

从语言直觉到计算模型

汉语自动分词

-从直觉到计算模型

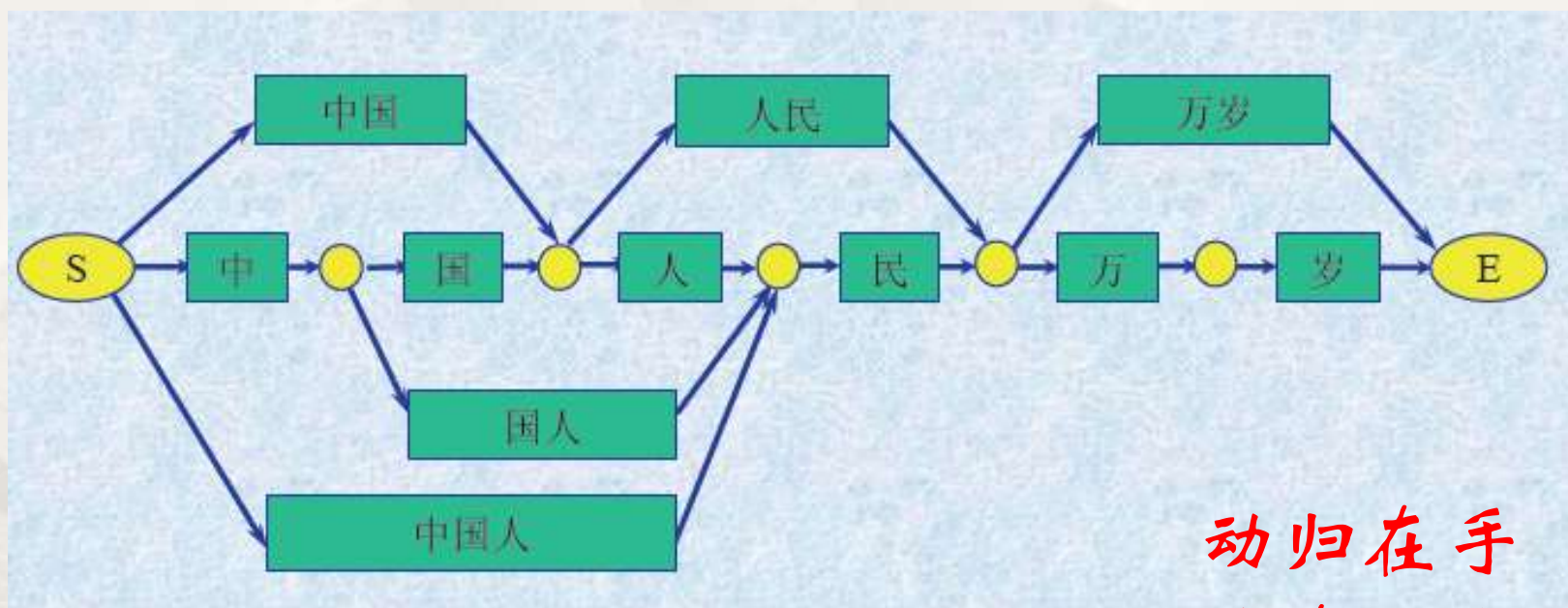
杨沐昀

哈工大教育部-微软语言语音重点实验室

MOE-MS Joint Key Lab of NLP and Speech (HIT)

上一讲回顾

- * 分词问题的空间：全切分有向图



动归在手
路在何方

- * 难点：歧义
- * 问题本质：最优路径的计算

内容提要

- * 基于N元文法的分词（MM）
- * 基于HMM的分词/词性标注一体化(模型)
- * 由字构词的汉语分词方法
- * 汉语分词方法的后处理方法

课下阅读：

- * 未登录词的识别
- * 数据平滑

统计方法**
机器学习
规则方法*

基于N元文法的分词

* N元语法的模型推导

#Text表示输入文本，可以是一句话

$$\hat{Seg} = \arg \max_{Seg} P(Seg | Text)$$

$$= \arg \max_{Seg} \frac{P(Text | Seg) P(Seg)}{P(Text)}$$

$$\propto \arg \max_{Seg} P(Text | Seg) P(Seg)$$

$$= \arg \max_{Seg} P(Seg)$$

对比推导结果：
为何不直接假设？

?

?

基于N元文法的切分排歧

- * Seg简写为S, 含有n个词: $\{w_1, w_2, \dots, w_n\}$

$$p(S) = p(w_1^n) = p(w_1) \cdot \prod_{i=2}^n p(w_i | w_1^{i-1})$$

- * MM(马尔可夫模型/过程): 有限历史假设, 仅依赖前n-1个词
 - * 一种最简化的情况: 一元文法/uni-gram

$$P(S) = p(w_1) \cdot p(w_2) \cdot p(w_3) \dots p(w_n)$$

#连乘的代码实现?

基于N元文法的切分排歧

- * 采用一元语法 (uni-gram)
 - * 等价于最大频率分词
 - * 即把切分路径上每一个词的概率相乘得到该切分路径的概率
 - * 把词概率的负对数理解成路径“代价”，输出结果就是整体代价最“小”分词序列
 - * 正确率可达到92%
 - * 简便易行，效果一般好于基于词表的方法

基于N元文法的切分排歧

❖ 采用二元语法(bi-gram): 性能进一步提高

$$p(S) = p(w_1) \cdot p(w_2 | w_1) \cdot p(w_3 | w_2) \cdots p(w_n | w_{n-1})$$

- * 更大的n: 对下一个词出现的约束性信息更多, 更大的辨别力。
- * 更小的n: 出现的次数更多, 更可靠的统计结果, 更高的可靠性。

❖ 参数空间

词表=20,000

#解决参数爆炸的策略? 得失?

n	n-gram的个数
2 (bigrams)	400,000,000
3 (trigrams)	8,000,000,000,000
4 (4-grams)	1.6×10^{17}

基于N元文法的切分排歧

* 等价类映射：降低语言模型参数空间

- * 绝大多数历史不会出现在训练数据中。
- * 将历史 $\omega_1\omega_2 \dots \omega_{i-1}$ 映射到等价类 $E(\omega_1\omega_2 \dots \omega_{i-1})$ ，其中等价类的数目远小于全部历史的数目。
- * 假设： $p(\omega_i|\omega_1 \dots \omega_{i-1}) = p(\omega_i|E(\omega_1\omega_2 \dots \omega_{i-1}))$ ，则自由参数的数目会大大减少
- * 思考题：等价类的依据：必须符合语言学的分类吗？

* 数据平滑（smoothing）：保持模型的辨别能力

- * 调整最大似然估计结果，更准确的估计未见事件
- * 提高低概率事件，降低高概率事件，概率分布更均匀。
- * 课下专题阅读：统计自然语言处理，宗成庆，5.3节

内容提要

- * 基于N元文法的分词（MM）
- * 基于HMM的分词/词性标注一体化(模型)
- * 由字构词的汉语分词方法
- * 汉语分词方法的后处理方法

课下阅读：

- * 未登录词的识别
- * 数据平滑

基于HMM的分词/词性标注一体化

* 词的句法类别

* 词性集合：

- * 名词、动词、形容词、副词、介词、助动词
- * 开放词类(Open Class)和封闭词类(Closed Class)

- * 可称为：语法类、句法类、POS标记、词类等

* 词的兼类现象

* 例如

- * 扛人 = 动词
- * 一扛衬衫 = 量词

* 词性标注

- * 确定每个词在特定的句子中词性

基于HMM的分词/词性标注一体化

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

基于HMM的分词/词性标注一体化

- POS歧义(在Brown语料库中)

无歧义的词(1 tag): 35,340个

有歧义的词 (2-7 tags): 4,100个

2 tags	3,760
3 tags	264
4 tags	61
5 tags	12
6 tags	2
7 tags	1

基于HMM的分词/词性标注一体化

- 输入：待处理句子S

- 输出：S的

- 词序列 $W = w_1, w_2 \dots w_n$
- 词性序列 $T = t_1, t_2 \dots t_n$

- 提示

- W可以代表S
- 分词结果即观测序列
- 词性序列是状态序列

- 公式推导

$P(T | S)$ --将词性作为求解目标

$$\propto P(S | T) * P(T)$$

$$\Rightarrow P(W | T) * P(T)$$

$P(T) = P(t_1, t_2 \dots t_n)$ --n元文法

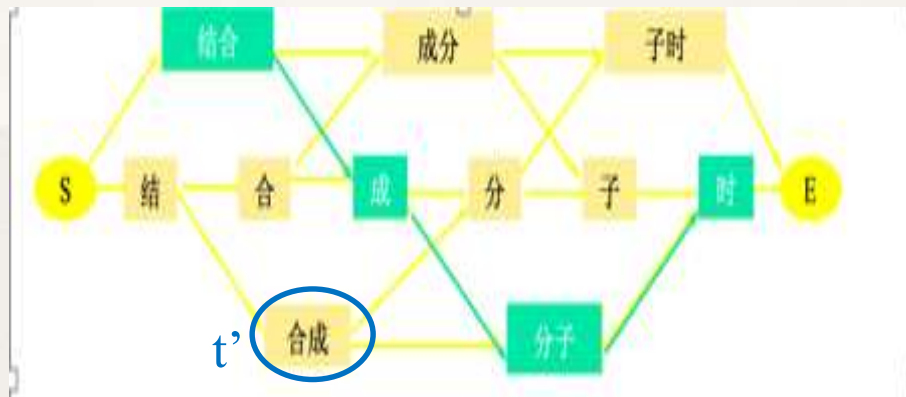
$$\propto P(t_1 | s_{\text{begin}}) * P(t_2 | t_1) * \dots * P(t_n | t_{n-1}) * P(s_{\text{end}} | t_n)$$

