隐马尔科夫模型和词类标注

常宝宝
北京大学计算语言学研究所
chbb@pku.edu.cn

主要内容

- **词类标注及词类划分的语言学背景**
  - 划分词类的标准
  - 常见词类标记集
  - 兼类词和词类排歧
- 隐马尔可夫模型
  - 向前算法
  - 韦特比算法
  - Baum-Welch算法
- 隐马尔可夫模型和词类标注

什么是词类标注

- 什么是词类标注？

  判定自然语言句子中的每个词的词类并给每个词赋以词类标记。

  例如：

  - book that flight.
    book/VB that/DT flight/NN ./.

  - 这 份 特区 政府 的 报告 长 达 2 0 页 。
    这/r 份/q 特区/n 政府/n 的/u 报告/n 长/a 达/v 2 0 /m 页/q 。/w

VB-动词 DT-限定词 NN-名词 .-句号

r-代词 q-量词 n-名词 u-助词 a-形容词 v-动词 m-数次 w-标点

词类的划分标准

- **形态标准**

  *Words that function similarly with respect to the **affixes** they take (their **morphological properties**) are grouped into classes.*

- **分布标准**

  *Words that function similarly with respect to what can occur **nearby** (their "**syntactic distributional properties**") are grouped into classes.*

- 意义标准(×)

  *While word classes do have tendencies toward **semantic coherence** (nouns do in fact often describe "people, places or things", and adjectives often describe properties), this is not necessarily the case, and in general **we don't use the semantic coherence as a definition criterion for part-of-speech**.*

英语中词的分类

- 英语词类
  - ***preposition***, ***determiner***, *pronoun*, ***conjunction***, *nouns*, *verbs*, *adjectives*, *adverbs*, *numeral*, *interjection*
- *closed class* and *open class*
  - *Closed classes are those that have relative fixed membership, in which new words are rarely coined.*
- *function word* and *content word*

汉语中词的分类

- 汉语中词的分类依据
  - 缺乏形态，形态特征不能用作分类依据。
  - 词的分布特征，或者说词的语法功能
  - 汉语中划分词类也不用意义作为分类依据。
    概念相近的词，语法性质未必相同，例：战争(名词)、战斗(动词)
- 词的语法功能:词在句法结构里所能占据的语法位置
  - 词在句法结构中充当句法成分的能力
  - 词与某类词或某些词组合成短语的能力
- 虽然不能根据意义对词进行分类，但按照分布特征同属一类的词，意义上也常有共性。
  - 名词通常表示事物的名称、动词通常表示动作和行为、形容词表示事物的性质和状态。

汉语中词的分类

- 实词和虚词
  - 从功能上看，实词可以充当主语、谓语和宾语。虚词则不可以。
  - 从意义上看，实词有实在的意义，表示事物、动作、行为、变化、性质、状态、处所、时间等。虚词基本只起语法作用，本身多无实在意义。
  - 从数量上看，实词多为开放类，虚词多为封闭类。
- 体词和谓词
  - 实词通常可进一步分成体词和谓词。体词可以做主语和宾语。谓词主要做谓语。

汉语中词的分类

- 体词
  - 名词(1)、处所词(2)、方位词(3)、时间词(4)、区别词(5)、数词(6)、量词(7)、代词(8)
- 谓词
  - 动词(9)、形容词(10)
- 虚词
  - 副词(11)、介词(12)、连词(13)、助词(14)、语气词(15)
- 拟声词(16)、感叹词(17)

朱德熙，《语法讲义》，商务印书馆，1982

汉语中词的分类

- 为什么说一个词是形容词？
  - 可以用作主谓结构中的谓语，但不能带真宾语。
    - 例：长江比黄河**长**
  - 可以受“很”这类程度副词修饰。例：很**长**、很**雄伟**、很**安静**
  - 可以作述补结构中的补语。例：洗**干净**、捆**结实**
  - 直接或加“地”后作状中结构中的状语。例：**迅速**提高
  - 直接或加“的”后作定中结构中的定语。例：**美丽**人生
  - 可以用“a + 不 + a”的形式提问。例：**舒服**不**舒服**？
  - ……

俞士汶等，《现代汉语语法信息词典详解》，清华大学出版社，2003

汉语中词的分类争议

- 对汉语词类问题有兴趣，可进一步参考有关书籍。
- 由于汉语缺乏形态，词的类别不如英语等西方语言那样易于判别。汉语语言学家在汉语词类划分问题上一直有不同见解，经过长期争议，至今仍然存在多种看法，如：

  汉语无词类

  依句辩品、离句无品

- 利用计算机处理语言，需要进行词语的语法分类和代码化。需要建立面向信息处理用汉语词类体系并进行大规模词语归类实践。

英语词类标记集

- *Brown corpus tagset*
  - 87 *tags*
  - *Used for Brown Corpus* (1-million-word,1963-1964, Brown University)
  - TAGGIT program
- *Penn treebank tagset*
  - 45 *tags*
  - *Used for Penn treebank, Brown Corpus, WSJ Corpus*
  - *Brill tagger*
- *UCREL's C5 tagset*
  - 61 *tags*
  - *Used for British National Corpus (BNC)*
  - *Lancaster CLAWS tagger*

英语词类标记集

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 treebank POS tagset

(45 *tags*)

汉语词类标记集

- 北京大学《人民日报》语料库词类标记集
  - 规范2001
    - 约40+个词类标记
    - 用于标注《人民日报》语料库
  - 规范2003
    - 扩充至106个词类标记
- 国家语委语用所词类标记集
  - ??个词类标记
  - 参见语委用用所《信息处理用现代汉语词类及词性标记集规范》
- 其它词类标记集

汉语词类标记集

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

北大《人民日报》标注语料库词类标记集(40+ tags)

为了处理真实语料，汉语词类标记集中通常包含一些非功能分类的标记，例如：成语、习用语、简称略语；也包含一些语素、前接成份、后接成份等比词小的标记。

一次 深入 的 **考察** (vn)
予以 严肃 **处理** (vn)

