**第八章读书报告**

09118223 吴亦珂

读书进度：第八章读完

1. **问题列表**

（我提出）P21页8.6中label bias、observation bias应该是什么意思？下面举的例子是如何说明这一点吗？

讨论结果：根据后面的to,TO标签之前更多的是NN，而不是MD，因此标签给的是NN。但是如果从，<s>的角度考虑，其之后应该是MD的概率更大，而且这个概率接近1，最终导致预测结果为MD，导致了一个错误的预测。而强信息to的标签必须是TO，则由于已经用于预测出其标签为TO，因此导致第一种解释被忽视。

（我提出）8.7中说的rich languages具体指什么问题吗？是指某些语言的含义比英语丰富，导致tag的种类非常多吗？

讨论结果：不同语言蕴含的信息不同，因此如果是其他语言，可能要标注的信息就会非常多，导致tag种类增加。

（我提出）将原来的bigram拓展为trigram为什么要增加tn+1？

讨论结果：本来应该不存在以及的，但是为了增加预测的准确性，如哪些单词出现在句首或者句末，另一方面也可以使得分类器知道句子的结束位置。

（别人提出）beam search的优势和不足是什么？还可以用在哪些应用中？

讨论结果：beam search就是每次只选择概率最大的k个（人为设定），相比于穷举策略，不会存储所有可能，但相较于贪心算法，多了一些情况，相当于两者的折中。Beam Search多用在一些大型系统中，比如机器翻译系统，语音识别系统等，因为这些系统中的数据集可能非常大，而且结果也没有唯一正确的解，系统用最快的方式找到最接近正确的解才是系统的目标。

（别人提出）HMM,MEMM,CRF三者之间有何异同？

讨论结果：HMM模型是对转移概率和表现概率直接建模，统计共现概率。MEMM模型是对转移概率和表现概率建立联合概率，统计时统计的是条件概率，但MEMM容易陷入局部最优，是因为MEMM只在局部做归一化。CRF模型中，统计了全局概率，在 做归一化时，考虑了数据在全局的分布，而不是仅仅在局部归一化，这样就解决了MEMM中的标记偏置（label bias）的问题。CRF没有HMM那样严格的独立性假设条件，因而可以容纳任意的上下文信息。同时，由于CRF计算全局最优输出节点的条件概率，它还克服了最大熵马尔可夫模型标记偏置（Label-bias）的缺点。CRF是在给定需要标记的观察序列的条件下，计算整个标记序列的联合概率分布，而不是在给定当前状态条件下，定义下一个状态的状态分布。但是CRF模型较为复杂存在训练代价高，复杂度高的问题。

