*a_{24}b_3(形),$
$v_2(3)*a_{34}b_3(形),$
$v_2(4)*a_{44}b_3(形),$
$v_2(5)*a_{54}b_3(形)$
$\}$
$= X$

$bt_3(4)=Y$

……

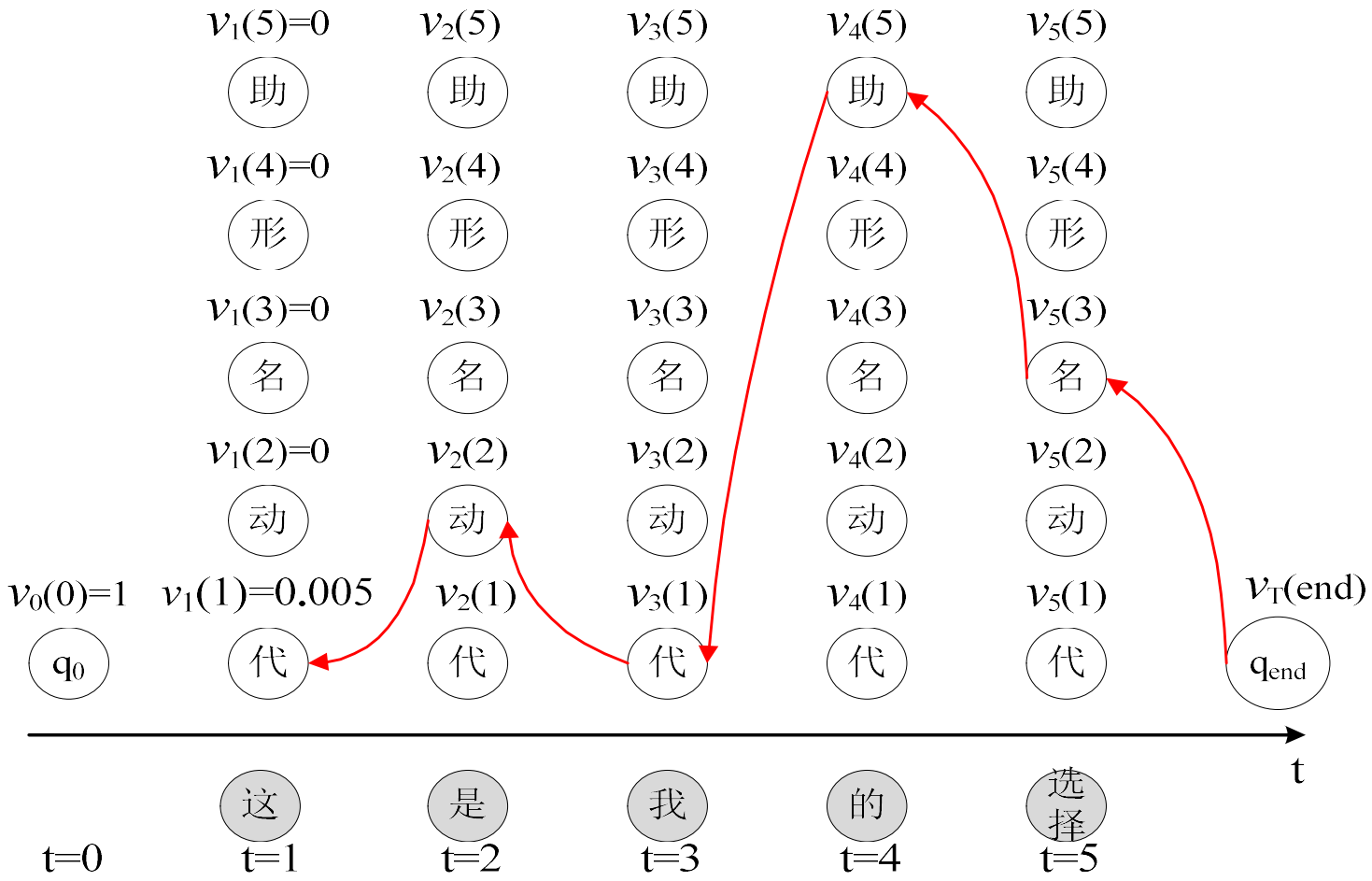
■前向计算



■T时刻之前的时刻有5个状态，T时刻只有一个状态 q_{end}：

$$v_T(end)=\max\{$$
$$v_5(1)*a_{1,}q_{end},$$
$$v_5(2)*a_2,q_{end},$$
$$v_5(3)*a_{3,}q_{end},$$
$$v_5(4)*a_{4,}q_{end},$$
$$v_5(5)*a_5,q_{end}$$
$$\}$$
$$= XX$$
$$bt_T(end)=3$$

反向回溯



■Viterbi算法总结

■初始化：

$$v_1(j) = a_{0j} b_j(o_1) \qquad 1 \le j \le N$$

$$bt_1(j) = 0$$

■递推：

$$v_t(j) = \max_{1 \le i \le N} v_{t-1}(i)\, a_{ij} b_j(o_t) \qquad 1 \le j \le N, 1 < t < T$$

$$bt_t(j) = \operatorname{argmax}_{1 \le i \le N} v_{t-1}(i)\, a_{ij} b_j(o_t) \qquad 1 \le j \le N, 1 < t < T$$

■回溯：

- ■最优概率为：$P^* = \max_{1 \le i \le N} v_{T-1}(i)\, a_{iq_{end}}$
- ■最优路径(回溯第一开始点)：$qT^* = \operatorname{argmax}_{1 \le i \le N} bt_T(i)$
- ■按回溯点回溯到起点q_0，获得最优状态序列。

- 基于HMM的词性标注
- 例：求The plan flies的最优POS序列
- 参数：

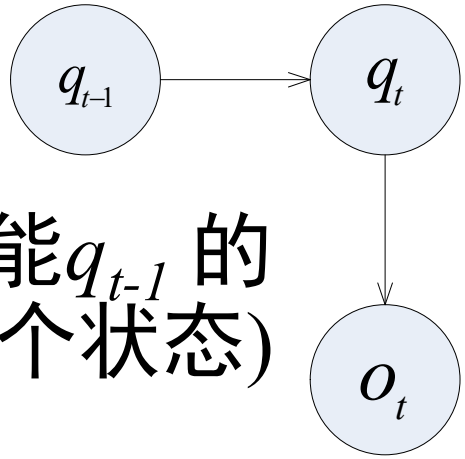
初始	N	V	DET
q0	0.5	0	0.5
转移	**N**	**V**	**DET**
N	0.2	0.8	0
V	0.6	0	0.4
DET	1	0	0

