自然语言处理中的N-gram模型的Smoothing算法

在使用N-gram模型时可能引入稀疏数据问题。假设我们词表的大小是50万，则要覆盖所有的bigram情况，需要至少2500亿个词的语料，参数必然也是这个数量级；对于trigram(n=3)以及更大的n，还会更大，但是很多的词不会相邻出现，即大部分P() = 0 (稀疏)，另外，还有很多训练语料中不存在的词（OOV, out-of-vocabulary）的词。所以，如果训练语料数量不够大，或者词表不够全，得到的语言模型容易出现过拟合。

这时的解决方法，称之为平滑技术(smoothing)或者减值(discount)，其主要策略是把在训练样本中出现过的时间的概率适当减小，然后把减小得到的概率密度分配给训练预料中没有出现过的事件。实际中平滑算法有很多种：

* Additive smoothing
* Good-Turing estimate
* Jelinek-Mercer smoothing(interpolation)
* Katz smoothing( backoff )
* Witten-Bell smoothing
* Absolute discounting
* Kneser-Ney smoothing

**差值(Interpolation) vs 回退(backoff)**

先来理解下平滑方法的两种思想，一个是差值，简单来讲，就是把不同阶的模型结合起来，另一种是回退，直观的理解，就是说如果没有3gram，就用bigram，如果没有bigram，就用unigram，两者的区别：

* 差值和回退都涉及到来自较高和较低阶模型的信息。
* 主要区别：在决定非零计数的n-gram概率时，差值模型使用了低阶模型的信息，而回退模型则不使用。
* 相同点：在决定没有出现过的（零计数）N-gram时，两者都使用了低阶模型的信息。

**Additive smoothing**

**Add-one smoothing**

也叫做Laplace smoothing，它是最简单，最直观的一种平滑算法。既然希望出现过的N-Gram的概率不再是0，那就不妨规定任何一个N-gram在训练模型中至少出现一次，也就是规定没有出现过的N-gram在训练语料中出现了一次。

