COMP9414：人工智能第5b讲。语言模型

韦恩-沃布克
电邮：w。wobcke@unsw。edu。au

概率语言模型

- 基于从大型文本/语音语料库中得出的统计数据
  - 布朗语料库（1960年代）--100万字
  - 佩恩树库（1980年代）--700万字
  - 北美新闻（1990年代）--3.5亿字
  - IBM - 10亿字

本讲座

- 语篇标签化
  - n-gram模型
  - 隐马尔科夫模型
  - 维特比算法
- 词义歧义
  - 互惠信息
  - 基于类的模型

  - 谷歌、脸书和微软 - 数以万亿计的文字
- 与认为语言能力基于（先天）知识的观点相反的是
- 我的想法是语言能力是可以学习的。.有足够的数据.. .

潘恩树库标签集

Tag	Description	Example	Tag	Description	Example
CC	coord. conjunction	*and, or*	RB	adverb	*extremely*
CD	cardinal number	*one, two*	RBR	adverb, comparative	*never*
DT	determiner	*a, the*	RBS	adverb, superlative	*fastest*
EX	existential there	*there*	RP	particle	*up, off*
FW	foreign word	*noire*	SYM	symbol	*+, %*
IN	preposition or sub-conjunction	*of, in*	TO	"to"	*to*
JJ	adjective	*small*	UH	interjection	*oops, oh*
JJR	adject., comparative	*smaller*	VB	verb, base form	*fly*
JJS	adject., superlative	*smallest*	VBD	verb, past tense	*flew*
LS	list item marker	*1, one*	VBG	verb, gerund	*flying*
MD	modal	*can, could*	VBN	verb, past participle	*flown*
NN	noun, singular or mass	*dog*	VBP	verb, non-3sg pres	*fly*
NNS	noun, plural	*dogs*	VBZ	verb, 3sg pres	*flies*
NNP	proper noun, sing.	*London*	WDT	wh-determiner	*which, that*
NNPS	proper noun, plural	*Azores*	WP	wh-pronoun	*who, what*
PDT	predeterminer	*both, lot of*	WP$	possessive wh-	*whose*
POS	possessive ending	*'s*	WRB	wh-adverb	*where, how*
PRP	personal pronoun	*he, she*			

新南威尔士大学
©W.Wobcke等人，2019-2022年

语篇标签化

- DT大/JJ陪审团/NN评论/VBD在/IN一个/DT数字/NN的/IN其他/JJ主题/NNS ./。
- 那里/EX是/VBP 70/CD儿童/NNS那里/RB
- 初步/JJ发现/NNS是/VBD报告/VBN在/IN今天/NN's/POS新/NNP英格兰/NNP杂志/NNP的/IN医学/NNP ./。

为什么这很难？

含糊不清，比如说背部

- 盈利增长处于次要地位/JJ
- 后面的一个小建筑/NNN
- 绝大多数参议员支持/VBP该法案
- 戴夫开始向后退/VB走向门口
- 使国家能够回购/偿还债务
- 那时候我21岁了/RB

概率论的表述

- 事件。词w的出现，带有标签t的词的出现
- 给定单词序列$w_1, ---, w_n$， 选择$t_1, ---, t_n$， 以便 - $P(t_1, --, t_n | w_1, --, w_n)$是最大的。
- 应用贝叶斯规则
  - $P(t1,--,tn|w1,--,wn) = \frac{P(w1,--,wn|t1,--,tn) \cdot P(t1,---,tn)}{P(w,--,wn)}$
  - 因此最大化$P(w_1, --, w_n | t_1, --, t_n) \cdot P(t_1, --, t_n)$

单元模式

最大化$P(w_1, --, w_n | t_1, --, t_n) \cdot P(t_1, --, t_n)$

- 应用独立假设
  - $P(w_1, --, w_n | t_1, --, t_n) = P(w_1 | t_1) \cdot --- .P(w_n | t)_n$
  - 由t产生的词w的概率与上下文无关
  - $P(t_1, --, t_n) = P(t_1) \cdot --- .P(t)_n$
  - 标签序列的概率与顺序无关
- 估计概率
  - $P(w|t)$ = #(w与标签t一起出现)/#(有标签t的词)
  - $P(t)$ = #(带有标签t的词)/#词
  - 选择使$\Pi P(t_i | w_i)$最大化的标签序列。
  - 为每个词选择最常见的标签