$s_{\text{begin}}/s_{\text{end}}$: 句首/句尾

基于HMM的分词/词性标注一体化

- 公式推导（续）

$$P(W|T) = P(w_1, w_2 \dots w_n | t_1, t_2 \dots t_n) \\ = P(w_1 | t_1) * P(w_2 | t_2) * \dots * P(w_n | t_n)$$



注意：每个节点的内部词和词性有关系，而节点间词彼此独立

- 最终公式：

$$T_{Best} = \arg \max_T \prod_{i=1}^n P(w_i | t_i) P(t_i | t_{i-1})$$

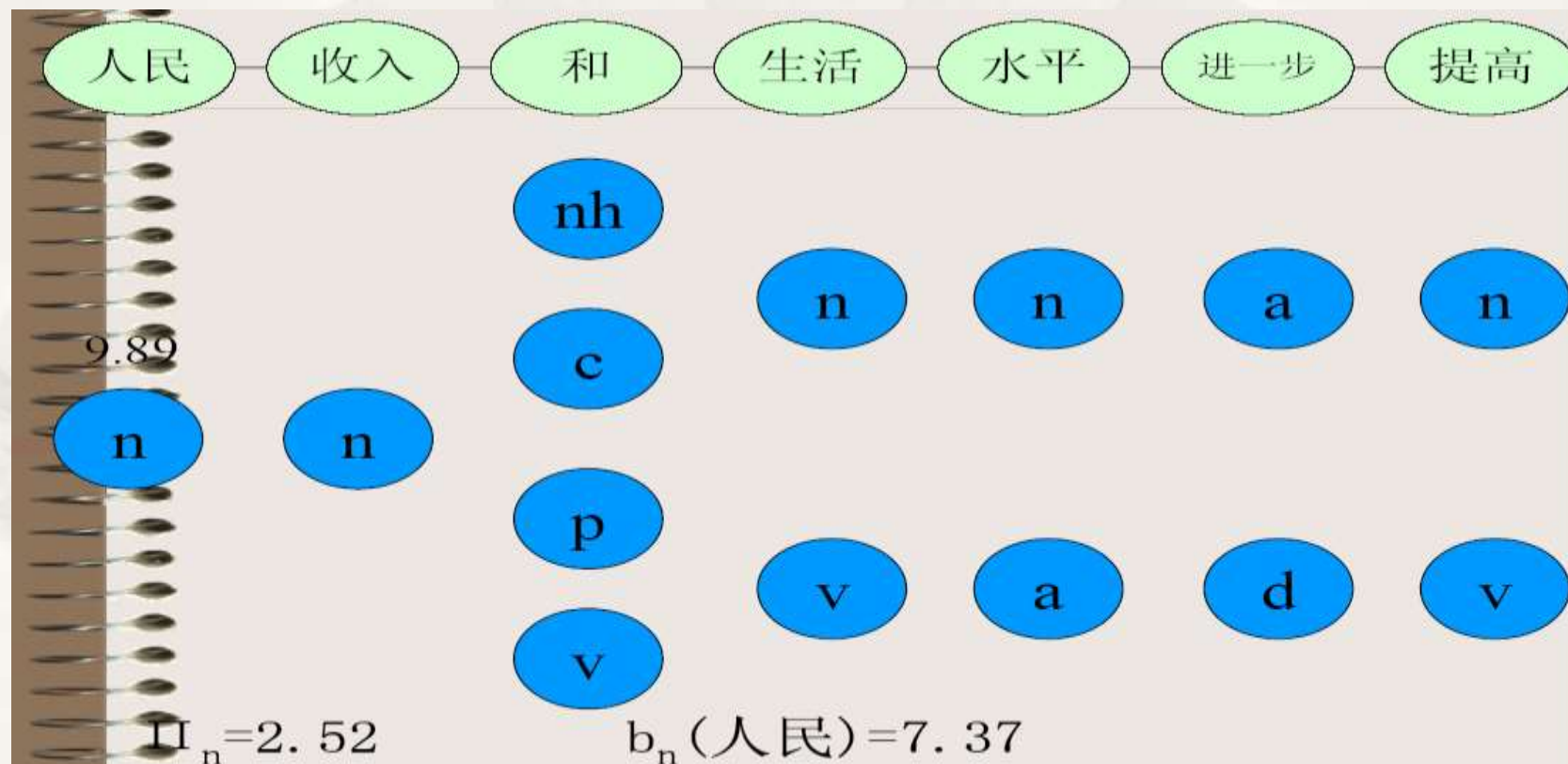
思考题：给定 $P(W, T|S)$ 能够推导出上面的结果？

基于HMM的分词/词性标注一体化

- * 考虑模型的初步应用
 - * 给定大规模标记语料库
 - * 完成HMM的参数估计：MLE