1. **读书收获**

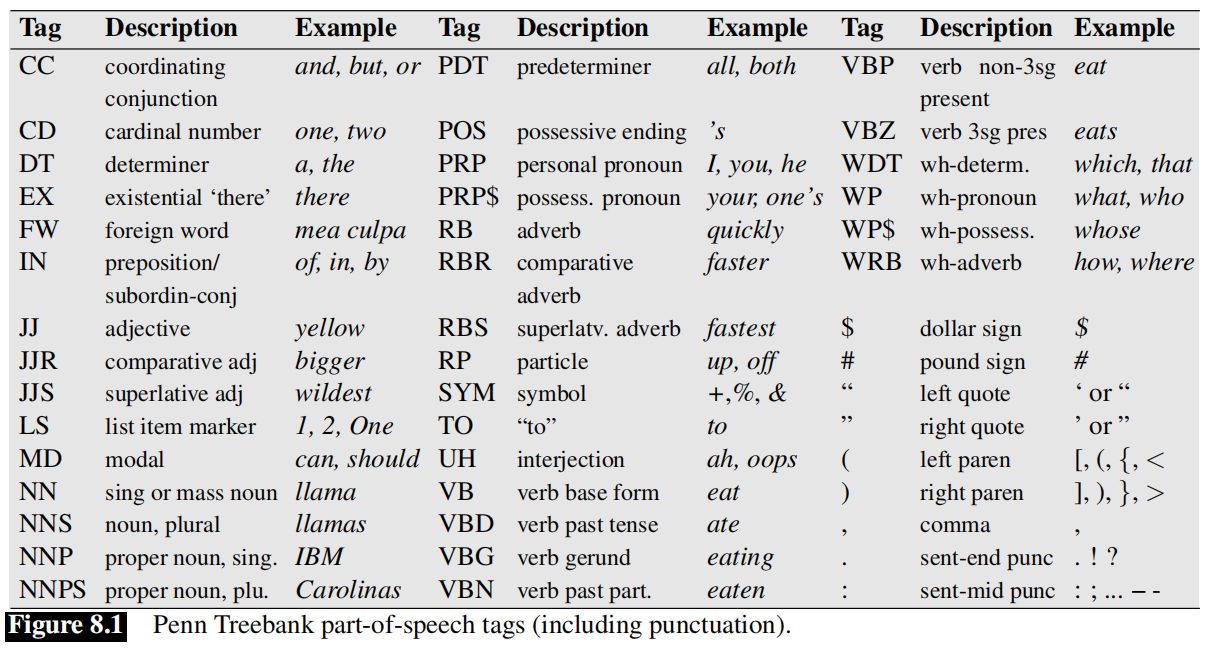
**8.1 (Mostly) English Word Classes**

主要介绍了英语中的各种词性。可以将英语中的词性分为opened class,以及closed class。

Opened class主要分为名词，动词，形容词和副词。

Closed class主要分为介词，小品词，限定词，连词，代词，助动词，数词。

**8.2 The Penn Treebank Part-of-Speech Tagset**

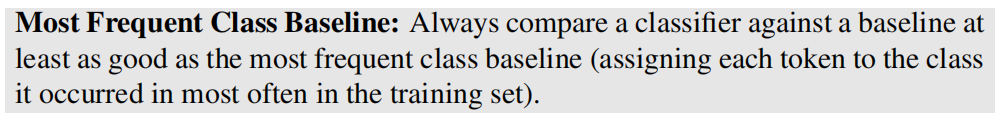


**8.3 Part-of-Speech Tagging**

词性标注：输入一序列单词和一个标签集合，输出是一序列标签。

标注是一个消除歧义的工作，有一些词语是有歧义的，可能会有多个标签。有歧义的单词只有14%—15%，但是很多都是常见的单词，因此55%—67%的单词在文章中都是有歧义的。

