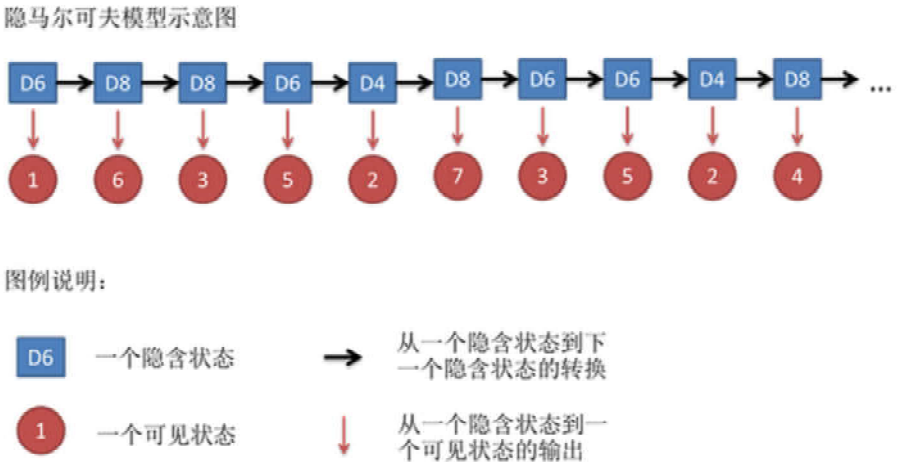
HMM词性标注

Friday, April 17, 2020 12:43 PM

HMM对词性标注建模



模型变量和参数

隐状态： 词性 （K个）

观测值： 词汇（V个）

初始概率：pi， 句首词性的概率分布， 维度(K, 1)

状态转移矩阵：A，从当前词性转移到其他词性的概率分布，维度（K，K）

发射概率矩阵：B，给定词性下词汇的概率分布，维度 (K, V)

当是"。"的时候，就终止

参数学习

非监督：EM

有监督：最大似然，统计计算

$A[i][j] = N_{ij} / N_i$

$B[i][v] = N_{iv} / N_i$

$Pi[i] = N_{0i} / N_0$

一个例子

语料：

你/(代), 是/(动), 程序员/(名)

我/(代), 是/(动), 程序员/(名)

我们/(代) 都/(副) 是/(动) 程序员/(名)

统计结果和参数估计：

观测/词汇集合：{你，我，我们，是，程序员，都}

隐状态/词性集合: {代，动，名，副}

初始概率 pi：

代	动	名	副
3/3=1	0/3=0	0/3=0	0/3=0

状态转移矩阵 A：

	代	动	名	副
代	0/3=0	2/3=0.6667	0/3=0	1/3=0.3333
动	0/3=0	0/3=0	3/3=1	0/3=0
名	0/3=0	0/3=0	0/3=0	0/3=0
副	0/1=0	1/1=1	0/1=0	0/1=0

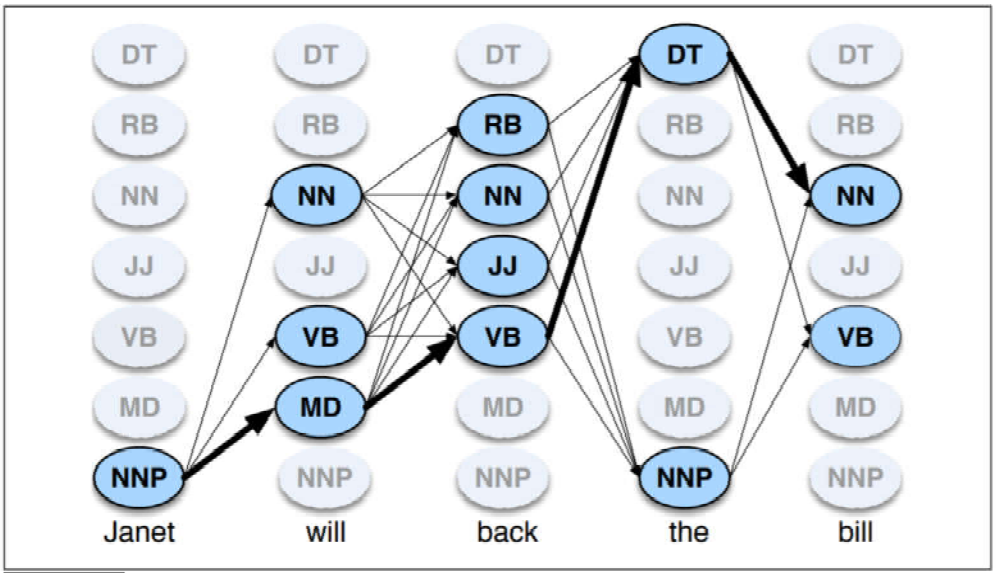
发射概率矩阵 B：

	你	我	我们	是	程序员	都
代	1/3=0.3333	1/3=0.3333	1/3=0.3333	0/3=0	0/3=0	0/3=0
动	0/3=0	0/3=0	0/3=0	3/3=1	0/3=0	0/3=0
名	0/3=0	0/3=0	0/3=0	0/3=0	3/3=1	0/3=0
副	0/1=0	0/1=0	0/1=0	0/1=0	0/1=0	1/1=1

推断：

给定句子，求出最可能的词性标注序列。

Viterbi算法（动态规划）



求概率最大的路径，定义动态规划问题：

1. dp[i][t]，二维数组，表示从句子开始到第t个词并且此时词性是i的最佳路径得分。
2. 递归计算：

   dp[i][t] = max{dp[i'][t-1] * A[i'][i]} * B[i][w_t]; (i' 遍历所有可能词性)

   为防止连乘产生数值问题，用log似然:

   dp[i][t] = max{dp[i'][t-1] + log A[i'][i]} + log B[i][w_t]
3. 初始值：

   dp[i][0] = log pi[i] + log B[i][w_0]
4. 计算顺序，从t = 1开始，从左往右，从上到下。

O(T*K^2), O(T*K)