 - * 一个具体的分词+词性标注是如何实现的？

Viterbi搜索——例子

分词“词图”中的某段局部路径（图中代价为概率的负对数）



本例出处待考，谨在此致谢！

人民 收入 和 生活 水平 进一步 提高

nh

n

n

a

n

9.89

20.02

c

n

n

p

v

a

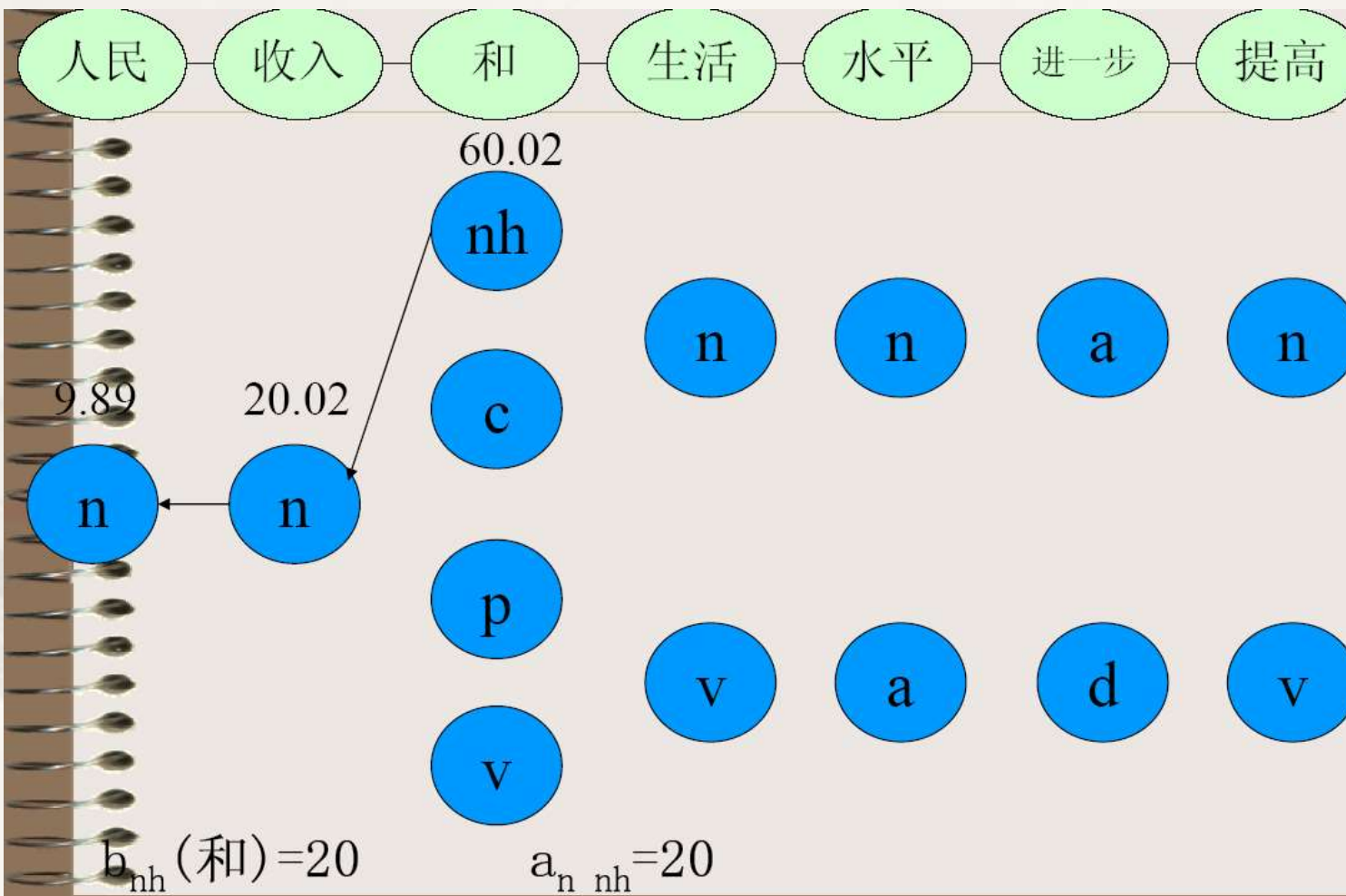
d