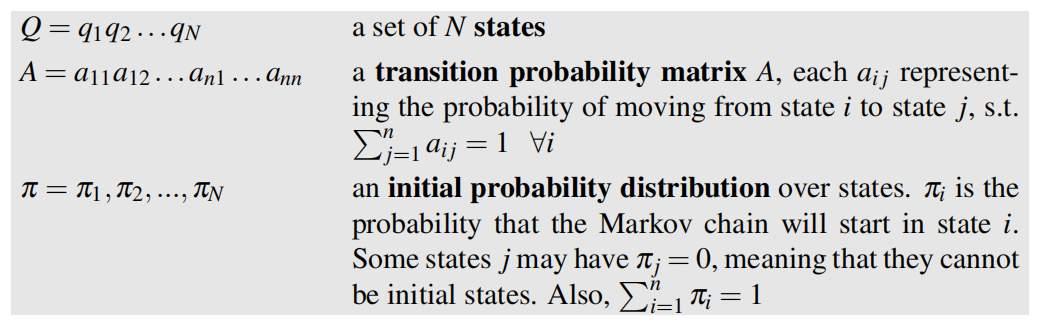
最简单的消除歧义的方法是选择在训练语料库中最常的标签。



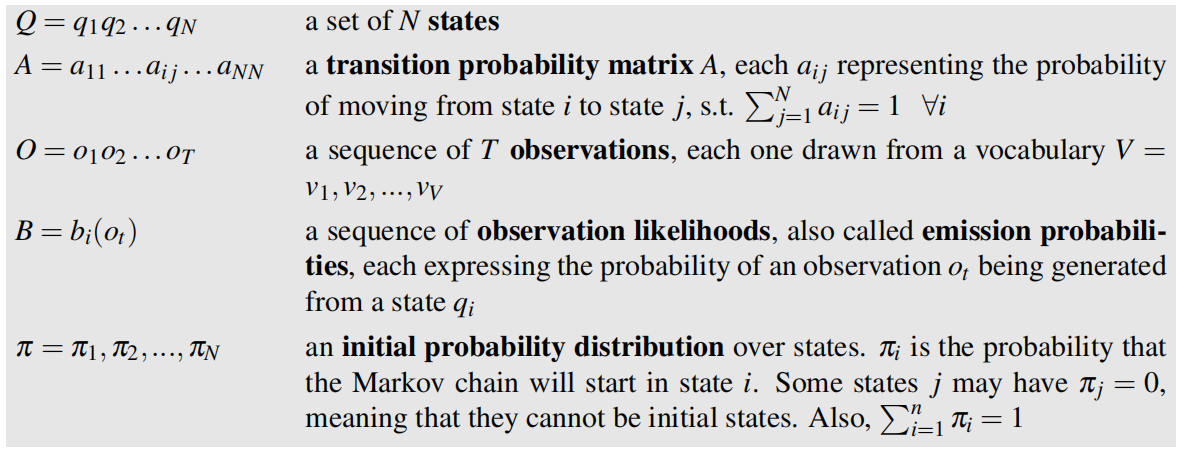
**8.4 HMM Part-of-Speech Tagging**

马尔可夫链：马尔可夫假设：对未来进行预测的时候，只取决于现在的状态而与过去的状态无关。



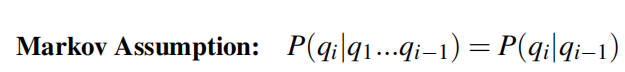


隐马尔可夫模型：



进行了两个假设：

现在的状态仅仅取决于之前的状态：

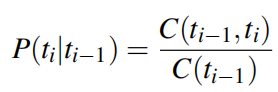


条件独立假设：

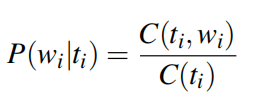


HMM tagger的组成部分：

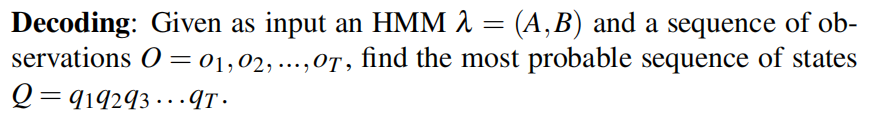
A：转移概率矩阵



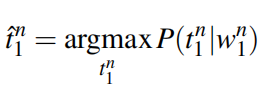
B：输出概率矩阵：



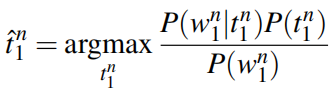