研究 思路 (vn)

他 **讽刺** 说 (vd)
主任 **强调** 指出 (vd)

维护 环境 的 **整洁** (an)
交通 **安全** (an)

认真 学习 (ad)
深入 研究 (ad)

兼类问题

- 如果同一个词具有不同词类的语法功能，则认为这个词兼属不同的词类，简称兼类。
- 例一

  (4a) 买了一束**花**　　(4b) **花**了很多时间
  (5a) 开了一个**会**　　(5b) **会**拉小提琴
  (6a) 桌子上有两封**信**　　(6b) 别**信**他的话
  (7a) 选举他当**代表**　　(7b) 他**代表**我们发言
  在(a)组中是名词，在(b)组中是动词。

- 例二

  (1a) **共同**完成一些任务　　(1b) 我们的**共同**愿望
  (2a) **自动**控制这个开关　　(2b) 方便的**自动**步枪
  (3a) **定期**检查机器　　(3b) 一笔**定期**存款
  在(a)组中，是副词、在(b)组中是区别词。

兼类问题

- English data, from Brown corpus:
  - 11.5 percent of the lexicon is ambiguous as to part-of-speech (types)
  - 40 percent of the words in the Brown corpus are ambiguous (tokens)
- Degree of ambiguity (No. tags per word)
  - 1 tags 35340
  - 2-7 tags 4100 total: 39440
  - 2 tags 3760
  - 3 tags 264
  - 4 tags 61
  - 5 tags 12
  - 6 tags 2
  - 7 tags 1

兼类问题

- 《现代汉语语法信息词典》数据(1997年版)
  - 总词数 55191
  - 2-5 tags 1624 2.94%
  - 2 tags 1475 2.67%
  - 3 tags 126 0.23%
  - 4 tags 20 0.04%
  - 5 tags 3 0.01%
- 例：
  - 和 c-n-p-q-v
  - 光 a-d-n-v

词类自动标注

- 对于兼类词，词类标注程序应根据上下文确定兼类词在句子中最合适的词类标记。(难点所在)

  例如：

  - book VB or NN
    book that flight.
    book/VB that/DT flight/NN ./.

  - 报告 v or n
    这 份 特区 政府 的 报告 长 达 ２ ０ 页 。
    这/r 份/q 特区/n 政府/n 的/u 报告/n 长/a 达/v ２ ０/m 页/q 。/w

词类自动标注

- 词类自动标注的方法
  - 基于规则的词类标注 (早期)
  - 基于统计的词类标注
    - 基于隐马尔可夫模型
    - 基于条件随机场模型
    - 基于深度学习模型

- 词类自动标注 和 隐马尔可夫模型
  - 规则方法转向统计方法的起点

主要内容

- 词类标注及词类划分的语言学背景
  - 词类的划分标准
  - 常见词类标记集
  - 兼类词和词类排歧
- **隐马尔可夫模型**
  - 向前算法
  - 韦特比算法
  - Baum-Welch算法
- 隐马尔可夫模型和词类标注

隐马尔科夫模型

- 隐马尔科夫模型(Hidden Markov Model, HMM)
- 马尔科夫模型的一种扩充
- 在自然语言处理领域应用广泛
  - 词类自动标注
  - ……