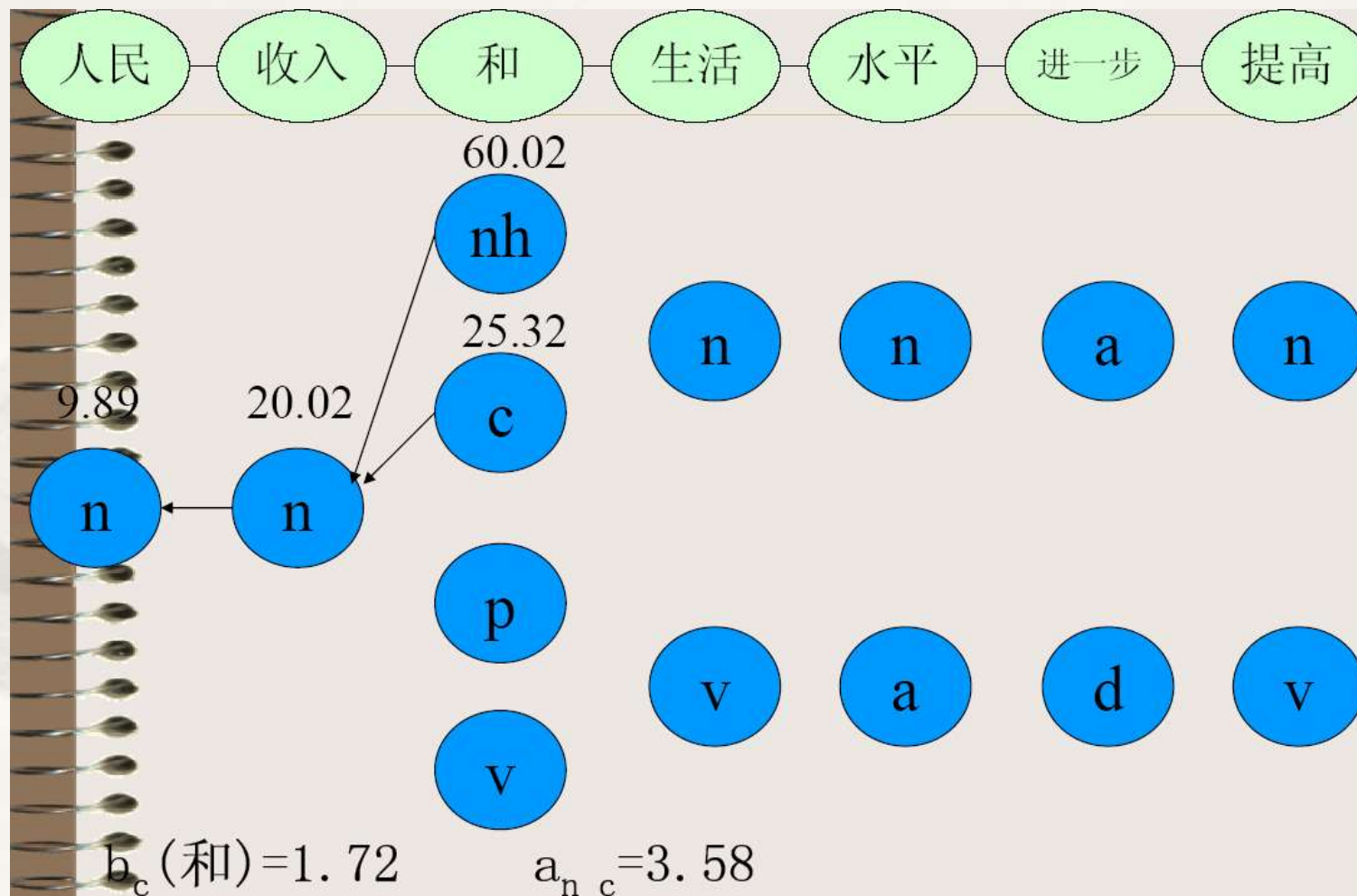
v

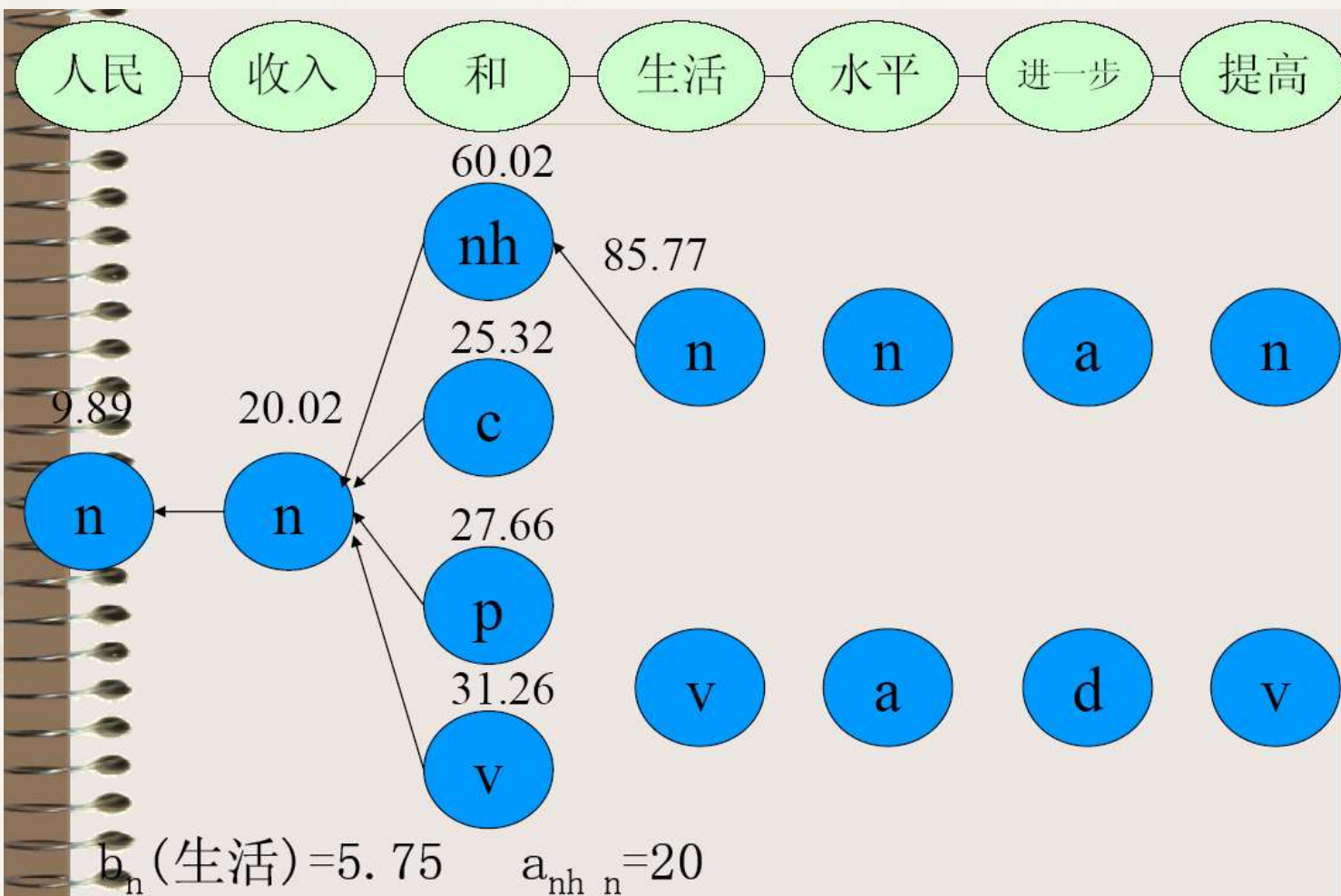
v

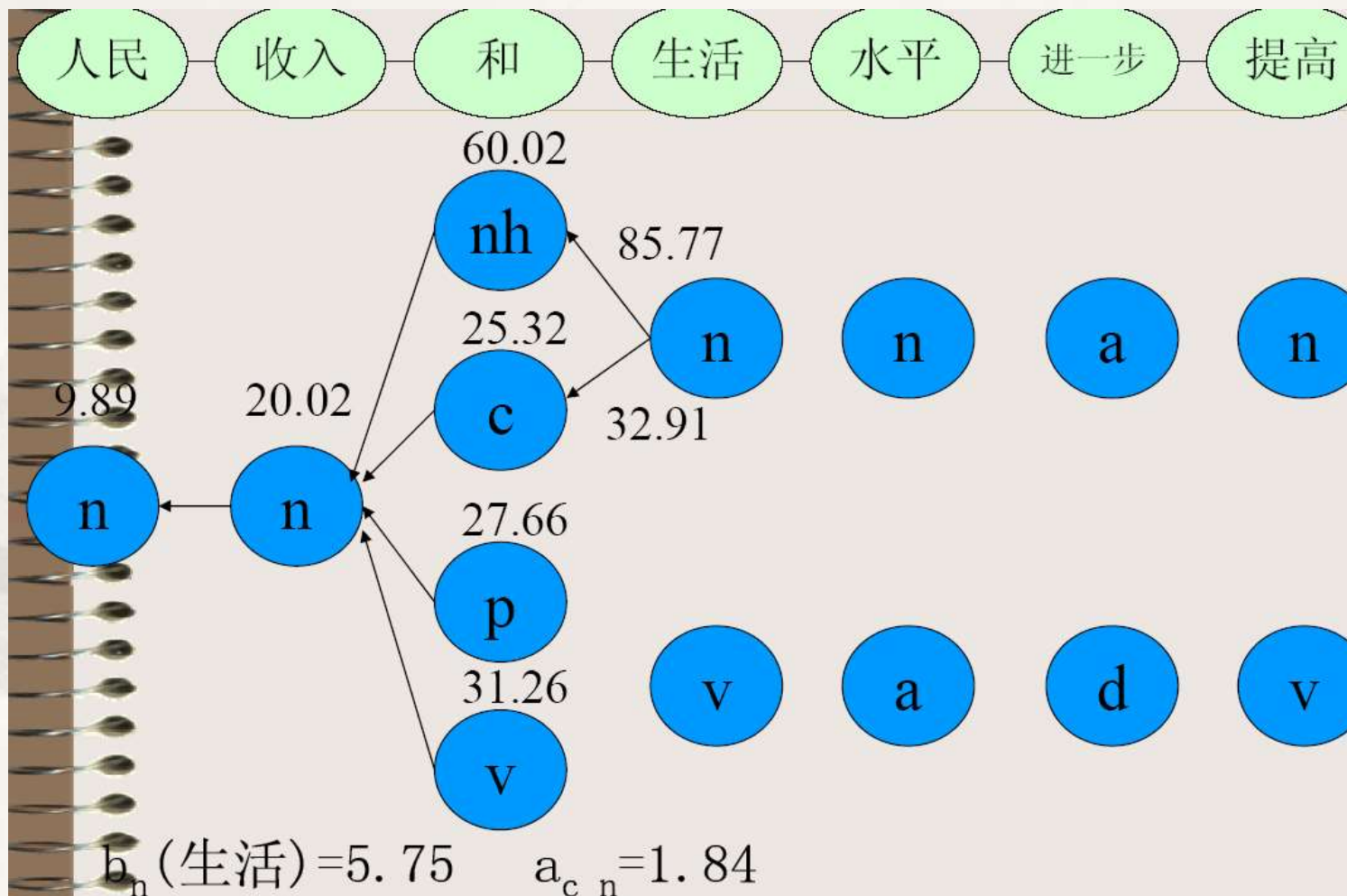
$b_n(\text{收入}) = 6.98$

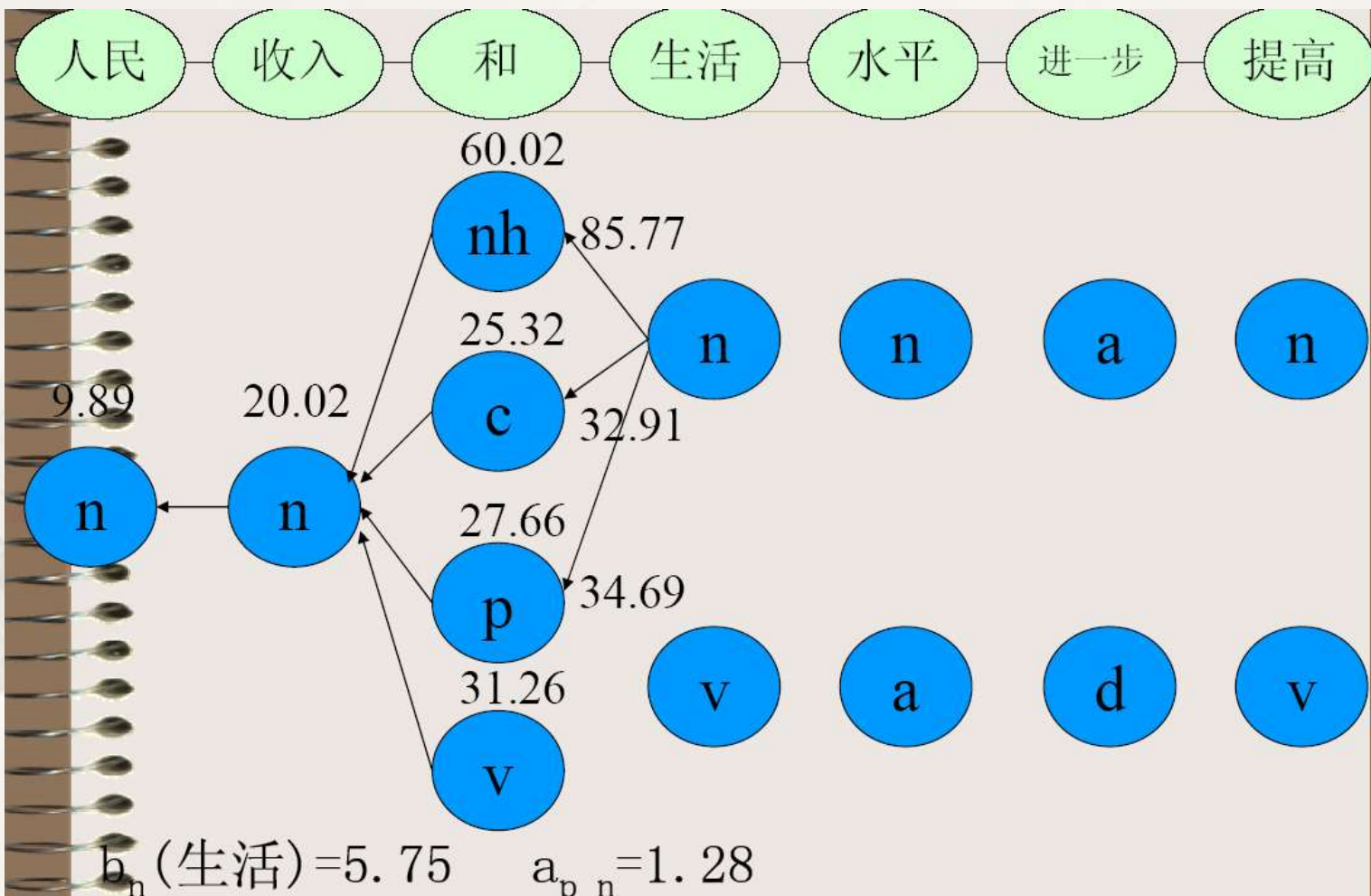
$a_{nn} = 3.15$