HMM tagging as decoding：



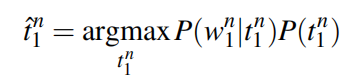
用数学语言表示：



根据贝叶斯公式展开：

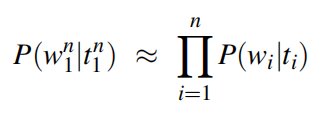


考虑到分母是一样的：

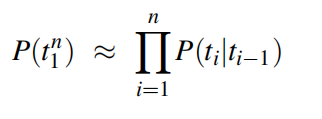


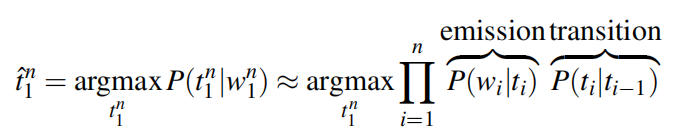
HMM tagger做了两个简单的假设（和HMM一致）：

一个单词出现的概率只取决于它自己的标签，和相邻的单词和标签独立：

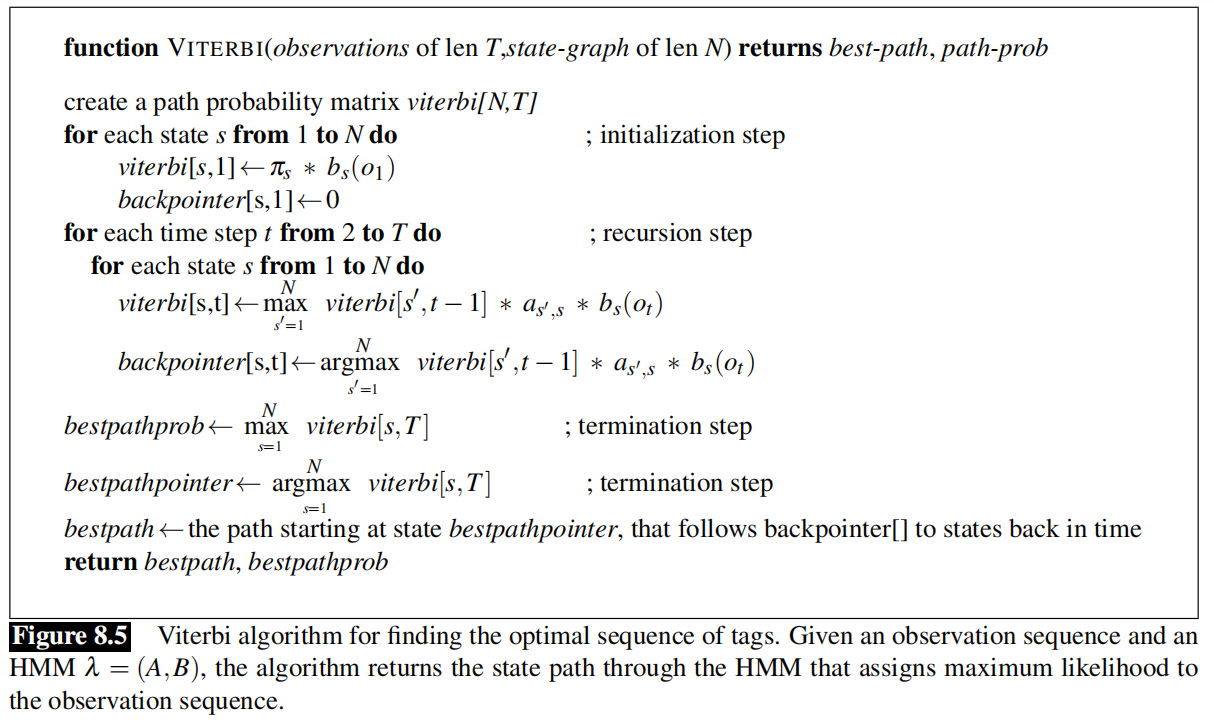


Bigram的假设，一个标签的概率只取决于前一个标签：



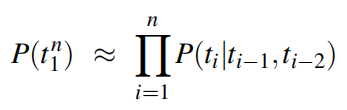


维特比算法：



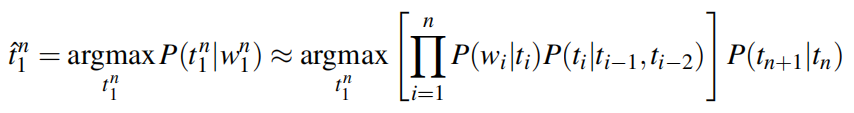
Extending the HMM Algorithm to Trigrams ：