马尔科夫模型

- 设S是状态集

  $$S = \{1,2,\cdots,n\}$$

  X 在 t 时刻所处的状态为q_t, 其中 $q_t \in S$，若有：

  $$P(q_t|q_{t-1},q_{t-2},\dots) = P(q_t|q_{t-1})$$

  则随机序列 X 构成一阶马尔科夫链。

- 若有$P(q_t = j|q_{t-1} = i) = P(q_s = j|q_{s-1} = i)$

  则随机序列称为时间齐次马尔可夫链。

- 令 $a_{ij} = P(q_t = j|q_{t-1} = i), 1 \le i,j \le n$，则有：

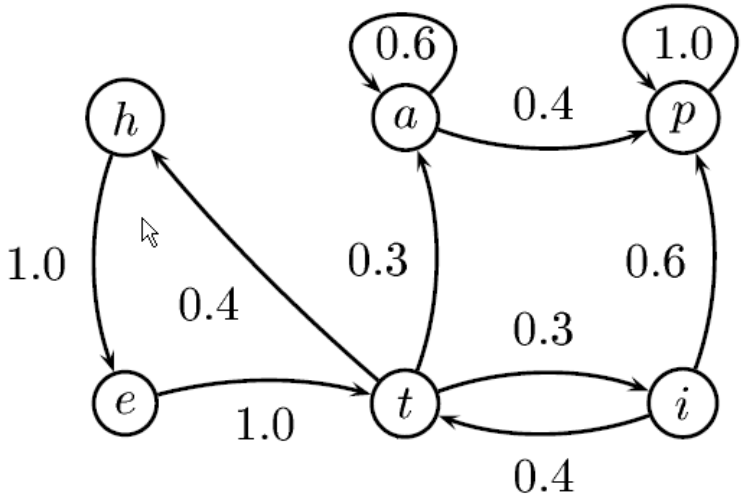
  $a_{ij} \ge 0$

  $\sum_{j=1}^{n} a_{ij} = 1$

马尔科夫模型

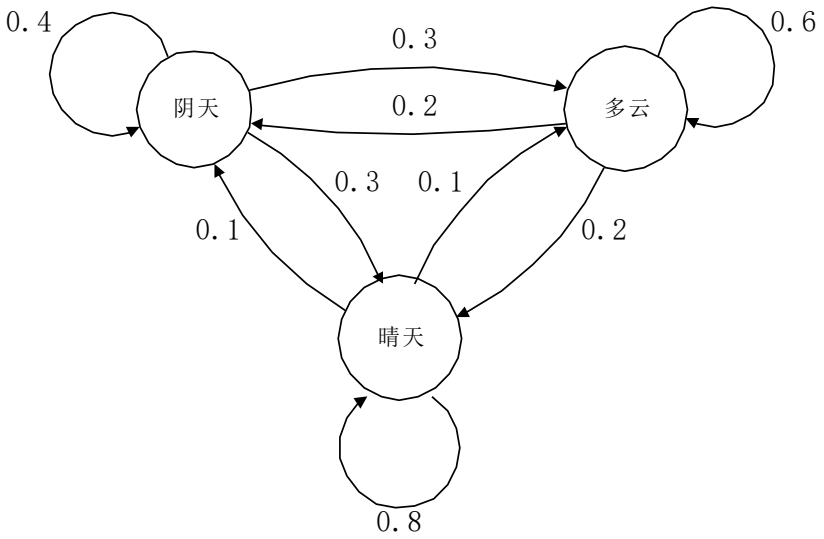
- 一阶马尔科夫模型是一个三元组(S, A, π)，S是状态集，A是状态转移概率矩阵，其元素a_{ij}代表从状态i转移到状态j的概率 , π是初始状态概率, 其元素π_i代表初始时刻处在状态i的概率。

- 状态转移关系也可用状态转换图来表示



马尔科夫模型举例

- 天气的变化，三种状态{1(阴天), 2(多云), 3(晴天)}
- 今天的天气情况仅和昨天的天气状况有关。
- 根据对历史数据的观察得到下列状态转移关系。



$$A = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$

马尔科夫模型

- 状态和输出是一对一的关系
  - 状态1输出阴天
  - 状态2输出多云
  - 状态3输出晴天
- 根据观察到的输出序列可唯一确定状态转换序列
  - 给定天气状况的观察序列。
    (晴 晴 晴 阴 阴 晴 云 晴)
    则可确定状态转换序列为
    (3, 3, 3, 1, 1, 3, 2, 3)

坛子与小球

一个房间中，有 N 个坛子，每个坛子中装有M 种不同颜色的小球。

一个精灵在房间中随机地选择一个坛子，从这个坛子中随机选择一个小球，把小球的颜色报告给房间外面的人员记录下来作为观察值。

精灵然后把球放回到坛子中，以当前坛子为条件再随机选择一个坛子，从中随机选择一个小球，并报告小球的颜色，然后放回小球，如此继续…，随着时间的推移，房间外的人会得到由这个过程产生的一个小球颜色的序列。

坛子与小球

- 令坛子对应状态，令小球颜色对应状态的输出
- 可用一阶马尔科夫过程来描述坛子的选择过程
- 在马尔科夫过程中，每个状态只有一个输出，但在坛子和小球的问题中。可从每个坛子中拿出不同颜色的小球。状态和输出之间不是一一对应关系
- 给定一个观察序列(不同颜色的小球序列)，不能直接确定状态转换序列(坛子的序列)
- 选择坛子的过程(状态转移过程)被隐藏起来了

隐马尔科夫模型

- 隐马尔可夫模型λ可以表示为一个五元组(S, V, A, B, π)
  - S是状态集合
    $S = \{1, 2, 3, \dots, N\}$ (状态n对应坛子n)
  - V 是输出符号集合
    $V = \{v_1, v_2, \dots, v_M\}$ (v_1对应红色小球)
  - A 是状态转移矩阵，N 行 N 列。
    $A = [a_{ij}]$
    $a_{ij} = P(q_t = j \,|\, q_{t-1} = i), \qquad 1 \le i, j \le N$

隐马尔科夫模型

- B是输出符号的概率分布。
  $B = \{b_j(k)\}$
  $b_j(k)$ 表示在状态j时输出符号v_k的概率
  $b_j(k) = P(v_k|j), 1 \le k \le M, 1 \le j \le N$
- π 是初始状态概率分布$\pi = \{\pi_i\}$
  $\pi_i = P(q_1 = i)$ 表示时刻1选择某个状态的概率。

- 隐马尔可夫过程是一个双重随机过程，其中一重随机过程不能直接观察到，通过状态转移概率矩阵描述。另一重随机过程输出可以观察到的观察符号，由输出概率矩阵定义。

隐马尔科夫模型是生成模型

- 可以把隐马尔可夫模型看做一个符号序列的生成装置，按照一定的步骤，隐马尔可夫模型可以生成下面的符号序列：

$$O = (o_1\ o_2\ o_3\ \ldots\ o_T)$$

1. 令 $t = 1$，按照初始状态概率分布 π 选择一个初始状态$q_1 = i$。

$$i \sim \pi$$

2. 按照状态i输出符号概率分布 $b_i(k)$选择一个输出值 $o_t = v_k$ 。

$$v_k \sim b_i(k)$$

3. 按照状态转移概率分布a_{ij}选择一个后继状态$q_{t+1} = j$ 。

$$j \sim a_{ij}$$

4. 若 $t < T$，令 $t = t + 1$，并且转移到算法第2步继续执行，否则结束。

抛掷硬币

- 三枚硬币，随机选择一枚，进行抛掷，记录抛掷结果。可以描述为一个三个状态的隐马尔科夫模型λ。

  $\lambda = (S, V, A, B, \pi)$，其中

  $S = \{1, 2, 3\}$

  $V = \{H,T\}$

  A 如下表所示

	1	2	3
1	0.9	0.05	0.05
2	0.45	0.1	0.45
3	0.45	0.45	0.1

  B 如下表所示

	1	2	3
H	0.5	0.75	0.25
T	0.5	0.25	0.75

  $\pi = \{1/3, 1/3, 1/3\}$

抛掷硬币

- 问题一：
  给定模型，观察到下列抛掷结果的概率是多少？
  $O = (H\ H\ H\ H\ T\ H\ T\ T\ T\ T)$
- 问题二：
  给定模型，若观察到上述抛掷结果，最可能的硬币选择序列(状态转换序列)是什么？
- 问题三：
  若上述模型中的状态转移矩阵 A、状态输出概率 B 和初始状态分布 π 均未知，如何根据观察序列学习？