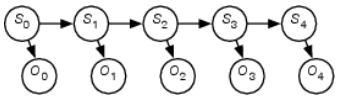
- 准确率约为90%--
  但每个句子中仍有≈1个单词的错误。

新南威
尔士大
学
©W.Wobcke等人，2019-
2022年

马尔科夫链

- 贝叶斯网络
  - $P(S_0)$ 指定初始条件
  - $P(S_{i+1} \mid S_i)$指定动态（如果每个i都相同，则为静止）。
- 独立性假设
  - $p(s_{i+1} \mid s_0, --, s_i) = p(s_{i+1} \mid s_i)$
  - 过渡概率仅取决于当前状态S_i - 与达到该状态的历史无关 $S_0, --, S_{i-1}$
  - 鉴于现在，未来是独立于过去的。

隐马尔科夫模型



- 贝叶斯网络
  - $P(S_0)$ 指定初始条件
  - $P(S_{i+1} \mid S_i)$指定动态
  - $P(O_i \mid S_i)$指定 "观察"。
- 独立性假设
  - $P(S_{i+1} \mid S_0, --, S_i) = P(S_{i+1} \mid S_i)$ （马尔可夫链）。
  - $p(o_i \mid s_0, --, s_{i-1}, s_i, o_0, --, o_{i-1}) = p(o_i \mid s_i)$

Bigram模型

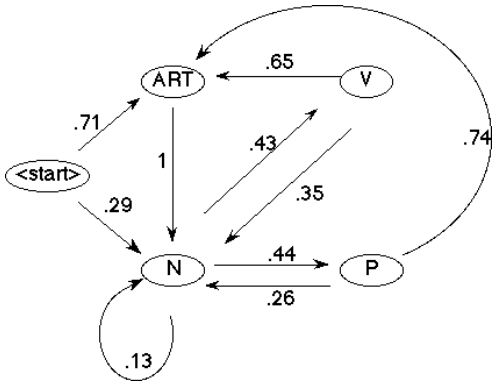
最大化$P(w_1, --, w_n \mid t_1, --, t_n)$。$P(t_1, --, t_n)$

- 应用独立假设（马尔可夫假设）。
  - $P(w_1, --, w_n \mid t_1, --, t_n) = \Pi P(w_i \mid t_i)$
  - 观察（词）只取决于状态（标签）。
  - $P(t_1, --, t_n) = P(t_n \mid t_{n-1})$。$---.P(t_1 \mid \phi)$，其中$\varphi$ = 开始
  - 大图模型：状态（标签）只取决于先前的状态（标签）。
- 估计概率
  - $P(t_i \mid t_j) = \#((t_j, t_i)$发生)/#($t_j$ 开始一个大数)
  - 选择使$\Pi P(w_i \mid t_i)$最大化的标签序列。$P(t_i \mid t_{i-1})$
  - 由有限状态机生成的语篇

用于POS标签的马尔科夫模型

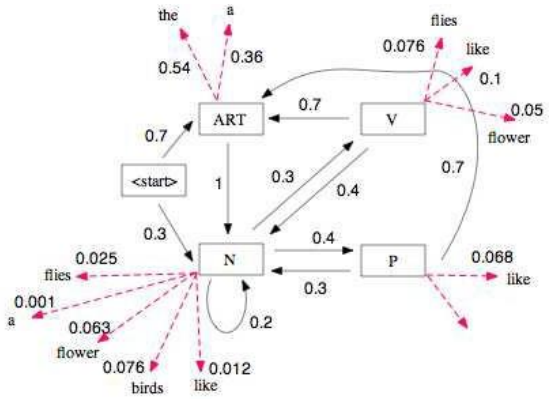
- 观察（词）只取决于当前状态（标签）。

过渡概率定义静止分布

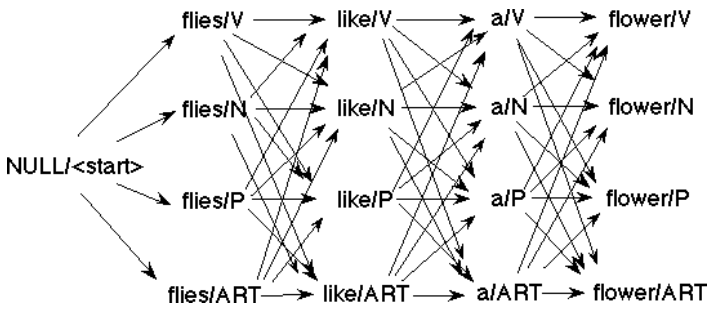


新南威尔士大学
©W.Wobcke等人，2019-2022年

用于POS标签的隐马尔可夫模型



计算概率



例子

w	*P*(*w*|*ART*)	*P*(*w*|*N*)	*P*(*w*|*P*)	*P*(*w*|*V*)
a	0.36	0	0	0
苍蝇	0	0.025	0	0.076
花	0	0.063	0	0.05
喜欢	0	0.012	0	0.1

