读书报告

09118119 黄一凡

2020 年 8 月 31 日

1 别人提出的问题

1. beam search 的优势和不足是什么，还可以用在哪些应用中？

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。

2. HMM，MEMM，CRF 三者之间有何异同？

首先，CRF 是判别模型，对问题的条件概率分布建模，而 HMM 是生成模型，对联合概率分布建模。可以将 HMM 模型看作 CRF 模型的一种特殊情况，即所有能用 HMM 解决的问题，基本上也都能用 CRF 解决，并且 CRF 还能利用更多 HMM 没有的特征。CRF 可以用前一时刻和当前时刻的标签构成的特征函数，加上对应的权重来表示 HMM 中的转移概率，可以用当前时刻的标签和当前时刻对应的词构成的特征函数，加上权重来表示 HMM 中的发射概率。所以 HMM 能做到的，CRF 都能做到。另外，CRF 相比 HMM 能够利用更加丰富的标签分布信息。

3. 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？

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)，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。

4. 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？

直观上是一种补充的处理，因为根据 8.24 的等式，是不存在 t-1 和 t0，tn+1 位置对应的单词的，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，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。

2 读书计划

本周所读： 《Speeach and Language Processing》Chapter 8

下周计划： 《Speeach and Language Processing》Chapter 9