Part-of-speech tagging

实验说明文档

姓名：谭逸佳

学号：2015211207

任务定义：

利用给定的已经标注语料，得到一个词性标注模型，对测试集中数据进行标注，并分析结果(Predicate a POS tags for each word in sentences.)。

方法描述:

词性标注采用隐马尔可夫模型(Hidden Markov Model)，将给定语料中的80%作为测试集，20%作为训练集得到标注模型。将输入的未标注的词语看做**观测状态序列**，词性看做**隐藏序列**，在训练集中通过对每个词性进行分析，得到其状态转移概率和到可见观测状态的概率(发射概率)。训练结束后，将测试句子的词语作为输入，首先计算第一个单词出现概率最大的状态作为初始状态，采用 Viterbi 算法，当前状态序列得到的概率始终为最大概率，句子序列每分析一个单词都得到一个**最大的隐藏序列**(词性标注)概率，一直到句子分析结束，最后得到的状态序列即为使得句子词性标注准确度最高的结果。

输入与输出:

以某一个测试句子为例，先从输入句子的第一个词“戏曲”进行分析，从每个词性的发射概率得到使得“戏曲”出现概率最大的状态下的概率作为初始概率，当前词性标注作为**初始状态**进行下一步分析。分析下一个词语“学校”，我们需要 1 状态下的标注 S 到每一个标注 S‘的概率 P_1，以及标注 S’到词语“学校”的发射概率 P_2，计算 P_1* P_2的结果并找到**最大值**，此时的状态 2 下的标注为 S’，以此类推，得到所有的标注状态。

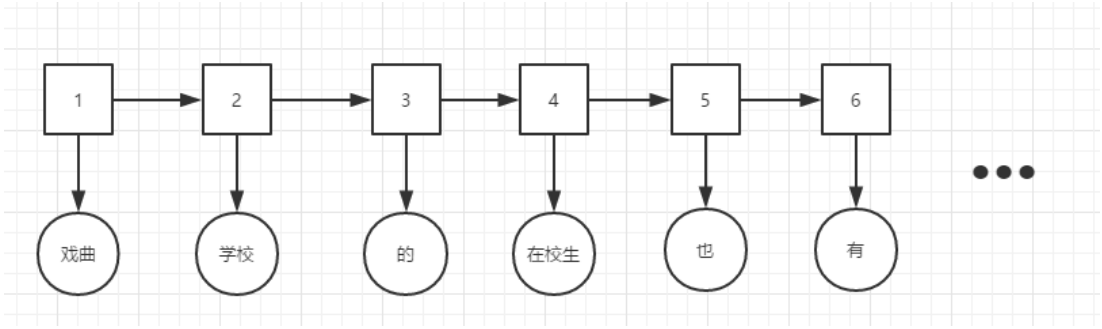

图 1：输入词语的转移状态与发射状态

以词性"n"为例，计算得到该词性的转移概率为（部分）

{'ud': 0.08065424586126542, 'wp': 0.0013197565904999613, 'wkz': 0.00684175069301026, 'm': 0.016913867935989046, 'wu': 0.036914530575281344, 'n': 0.16750969109964992, 'nrf': 0.01826675649110407, 'wky': 0.0103482169481043, 'k': 0.003384982384838813, 'vn': 0.054491037803571625, 'n]nt': 0.01090041635835533, 'rz': 0.004102841618165152, 'wd': 0.11605575005245894, 'f': 0.03439097927043414, 'd': 0.04270158039471214, 'v': 0.10496758589461827, 'a': 0.01770903508675053, 'wj': 0.06888687642881597, 'mq': 0.0005411554220460092,……

发射概率为（部分）

'[黑岛镇': 3.875413215934149e-06, '[粟子房镇': 3.875413215934149e-06, '永记': 3.875413215934149e-06, '令安': 3.875413215934149e-06, '秉广': 3.875413215934149e-06, '向军': 3.875413215934149e-06, '19980114-05-011-001': 3.875413215934149e-06, '调流': 3.875413215934149e-06,……