发射	the	plan	fly	…
N	0	0.2	0.1	…
V	0	0.1	0.3	…
DET	0.5	0	0	…

■基于HMM，还可以计算句子概率：评估问题

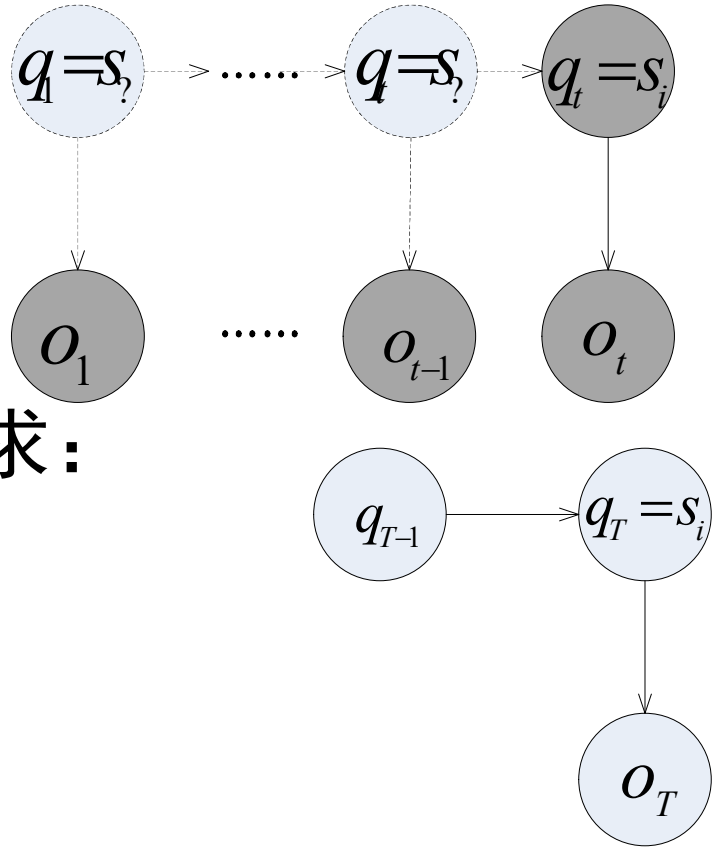
评估问题

- 给定 $O = (o_1, o_2, \ldots, o_T)$ 和 HMM $\lambda = (A, B, \pi)$，计算 $P(O \mid \lambda)$
- o_t的两步生成过程
  - 状态转移： $q_{t-1} \rightarrow q_t$ (依据转移阵A，可能q_{t-1} 的每一个可能状态取值都可以生成 q_t 的一个状态)
  - 输出： $q_t \rightarrow o_t$ 依据发射矩阵B
- 为此定义前项变量：



■前向变量

$$\alpha_t(j) = P(o_1, o_2 \ldots o_t, q_t = s_j \mid \lambda)$$



■为何？

■当$t=T$时，有：

■对所有可能的s_j求和即为所求：

$$\alpha_T(j) = P(o_1, o_2 \ldots o_T, q_T = s_j \mid \lambda)$$

$$\sum_{j=1}^{N} \alpha_T(j) = \sum_{j=1}^{N} P(o_1, o_2 \ldots o_T, q_T = s_j \mid \lambda)$$

$$= P(o_1, o_2 \ldots o_T \mid \lambda)$$

$$= P(O \mid \lambda)$$

- 直接求$\alpha_T(j)$难，要用$\alpha_t(j)$的递推关系:
  - 初始化:

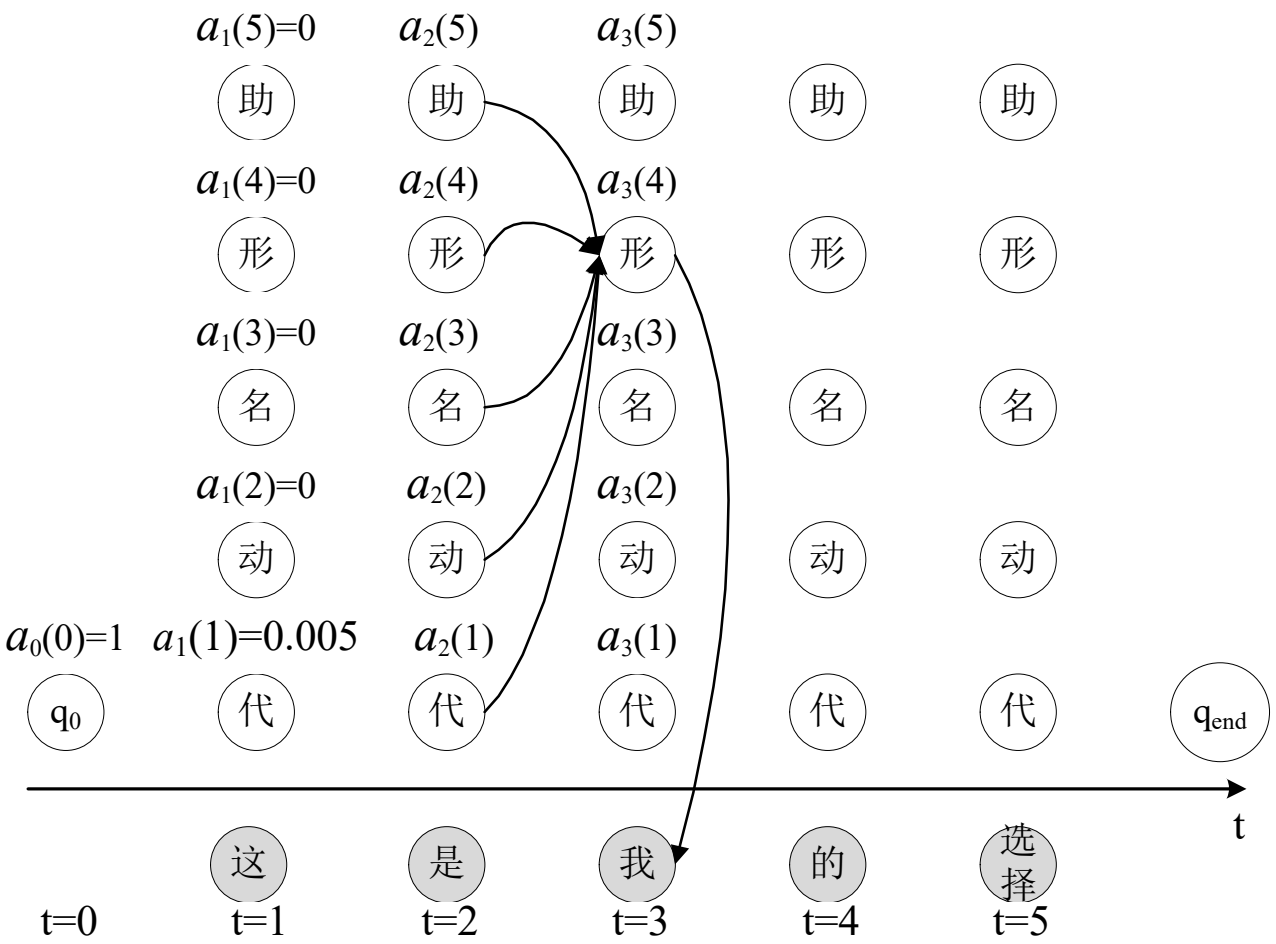
$$\alpha_1(j) = P(o_1, q_1 = s_j) = a_{0j} b_j(o_1), \quad 1 \le j \le N$$

  - 有递推关系:

$$\alpha_t(j) = \sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij} b_j o(t), \quad 1 \le j \le N,\ 1 < t \le T$$

