词性标注

给定单词 w，找到对应的词性 z

1. 问题概述

已知（S）

Sentence = $w_1\ w_2\ w_3\ w_4\ w_5$ 单词

$z_1\ z_2\ z_3\ z_4\ z_5$ 词性

求 S' = $w_1'\ w_2'\ w_3'\ w_4'\ w_5'$ 求标注词性

Noise channel model:

$$p(z|s) = p(s|z)\,p(z)$$

argmax　translation model　language model

$$= p(w_1 w_2 w_2 \cdots w_N \mid z_1 z_2 \cdots z_n)\, p(z_1 z_2 \cdots z_n)$$

$$= \prod_{i=1}^{n} p(w_i|z_i) \cdot p(z_1)\, p(z_2|z_1) \cdots p(z_n|z_{n-1})$$

edimission probability：给定一个词性，在该 z_i 下单词出现概率是多少。

$p(z_1)$ 该词性出现在句首的概率

$p(z_n|z_{n-1})$ bigram

$$\hat{z} = \arg\max p(z|s)$$

$$= \arg\max \prod_{i=1}^{n} p(w_i|z_i)\, p(z_1) \cdot \prod_{i=1}^{n} p(z_n|z_{n-1})$$

$$= \arg\max \log(\quad\quad)$$

$$= \arg\max \sum \log \underbrace{p(w_i|z_i)}_{A} + \log \underbrace{p(z_1)}_{\lambda} + \sum \log \underbrace{p(z_n|z_{n-1})}_{B}$$

↓ 也就是求三个概率

$\theta = \{A, B, \lambda\}$

Step 1. 求 A, B, λ

Step 2. 求 z.（Viterbi Arg）

一般情况下 $p(w_1 w_2 w_3 | z_1 z_2 z_3) \neq p(w_1|z_1)p(w_2|z_2)p(w_3|z_3)$

我们在此简化为 $p(w_1|z_1)p(w_2|z_2)p(w_3|z_3)$

（也就是假设当前词 w_i 只依赖于 z_i）

2. 三个所需要求的概率

$$\hat{z} = \arg\max p(z|s)$$

$$= \arg\max \prod_{i=1}^{n} p(w_i|z_i)\, p(z_1) \cdot \prod_{i=1}^{n} p(z_n|z_{n-1})$$

$$= \arg\max \log(\quad\quad)$$

$$= \arg\max \sum \log \underbrace{p(w_i|z_i)}_{A} + \log \underbrace{p(z_1)}_{\pi} + \sum \log \underbrace{p(z_n|z_{n-1})}_{B}$$

1. A $p(w_i|z_i)$

$A = \begin{bmatrix} & \\ & \end{bmatrix}$ N × M

N: tags nr.

M: words nr.

给定一个假设 tag 是 vb.

[apple, study, ·····]

词库

在给定动词情况下，每个单词出现概率.