隐马尔科夫模型的三个问题

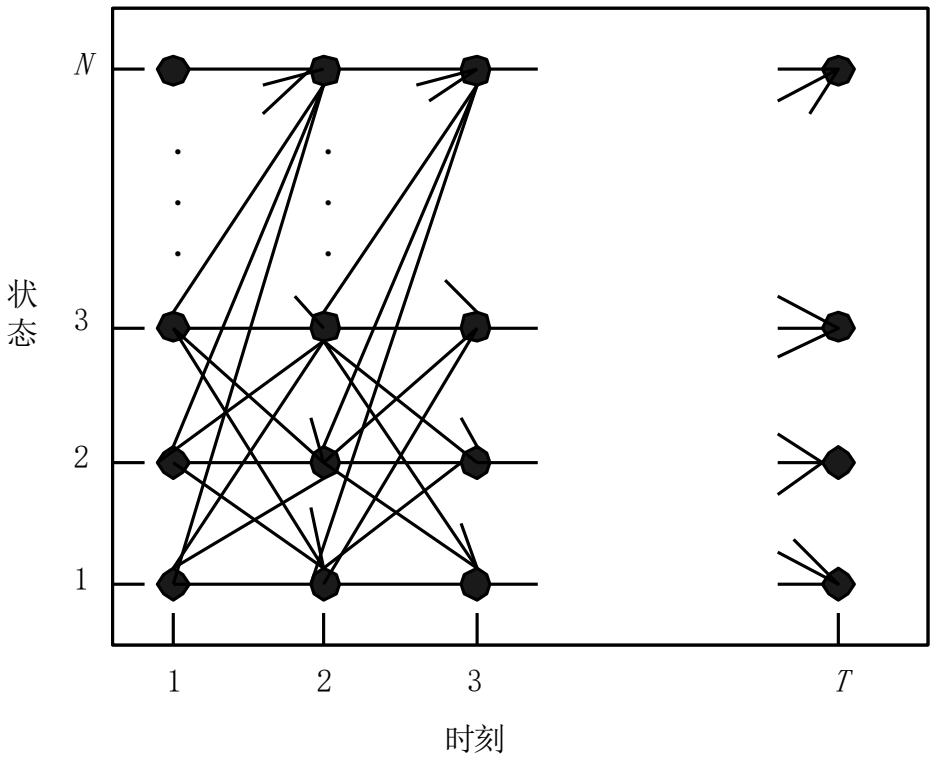
- 给定HMM $\lambda = (A, B, \pi)$和 观察序列$O = (o_1 o_2 \dots o_T)$
  怎样计算观察序列的概率$P(O|\lambda)$？

- 给定HMM $\lambda = (A, B, \pi)$和 观察序列$O = (o_1 o_2 \dots o_T)$
  怎样找到最有可能生成观察序列的状态转换序列$q = (q_1 q_2 \dots q_T)$？

- 在模型参数未知或不准确的情况下，怎样根据观察序列$O = (o_1 o_2 \dots o_T)$求得模型参数或调整模型参数？
  按照最大似然估计原则，如何确定一组模型参数，使得$P(O|\lambda)$最大

问题1: 估算观察序列概率

- 观察序列可由任何状态转换序列产生。
- 要计算一个观察序列的概率值，就必须考虑所有可能的状态转换序列



- 生成观察序列$O = (o_1 o_2 \ldots o_T)$的所有可能的状态转换序列

- 硬币抛掷，三枚硬币(1、2、3)，四次抛掷得到HHTT
- 共有多少种不同的抛掷方法(状态转移路径)？

$$3 \times 3 \times 3 \times 3 = 81$$

- 若转移路径是 1-2-2-3，抛掷得到HHTT的概率是多少？
  - 选择转移路径1-2-2-3的概率是多少？
  - 选择转移路径1-2-2-3，且抛出结果HHTT的概率是多少？
- 四次抛掷得到HHTT的概率是多少？

估算观察序列概率

- 给定λ，计算$P(O)$

$$P(O) = \sum_{q} P(O, q)$$

- 如何计算$P(O, q)$

$$P(O, q) = P(O|q)P(q)$$

$$P(q) = \pi_{q_1} a_{q_1 q_2} a_{q_2 q_3} \dots a_{q_{T-1} q_T}$$

$$P(O|q) = b_{q_1}(o_1) b_{q_2}(o_2) \dots b_{q_T}(o_T)$$

则O和q 的联合概率为：

$$P(O, q) = \pi_{q_1} b_{q_1}(o_1) a_{q_1 q_2} b_{q_2}(o_2) \dots a_{q_{T-1} q_T} b\_q_T(o_T)$$

- 观察序列概率$P(O)$

$$P(O) = \sum_{q} P(O, q) = \sum_{q_1 \dots q_T} \pi_{q_1} b_{q_1}(o_1) a_{q_1 q_2} b_{q_2}(o_2) \dots a_{q_{T-1} q_T} b\_q_T(o_T)$$

估算观察序列概率

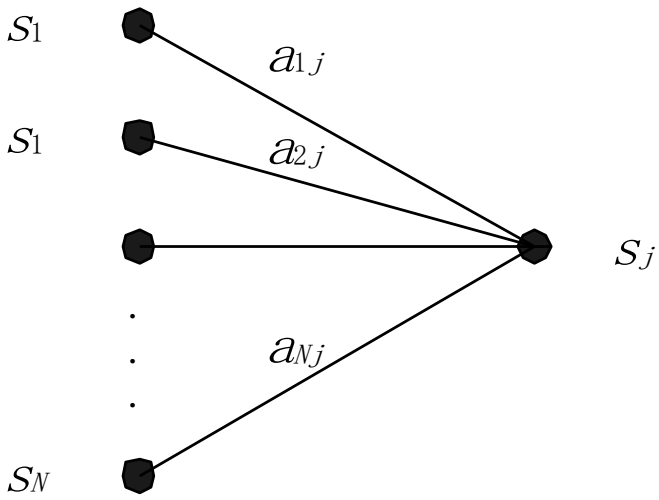
- 可以通过穷举所有状态转换序列的办法计算观察序列O的概率。
- 实际上，这样做并不现实。
  - 可能的状态转换序列共有N^T个。
  - 需要做$(2T-1)N^T$次乘法运算，N^T-1 次加法运算。
  - 若$N=5$，$T=100$，则
    $(2\times100-1)\times5^{100}\approx10^{72}$
- 需要寻找更为有效的计算方法。