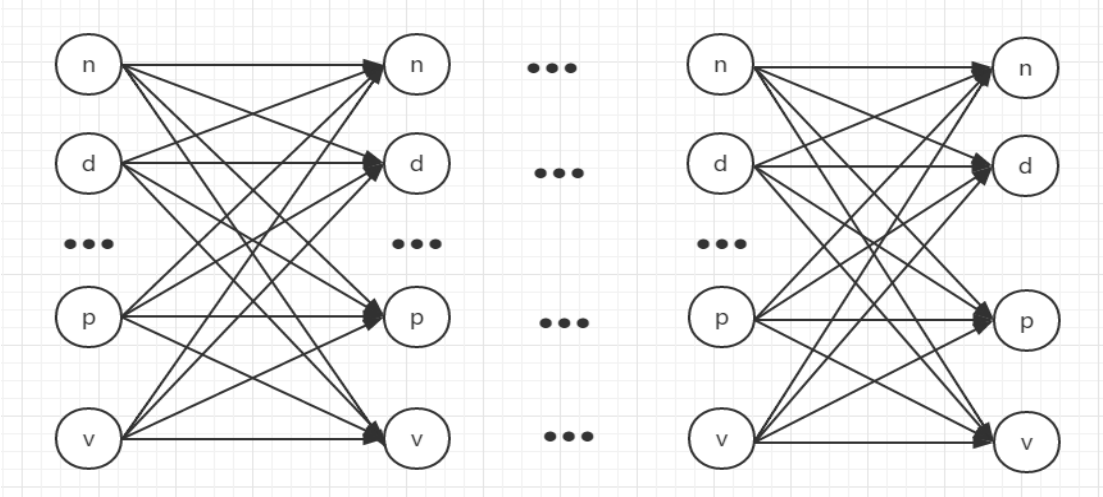

图 2：词性转移组合

word	test	result
19980124-10-003-004	m	jb]nz
戏曲	n	n
学校	n	n
的	ud	ud
在校生	n	n
也	d	d
有	vx	v
机会	n	n
到	v	v
境外	s	s
进行	vx	vx
文化	n	n
交流	vn	vn
。	wj	wj
如	v	v
[中国	ns	ns
京剧	n	n
艺术	n	n
基金会	n]nt	n
在	p	p
１９９４年	t	t
组织	v	v
[北京市	ns	ns
戏曲	n	n
学校	n]nt	n
、	wu	wu
[天津市	ns	ns
艺术	n	n
学校	n]nt	n
、	wu	wu
[上海市	ns	ns
[上海市	ns	ns
戏曲	n	n
学校	n]nt	n
的	ud	ud
部分	m	n
优秀	a	a
学生	n	n
赴	v	v
香港	ns	ns
演出	v	v
,	wd	wd
并	c	c
取得	v	v
成功	an	a
。	wj	wj
演员	n	n
们	k	k
在	p	p
香港	ns	ns
的	ud	ud
轰动	vn	vn
又	d	d
引起	v	v
台湾	ns	ns
文化界	n	n
的	ud	ud
关注	vn	vn
。	wj	wj
１９９５年	t	t
９月	t	t
他们	rr	rr
应邀	vi	vi
赴	v	v
台湾	ns	ns
演出	v	v
,	wd	wd
小	a	a
演员	n	n
们	k	k
规范	a	v
熟悉	v	v
的	ud	ud
技艺	n	n
,	wd	wd
又	d	d
征服	v	v
了	ul	ul
台湾	ns	ns
同胞	n	n
。	wj	wj
我们	rr	rr
还	d	d
应{ying4}	v	v
[台湾	ns	ns
复兴	nz	vn
剧校	n]nt	n
邀请	v	v
多次	mq	mq
派出	v	v
教员	n	n
赴	v	v
台	jb	n
教学	vi	n
,	wd	wd
不但	c	c
培养	v	v
了	ul	ul
艺术	n	n
人才	n	n
,	wd	wd
同时	c	c
加强	v	v
了	ul	ul
两岸	n	n
同胞	n	n
在	p	p
艺术	n	n
上{shang5}	f	f
的	ud	ud
交流	vn	vn
,	wd	wd
增进	v	v
了	ul	ul
友谊	n	n
。	wj	wj

图 3：词性标注的测试结果与标准结果的比对（部分）

经过训练得到的模型在测试集上得到的每个句子的平均准确率(Accuracy)为

Accuracy:0.7855756237359307

经过训练得到的模型在语料中所有句子上得到的平均准确率(Accuracy)为

Accuracy:0.8765127325092565

结果分析与性能评价:

以图 3 中结果为例，在第一个 19980124-10-003-004 词语中，正确标注为“m”，而模型却将其错误地标注为“jb]nz”。原因是在整个训练语料中，“jb]nz“标注仅出现了一次“英/jb]nz”，所以当计算发射概率时，对每一个标注状态的发射概率进行**平滑处理**(Laplace Smoothing)后，“jb]nz“的分母较小，并且 19980124-10-003-004 词语在训练集中从未出现，平滑后分子均为 1，所以使得标注 jb]nz“得到 19980124-10-003-004 的概率最大，造成了错误。

另外一种错误情况例如图 2 中把词性本该为“n]nt”的“基金会”标注成了“n”。原因是在训练语料中，“基金会”这个单词本身有“n]nt”和“n”两种词性，由于训练语料的客观原因，造成转移概率与发射概率乘积在两种词性中大小关系有所不同，可以说，造成标注错误的很多结果都是由于**该词语在语料存在多种词性，而训练语料内容不够充分，使得标注结果有一定的特殊性，存在误差。**将词性为“n]nt”的“学校”标注成了“n”、将词性为“n]nt”的“剧校”标注成了“n”、词性为“nz”的“复兴”标注成了“vn”都是由于此类原因。

当模型应用于语料中所有句子时，由于模型由其中部分语料训练得到，所以对于已存在的句子的预测性增强，词性标注的准确率也会相应提高。

若将每个词语出现次数最多的词性作为测试集中的标注，得到准确率

Accuracy:0.88152388357654

原因是测试集中词语内容与训练集相差不大，大部分词性都**没有发生变化**，所以采用频率最高的词性会有很高的准确率，但当词语的词性变化较多时，这种方法准确率会下降。

源码运行环境:

Python3.6