任务描述

Part-of-speech tagging

- This data set contains one month of Chinese daily which are segmented and POS tagged under Peking Univ. standard.
- Project ideas:
  - Design a sequence learning method to predicate a POS tags for each word in sentences.
  - Use 80% data for model training and other 20% for testing (or 5-fold cross validation to test learner's performance. So it could be interesting to separate dataset.)

输入输出

输入：做好词性标注的单词序列，用于测试的单词序列
输出：输出标注结果

方法描述

汉语词性标注方案主要有基于统计模型的词性标注方法、基于规则的词性标注方法和统计方法与规则方法相结合的词性标注方法。 本方案采用隐马尔科夫模型。隐马尔科夫模型是一种生成式模型，本方案中模型的观测序列为单词序列，状态序列为词性序列。

预处理

由于注音的出现没有规律可循，将其删除。

训练

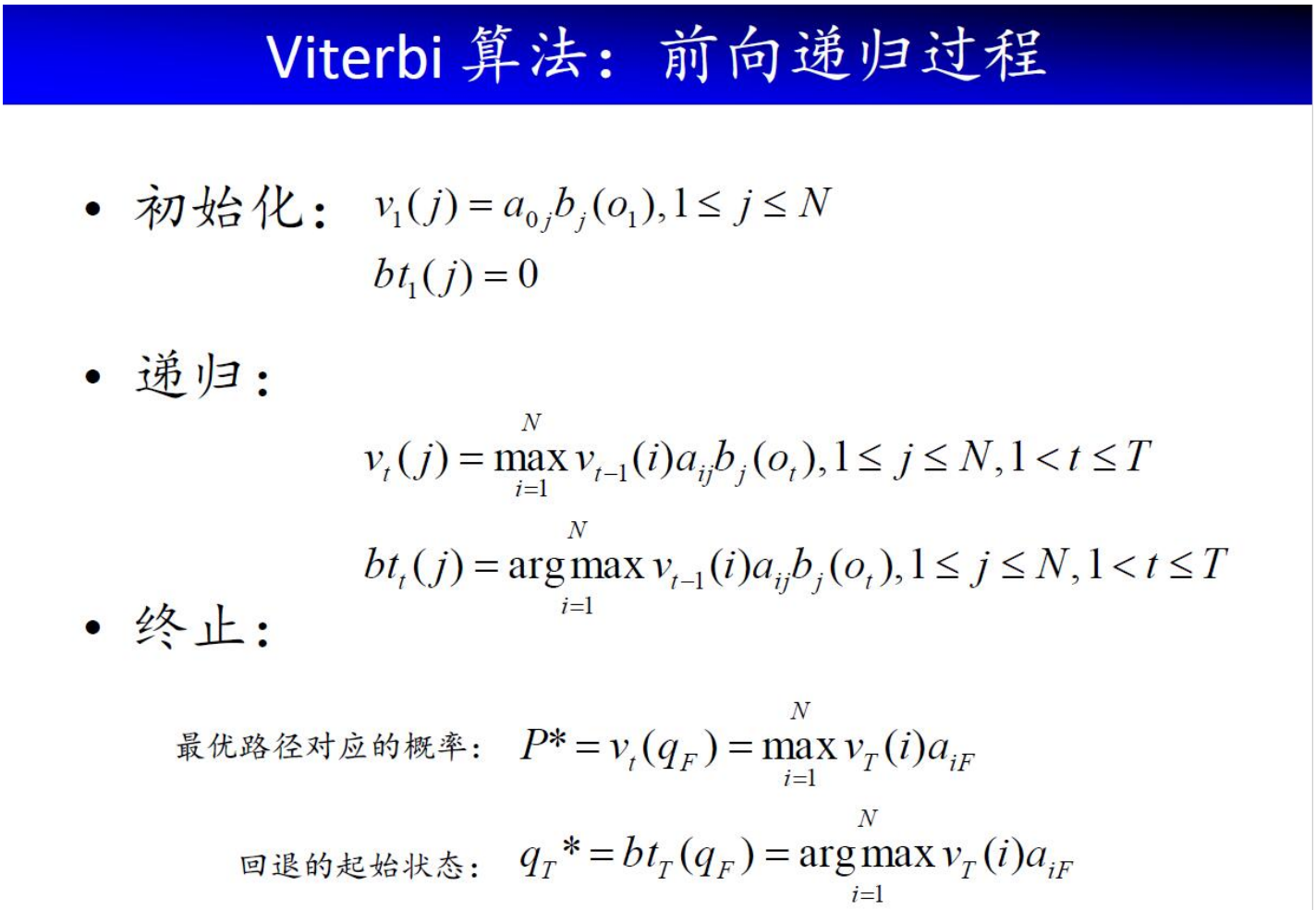
将北大语料库中的分词并进行过词性标注的长达一个月的数据作为训练集。训练得到转移概率矩阵A，发射矩阵B和开始概率矩阵π。

参数	意义	训练方法
A	词性转移概率矩阵	通过最大似然概率估算词性X转移到词性Y的概率
B	处于状态sj并且输出观察值wk的次数的概率	统计每一个词性发射到各单词的频次

π	在 t = 1 时处于状态si的次数的概率	统计各词性作为句子开头的概率

Viterbi算法

Viterbi算法公式如下：



其中

参数	含义	使用
a_{ij}	从状态i到状态j的转移概率。	从一个单词转移到另一个单词的概率
$b_{j(o_t)}$	由状态i产生观测ot的发射概率。	单词i带有词性ot的概率
$v_t(j)$	t时刻时到达状态j的viterbi路径的概率。	初始化时将$v_1(j)$作为单词j作为句子开头的概率

优化

Viterbi算法中的递归在连乘过程中会不断趋近于0，造成下溢。我使用$\log_2$函数对其进行单调性的等价。

结果分析

...
输入序列：那/rz 音韵/n 如/v 轻柔/a 的/ud 夜风/n ，/wd
输出序列：那/rz 音韵/n 如/v 轻柔/a 的/ud 夜风/n ，/wd
成功率：1.0
输入序列：惊/v 溅/v 起/vq 不可言传/i 的/ud 天籁/n 。/wj
输出序列：惊/v 溅/v 起/vq 不可言传/i 的/ud 天籁/n 。/wj
成功率：1.0
输入序列：怀/Ng 揣/v 这/rz 如泣如诉/i 的/ud 呵护/vn ，/wd
输出序列：怀/v 揣/v 这/rz 如泣如诉/n 的/ud 呵护/vn ，/wd
成功率：0.714
输入序列：才/d 发觉/v 已/d 迷失/v 了/ul 来路/n 。/wj
输出序列：才/d 发觉/v 已/d 迷失/v 了/ul 来路/n 。/wj
成功率：1.0
总成功率：0.9285660522923782

成功率较为理想。

源码运行环境

OS: Windows 10 Version 1709
Python SDK Version: 3.6.3
IDE: PyCharm 2017.3.2