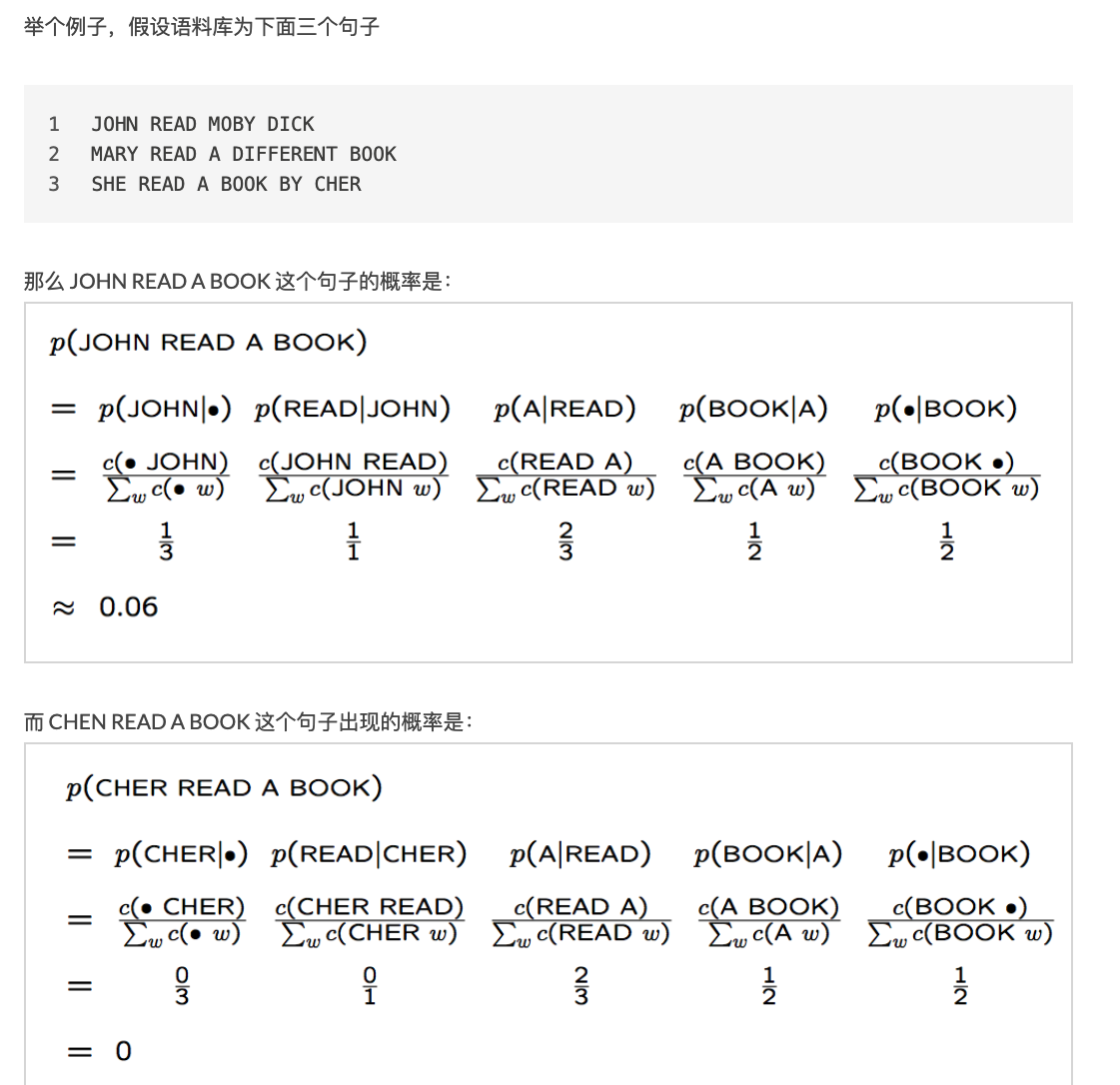
下面以bigram model为例给出add one平滑的模型：

MLE Estimate(最大似然评估)：

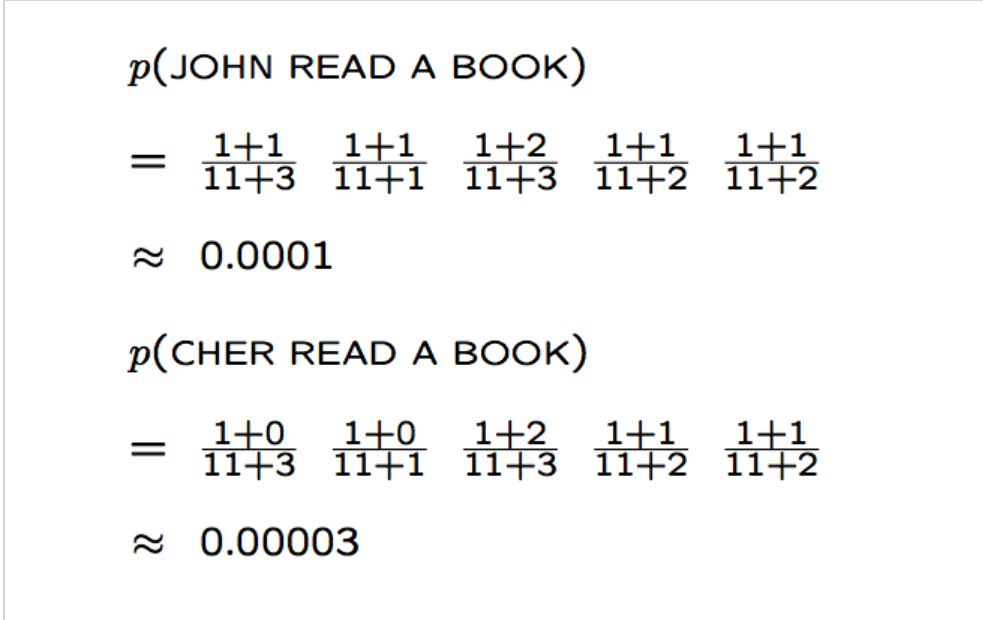
Add one Estimate:

这里的V是所有的可能的不同的N-gram数量，通常V={w:c(w) >0}{UNK}，这里是bigram，所以V可以理解为词表的大小。

下面是个例子：



加1平滑后，概率就变成了：



加1平滑通常情况下是一种很糟糕的算法，与其他平滑方法相比，然而我们可以把加1用在其他任务中，如文本分类，或者非零计数没那么多的情况下。

**Additive smoothing**

对加1平滑的改进就是把1变成了，且0<1，公式如下：

**Good-Turing Smoothing**

假设语料中出现了r次的词有个(出现r次的词的集合大小)，语料大小为N，则N=。当r较小时，极大似然估计可能不准确，同时我们也要考虑一下那些没有出现的词，即r=0的词，所以我们需要给所有的r打一个折扣：

此时，易得。分析可知，r越大，越小，所以，一般情况下。

Good-Turing Smoothing的核心思想是：用观察技术较高的N-gram数量来重新估计概率大小，并把它指派给那些具有零计数或较低计数的N-gram，一般我们可以使用已经出现过一次的事物的数量帮助估计从来没有见过的事物的数—也就是Things seen once。

下面我们用一个例子来进行说明，假设你在钓鱼，然后抓到了18条鱼，种类如下：10 crap, 3 perch, 2 whitefish, 1 trout, 1 salmon, 1 eel，那么，下一条鱼是trout的概率是多少？这个显然很好计算，因为trout出现过，答案显然是1/18。

但是，下一条鱼是新品种的概率是多少，如果不考虑其他，那么概率就是0，但是如果我们采用Things seen once的概率来估计新事物，那么他的概率就是3/18。

在包括新品种的基础上，下一条鱼是trout的概率是多少？显然是要小于1/18的，那么这是怎么估计的呢？

在Good Turing下，对每一个计数r，我们做一个调整，变成，公式如下，其中表示出现过r次n-gram的数量：

然后，我们就可以有：

所以，c=1时，也就是出现r=1次的n-gram：

**Jelinek-Mercer smoothing(Interpolation差值)**

设想对于一个trigram的模型，我们要统计语料库中的“like Chinese food”出现的次数，结果他没有出现，则计数为0.在回退策略中，将会试着用低阶gram来进行替代，也就是用“Chinese food”出现的次数来代替。

在使用差值算法时，我们把不同阶别的n-gram模型线性加权组合后再来使用。简单线性插值(Simple Linear Interpolation)可以用下面的公式来定义：

其中。可以根据凭据经验设定，也可以通过应用某些算法确定，例如EM算法。