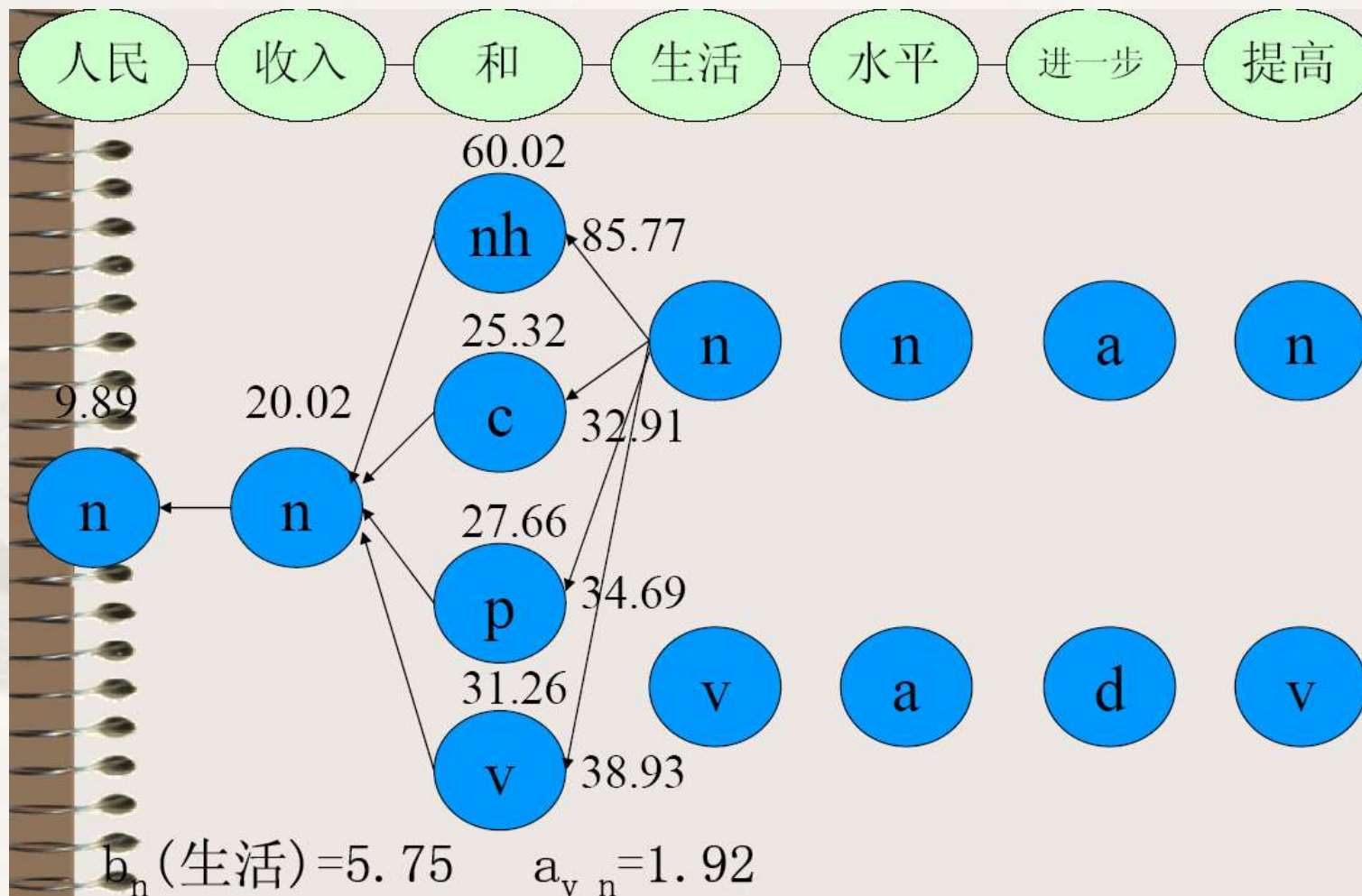




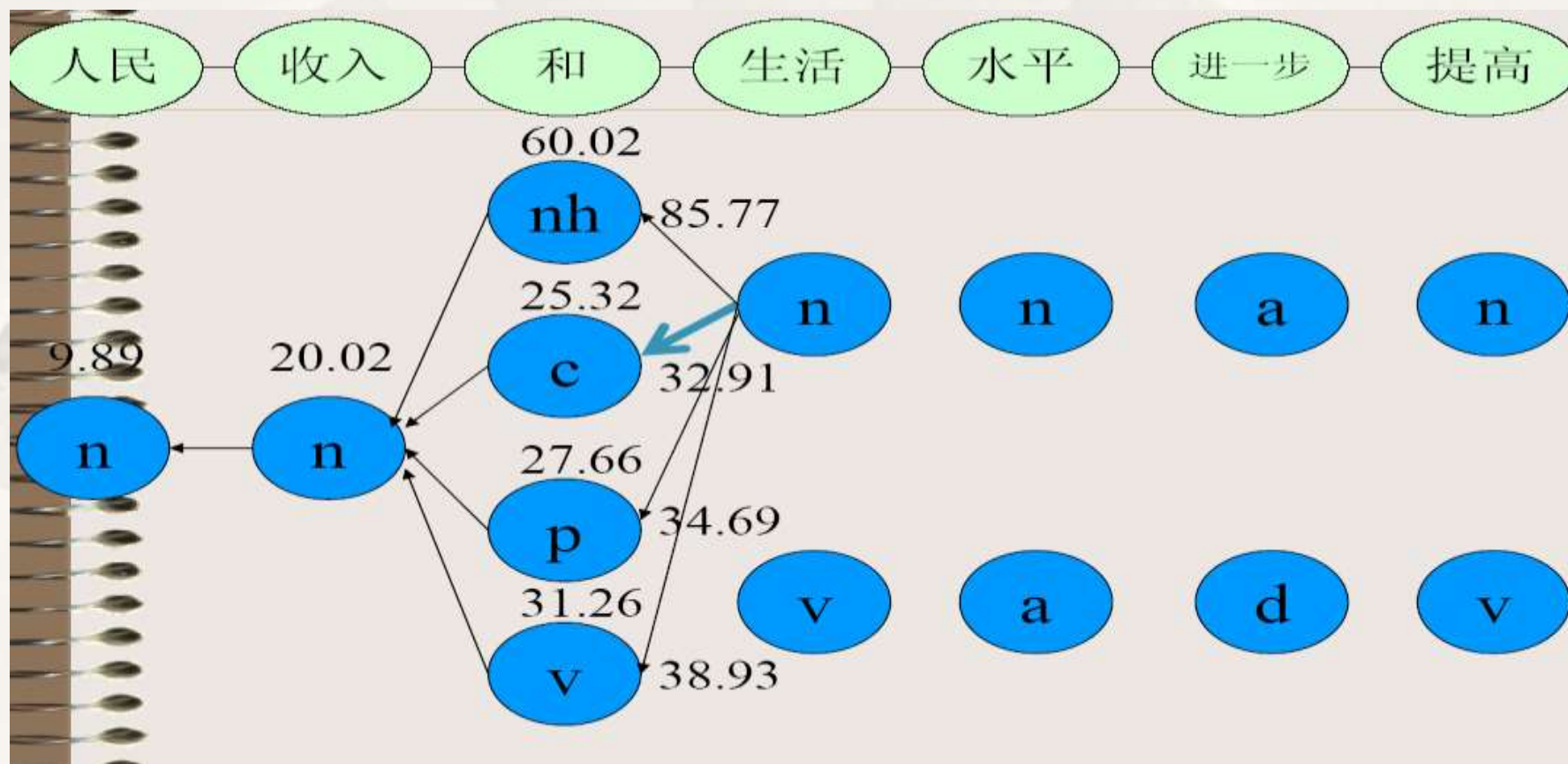


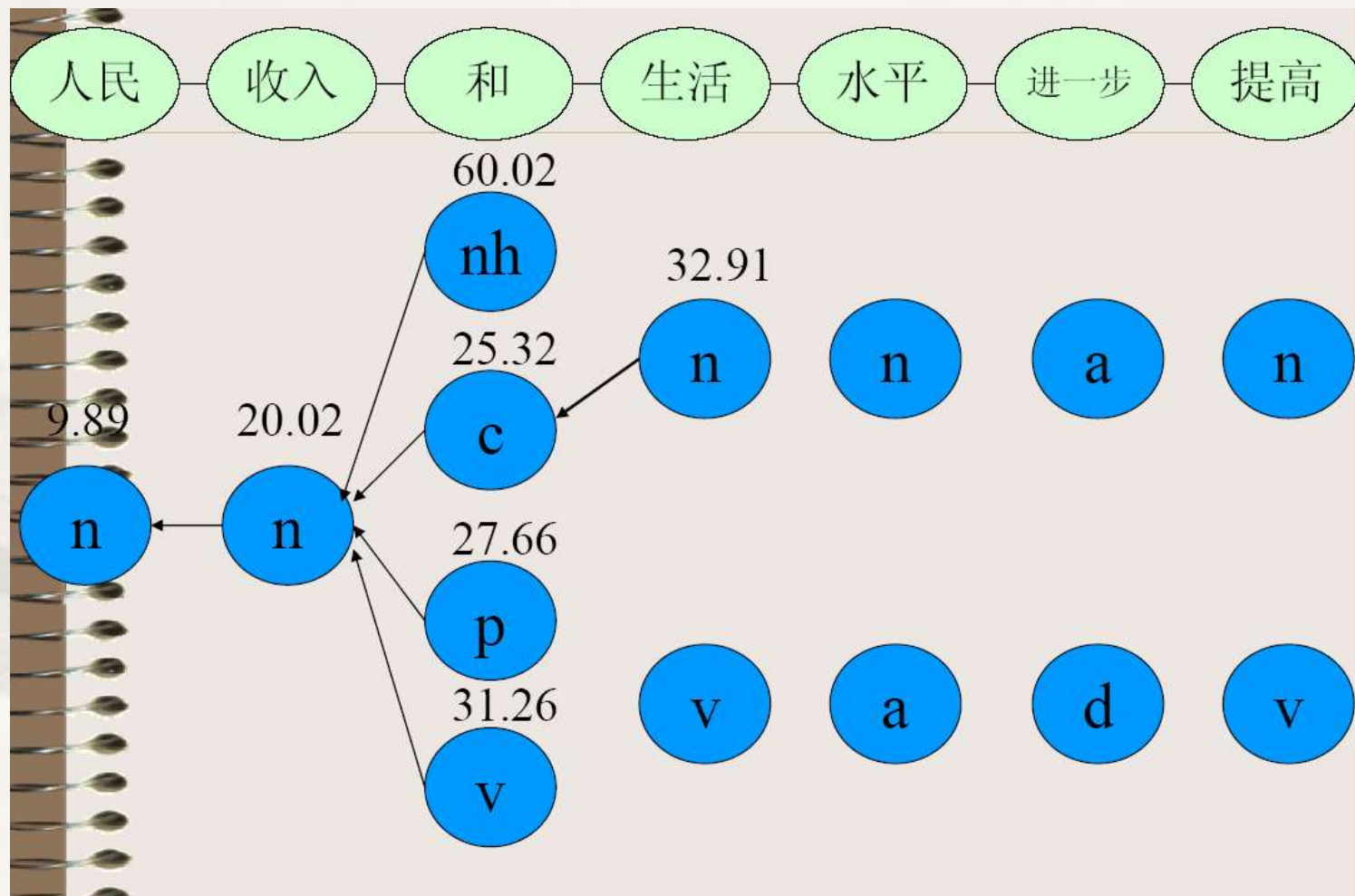


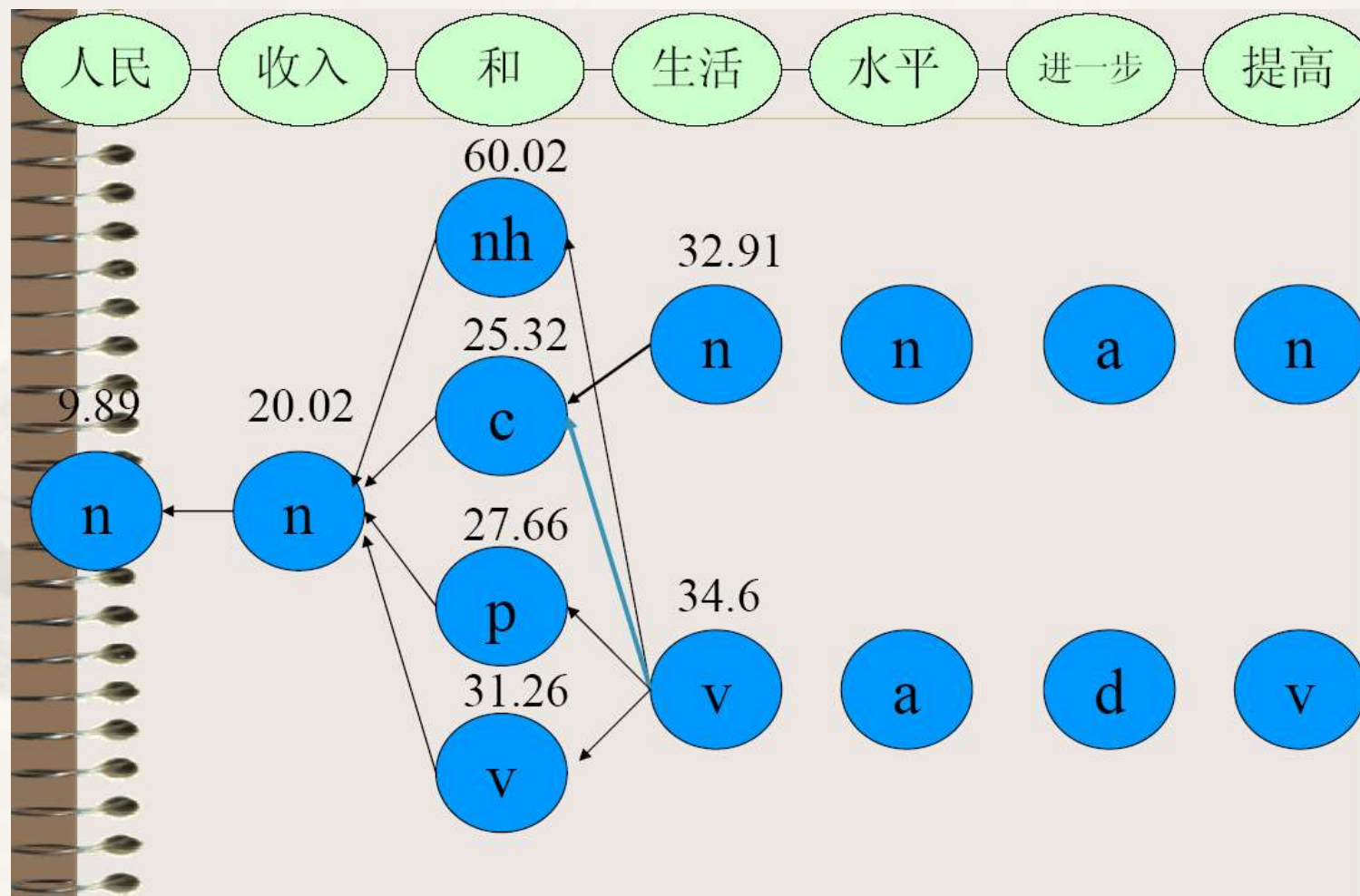


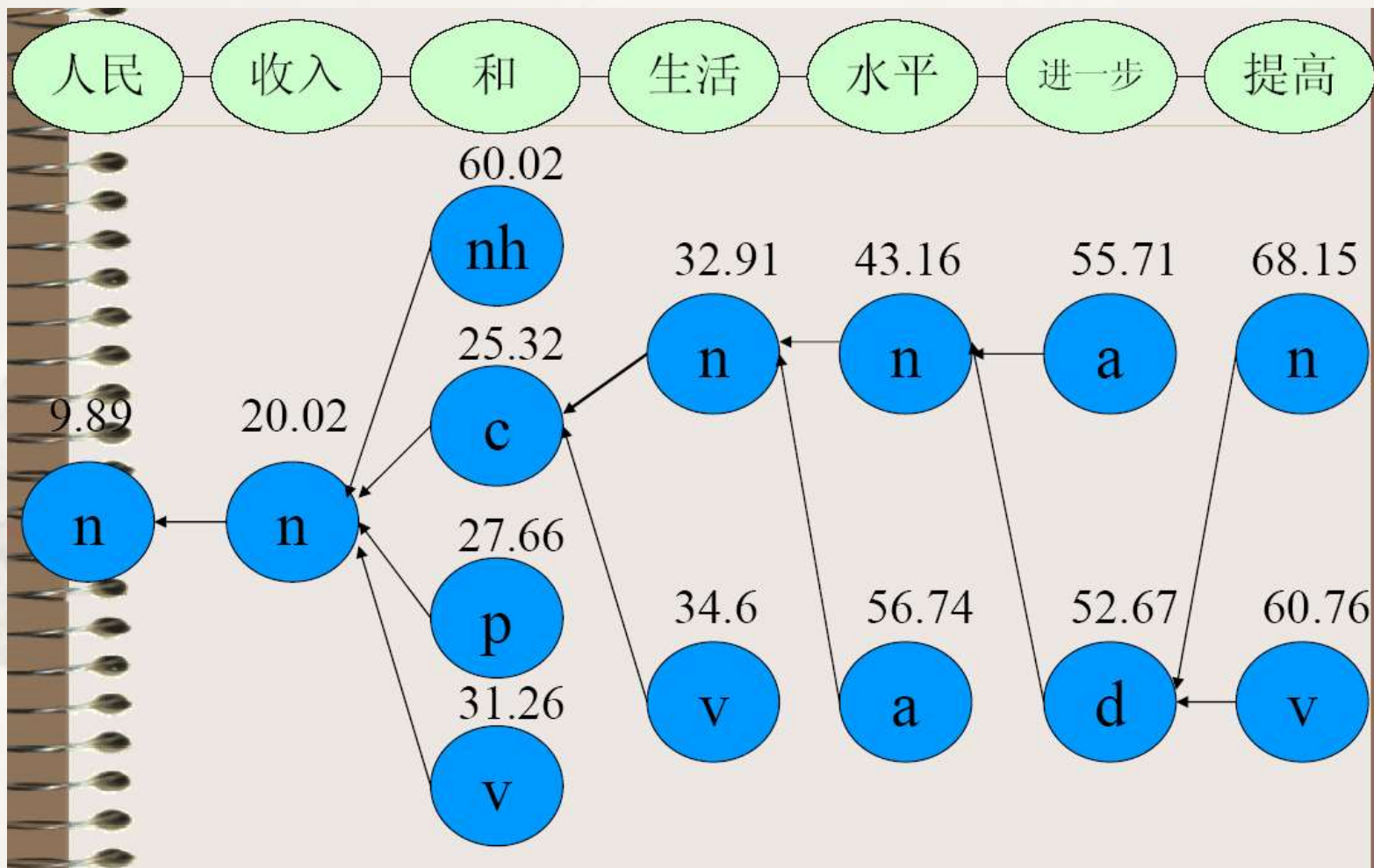


记录“和→生活”的最佳路径是“c-n”