2. π $p(z_1)$

N = nr. tags

一个句子以某个词性开始的 P

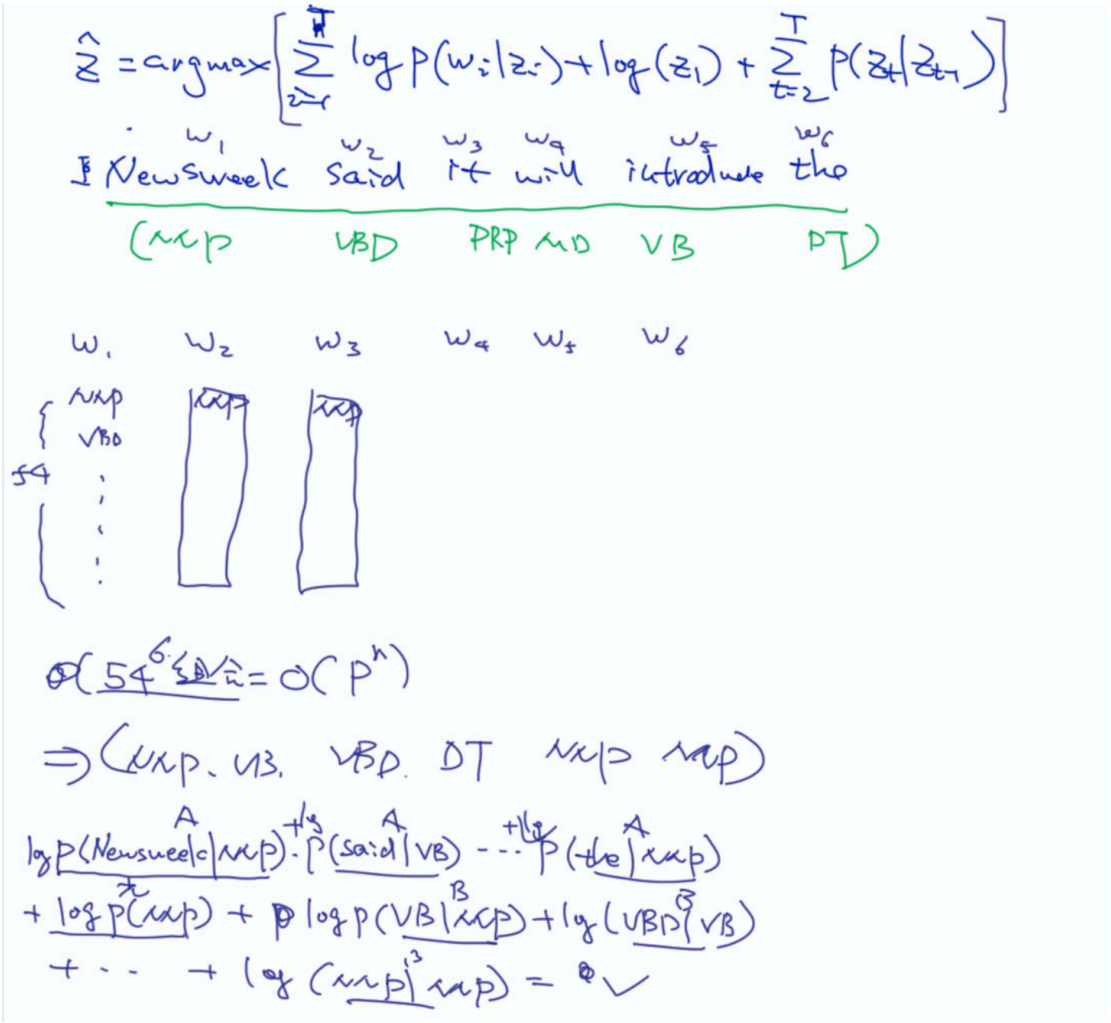
3. B $p(z_n|z_{n-1})$ tags

$B = \begin{bmatrix} & \\ & \end{bmatrix}$ N × N

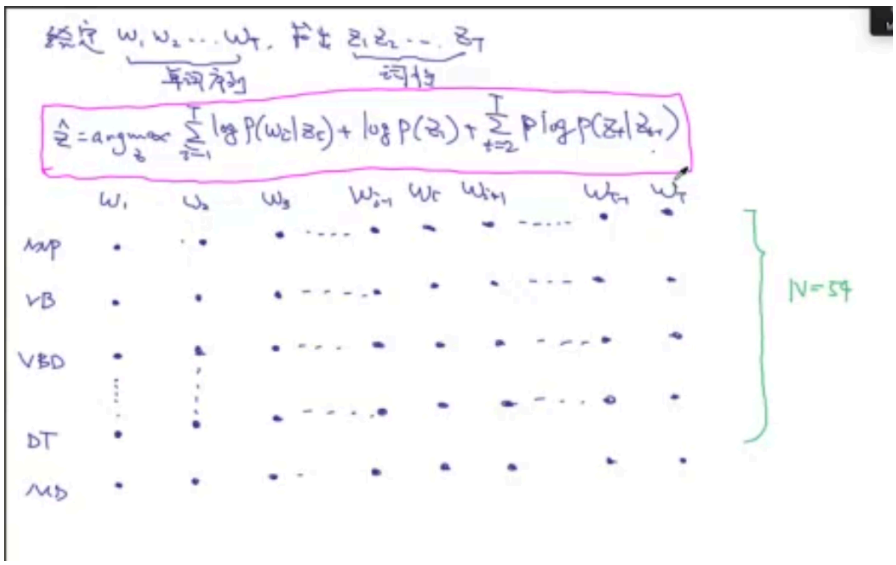
3.