  - 推导:
    - ?
  - 图示见下页

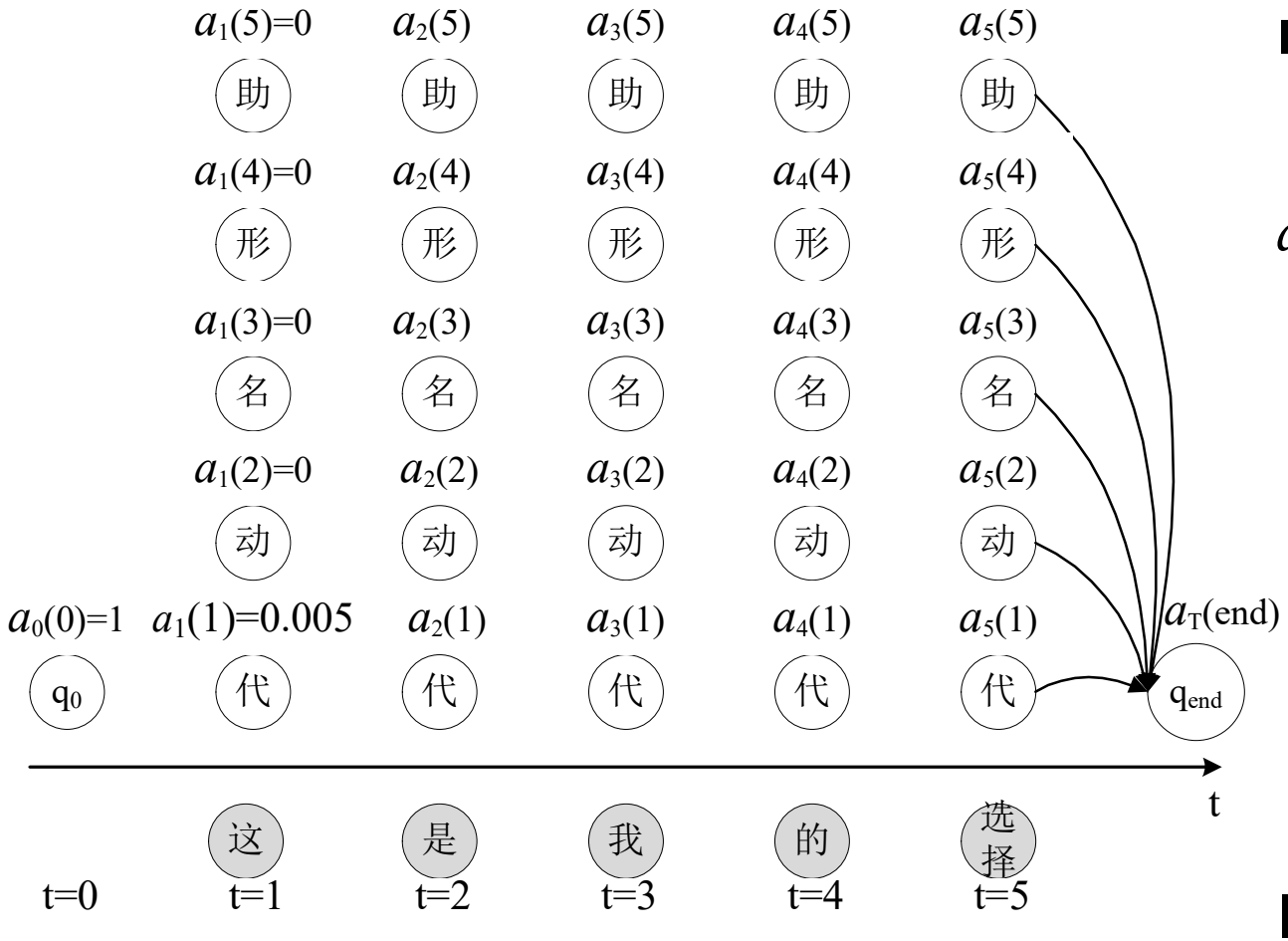
前向变量--递推关系图示



■3时刻之前的2时刻有5个状态，以下以3时刻的第二个状态为例：

$$a_3(4)=\text{Sigma}\{ a_2(1)*a_{14}b_3(形), a_2(2)*a_{24}b_3(形), a_2(3)*a_{34}b_3(形), a_2(4)*a_{44}b_3(形), a_2(5)*a_{54}b_3(形) \} = X$$

■前向计算



■T时刻之前的时刻有5个状态，T时刻只有一个状态 q_{end}：

$$a_T(end)=Sigma\{$$
$$a_5(1)*a_{1,}q_{end},$$
$$a_5(2)*a_2,q_{end},$$
$$a_5(3)*a_{3,}q_{end},$$
$$a_5(4)*a_{4,}q_{end},$$
$$a_5(5)*a_5,q_{end}$$
$$\}$$
$$= XX$$

■此即为观测序列的概率！

基于后向变量的评估问题求解

■类似定义后向变量:

$$\beta_t(j) = P(o_{t+1}, o_{t+2} \ldots o_T \mid q_t = j, \lambda)$$

■递推公式:

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j o(t+1) \beta_{t+1}(j) \qquad 1 \le i \le N,\ 1 \le t < T$$

$$P(O/\lambda) = \beta_1(q_0) = \sum_{j=1}^{N} a_{0j} b_j(o_1) \beta_1(j)$$

■综合前后向：

$$P(O \mid \lambda)$$

$$= \sum_{i=1}^{N} P(O, q_t = i \mid \lambda)$$

$$= \sum_{i=1}^{N} \alpha_t(i)\beta_t(i)$$

参数估计 A $(\hat{a}_{ij})_{N\times N}$、 B $(\hat{b}_i(v_k))_{N\times M}$

- 有训练数据是的估计:MLE
- 无训练数据时的估计
  - Baum-Welch法（等价于Estimation-Maximization法）
  - 核心思想:构建一个逐步迭代求精的过程
    - 估计一个A,B
    - 在已知A,B下依据某个方式重估A,B

■构建

$$\xi_t(i,j)$$

$$= P(q_t = i, q_{t+1} = j \mid O, \lambda)$$

$$= \frac{\alpha_t(i) a_{ij} b_{t+1}(j) \beta_{t+1}(j)}{\sum_{i=1}^{N} \alpha_t(i) \beta_t(i)}$$

■另一方面

$$\hat{a}_{ij}$$

$$= \frac{\text{从状态i转移到j的次数}}{\text{从状态i转出的次数}}$$

$$= \frac{\text{从状态i转移到j的期望次数}}{\text{从状态i转出的期望次数}}$$

$$= \frac{\sum_t \xi_t(i,j)}{\sum_j \sum_t \xi_t(i,j)}$$

$$\hat{b}_i(v_k)$$

$$= \frac{\text{从状态i发射输出}v_k\text{的次数}}{\text{在状态i的次数}}$$

$$= \frac{\text{从状态i发射输出}v_k\text{的期望次数}}{\text{在状态i的期望次数}}$$

$$= \frac{\sum_{t=1:s.t.o_t=v_k}^{T} \sum_j \xi_t(i,j)}{\sum_j \sum_t \xi_t(i,j)}$$

- 算法：
- 初始化一个A，B,
- k=0
- 进行如下循环：
- { E-step：由第k个A，B求 $\xi_t(i,j)$
- M-step：由 $\xi_t(i,j)$求第k+1个A，B }
- 终止条件: 第k+1个A，B与第k个A，B相比差别在预定范围。
- 输出此时的A，B即为所求。

总结：HMM三问题求解关键

- 模型参数估计

$$\xi_t(i,j) = P(q_t = i, q_{t+1} = j \mid O, \lambda)$$

- 解码问题，最优状态序列

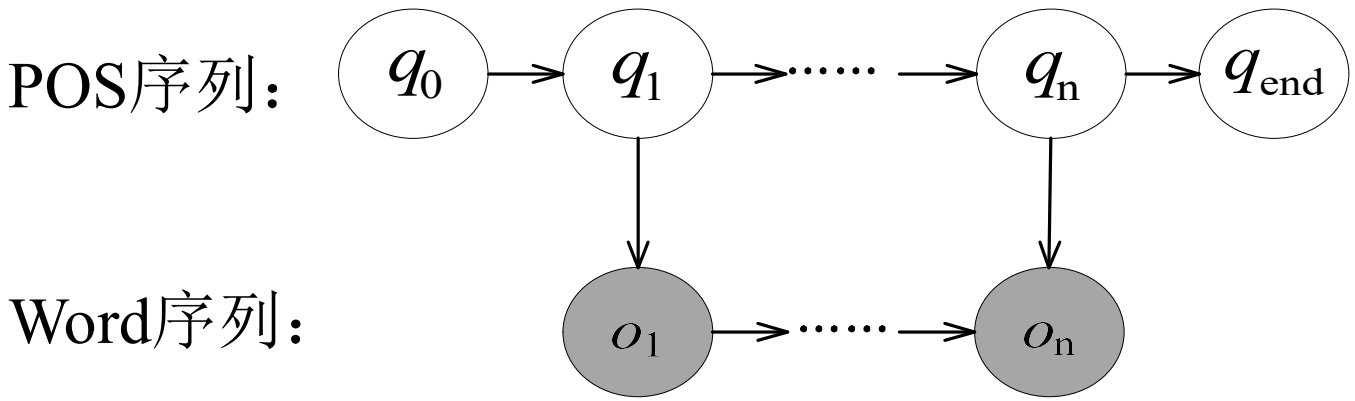
$$v_t(i) = \max_{1 \le j \le N} P(o_1, o_2 ... o_t, q_1, q_2, .., q_{t-1} = s_j, q_t = s_i \mid \lambda)$$