向前算法(Forward Algorithm)

- **向前变量**$\alpha_t(i)$
  $\alpha_t(i) = P(o_1 o_2 \ldots o_t, * \; q_t = i)$
- $\alpha_t(i)$含义:给定模型λ，时刻t，处在状态i，部分观察序列为$o_1 o_2 \ldots o_t$的概率
- $\alpha_1(i) = \pi_i b_i(o_1), 1 \le i \le N$
- 已知$\alpha_t(i)$ (1≤i≤N)，计算$\alpha_{t+1}(j)$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(o_{t+1})$$

$$1 \le t \le T-1, 1 \le j \le N$$



向前算法

1. 初始化

$$\alpha_1(i) = \pi_i b_i(o_1), 1 \le i \le N$$

2. 迭代计算

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) a_{ij}\right] b_j(o_{t+1}), 1 \le t \le T-1, 1 \le j \le N$$

3. 终止

$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

- 计算量
  - $N(N+1)(T-1)+N$次乘法
  - $N(N-1)(T-1)+(N-1)$次加法
  - 若$N = 5$，T=100，则大约需要5000次运算

计算实例

- 抛掷硬币问题，计算观察到($H\ H\ T$)的概率。

$\alpha_t(i)$	H	H	T	$P(H\ H\ T \mid \lambda)$
1	0.16667	0.15000	0.08672	
2	0.25000	0.05312	0.00684	0.11953
3	0.08333	0.03229	0.02597	

向后算法(Backward Algorithm)

- **向后变量** $\beta_t(i)$

$$\beta_t(i) = P(o_{t+1}o_{t+2} \dots o_T, q_t = i * | q_t = i)$$

- $\beta_t(i)$的含义: 给定模型λ，时刻t，从状态i出发，并且部分观察序列为$o_{t+1}o_{t+2} \dots o_T$的概率。

- $\beta_T(i) = 1 \ (1 \le i \le N)$

- 已知$\beta_{t+1}(j) \ (1 \le j \le N)$，计算$\beta_t(i)$

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)$$

$$1 \le t \le T-1, 1 \le j \le N$$

向后算法

1. 初始化

$$\beta_T(i) = 1\ (1 \le i \le N)$$

2. 迭代计算

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

3. 终止

$$P(O|\lambda) = \sum_{i=1}^{N} \pi_i b_i(o_1) \beta_1(i)$$

计算实例

- 抛掷硬币问题，计算观察到$(H\ H\ T)$的概率

	H	H	T		$P(H\ H\ T\|\lambda)$
$\beta_t(i)$	$\pi_i b_i(H)\beta_1(i)$	$\beta_1(i)$	$\beta_2(i)$	$\beta_3(i)$	
1	0.04203	0.25219	0.50000	1.00000	
2	0.05074	0.20297	0.58750	1.00000	0.11953
3	0.02676	0.32109	0.41250	1.00000	

问题2:求解最佳状态转换序列

- 隐马尔可夫模型的第二个问题:
  计算能最好解释观察序列的状态转换序列。
- 理论上，可以通过枚举所有的状态转换序列，并对每一个状态转换序列q计算$P(O,q|\lambda)$，能使$P(O,q|\lambda)$取最大值的状态转换序列q^*就是能最好解释观察序列的状态转换序列，即：

$$q^* = \underset{q}{\operatorname{argmax}} P(O,q|\lambda)$$

- 需要更有效率的计算方法

韦特比算法(Viterbi Algorithm)

- **韦特比变量** $\delta_t(i)$

$$\delta_t(i) = \max_{q_1 \cdots q_{t-1}} P(q_1 q_2 \ldots q_{t-1} q_t = i, o_1 o_2 \ldots o_t | \lambda)$$

- $\delta_t(i)$的含义: 时刻t处于状态i，观察到$o_1\ o_2\ o_3\ \ldots\ o_t$的最佳状态转换序列是$q_1 q_2 \ldots q_t$的概率。
- $\delta_1(i) = \pi_i b_i(o_1)$，$1 \le i \le N$
- 若$\delta_t(i)(1 \le i \le N)$已知，如何计算$\delta_{t+1}(i)$?

$$\delta_{t+1}(j) = \left[\max_i \delta_t(i) a_{ij}\right] b_j(o_{t+1})$$

- 如何记录路径? 设定T个数组$\psi_1(N),\ \psi_2(N),\ \ldots \psi_T(N)$

$\psi_t(i)$ 记录在时刻t到达状态i的最佳状态转换序列t-1时刻的状态。

韦特比算法

1. 初始化

$\delta_1(i) = \pi_i b_i(o_1), 1 \le i \le N$

$\psi_1(i) = 0$

2. 迭代计算

$\delta_t(j) = \max_i \left[\delta_{t-1}(i) a_{ij}\right] b_j(o_t)$

$\psi_t(j) = \operatorname{argmax}_i \delta_{t-1}(i) a_{ij}$

3. 终止

$P^* = \max_i \delta_T(i)$

$q_T^* = \operatorname{argmax}_i \delta_T(i)$

4. 求解最佳路径

$q_t^* = \psi_{t+1}(q_{t+1}^*), t = T-1, T-2, \dots, 1$

计算实例

- 抛掷硬币问题，观察到($H\ H\ T$)，寻找产生该观察序列的最佳路径以及最佳路径的概率。

$\delta_t(i)$	H	H	T	P^*
1	0.16667	0.07500	0.03375	
2	0.25000	0.02812	0.00316	0.03375
3	0.08333	0.02812	0.00949	