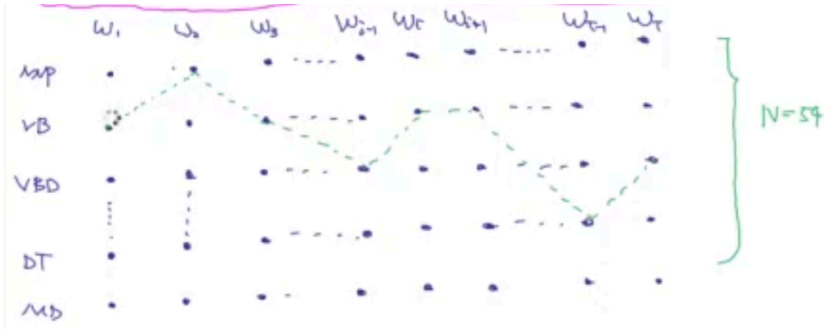
见 jupyter9 pos tagger

4. Viterbi Alg

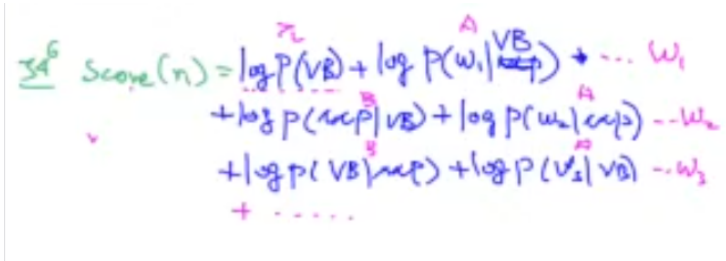


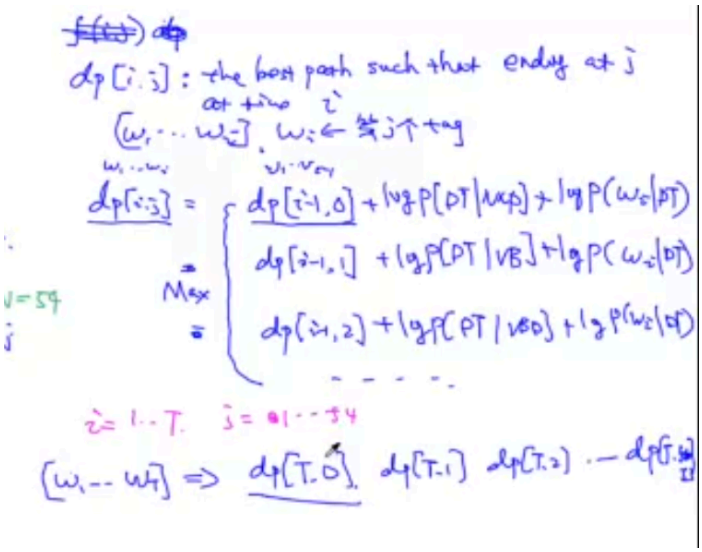
如果每个单词 w，都去遍历一遍所有的 tag，计算量太大。
Viterbi 算法找出来的是全局最优解。