- 评估问题：哪组参数合适，观测序列概率估计

$$\alpha_t(j) = P(o_1 ... o_{t-1}, o_t, q_t = s_j \mid \lambda)$$

回顾与比较：MC(n-gram) vs HMM

	隐状态	观测
MC(N-gram)	无	有，观测间有马氏性
HMM	有，状态间有马氏性	有，观测间无之间关联，当前观测由当前状态决定



- 基于HMM模型中所用知识：
  - POS之间的转移概率
  - POS到词的发射概率
- 可能不够：例如

- 例子：包的词性
  - 请/把/这/些/文件/N 包/V 好/。
  - 本/网络/限制/用户/发/ 文件/N 包/N 。
  - 这/个/软件/开发/V 包/V 在/我/身上
  - 可以/下载/免费/ 开发/V 包/N 。
- HMM用的局部信息难以区别上例
- 利用更多信息的能力？更多的上下文：把+V、发+N的可能信息的帮助
  - 增加马氏性的阶数：有效，但是，有时只是远距离某个特定词有关，高阶增加复杂性的同时也增加了噪声。

- 极大熵(MaxEnt)模型是一个能综合丰富上下文特征的模型，但是是一个分类模型，每个样本的标签是独立的，而序列标注可以说是序列分类问题，标签间存在关联约束， MaxEnt不能建模这种序列之间的关联
- MaxEnt + Markov模型

- MaxEnt
  - 基于多特征建模条件概率

$$p(y \mid x) = \frac{1}{Z_\lambda(x)} \exp(\sum_i \lambda_i f_i(x, y))$$

$$Z_\lambda(x) = \sum_y \exp(\sum_i \lambda_i f_i(x, y))$$

■MEMM-序列标注

$$\hat{T} = \arg\max p(t_1, t_2, \ldots, t_n / w_1, w_2, \ldots, w_n)$$

$$\hat{T} \approx \arg\max \prod_i p(t_i \mid w_i, t_{i-1})$$

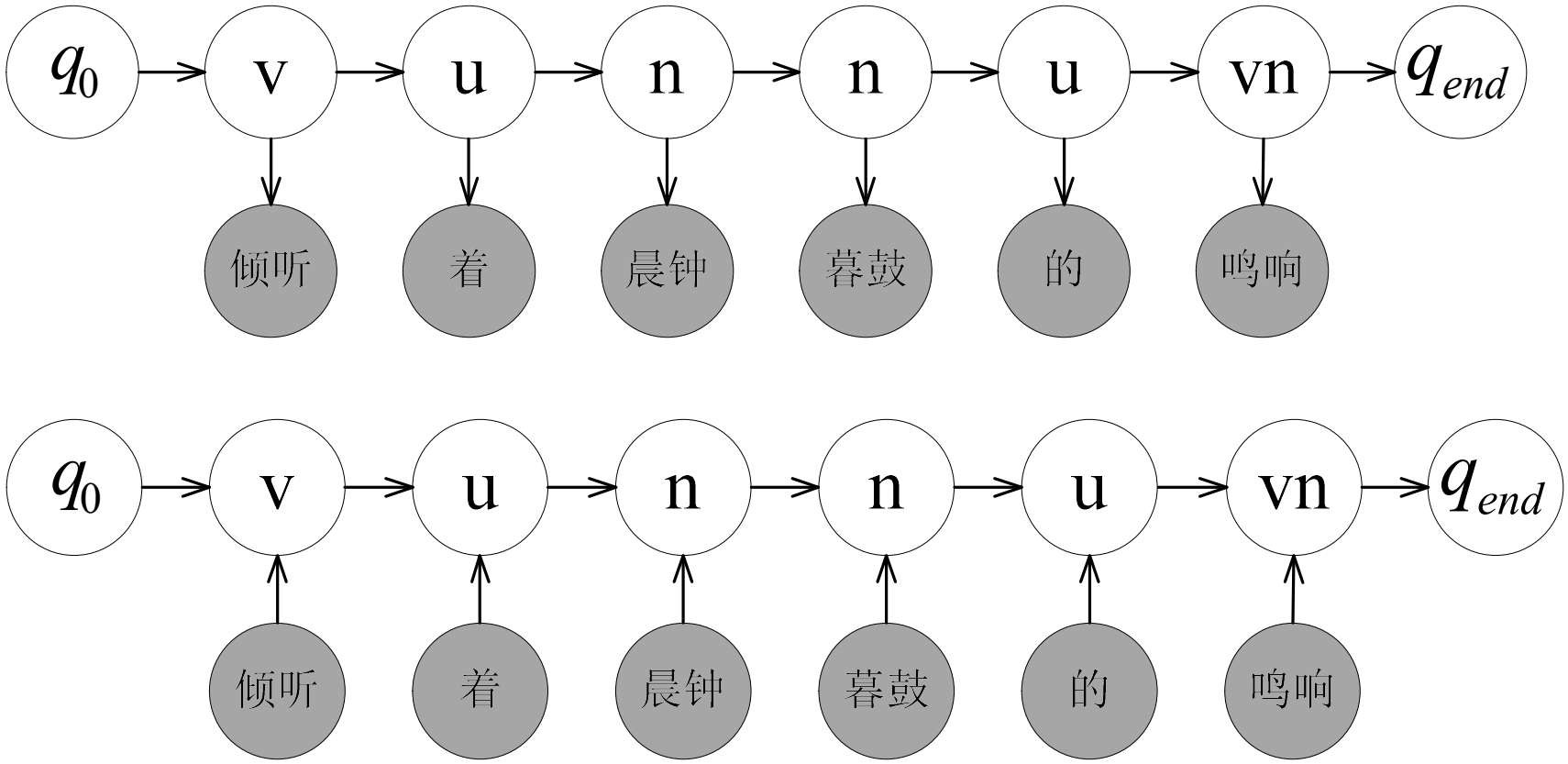
■其中的条件概率 $p(t_i \mid w_i, t_{i-1})$ 采用MaxEnt来进行建模

$$p(t_i \mid w_i, t_{i-1}) = \frac{1}{Z} \exp(\sum_j w_j f_j(w, t))$$

■HMM vs MEMM

HMM $\hat{T} = \arg\max \prod_i p(word_i / tag_i) \prod_i p(tag_i \mid tag_{i-1})$