$\psi_t(i)$	$\psi_1(i)$	$\psi_2(i)$	$\psi_3(i)$	q^*
1	0	1	1	
2	0	3	3	1
3	0	2	2	

- 最佳状态转换序列为1 1 1

问题3:参数学习

- 根据观察序列$O = (o_1 o_2 \ldots o_T)$求得模型参数或调整模型参数
- 原则：最大似然估计
  如何确定一组模型参数使得$P(O|\lambda)$最大？
- 隐马尔科夫模型的前两个问题均假设模型参数已知，第三个问题是模型参数未知，求最佳模型的问题。
  三个问题中最为困难的问题。

有指导的参数学习(supervised learning)

- 在模型(λ)未知的情况下，如果给定观察序列的同时，也给定了状态转换序列，可通过有指导方法学习模型参数

  H/1 *H*/1 *T*/1 *T*/2 *H*/3 *T*/3
  T/2 *H*/1 *T*/2 *H*/3 *H*/3 *H*/1

- 优点：参数学习简单，效果好
- 缺点：状态信息未知时无法使用，或需要人工标注状态信息，代价高
- 在NLP中，无指导学习效果不佳时，采用有指导学习

无指导的参数学习(unsupervised learning)

- 在模型(λ)未知的情况下，如果仅给定观察序列，此时学习模型的方法被称做无指导的学习方法。
- 对于隐马尔科夫模型而言，采用无指导学习方法，没有解析方法
- 首先给定一组不准确的参数，再通过反复迭代逐步求精的方式调整模型参数，最终使参数稳定在一个可以接受的精度。
- 利用无指导学习方法估计隐马尔科夫模型参数，不能保证求得最优模型，能保证得到局部最优模型。

直观的想法

- 给定一组初始参数($A\ B\ \pi$)
- 由于没有给定状态转换序列，无法计算状态转移频率、状态输出频率以及初始状态频率。
- 假定任何一种状态转换序列都可能
- 对每种状态转换序列中的频次加权处理，计算状态转移、状态输出、以及初始状态的期望(频数)
- 利用计算出的期望(频数)更新A、B和π

直观的想法

- 权值如何选择？对状态转换序列q而言，选择$P(q|O)$

  例：三枚硬币(1、2、3)，抛掷四次抛掷得到HHTT

$$\left.\begin{array}{ll} 1111 & p(1111|HHTT) * c_{q1}(1,1) \\ 1112 & p(1112|HHTT) * c_{q2}(1,1) \\ 1113 & p(1113|HHTT) * c_{q_3}(1,1) \\ 1121 & p(1121|HHTT) * c_{q4}(1,1) \\ \ldots & \ldots \end{array}\right\} \oplus\rightarrow \hat{c}(1,1)$$

同理，计算

$$\hat{c}(1,*)、\hat{c}(1,H) \ldots$$

直观的想法

- $\hat{c}(i,j)$代表$i \to j$在状态转移路径上出现次数的期望。

$$\hat{c}(i,j) = \sum_{q} P(q|O) \cdot c(i,j,q) = \mathbb{E}_{P(q|O)}[c(i,j,q)]$$

- 利用期望频次代替频次进行计算
- 理论上可行，现实不可行
  要考虑所有的状态转移路径
  需要多次迭代，问题更为严重

- 需要更为有效的算法，即Baum-Welch算法

直观的想法

- 对于时刻t和时刻$t+1$，出现转移(i,j)的期望如何计算？

- 在所有的路径中，选择出$q_t = i$且$q_{t+1} = j$的路径，设满足这样条件的路径集合是：

$$Q = \{q | q_t = i, q_{t+1} = j\}$$

- 则对于时刻t和时刻$t+1$，出现转移(i,j)的期望为：

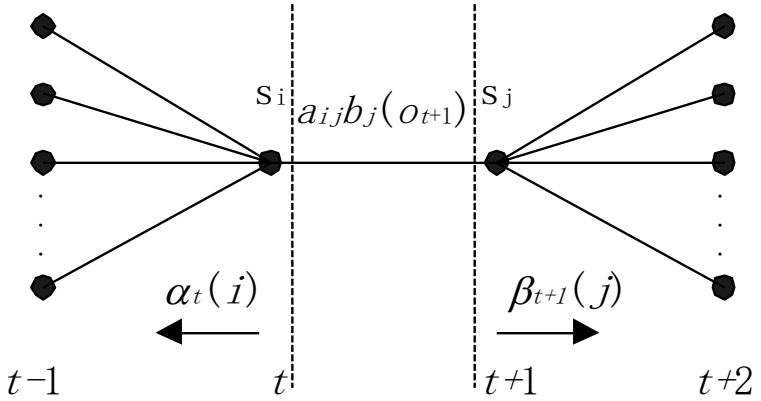
$$\xi_t(i,j) = \sum_{q \in Q} P(q|O)$$

$$= \sum_{q \in Q} \frac{P(q,O)}{P(O)} = \frac{\sum_{q \in Q} P(q,O)}{P(O)}$$

Baum-Welch Algorithm

- 定义变量$\xi_t(i,j)$
  $\xi_t(i,j) = P(q_t = i, q_{t+1} = j | O, \lambda)$

- $\xi_t(i,j)$含义: 给定模型λ和观察序列O，在时刻t处在状态i，时刻t+1处在状态j的期望



$$\xi_t(i,j) = \frac{P(q_t = i, q_{t+1} = j, O|\lambda)}{P(O|\lambda)} = \frac{\alpha_t(i)a_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{P(O|\lambda)}$$

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

Baum-Welch Algorithm

- 定义变量$\gamma_t(i)$，表示在给定模型以及观察序列的情况下，t时刻从状态i出发的转换的期望

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j)$$

即：$\hat{c}_t(i,*)$

- 考虑所有时刻，从状态i出发的转换的期望

$$\sum_{t=1}^{T-1} \gamma_t(i)$$

即：$\hat{c}(i,*)$

- 考虑所有时刻，从状态i到状态j的转换的期望

$$\sum_{t=1}^{T-1} \xi_t(i,j)$$

即：$\hat{c}(i,j)$

Baum-Welch Algorithm

- π, A, B可估计如下

$\bar{\pi}_i = \gamma_1(i)$ $t = 1$时处在状态i的期望

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

从状态i到状态j的转换的期望除以从状态i出发的转换的期望

$$\bar{b}_j(k) = \frac{\sum_{t=1}^{T} \gamma_t(j) \times \delta(o_t, v_k)}{\sum_{t=1}^{T} \gamma_t(j)}$$