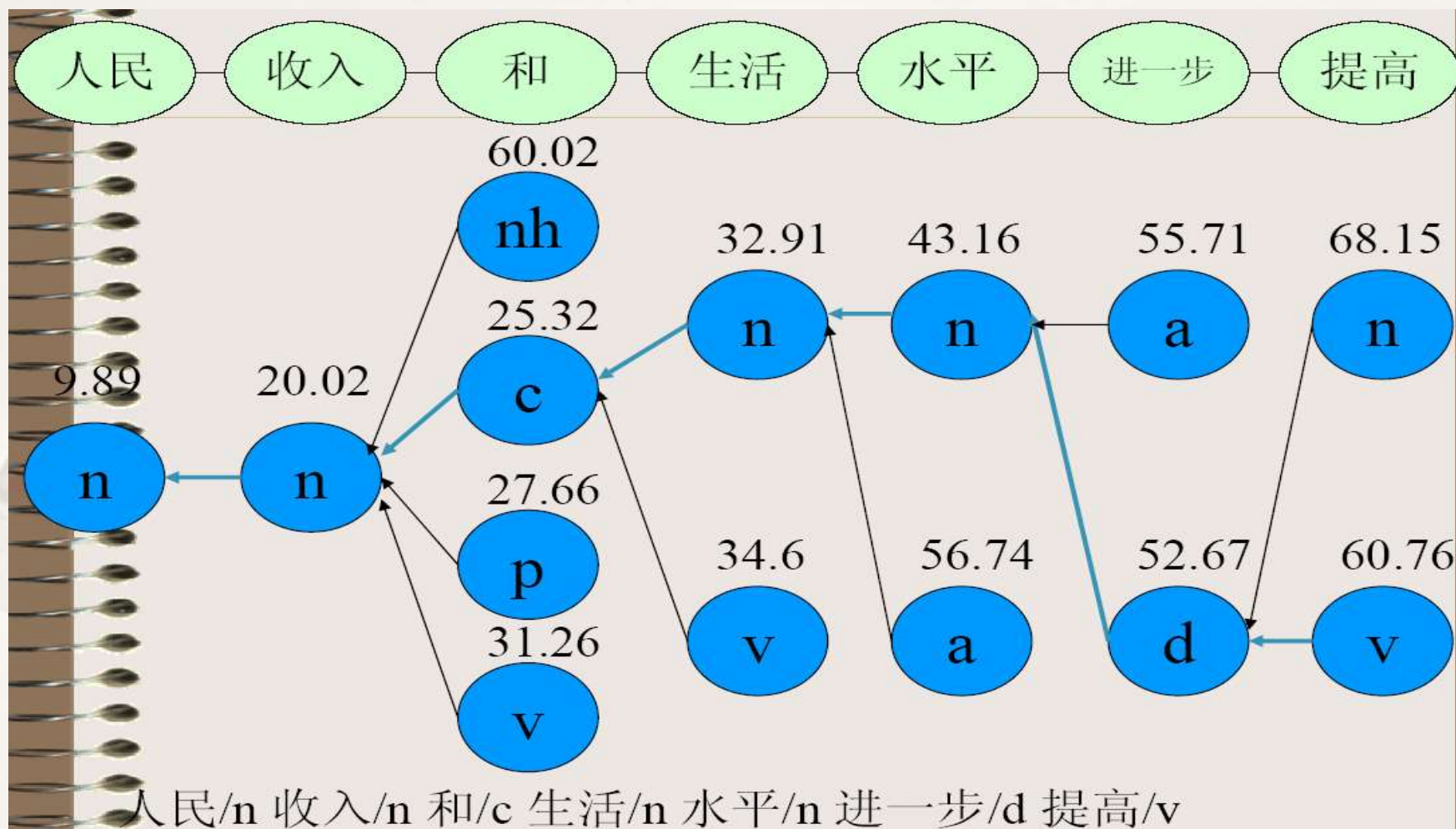








提取所有的“最优路径”记录，确定最优标记状态



内容提要

- * 基于N元文法的分词（MM）
- * 基于HMM的分词/词性标注一体化(模型)
- * 由字构词的汉语分词方法
- * 汉语分词方法的后处理方法

课下阅读：

- * 未登录词的识别
- * 数据平滑

由字构词的汉语分词方法

- * 基本思路

- * 分词过程：一个字的分类问题；
- * 每个字在词语中属于一个确定位置

- * 每个字一定处于下面4个状态(词位)之一

- * 词首 (B)
- * 词中 (M)
- * 词尾 (E)
- * 单独成词 (S)

这里的“字”不仅限于汉字，也可以指标点符号、外文字母、注音符号和阿拉伯数字等任何可能出现在汉语文本中的文字符号。

由字构词的汉语分词方法

* 示例分析：分词结果（1）-> 字标注形式（2）

* （1）上海 / 计划 / 到 / 本 / 世纪 / 末 / 实现 / 人均 / 国内 / 生产 / 总值 / 五千美元 / 。 /

* （2）上 / B 海 / E 计 / B 划 / E 到 / S 本 / S
世 / B 纪 / E 末 / S 实 / B 现 / E 人 / B 均 / E
国 / B 内 / E 生 / B 产 / E 总 / B 值 / E 五 / B
千 / M 美 / M 元 / E 。 / S

由字构词的汉语分词方法

- * 字的标注过程中，对所有的字根据预定义的特征进行词位特征学习，获得一个概率模型
- * 常用的两类特征
 - * 字本身
 - * 词位（状态）的转移概率
- * 在待切分字串上，根据字与字之间的结合紧密程度，得到一个词位的分类结果

这里说的至少是HMM

如何估计？

由字构词的汉语分词方法

生成方法与判别方法

* 生成方法

- * 以词为单位，基于bayes公式，Language Model/HMM等，最后决定生成的句子序列

$$W_{Seq}^* = \operatorname{argmax}_{W_{Seq}} P(W_{Seq} | c_1^n) * P(c_1^n)$$

c_1^n : 组成输入句子的 n 个字的标记 (词位) ;

* 判别方法

- * 注意条件概率的不同

$$P(t_1^n | c_1^n) = \prod_{i=1}^n P(t_i | t_1^{i-1}, c_1^n) \longrightarrow \prod_{i=1}^n P(t_i | t_1^{i-1}, c_{i-2}^{i+2})$$

t_k 表示第 k 个字的词位，即 $t_k \in \{B, M, E, S\}$ ， C_i 代表第 i 个字

由字构词的汉语分词方法

* 关于 c_{k-2}^{k+2}

- * 通常情况下，使用基于字的判别式模型时需要在当前字的上下文中开一个 w 个字的窗口（一般取 $w=5$ ，前后各两个字），在这个窗口抽取分词相关的特征

* 其他常见特征模板：北京奥运会

- * (a) c_k ($k=-2,-1,0,1,2$) $c_{-2}=\text{北}, c_{-1}=\text{京}, c_0=\text{奥}, c_1=\text{运}, c_2=\text{会}$

- * (b) $c_k c_{k+1}$ ($k=-2,-1,0,1$)

$c_{-2}c_{-1}=\text{北京}, c_{-1}c_0=\text{京奥}, c_0c_1=\text{奥运}, c_1c_2=\text{运会}$

- * (c) $c_{-1} c_1$

$c_{-1}c_1=\text{京运}$

由字构词的汉语分词方法

- * 得到对应的特征之后，利用常见的机器学习模型建模求解
 - * 感知机、最大熵、条件随机场、支持向量机
- * 训练数据格式基本都符合如下形式：

[illegible]

由字构词的汉语分词方法

- * 由字构词的分词技术的优势
 - * 简化了分词系统的设计
 - * 文本中的词表词和未登录词都是用统一的字标注过程来实现的，分词过程成为字重组的简单过程。
 - * 既可以不必专门强调词表词信息，也不用专门设计特定的未登录词识别模块

内容提要

- * 基于N元文法的分词（MM）
- * 基于HMM的分词/词性标注一体化(模型)
- * 由字构词的汉语分词方法
- * 汉语分词方法的后处理方法

课下阅读：

- * 未登录词的识别
- * 数据平滑

汉语分词方法的后处理方法

- * 马尔可夫模型遇到的问题：

- * 马尔可夫假设对于自然语言语法结构的很多属性来说太粗糙
- * 为什么不采用更精巧的模型？ 四元或更高阶...
 - * 不可行，需要大量的参数
 - * 不得不做一些平滑或差值
 - * 难度随模型复杂度而加剧

- 基于转换错误驱动的标注学习

- 可以利用更大的词汇和语法结构规则
- 标注可以建立在词语或更多的上下文上
- 编码了词语和标记之间复杂的依存关系
- 决策量比估计大量的马尔可夫模型的参数要少一个级别

汉语分词方法的后处理方法

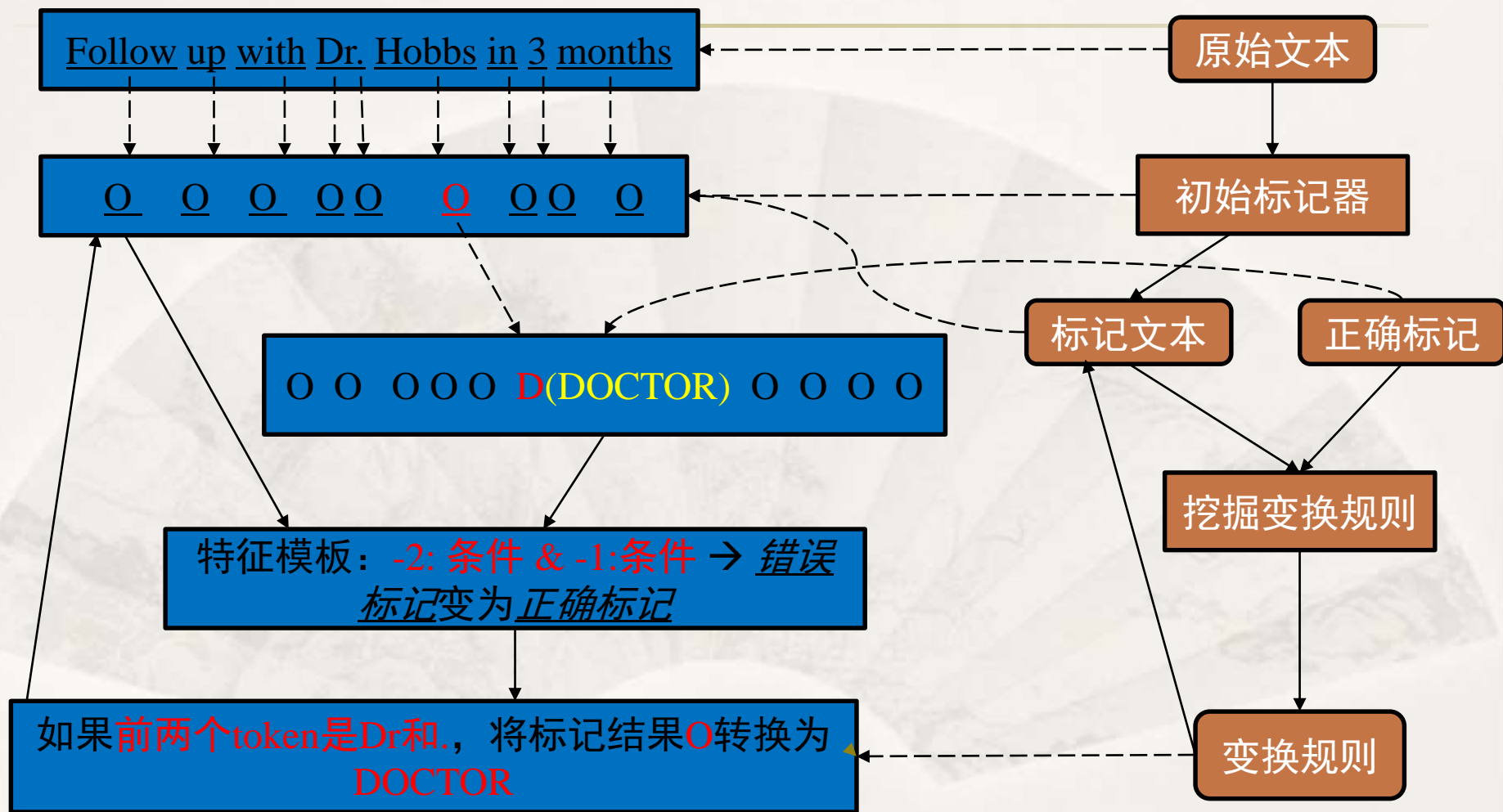
* 两个重要组成部分：

- * 允许的误差校正转换的详细说明
- * 学习算法

* 输入数据：一个已经标注好的语料库，* 一个词典(不是必须的)

- * 用最常见的标记来标注训练语料库中的每个词
 - * 需要词典的原因，不是一般的，可以理解为一个初始标注器
- * 构建一个转换的排序表，把初始的标注转化为接近正确的标注
- * 通过再次初始化来选择每个词最常用的标记，并应用转换
- * 得到一个可以用来标注新的文本的排序表

* 基于转换错误驱动的规则学习方法



Brill E. Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging[J]. CL, 1995, 21(4): 543-565.