MEMM $\hat{T} = \arg\max \prod_i p(tag_i \mid word_i, tag_{i-1})$

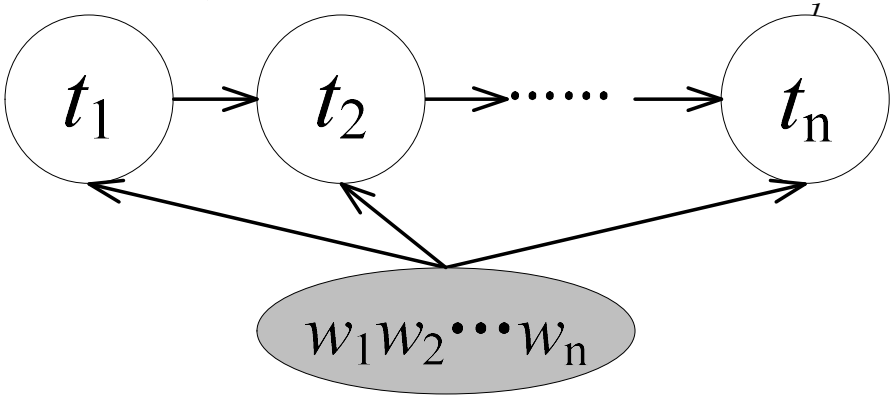


- MEMM解码
  - 可用与HMM中类似的Viterbi算法。
- MEMM参数估计
  - 可用和MaxEnt中类似的方法(L-BFGS…)。

- MEMM的问题：使用更多上下文信息
  - 后面的信息也有价值
    - 把这个 包/V 一下
    - 把这个 包/N 给/V 我

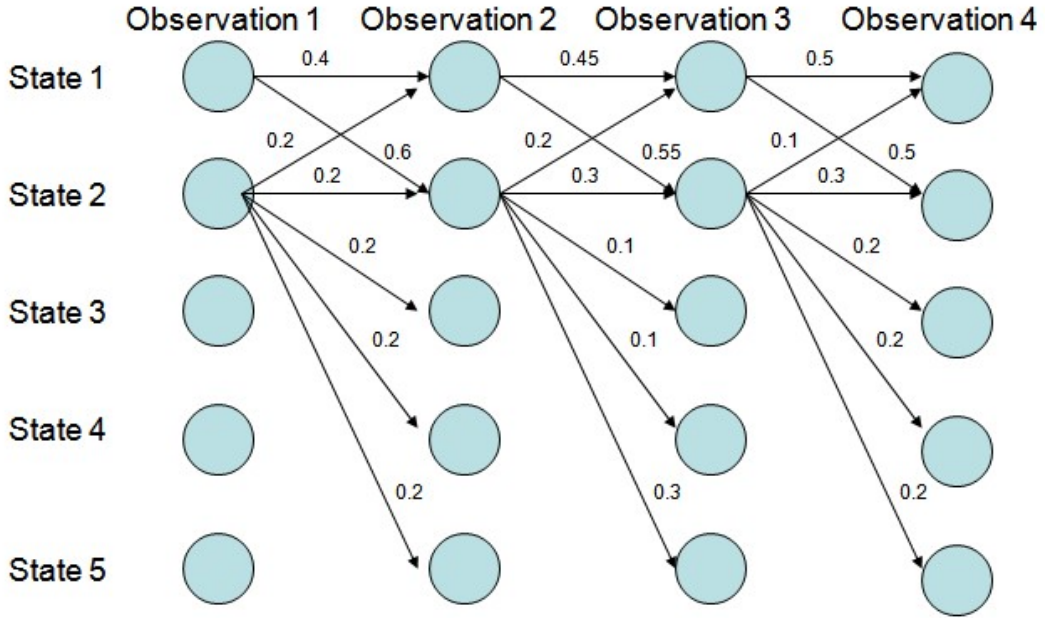
$$\hat{T} \approx \arg\max \prod_i p(t_i \mid t_{i-1}, w_i)$$

$$\hat{T} \approx \arg\max \prod_i p(t_i \mid t_{i-1}, w_1, w_2 \ldots w_n)$$



■MEMM的问题：模型偏置

- ■最优：1-1-1-1
- ■MEMM：1-2-2-2



■局部归一化

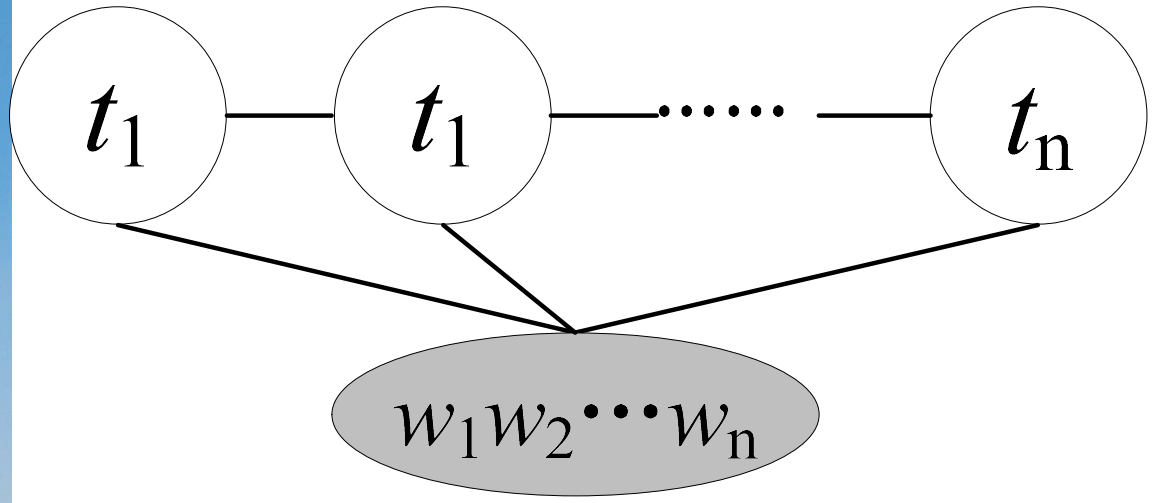
$$\hat{T} = \arg\max \prod_i p(t_i \mid t_{i-1}, w_1, w_2 \ldots w_n)$$

$$= \arg\max \prod_i \frac{1}{Z} \exp\left(\sum_i \lambda_i f_i(t_{i-1}, t_i, w_1, w_2 \ldots w_n)\right)$$

$$Z = \sum_{t_{i-1}, w_1, \ldots w_n} \exp\left(\sum_i \lambda_i f_i(t_{i-1}, t_i, w_1, \ldots w_n)\right)$$

■条件随机场(Conditional Random Field:CRF)

$$P(t_1,..t_n \mid w_1,...,w_n;\lambda) = \frac{1}{Z(w_1,...,w_n)} \exp(\sum_i \lambda_i f_i(t_1,...,t_i,w_1,w_2...w_n)$$



CRF的问题

■特征可以很丰富：上下文及其组合均可以用

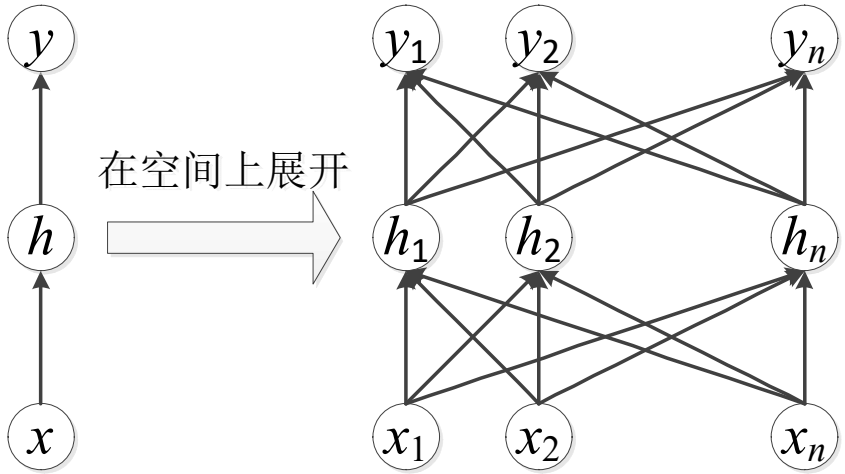
$$f_1 = \begin{cases} 1 & w_1 = Can \ \wedge \ pos_1 = AUX \\ 0 & other \end{cases} \qquad f_2 = \begin{cases} 1 & pos_1 = v \ \wedge \ w_n = ? \\ 0 & other \end{cases}$$

