**读书报告**

王贵涛

自然语言处理的第八章主要介绍了词性标注。

**英语单词分类（English Word Classes）**

词类可以分为两大类：封闭类和开放类。封闭类是指成员相对固定的类。四种主要的开放类有：名词、动词、形容词和副词。开放类名词分为专有名词和普通名词。

**宾州树库词性标签集（The Penn Treebank Part-of-Speech Tagset）**

标签集在每个单词后加上标签来表示，标签由斜线分隔。标签集是由一个更大的标签集折叠而来的，丢失了一些潜在区别。不过penn标签集是为树库设计的，因此忽略的语法信息可以从解析树中恢复。

**词性标注（Part-of-Speech Tagging）**

词性标注是为每个单词分配一个标签，输入单词序列和标签集，输出每个单词的标签。

加标签是一项消除歧义的任务，许多单词是模棱两可的，可能有不同的词性。

尽管如此，许多单词很容易消除歧义，因为它们每个标签的概率不同，只需选取概率最大的标签。将最频繁分类作为基线，将其他分类器与基线作比较。

**隐马尔可夫模型词性标注（HMM Part-of-Speech Tagging）**

HMM是一种概率序列模型，给定一个序列，它计算可能的标签序列概率分布，并选择最佳的标签序列。

对于马尔可夫模型来说，预测将来只基于现状，而过去没有影响。

许多情况下，我们感兴趣的事件是隐藏的，并不能直接看到。隐马尔可夫模型讨论的是观测到的事件和隐藏事件。

HMM有两个组成部分，A概率和B概率。A矩阵包含标签转移矩阵，给出标签转移的概率，用两个标签相接的频率除以第一个标签的次数计算。B矩阵表示给定一个标签，与一个单词相关的概率，依然用频率计算。

给定一个观测序列，确定与观测序列相对应的隐藏序列称为解码（decoding）。

HMM标签法作了两个假设：一是单词出现的概率只依赖于自己的标签，与邻近的单词和标签无关，二是一个标签的出现只依赖于前一个标签，与整个标签序列无关。

HMMs的译码算法称为维特比算法（Viterbi algorithm），是一个动态规划算法。

HMM算法可以扩展到三元标签，略微提高性能，不过每次需要观测N²个隐藏状态。也可以增加序列结束标签，知道句子结束的位置。三元标签可能在测试集中出现训练集从未出现过的序列标签，因此可以使用线性插值算法，通过单元概率、双元概率和三元概率的加权和来替代。

在状态数增长非常大时，普通的维特比算法速度很慢。定向搜索可以解决复杂性的问题，我们不需要在每个时间点保留所有的状态，只需要保留最好的几个假设，可以是固定的状态数，或者状态数的百分比，或者设定一个概率阈值。

对于生词来说，专有名词通常首字母缩写，大写字母开头，复数名词以s结尾，过去分词以ed结尾，形容词以able结尾。未知单词不太可能是封闭词类，所以只能计算训练集频率小于10或开放类的单词的后缀概率。

**最大熵马尔可夫模型（Maximum Entropy Markov Models）**

虽然HMM精度很高，但需要处理生词，回退和后缀。将逻辑回归转换为一个有区别的序列模式，在连续的单词上运行它，将前一个单词的类作为下一个单词的特征，称为最大熵马尔可夫模型。

MEMM是判别模型，区别于HMM的生词模型，直接计算后验概率。但是MEMM不只计算单词和上一个状态，使用判别模型的原因是它更容易包含更多的特征。处理生词时，可以添加表达单词拼写或形状的特征。

单词形状特征将小写字母映射为x，大写字母映射为X，数字映射为d，保留符号来表示单词的抽象形式。

将逻辑回归模型转化为序列模型最简单的方法是建立局部分类器，对第一个词硬分类，对第二个词硬决策，以此类推，称为贪心译码算法，为每个单词选择最好的标签。

贪心算法无法使用未来决策，会导致精度大幅下降。我们可以同样使用维特比算法对MEMM进行解码，只需要将HMM公式中的先验概率和似然概率替换成直接后验概率，使用梯度下降法训练权值。

**双向性（Bidirectionality）**

HMM和MEMM模型都是从左向右运行，如果可以直接使用未来标签将带来好处。MEMM有定理上的弱点，称为标签偏差，忽略了一种信息来源，如前缀概率。

实现双向性的一种方法是切换到一个更强大的模型，如条件随机场（crf）。条件随机场是一个无向图模型，它并不计算给定时间的每个标签概率，而是计算一组相关特征。训练需要计算所有标签的总和，这使得crf的训练相当缓慢。

任何序列都可以使用多个通道转换为双向模型。

**形态丰富语言的词性标注（Part-of-Speech Tagging for Morphological Rich Languages）**

在处理具有丰富形态学的语言时，有必要增加标签算法。

高度变化的语言比英语有更多的信息，如格（主格、宾格、属格）和性别（阴性、阳性）。有时一个单词有多个词性，所以标记集往往要更大。