speech and language processing 第八章报告

09017244 郑健雄

1. 自己提出的问题的理解：
2. beam search的优势和不足是什么？还可以用在哪些应用中？

讨论后的理解：beam search的优势在于每次只考虑最大概率的k种情况，比起完全使用动态规划，其可以更加节省空间，而又可以减少完全贪心算法带来的随机性问题。beam search一般可以用在一些大型系统中，比如机器翻译系统，语音识别系统中，在这种应用中，一般需要很高的实时性，并且很难完美，所以应该尽快找到接近最优解的解作为输出结果。

1. HMM,MEMM,CRF三者之间有何异同？

讨论后的理解：从模型种类上划分，HMM根据状态转移概率进行预测，是生成模型，而CRF和MEMM都是通过样本或者状态建立判别函数，是判别模型。而从图角度来看，HMM和MEMM都是基于有向图的，而CRF是无向图。其中CRF和HMM都是全局最优模型，并且CRF更加灵活，应用范围更加广泛，有着HMM和MEMM不具备的优点。

1. 别人提出的问题的理解
2. 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？

自己的理解：这个可以理解成一种倾向性，参考书中的例子，可以看出P(TO|to,twill)这个概率十分的趋近于1，无论twill的类型是什么都是如此，那么twill的种类信息就会因为这种TO标签和to之间强的关联而被忽略，那么TO和NN这种类型的关联性就会被掩盖，而<s>标签更加倾向于MD类型，所以NN在整体的概率会低于MD,从而预测出错误的结果。

1. 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？

自己的理解：直观上应该可以理解成一种语言的复杂程度，比如POS的类型数量，未知词数量(比如中文组合出的各种词语)，组合模式，以及所有格，性别(gender，比如德语似乎会区分这种状态)，我的理解是它们会导致要标注的信息变得非常多，并且处理过程更加复杂，比如要标注gender，POS等等标签，所以也可以理解为处理的tag类型会变得很多。

1. 将原来的bigram拓展为trigram为什么增加tn+1就可以解决句子边界的问题？

自己的理解：这个直观上是一种补充的处理，因为根据8.24的等式，实际句子中是不存在t-1和t0，tn+1位置对应的单词的，其仅仅是个位置，但是这种边界信息又很重要，比如最后一个词语出现在句尾的概率大不大从而判断出现这种预测是否合理，比如结尾放了一个介词类型显然就不是很合理。以及初始概率计算时，第一个词的前两个位置如何处理。引入了这种边界标记的话，一方面可以帮助计算顺利进行，同时可以考虑词语出现在序列头部和尾部的概率信息，让其预测更加可靠

1. 读书计划

1、本周完成的内容章节：阅读第9章。

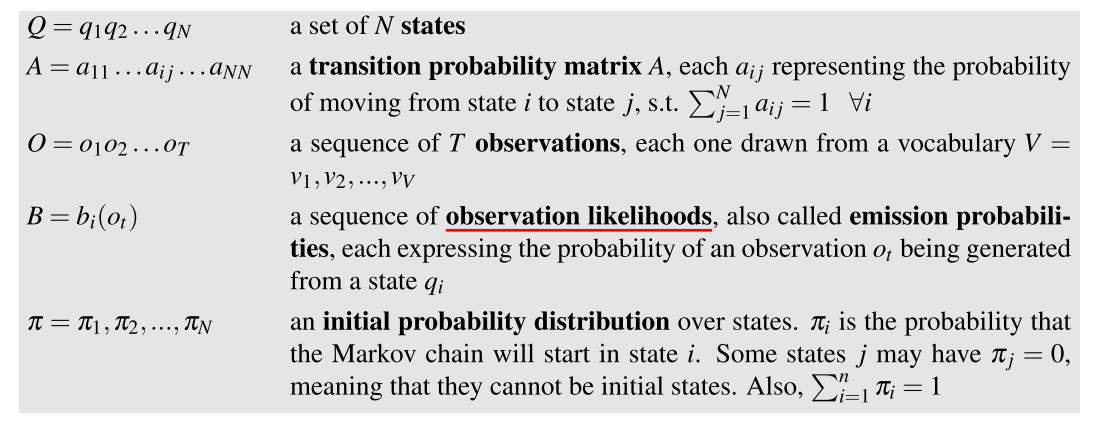
2、下周计划：阅读完speech and processing第10章内容。

四、读书摘要总结：

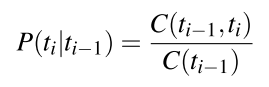
1. POS：part of speech，可以理解为句子中词语所属的类别，其对于理解句子结构有着非常重要的意义。一般而言，POS可以分为open class和closed class两种类型，并且比较经典的划分方式是根据英语的八种词类型(verb,adj)来划分。由于语言之间不同，所以这种划分并不适用于所有语言，但是其提供了一种基本的框架。实际的类型可以划分出更多的子类型，比如Penn Treebank Part-of-Speech Tagset就提供了对不同词对应的Tag类型。