■但是，都需要人工设计：特征工程

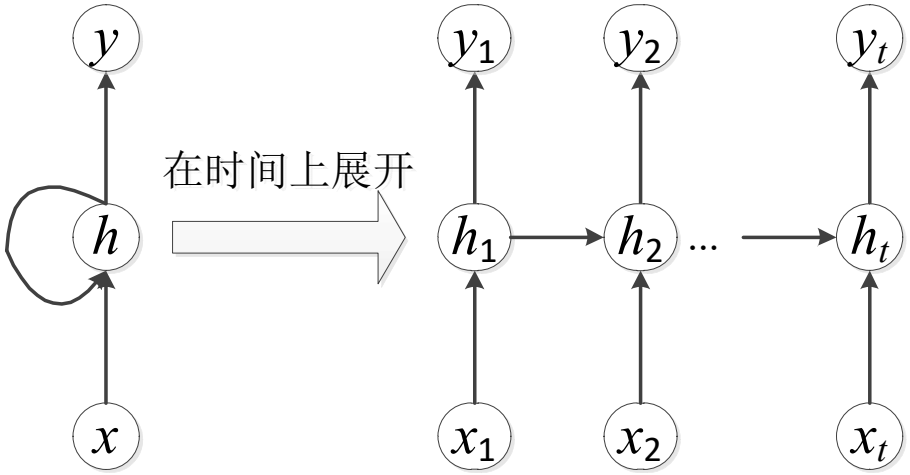
- 近年来，利用深层神经网络学习特征，之后再用于CRF进行序标

- RNN网络常用于提取句子序列的特征
- 循环神经网络 (RNN：Recurrent NN)
  - 简单RNN (SRN, Elman型RNN)：

RNN

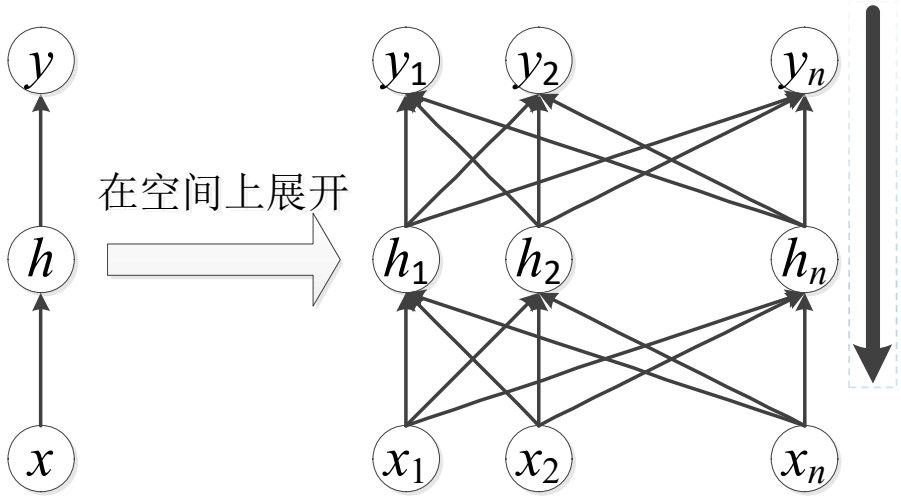


FNN　　FNN (在空间上展开)

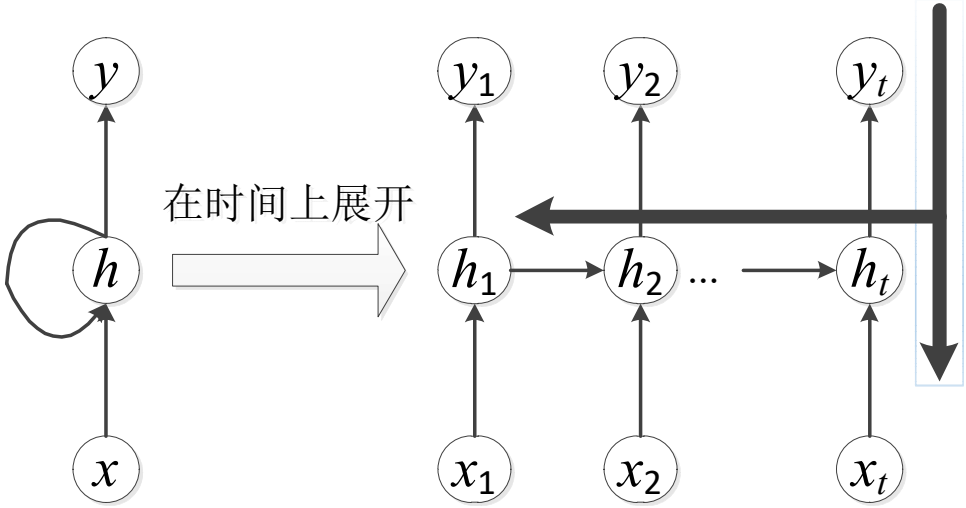


RNN　　RNN(在时间上展开)

RNN

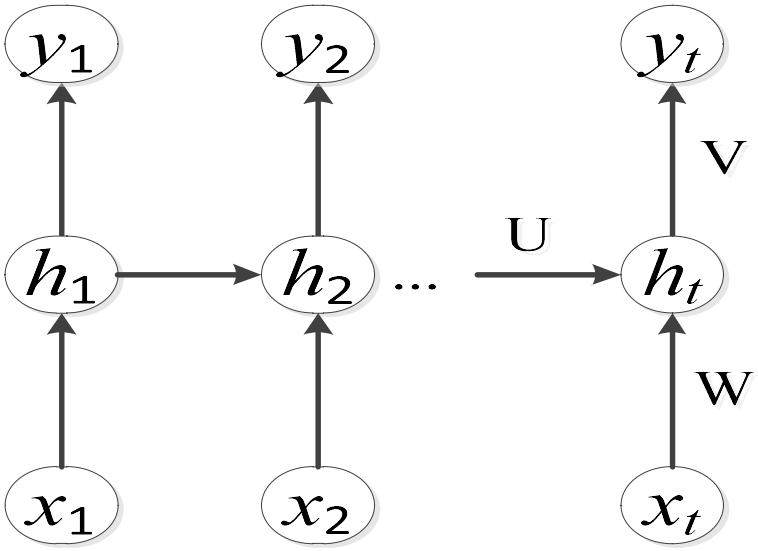


FNN　　FNN (在空间上展开)



RNN　　RNN(在时间上展开)

