**隐马尔可夫模型（1）**

词性标注的任务是：给定词序列W=w1,w2,…,wn，寻找词性标注序列T=t1,t2,…,tn，最大化概率为P(t1,t2,…,tn|w1,w2,…,wn)。

使用贝叶斯公式重新描述这个条件概率：

P(t1,t2,…,tn) P(w1,w2,…,wn|t1,t2,…,tn)/P(w1,w2,…,wn)

忽略掉分母P(w1,w2,…,wn)，同时做独立性假设，使用N元模型近似计算P(t1,t2,…,tn)。例如使用二元连接，则有：

|  |
| --- |
| <http://images.51cto.com/files/uploadimg/20110615/155533600.jpg> |
|  |

近似计算P(w1,w2,…,wn|t1,t2,…,tn)：假设一个类别中的词独立于它的邻居，则有：

|  |
| --- |
| <http://images.51cto.com/files/uploadimg/20110615/155619133.jpg> |
|  |

寻找最有可能的词性标注序列实际的计算公式为：

|  |
| --- |
| <http://images.51cto.com/files/uploadimg/20110615/155711663.jpg> |
|  |

这里把词w叫做显状态，词性t叫做隐状态。条件概率P(ti|ti-1)叫做转移概率，条件概率P(wi|ti)叫做发射概率。

抽象来说，基本的马尔可夫模型中的状态之间有转移概率。隐马尔可夫模型中有隐状态和显状态。隐状态之间有转移概率。一个隐状态对应多个显状态。隐状态生成显状态的概率叫做生成概率或者发射概率。在初始概率、转移概率以及发射概率已知的情况下，可以从观测到的显状态序列计算出可能性最大的隐状态序列，这个算法叫做维特比（Viterbi）算法。对于词性标注的问题来说，显状态是分词出来的结果--单词W，隐状态是需要标注的词性C。词性之间存在转移概率。词性按照某个发射概率产生具体的词。可以把初始概率、转移概率和发射概率一起叫做语言模型。因为它们可以用来评估一个标注序列的概率。采用隐马尔可夫模型标注词性的总体结构如图4-11所示。

|  |
| --- |
| <http://images.51cto.com/files/uploadimg/20110615/155755956.jpg> |
| 图4-11  词性标注中的隐马尔可夫模型 |

下面举例说明隐马尔可夫模型。假设只有词性：代词（r）、动词（v）、名词（n）和方位词（f）。

有如下一个简化版本的语言模型描述如下：

1. start: go(r,1.0) emit(start,1.0)
2. f: emit(来,0.1) go(n,0.9) go(end,0.1)
3. v: emit(来,0.4) emit(会,0.3) go(f,0.1)  
   go(v,0.3) go(n,0.5) go(end,0.1)
4. n: emit(会,0.1) go(f,0.5) go(v,0.3) go(end,0.2)
5. r: emit(他,0.3) go(v,0.9) go(n,0.1)

这里的start和end都是虚拟的状态，start表示开始，end表示结束，emit表示发射概率，go表示转移概率。语言模型中的值可以事前统计出来。中文分词中的语言模型可以从语料库统计出来。

这个语言模型的初始概率向量如表4-5所示。

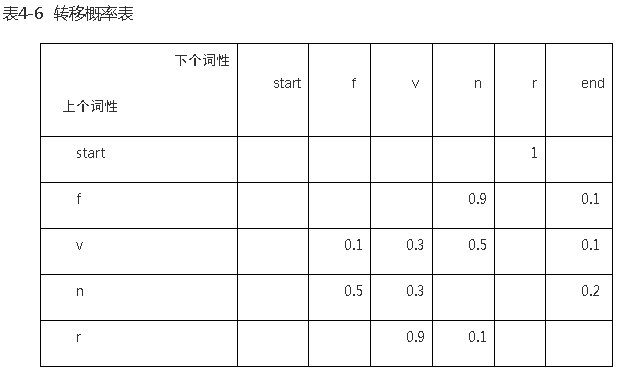
表4-5  初始概率表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | r | n | f | end |
| start1 | 1.0 | 0 | 0 | 0 |