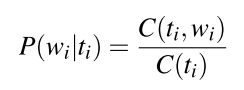
2. Part-of-Speech Tagging：由于人们希望研究句子，同时人工打标签的过于复杂，所以如果存在某种算法可以将输入的句子直接转化成为tag序列的话，就可以很好地提供学习的资料，并且用到其他NLP工作中去。tag的过程本质上是一个消除歧义的过程，提取出句子的本质构造。一种传统的思路是直接根据语料库中该单词最频繁的类型作为预测结果，这种方法有着不错的精确度。但是不太灵活。比较常用的方法有HMM,MEMM和RNN。

3. HMM：隐式马尔可夫模型，其基础是马尔可夫链，直观上可以理解为一个状态序列的每个元素仅仅依托于前一个元素，而与其他状态无关。称之为隐式是因为tag类型是隐藏变量，而观察到的是句子序列，所以需要改进，建立隐变量和句子元素之间的直接关系，所以其模型基础要素如下：

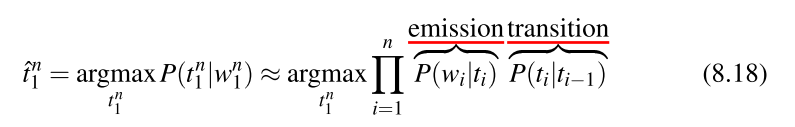


基本的概率估计还是使用频率：

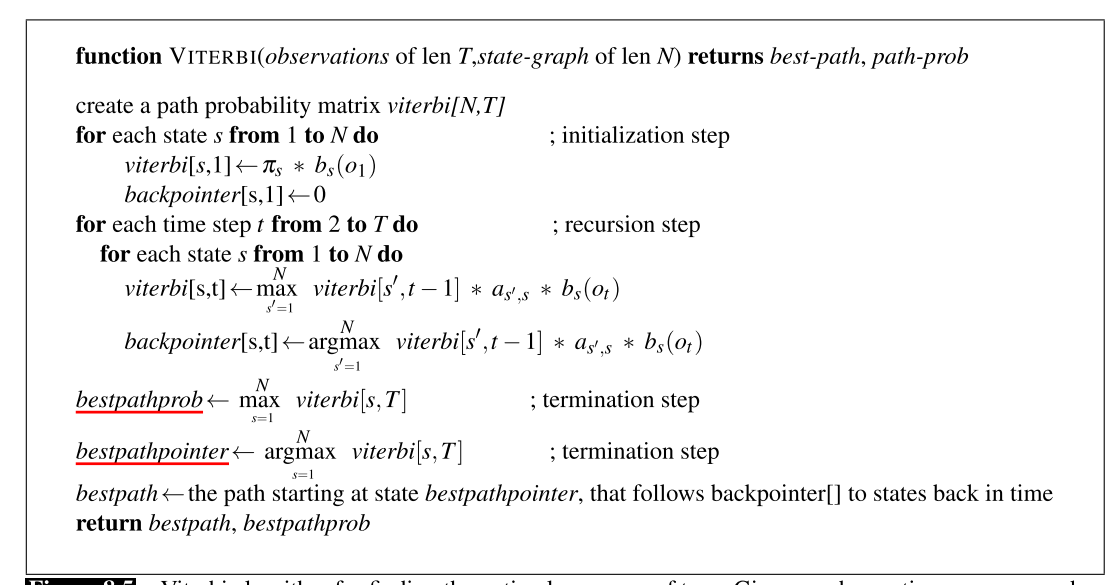




HMM根据一个句子，预测出最可能的tag序列其实是一种decoding的过程，其融合了贝叶斯的思想。



通过转换，可以看出，其通过构建词语和类别的直接预测关系以及马尔可夫链的序列预测关系，将马尔可夫链转换成了HMM。由于计算HMM的最优结果需要计算每一步的每种类别的可能性，以及前后状态的联系，所以使用基于动态规划的维持比算法来进行处理。



其可以通过穷举和记录得出最优解和路径。而一般情况下，可以用beam search减少搜索量，提升效率。HMM也可以扩充为trigram可用的情况。除此之外，MEMM也有着类似的思路。

4. Bidirectionality与Morphological Rich Languages：

考虑到序列总是从一个方向到另一个方向，这往往会忽视某个方向的信息，而这些信息对于预测结果有着很重要的效果。CRF和双向的RNN都可以对这种情况进行处理。而形态学上的rich language也是一个比较难处理的问题，有些语言由于形态，类型，性别，时态以及组合和未知词的问题，需要标注的特征过多，这也是一个需要处理的问题。

总结：POS对于理解句子的结构以及其含义有着重要的效果，其可以抽取出句子的基础结构，从而在翻译，实体识别等应用中起到重要的作用，其学习过程本质上是一种标注过程，而标注算法也可以用于识别entity等，这类算法有着广阔的应用场景。比较常用的模型有HMM,MEMM,RNN等，都有着较好的效果，同时这些算法也可以用来进行诸如机器翻译，人物识别，实体识别等工作，并且可以与其他技术结合，用在推荐系统等实际应用中。