将原来依赖于前一个标签的假设改为两个标签。

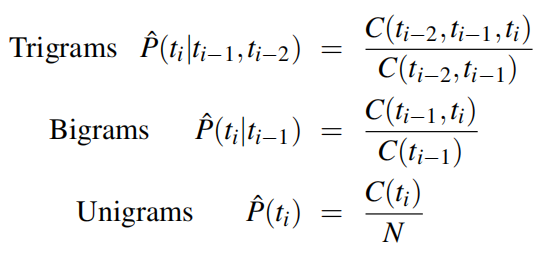


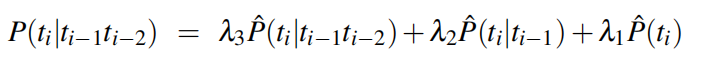
对维特比算法有较大的改动，需要考虑之前两部的标签这就需要而不是N个隐藏状态。

并且还需要让算法知道句子的结束，因此增加：



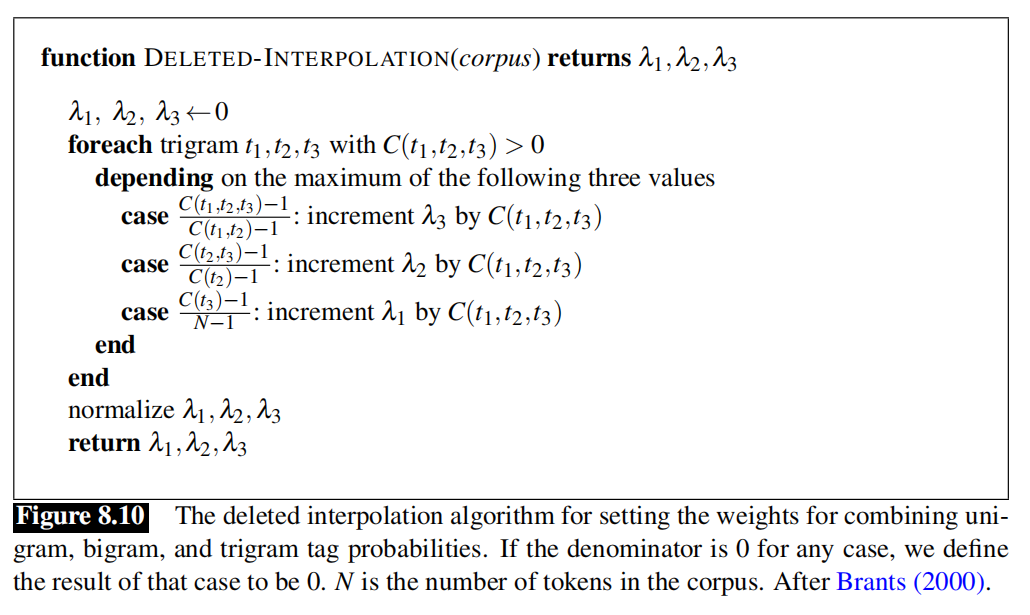
对于该模型，一个很严重的问题就是数据的稀疏性，因此可以采用线性插值的方法解决：





Beam search：

考虑到维特比算法效率较低的问题，因此采用束搜索的方法。在每一步只选择最好的几个状态进一步进行迭代，其余的不进行考虑。



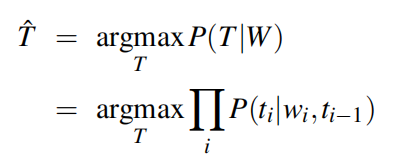
Unknown words：

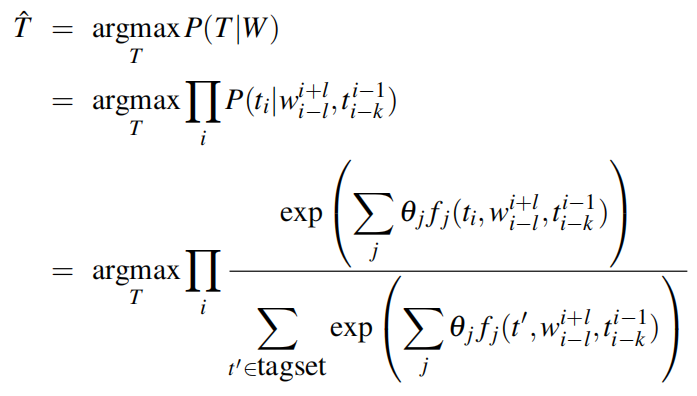
一种简单的方法是可以根据单词的一些特点进行判断，如首字母大写可以判断为专有名词。

还可以根据词语的后缀，如ed结尾应该是动词。

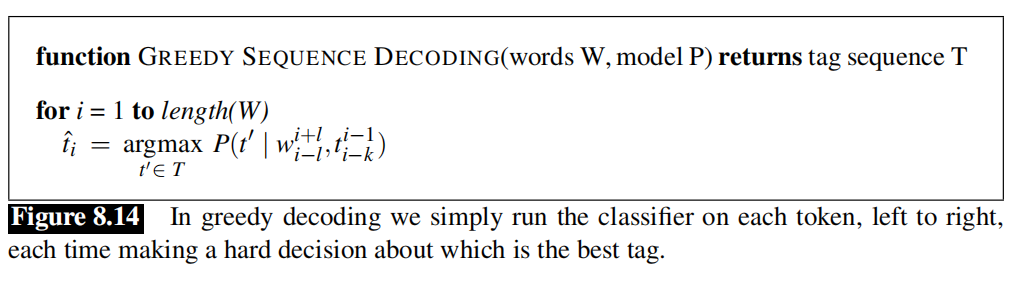
**8.5 Maximum Entropy Markov Models**

MEMM模型直接计算P(T|W)的值，相当于是一个判别模型。





可以使用这个公式从左到右依次计算出最大概率的标签，每个单词都赋予唯一的一个类，也就是使用贪心算法：



贪心算法虽然很快，但是很大的缺点是，没有用到之后决策的信息，因此可以采用维特比算法。

**8.6 Bidirectionality**

HMM以及HEMM的问题是计算仅仅是从左向右的，尽管维特比可以间接使用之后决策的信息，但是如果能实现直接使用知乎决策的信息将会更好。

可以采用条件随机场(CRF)，或者对现有方法进行改进，如使用bidirectional版本的维特比算法。并且任何一个序列模型可以改为bidirectional model，可以先从左向右执行一遍算法，在从右向左执行一遍。