这个初始概率的意思是，代词是每个句子的开始。

转移概率矩阵如表4-6所示。

表4-6  转移概率表



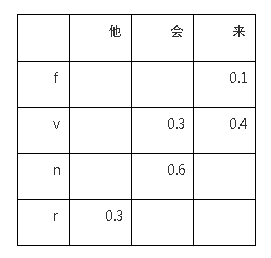
例如第3行表示动词后是名词的可能性比较大，仍然是动词的可能性比较小，所以上个词性是动词，下一个词性是名词的概率是0.5；而上个词性是动词，下一个词性还是动词的概率是0.3。

以"他会来"这句话作为例，分词后的输入是：[start] [他] [会] [来] [end]。考虑到某些词性更有可能作为一句话的开始，有些词性更有可能作为一句话的结束，这里增加了开始和结束的虚节点[start]和[end]。这句话的转移概率图如图4-12所示。

|  |
| --- |
| <http://images.51cto.com/files/uploadimg/20110615/155948725.jpg> |
| 图4-12  转移概率图 |

发射概率（混淆矩阵）如表4-7所示。

表4-7  发射概率表



以发射概率表的第二行为例：如果一个词是动词，那么这个词是"来"的概率比"会"的概率大。例句的发射概率图如图4-13所示。

|  |
| --- |
| <http://images.51cto.com/files/uploadimg/20110615/160026944.jpg> |
| 图4-13  发射概率图 |

每个隐状态和显状态的每个阶段组合成一个如图4-14所示的由节点组成的二维矩阵。

|  |
| --- |
| <http://images.51cto.com/files/uploadimg/20110615/160101286.jpg> |
| 图4-14   维特比求解格栅 |

**隐马尔可夫模型（2）**

采用动态规划的方法求解最佳标注序列。每个词对应一个求解的阶段，当前节点概率的计算依据是：

上一个阶段的节点概率；

上一个阶段的节点到当前节点的转移概率；

当前节点的发射概率。

动态规划的思想产生了维特比算法。维特比求解方法由两个过程组成：前向累积概率计算过程和反向回溯过程。前向过程按阶段计算。从图上看就是从前向后按列计算，分别叫做阶段"start"、"他"、"会"、"来"、"end"。

在阶段"start"计算：

Best(A) = 1

在阶段"他"计算：

Best(B) = Best(A)   P(r|start)   P(他|r) = 1 1 0.3=0.3

在阶段"会"计算：

Best(C)=Best(B)   P(v|r)   P(会|v) = 0.3 0.9 0.3= 0.081

Best(D)=Best(B)   P(n|r)   P(会|n) = 0.3 0.1 0.6= 0.018

在阶段"来"计算：

Best(E) = Max [Best(C)  P(v|v), Best(D)  P(v|n)]   P(来|v) = 0.081 0.3 0.4= 0.00972

Best(F) = Max [Best(C)  P(f|v), Best(D)  P(f|n)]   P(来|f)= 0.081 0.1 0.1= 0.00081

在阶段"end"计算：

Best(G) = Max [Best(E)  P(end|v), Best(F)  P(end|f)]   P(</s>|end)= 0.00972 0.1 1= 0.000972

执行回溯过程发现最佳隐状态（粗黑线节点），如图4-15所示。

G的最佳前驱节点是E，E的最佳前驱节点是C，C的最佳前驱节点是B，B的最佳前驱节点是A。所以猜测词性输出：他/r  会/v  来/v。这样消除了歧义，判断出[会]的词性是动词而不是名词，[来]的词性是动词而不是方位词。

|  |
| --- |
| <http://images.51cto.com/files/uploadimg/20110615/160325868.jpg> |
| （点击查看大图）图4-15  维特比求解过程 |

参考资料：

<http://book.51cto.com/art/201106/269055.htm>