当$o_t = v_k$时，$\delta(o_t, v_k) = 1$

当$o_t \neq v_k$时，$\delta(o_t, v_k) = 0$

在状态j观察到v_k的期望

处在状态j的期望

Baum-Welch Algorithm

- 利用上述结论，即可进行模型估算
- 选择模型参数初始值，初始值应满足条件：

$$\sum_{i=1}^{N}\pi_i = 1, \sum_{j=1}^{N} a_{ij} = 1, 1 \le i \le N, \sum_{k=1}^{M} b_j(k) = 1, 1 \le j \le N$$

- 将初始值代入前面的公式中，计算一组新的参数$\bar{\pi}, \bar{A}, \bar{B}$
- 将新的参数代入，再次计算更新的参数。
- 如此反复，直到参数收敛。

Baum-Welch Algorithm

- Baum-Welch算法是一种EM算法。
- E-step:
  - 计算$\xi_t(i,j)$和$\gamma_t(i)$
- M-step:
  - 估计模型$\bar{\lambda}$
- 终止条件

$$\left|\log P\left(O\middle|\bar{\lambda}\right) - \log P(O|\lambda)\right| < \epsilon$$

Baum-Welch Algorithm

- Baum等人证明要么估算值$\bar{\lambda}$和估算前的参数值λ相等，要么估算值$\bar{\lambda}$比估算前的参数值λ更好的解释了观察序列O。
- 参数最终的收敛点并不一定是一个全局最优值，但一定是一个局部最优值。

L.R.Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech recognition, Proc. IEEE, 77(2): 257-286, 1989

隐马尔科夫模型的实现

- 浮点溢出问题
  - 对于韦特比算法，采用取对数的方式
  - 对于Baum-Welch算法，采用放大因子
  - 对于向前算法采用放大因子以及取对数的方式。

$$c_t = \frac{1}{\sum_i \tilde{\alpha}_t(i)}$$

$$\tilde{\alpha}_t(i) = \left(\prod_{\tau=1}^{t} c_t \right) \alpha_t(i)$$

主要内容

- 词类标注及词类划分的语言学背景
  - 词类的划分标准
  - 常见词类标记集
  - 兼类词和词类排歧
- **隐马尔可夫模型**
  - 向前算法
  - 韦特比算法
  - Baum-Welch算法
- **隐马尔可夫模型和词类标注**

基于隐马尔科夫模型的词类标注

- HMM状态集　　词类标记集
- HMM输出符号集　　词表
- 如何根据观察到的词串(句子)，求解最可能的词类标记序列(状态转换序列)。　　Viterbi算法
- 模型参数
  - $p(t_i|t_{i-1})$　　词类转移概率
  - $p(w_i|t_i)$　　词类t_i生成词w_i的概率
  - $p(t) = p(t|<bos>)$　　词类t出现在句首的概率

$$\hat{t}_1^n = \underset{t_1^n}{\operatorname{argmax}} P(t_1^n|w_1^n) = \underset{t_1^n}{\operatorname{argmax}} \prod_{i=1}^{n} p(w_i|t_i)p(t_i|t_{i-1})$$

基于隐马尔科夫模型的词类标注

- 参数学习
  - 可采用有指导的学习方法
  - 需要预先准备带词类标记的语料库
    - 例如，1998年1月《人民日报》标注语料库
  - 也可以采用无指导学习，例如用Baum-Welch算法
- 最大似然估计

$$p(t_i|t_{i-1}) = \frac{c(t_{i-1}, t_i)}{c(t_{i-1})}$$

$$p(w_i|t_i) = \frac{c(w_i, t_i)}{c(t_i)}$$

汉语词类标注实例

- 1998年1月《人民日报》标注语料
- 作为**动词**的“报告”(30次)

  1...５３岁的福塞特向总部**报告**说，负责热气球...

  2...将刘青山、张子善的严重犯罪事实**报告**党中央，...

  3...有关矿产资源情况，要每周向中央主要领导**报告**。

- 作为**名词**的“报告”(200次)

  1...在党的十五大**报告**中，江主席再次郑重地...

  2...**报告**认为，虽然日本政府为减少限制性贸易...

  3...国际金融协会发表资金流动**报告**...

？ 发生交通事故时，当事人应当迅速**报告**公安机关，听候处理...

汉语词类标注实例

$c(t_i,t_{i+1})$	…	a	ad	an	n	v	vn	…	Σ
…	…		…	…	…	…	…	…	…
a		800	8	127	10923	942	2267		34473
ad	…	76	34	0	3	5533	2	…	5933
an	…	10	5	47	238	257	218	…	2837
n	…	4047	1273	440	42491	32933	12508	…	312263
v	…	6924	855	735	42671	27142	4735	…	229776
vn	…	284	113	54	16021	2677	3165	…	42734
…	…		…						

$c(w_i,t_i)$	…	a	ad	an	n	v	vn	…	Σ
…	…	…	…	…	…	…	…	…	…
当事人	…	0	0	0	25	0	0	…	25
应当	…	0	0	0	0	340	0	…	340
迅速	…	50	116	1	0	0	0	…	167
报告	…	0	0	0	200	30	4	…	234
公安	…	0	0	0	188	0	0	…	188
机关	…	0	0	0	354	0	0	…	354
…	…	…	…		…	…	..	…	…