汉语分词方法的后处理方法

- * 转换的两个组成部分：



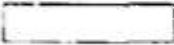
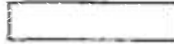







- * 一条重写规则

- * 形式： $t^1 \rightarrow t^2$, 表示：“用标记 t^2 来替换 t^1 ”。

- * 一个触发环境

- * Brill (1995a) 制定如下表所示的触发环境。

- * ‘*’表示潜在重写的位置，方框表示寻找触发的位置

方案	t_{i-3}	t_{i-2}	t_{i-1}	t_i	t_{i+1}	t_{i+2}	t_{i+3}
1				*			
2				*			
3				*			
4				*			
5				*			
6				*			
7				*			
8				*			
9				*			

汉语分词方法的后处理方法

在基于转换的标注中学习的一些转换的例子

源标记	目的标记	触发环境
NN	VB	前面的标记是 TO
VBP	VB	前面的三个标记之一是 MD
JJR	RBR	下一个标记是 JJ
VBP	VB	前面的两个词语之一是 <i>n't</i>

- 由标记触发的转换：

- 第一个转换：

指定名词在TO后面，应该被重新标记为动词

* go to **school**中的**school**：后面带有一些更专门的转换规则重新标记回NN

- 第二个转换：

具有相同原形和过去式形式的动词（如cut、put），前面有一个情态动词使得动词不可能被用做过去式

- 第三个转换：eg：重新标注**more** valuable player中的**more**

汉语分词方法的后处理方法

在基于转换的标注中学习的一些转换的例子

源标记	目的标记	触发环境
NN	VB	前面的标记是 TO
VBP	VB	前面的三个标记之一是 MD
JJR	RBR	下一个标记是 JJ
VBP	VB	前面的两个词语之一是 $n't$

- 由词触发的转换：
 - 第四个转换：

前面出现，如 don't、shouldn't 的词，（类似第二个转换），使得后面更可能是一个原形而不是过去式
- 词语触发环境可以建立在当前词或词语和词性标记的联合
 - 当前词是 w^i ，而下一个标记是 t^j 上

汉语分词方法的后处理方法

- * 基于转换的标注学习算法选择了最佳的转换，并且确定了它们的应用次序，其工作方式如下所示：

```
1  $C_0$  := corpus with each word tagged with its most frequent tag
3 for  $k := 0$  step 1 do
4    $v$  := the transformation  $u_i$  that minimizes  $E(u_i(C_k))$ 
6   if  $(E(C_k) - E(v(C_k))) < \epsilon$  then break fi
7    $C_{k+1} := v(C_k)$ 
8    $\tau_{k+1} := v$ 
9 end
10 Output sequence:  $\tau_1, \dots, \tau_k$ 
```


基于转换的标注学习算法。 C_i 指语料库标注的第 i 次迭代, E 指错误率

- * 第一行：用最常见的标记标注每个词
- * 第四行：每次迭代中，我们选择最可能减少错误率的转换
- * 通过标注过的语料库 C_k 中被错误标注的词语的数目来衡量错误率 $E(C_k)$
- * 当没有能够降低超过预先指定阈值 ϵ 大小的错误率的转换时将停止
- * 这是一个转换最优序列的贪心搜索过程

汉语分词方法的后处理方法

* 基于转换错误驱动的规则方法

- * 学习和标注在该方法种都是简单和直观的
- * 成功用于词性标注、句法分析、介词附着以及语义消歧
- * 经验上，没有出现过拟合现象
- * 可以被用来解决大部分后处理问题
- * 效率的提升优化，考验工程能力



Q & A!

中文未登录词识别

未登录词的类型

- * 命名实体 (Named Entity)

- * 汉语人名：李素丽 老张 李四 王二麻子
- * 汉语地名：定福庄 白沟 三义庙 韩村 河马甸
- * 翻译人名：乔治·布什 叶利钦 包法利夫人
- * 翻译地名：阿尔卑斯山 新奥尔良 约克郡
- * 机构名：方正公司 联想集团 国际卫生组织外贸部

- * 数字、日期词、货币等

- * 商标字号：非常可乐 乐凯 波导 杉杉 同仁堂

- * 专业术语：万维网 主机板 模态逻辑 贝叶斯算法

- * 缩略语：三个代表 五讲四美 打假扫黄 打非计生办

- * 新词语：卡拉OK 波波族 美刀 港刀

未登录词识别的依据

- * 内部构成规律（用字规律）
- * 外部环境（上下文）
- * 重复出现规律

中国人名的内部构成规律

- * 在汉语的未定义词中，中国人名是规律性最强，也是最容易识别的一类；
- * 中国人名一般由以下部分组合而成：
 - * 姓：张、王、李、刘、诸葛、西门、范徐丽泰
 - * 名：李素丽，张华平，王杰、诸葛亮
 - * 前缀：老王，小李
 - * 后缀：王老，赵总
- * 中国人名各组成部分用字比较有规律

中国人名的内部构成规律

- * 台湾出版的《中国姓氏集》收集姓氏5544个，其中，单姓3410个，复姓1990个，3字姓144个。
- * 中国目前仍使用的姓氏共737个，其中，单姓729个，复姓8个。
- * 根据我们收集的300万个人名统计：姓氏：974个，其中，单姓952个，复姓23个，300万人名中出现汉字4064个。

中国人名的内部构成规律

- * 中国人名各组成部分的组合规律
 - * 姓 + 名
 - * 姓
 - * 名
 - * 前缀 + 姓
 - * 姓 + 后缀
 - * 姓 + 姓 + 名（海外已婚妇女）

中国人名的上下文构成规律

* 身份词：

- * 前：工人、教师、影星、犯人
- * 后：先生、同志
- * 前后：女士、教授、经理、小姐、总理

* 地名或机构名：

- * 前：静海县大丘庄禹作敏

* 的字结构

- * 前：年过七旬的王贵芝

* 动作词

- * 前：批评，逮捕，选举
- * 后：说，表示，吃，结婚

中国人名识别的难点

- * 一些高频姓名用字在非姓名中也是高频字
 - * 姓氏：于，马，黄，张，向，常，高
 - * 名字：周鹏和同学，周鹏和同学
- * 人名内部相互成词，指姓与名、名与名之间本身就是一个已经被收录的词
 - * [王国]维、[高峰]、[汪洋]、张[朝阳]
- * 人名与其上下文组合成词
 - * 这里[有关]天培的壮烈；
 - * 费孝通向人大常委会提交书面报告
- * 人名地名冲突: 河北省刘庄

中文姓名识别方法

* 中文姓名识别方法

- * 姓名库匹配，以姓作为触发信息，寻找潜在的名字
- * 计算潜在姓名的概率估值及相应姓氏的姓名阈值，根据姓名概率评价函数和修饰规则对潜在的姓名进行筛选。

中国地名的识别

* 困难

- * 地名数量大，缺乏明确、规范的定义。
《中华人民共和国地名录》（1994）收集88026个，不包括相当一部分街道、胡同、村庄等小地方名称。
- * 真实语料中地名出现情况复杂。如地名简称、地名用词与其它普通词冲突、地名是其它专用名词的一部分，地名长度不一等。

未登录词识别的一般方法

- * 在统计方法中，未登录词识别的一种最通常的做法就是将识别问题转化成标注问题
- * 对于输入句子中的每个汉字，定义四个标记：
 - * 不属于未登录词O
 - * 未登录词首字B
 - * 未登录词尾字E
 - * 未登录词中间字I

将识别问题转化成标注问题

- * 如果能够把输入句子中的每个汉字都正确地按上述标记进行标注，那么未登录词的识别自然就解决了
- * 标注可以采用
 - * 隐马尔科夫模型 (HMM)
 - * 最大熵 (ME)
 - * 最大熵马尔科夫模型 (MEMM)
 - * 条件随机场 (CRF) 等

将识别问题转化成标注问题

- * 以人名识别为例，输入文本：
 - * 这是周恩来、邓颖超生前居住的地方
 - * 标注为：
 - * 这是周恩来、邓颖超生前居住的地方
 - * O O B I E O B I E O O O O O O O
 - * 两处标注为BIE的字串“周恩来”、“邓颖超”被识别为人名
- * 训练语料库为已经标注人名的语料库



End