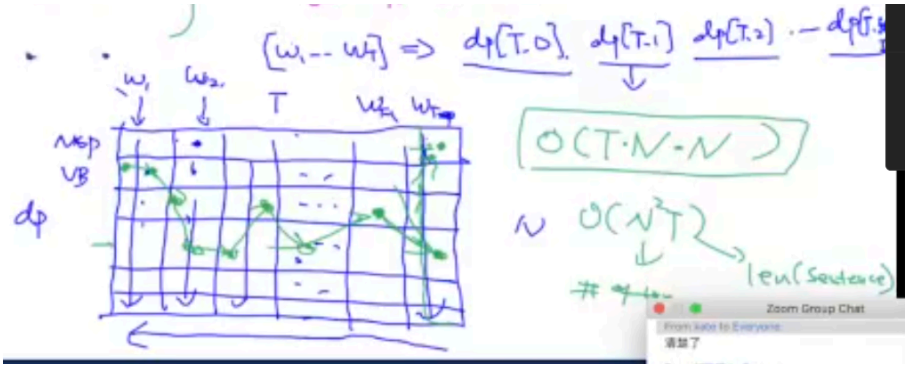

给定 $w_1 w_2 \dots w_T$，求 $z_1 z_2 \dots z_T$
词序列
词性
$\hat{z} = \arg\max_z \sum_{i=1}^{I} \log p(w_i|z_i) + \log p(z_1) + \sum_{t=2}^{I} p \log p(z_t|z_{t-1})$
w_1 w_2 w_3 w_{i-1} w_i w_{i+1} w_{T-1} w_T
RBP
VB
VBD
DT
MD
|V|=54


w_1 w_2 w_3 w_{i-1} w_i w_{i+1} w_{T-1} w_T
RBP
VB
VBD
DT
MD
|V|=54

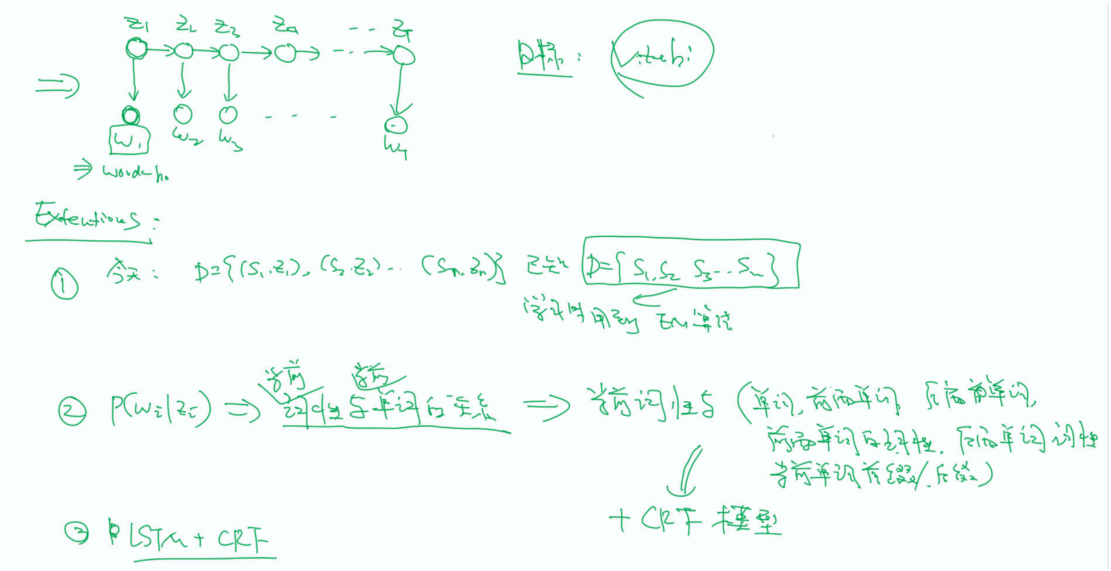
找路径，使得概率最小。


$Score(n_1) = \log p(VB) + \log p(w_1|VB) + \dots w_1$
$+ \log p(RBP|VB) + \log p(w_2|RBP) \dots w_2$
$+ \log p(VB|RBP) + \log p(w_3|VB) \dots w_3$
$+ \dots$


$dp[i,j]$: the best path such that ends at j at time i
$(w_1 \dots w_i)$, w_i 标注为第 j 个 tag
$dp[i,j] = \max_j \{ dp[i-1,0] + \log p[DT|RBP] + \log p(w_i|DT)$
$dp[i-1,1] + \log p[DT|VB] + \log p(w_i|DT)$
$dp[i-1,2] + \log p[DT|VBD] + \log p(w_i|DT)$
$i = 1 \dots T$ $j = 0 \dots 54$
$(w_1 \dots w_T) \Rightarrow dp[T,0], dp[T,1], dp[T,2] \dots dp[T,54]$


dp
O(T·N·N)
O(N²T)
len(Sentence)
Zoom Group Chat

参考 jupyer9


Extentions:
①
②
③ BiLSTM + CRF
+CRF 模型