P(苍蝇/N像/V a/ART花/N) =

=

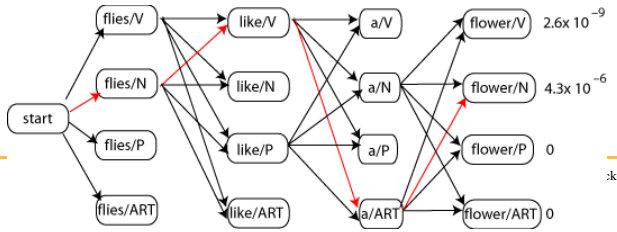
P(N|start)。*P*(flies|N)。*P*(V|N)。*P*(喜欢|V)。*P*(ART|V)。*P*(a|ART)。*P*(N|ART)。*P*(花|N)

= 0.29 x 0.025 x 0.43 x 0.1 x 0.65 x 0.36 x 1 x 0.063

最有可能的序列，即使*P*（苍蝇/V）>*P*（苍蝇/N）。

维特比算法

1. 向前扫过（一次一个词），只保存每个标签*t*的最可能序列（及其概率）$_i$ 的w_i
2. 选择最高概率的最终状态
3. 沿着链子往回走，提取标签序列



词义歧义

例子

我应该换掉那把愚蠢的锁，让你留下钥匙，如果我知道你会回来打扰我，只要一秒钟。

锁定 = - - -

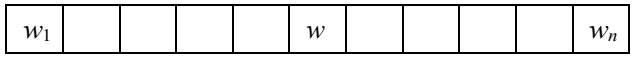
离开 = - - -

第二 = - - -

返回 = - - -

窗户

- 考虑在一个关于*w*的窗口中的共同发生情况



- 词的感觉应该与 "相关 "的词类共同出现
- 选择*w*的意义*s*，使$P(w \text{ as } s|w_1, - -, w_n)$最大化。
- 应用贝叶斯规则
  - 最大化$\frac{P(w1,--,wn|w \text{ as } s)。P(w为s)}{P(w_1,---,wn)}$
- 应用独立假设
  - $P(w_1, - -, w_n \mid \text{w as } s) = \Pi P(w_i \mid \text{w as } s)$

简单的（编造的）例子

词语	桥梁/结构	桥梁/牙齿	任何窗口
牙齿	1	10	300
悬挂	200	1	2000
的	5500	180	500 000
牙医	2	35	900
总数	5651	194	501 500

P（桥梁/结构）=5651/501 500 = 0.0113

P(桥/牙) = 194/501 500 = 3.87 x 10^{-4}

P(牙|桥/结构) = 1/5651 = 1.77 x 10^{-4}

P（牙|桥/牙） = 10/194 = 0.052

如果窗口包含 "牙齿"，则首选桥梁/牙齿

相互信息

$$MI(x, y) = \log \frac{P(x, y)}{^2P(x)。P(y)}$$

$$\log MI(sense(w), w) = \frac{N.\#(sense(w_1), w_2)}{}$$

- 估计概率。$P(w_i \mid \text{w as } s)$
  - #(w_i 在*w*为*s*周围的*n*个*词*窗口中)/#(*w*为*s*的窗口)

1 2

$^2\#(sense(w_1))$。

$\#(w_2)$

其中N是语料库的大小

- $MI = 0$：$sense(w_1)$和w_2是有条件的独立。

- $MI < 0$: $sense(w_1)$和w_2一起出现的次数少于随机次数
- $MI > 0$: $sense(w_1)$和w_2一起出现的次数多于随机出现的次数
- **增加相互信息就相当于假**设了独立性
- 选择w的感觉$s = \text{argmax}_{s \in senses(w)} \Sigma_{w_i \in window(w)} MI(s, w_i)$

基于类的方法

- 使用预定义的 "意义类"，例如WordNet、Wikipedia
  - 锁 → *机械装置* ←工具、曲柄、齿轮、 - - - -
  - 锁 → *身体部位* ← 头发、眼睛、手、 - - -
- 通过添加词的意义来计算词的意义的计数
- 把所有的$P(w$为$s)$仅仅看作是$P(s)$，即$P($任何具有s意义的词$)$。
- 优势
  - 减少空间和时间的复杂性
  - 降低数据稀疏度
  - 允许无监督的学习

总结

- 统计学（和神经网络）模型在许多任务上表现良好
  - 语料部分标记
  - 词义歧义
  - 对传统分析器的控制
  - 概率分析法
- 问题
  - 不切实际的简化假设（似乎有效）。
  - 要求有非常多的（标示）文本
  - 在（甚至是大型）文本语料库中的单词出现的稀疏程度

⚠ 随着时间的推移，词语使用的变化（例如，*参议员*奥巴马）。