RNN

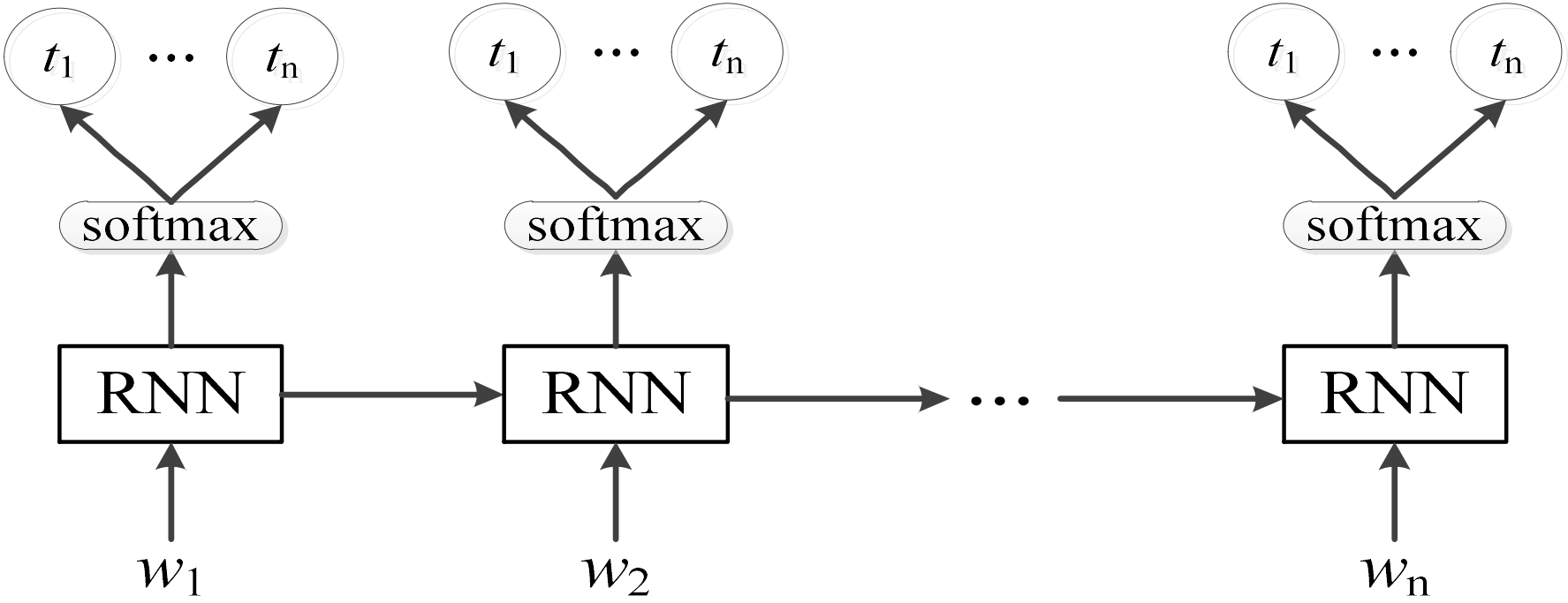


RNN(在时间上展开)

- 前向计算
  - t时刻：
  - $h_t = g(Uh_{t-1} + Wx_t + \mathrm{b})$
  - $y_t = f(Vh_t)$
    - $h_t = \tanh(Uh_{t-1} + Wx_t + \mathrm{b})$
    - $y_t = \mathrm{softmax}(Vh_t)$
- 反向误差反传
  - BPTT(back propagation through time)

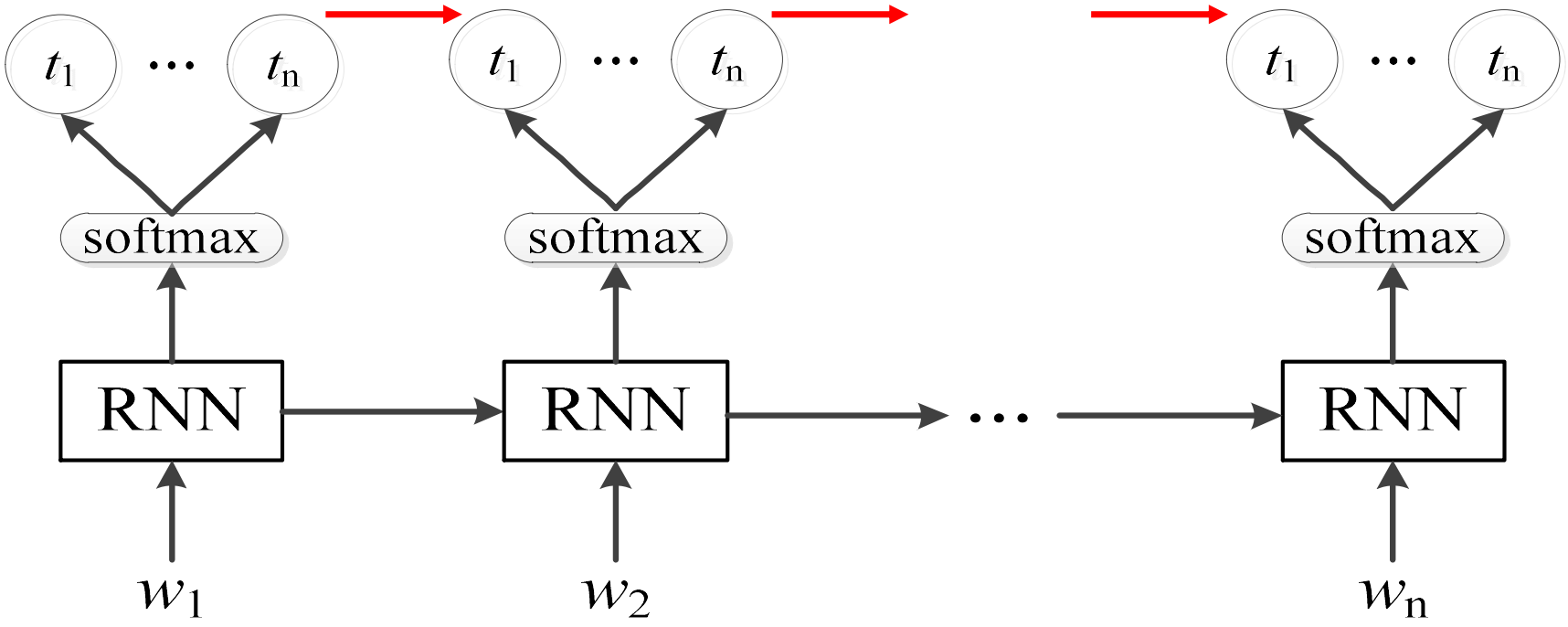
将RNN用于序列标注

- 利用输出进行分类获得标签，输出的维数定义为标签数
  - 对于POS标注任务，是POS数(N\V\...)，也可以用于NER等序列标注任务中(B-PER\I-PER\B-LOC\I-LOC\...)。



将RNN用于序列标注

■但是，每个标签是单独分类的结果，标签间的关系没有直接建模，可能出现矛盾标签序列，例如，对于NER任务： w_i标 B-PER，w_{i+1}标 I-LOC



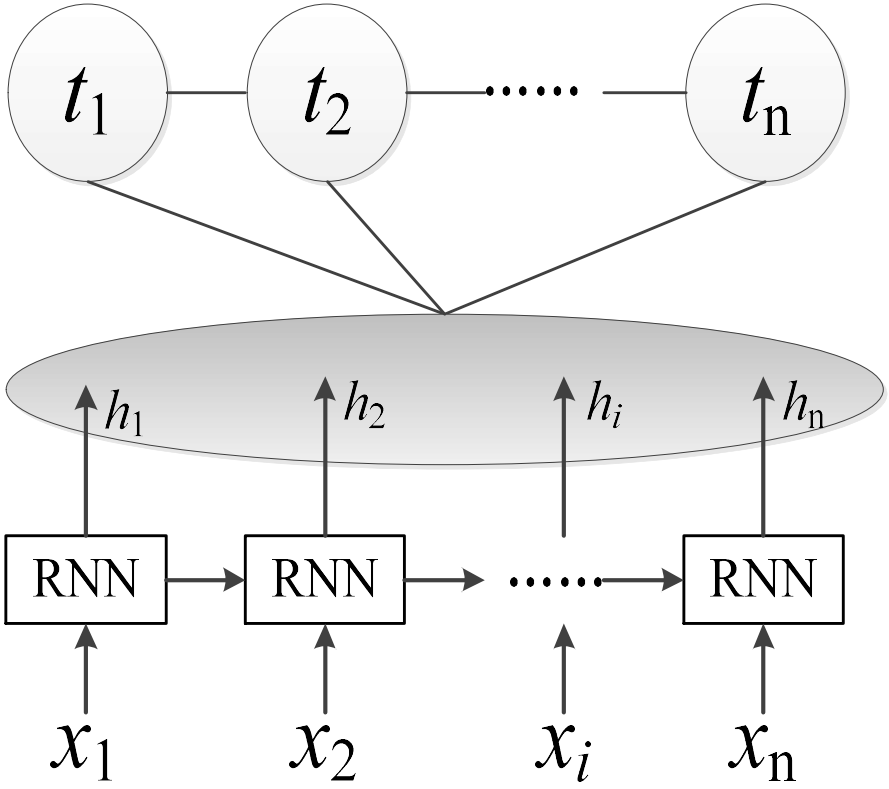
■增加标签关联信息

■例如(Lample2016NAACL-HLT：增加一个关联层)