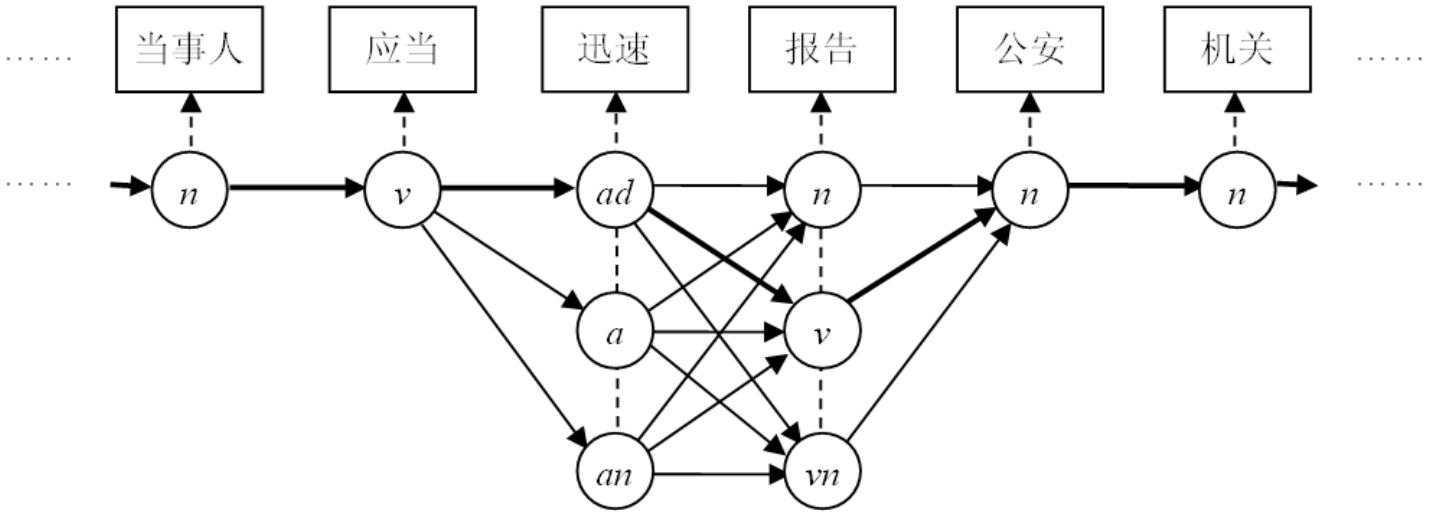
汉语词类标注实例

$p(t_{i+1}\|t_i)$	…	a	ad	an	n	v	vn	…	∑
…	…		…	…	…	…	…	…	…
a		0.0232065676	0.0002320657	0.0036840426	0.3168566704	0.0273257332	0.0657616105		
ad	…	0.0128097084	0.0057306590	0	0.0005056464	0.9325804820	0.0003370976	…	1
an	…	0.0035248502	0.0017624251	0.0165667959	0.0838914346	0.0905886500	0.0768417342	…	1
n	…	0.0129602290	0.0040766918	0.0014090686	0.1360743988	0.1054655851	0.0400559785	…	1
v	…	0.0301336954	0.0037210152	0.0031987675	0.1857069494	0.1181237379	0.0206070260	…	1
vn	…	0.0066457622	0.0026442645	0.0012636308	0.3749005476	0.0626433285	0.0740628071	…	1
…	…		…		…	…	…	…	…

汉语词类标注实例

$p(w_i\|t_i)$	…	a	ad	an	n	v	vn	…
…	…	…	…	…	…	…	…	…
当事人	…	0	0	0	0.0000800607	0	0	…
应当	…	0	0	0	0	0.0014797019	0	…
迅速	…	0.0014504105	0.0195516602	0.0003524850	0	0	0	…
报告	…	0	0	0	0.0006404857	0.0001305619	0.0000936023	…
公安	…	0	0	0	0.0006020566	0	0	…
机关	…	0	0	0	0.0011336598	0	0	…
…	…	…	…	…	…	…	..	…
Σ	…	1	1	1	1	1	1	…

汉语词类标注实例



P(... n v ad n n n ..., ... 当事人 应当 迅速 报告 公安 机关 ...)

=... × p(当事人|n) × p(v|n) × p(应当|v) × p(ad|v) × p(迅速|ad) × p(n|ad)

× p(报告|n) × p(n|n) × p(公安|n) × p(n|n) × p(机关|n) × ...

汉语词类标注实例

T	$\prod p(w_i\|t_i)\,p(t_i\|t_{i-1})$	$P(T\|$...应当 迅速 报告 公安...$)$
...v ad n n...	...p(ad|v)×p(迅速|ad)×p(n|ad)×p(报告|n)×p(n|n)...	3.2061059e-12
...v ad v n...	**...p(ad|v)×p(迅速|ad)×p(v|ad)×p(报告|v)×p(n|v)...**	**1.64503834e-9**
...v ad vn n...	...p(ad|v)×p(迅速|ad)×p(vn|ad)×p(报告|vn)×p(n|vn)...	8.6060396e-13
...v a n n...	...p(a|v)×p(迅速|a)×p(n|a)×p(报告|n)×p(n|n)...	1.20695769e-9
...v a v n...	...p(a|v)×p(迅速|a)×p(v|a)×p(报告|v)×p(n|v)...	2.8957414e-11
...v a vn n...	...p(a|v)×p(迅速|a)×p(vn|a)×p(报告|vn)×p(n|vn)...	1.0085986e-10
...v an n n...	...p(an|v)×p(迅速|an)×p(n|an)×p(报告|n)×p(n|n)...	8.2437876e-12
...v an v n...	...p(an|v)×p(迅速|an)×p(v|an)×p(报告|v)×p(n|v)...	2.4765193e-12
...v an vn n...	...p(an|v)×p(迅速|an)×p(vn|an)×p(报告|vn)×p(n|vn)...	3.0403464e-12

... 当事人/n 应当/v 迅速/ad 报告/v 公安/n 机关/n ...

未登录词

- 未登录词
  - 视作兼类词，可能是任何一个词类
  - 依照出现一次的词(hapax legomenon)的规律处理
    - 更可能是名词 不大可能是限定词等
    - 将出现一次的词的分布平均作为未登录词的分布
  - 对于英文等语言可以利用形态特性(词缀)、拼写特性判定(首字母大小写)

- 未登录词的词性标注是难点