$X = (x_1, x_2, \ldots, x_n) \in \mathrm{R}^{\mathrm{m}\times\mathrm{n}}$

$H = (H_{i,j}) \in \mathrm{R}^{k\times\mathrm{n}}$

$\mathrm{A} = (A_{i,j}) \in \mathrm{R}^{(k+2)\times(k+2)}$



■n：词表大小

■m：词向量维度

■k+2：实际标签数 + #{s,end}

■$A_{i,j}$：标签t_i转移到t_j的概率

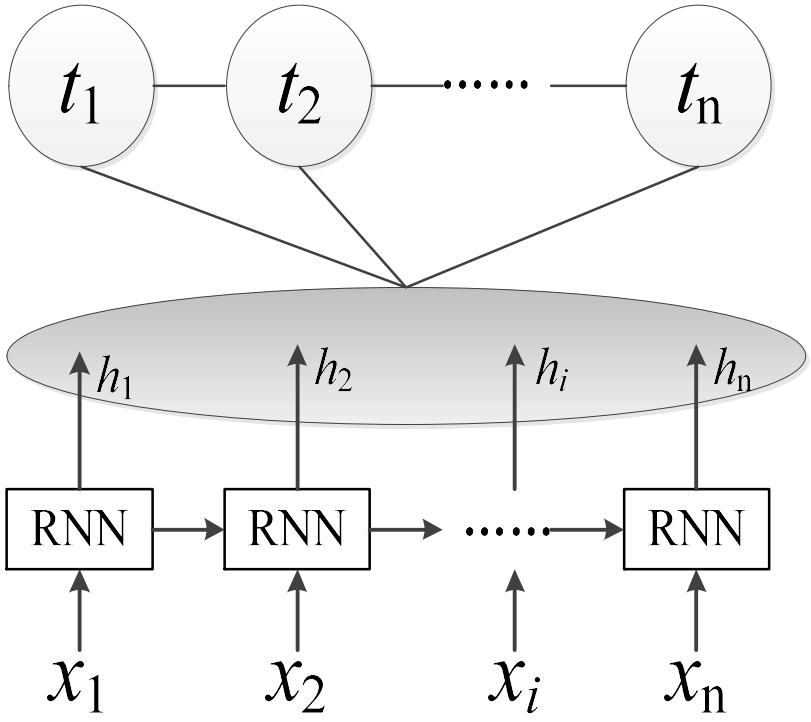
■$H_{i,j}$：i为第i个标签，j为第j个词，
为第j个词取到第i个标签的初始分

■定义一个得分函数建模标签关联：

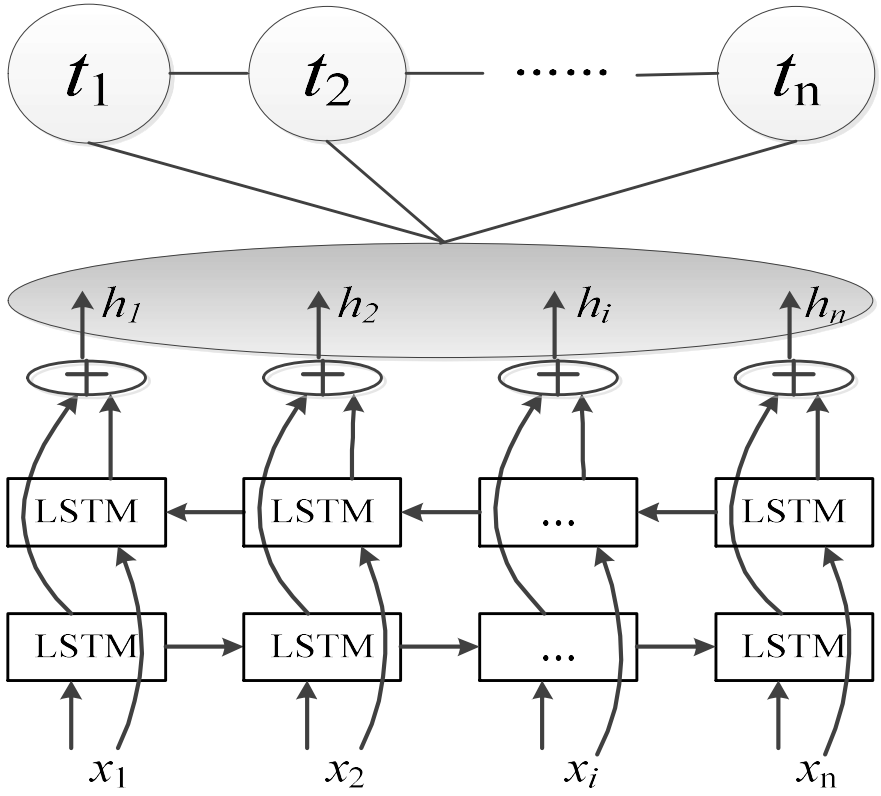
$$s\,(X, ti) = \sum_{i=0}^{n} A_{t_i,t_{i+1}} + \sum_{i=1}^{n} H_{t_i,i}$$

■基于此定义一个softmax分类

$$p\,(t_i|X) = \frac{e^{s(X,ti)}}{\sum_{t'} e^{s(X,t')}}$$



- RNN采用LSTM\GRU\BiLSTM等单元的RNN模型
- H层和CRF层间加一个隐层
- …



- 序列标注模型
  - HMM、MEMM、CRF、RNN、RNN-CRF…
- 序列标注模型能做什么？
  - 切分
  - 命名实体识别
  - POS tagging
  - 组块分析(后面会再提到)
  - 语义剖析(槽填充，后面会再提到)
  - ……

■应用举例：电子病历的结构化

主诉：右耳突发性听力下降，伴耳鸣

现病史：患者3天前感冒，出现右耳听力下降，伴有持续性右耳耳鸣，为高音调蝉鸣声，当时自行服用敏使朗、银杏叶片，效果欠佳。无发热、咳嗽、咳痰，今日患者出现头晕，呕吐一次，呕吐物为胃内内容，无视物旋转。

既往史：否认肝炎、结核、疟疾等传染病史，否认高血压、心脏病史，否认糖尿病、脑血管疾病病史，否认手术史，否认外伤史，否认输血史，否认药物、食物过敏史，预防接种史不详。

家族史：父亲已故，母亲健在，有1姐体健，家族无传染病及遗传病史

体格检查：体温36.3度，脉搏76次/分，呼吸18次/分，血压110/72mmHg，身高170cm，体重80kg。发育正常，营养良好，双侧膝、跟腱反射正常，双侧Babinski号阳性

名称	类别	状态
耳鸣	症状	有
感冒	疾病	有
听力下降	症状	有
高血压	疾病	无
心脏病	疾病	无
敏使朗	药物	有
音叉检查	检验	无
呼吸	指标	18次/分
体温	指标	36.3度
…	…	…

大纲

- 引言：短语结构问题
- 词性标注集(POS Tagset)
- 词性标注 (POS Tagging)
- 基于规则的词性标注方法
- 基于统计的词性标注方法
- 总结

总结

- 序列标注模型
  - HMM、MEMM、CRF…统计序列标注模型
    - 手工设计特征
  - 结合神经网络模型的序标模型
    - 神经网络自动提取多层次表示(特征)
    - 传统方法显式建模标签关系

谢谢！