N-Gram 及其平滑技术

报告人：李荣陆

E-Mail： lironglu@163.net

相关资料

- 吴立德等，《大规模中文文本处理》，“稀疏事件的概率估计”，pp.53~62 。
- 翁富良，王野翊，《计算语言学导论》，“概率语法”，pp.116~145 。

统计语言模型

- 假设一个句子 S 可以表示为一个序列 S=$w_1w_2\ldots w_n$，语言模型就是要求句子 S 的概率 P(S)：

$$p(S)=\prod_{i=1}^{n}p(w_i\,|\,w_1w_2\cdots w_{i-1})$$

这个概率的计算量太大，解决问题的方法是将所有历史 $w_1w_2\ldots w_{i-1}$ 按照某个规则映射到等价类 S($w_1w_2\ldots w_{i-1}$)，等价类的数目远远小于不同历史的数目，即假定：

$$p(w_i\,|\,w_1w_2\cdots w_{i-1})=p(w_i\,|\,S(w_1w_2\cdots w_{i-1}))$$

N-Gram 模型

- 当两个历史的最近的 *N-1* 个词（或字）相同时，映射两个历史到同一个等价类，在此情况下的模型称之为 N-Gram 模型。
- N-Gram 模型被称为一阶马尔科夫链。 N 的值不能太大，否则计算仍然太大。
- 根据最大似然估计，语言模型的参数：

$$p(w_i \mid w_1 w_2 \cdots w_{i-1}) = \frac{C(w_1 w_2 \cdots w_{i-1} w_i)}{C(w_1 w_2 \cdots w_{i-1})}$$

其中， $C(w_1w_2\ldots w_i)$ 表示 $w_1w_2\ldots w_i$ 在训练数据中出现的次数

平滑技术的引入（1）

- 传统的估计方法对于随机变量 £ 的 N 次独立观察的样本容量 N 有如下要求:

$$N>>K$$

其中 K 为随机变量能够取到的值的个数。

- 实际语言模型中往往无法满足这个要求。
- 例如：词性标注问题，共有 140 个可能的标记，考虑当前词前后两个词的影响的三阶模型。

$$K=140*140*140=2,744,000$$

给定一个 10 万词左右的人工标注训练集，即 N=100,00，可见训练数据显得非常不足。

平滑技术的引入（2）

- 假设 k 泛指某一事件， *N(k)* 表示事件 k 观察到的频数，极大似然法使用相对频数作为对事件 k 的概率估计：

$$p(k)=N(k)/N$$

- 在语言模型中，训练语料中大量的事件 *N(k)=0* ，这显然没有反映真实情况。我们把这个问题称为<u>数据稀疏</u>问题。
- 这种零值的概率估计会导致语言模型算法的失败，例如：概率值作为乘数会使结果为 0 ，而且不能做 *log* 运算。

计数等价类

- 根据对称性原理，事件除了出现次数之外不应具有细节特征，即所有具有相同计数 *r=N(k)* 的事件 k （事件出现的次数称为事件的计数）应当具有相同的概率估计值，这些计数相同的事件称为计数等价，$\sum_{r=1}^{R} r\cdot n_r = N$或$\sum_{r=0}^{R} p_r\cdot n_r = 1$ 将它们组成的一个等价类记为计数等价类 G_r。
- 对于计数为 r 的计数等价类，定义 n_r 为等价类中成员的个数， p_r 为等价类中事件的概率， R 是最大可能出现的计数次数，则

$$\sum_{r=1}^{R} r\cdot n_r = N \text{或} \sum_{r=0}^{R} p_r\cdot n_r = 1$$

交叉检验（1）

- 交叉检验就是把训练样本分为 *m* 份，其中一份作为保留部分，其余 *m-1* 份作为训练部分。训练部分作为训练集估计概率 p_r，保留部分作为测试集进行测试。
- 我们使用 C_r 表示保留部分中计数为 *r* 的计数等价类的观察个数。对于保留部分使用最大似然法对进行概率 p_r 进行估计，即使对数似然函数最大化：

$$F(p_0,\cdots,p_R)=\sum_{r=1}^{R}C_r\log p_r, \text{且} \sum_{r=0}^{R}n_r\cdot p_r=1$$

交叉检验（2）

- 使用拉格朗日乘子解决约束条件下的最大值问题，即

$$F(p_0,\cdots,p_R;\mu)=\sum_{r=0}^{R}C_r\log p_r-\mu(\sum_{r=0}^{R}n_r\cdot p_r-1)$$

- 对 p_r 求偏导，得到交叉检验估计：

$$p_r=\frac{1}{\mu}\frac{C_r}{n_r},\text{其中}\mu=\sum_r C_r=\text{保留部分语料的大小}$$

- 如果测试部分也作为保留部分的话，就是典型的极大似然估计：

$$C_r=r\cdot n_r\Rightarrow p_r=r/N$$

留一估计

- 留一方法是交叉检验方法的扩展，基本思想是将给定 *N* 个样本分为 *N-1* 个样本作为训练部分，另外一个样本作为保留部分。这个过程持续 N 次，使每个样本都被用作过保留样本。
- <u>优点：</u>充分利用了给定样本，对于 N 中的每个观察，留一法都模拟了一遍没有被观察到的情形。
- 对于留一方法，p_r的极大似然估计为：

$$p_r = \frac{1-n_R p_R}{N}\frac{(r+1)n_{r+1}}{n_r}, \text{其中} p_R = R/N, 0 \le r \le R-1$$

Turing-Good 公式

- 因为 $n_R p_R$ 与 1 相比一般可以忽略，留一估计公式可以近似为：

$$p_r = \frac{1}{N}\frac{(r+1)n_{r+1}}{n_r}, 其中 0 \le r \le R-1$$

- 留一估计可以利用计数 *r=1* 的事件来模拟未现事件，对于未现事件有如下估计：

$$n_0 p_0 = \frac{n_1}{N}$$

这个公式就是著名的 Turing-Good 公式。

空等价类

- 留一估计中要求么个 n_r 均不为 0 ，在实际问题中当 $r=5$ 时，这个要求通常都不能满足，即计数等价类 $G_1, \ldots, G_R$ 中存在空的等价类。这时按照出现次数进行排序：

$$0 = r(0) < r(1) < \cdots < r(l) < r(l+1) < \cdots < r(L) < R$$

- 对应的出现 $r(l)$ 次的事件的个数记为 $n_{r(l)}$，在进行留一估计时，使用下一个非空的等价类 $G_{r(l+1)}$ 代替可能为空的等价类 $G_{r(l)+1}$，留一估计公式变为：

$$p_{r(l)} = \frac{1 - n_R p_R}{N} \frac{r(l+1) n_{r(l+1)}}{n_{r(l)}}$$

式中对空的等价类没有估计概率，因为空等价类并没有对应任何有效事件。

Turing-Good 估计的优缺点和适用范围

- 缺点：（1）无法保证概率估计的“有序性”，即出现次数多的事件的概率大于出现次数少的事件的概率。(2)p_r 与 r/N 不能很好地近似，好的估计应当保证 $p_r<=r/N$ 。
- 优点：其它平滑技术的基础。
- 适用范围：对 0<r<6 的小计数事件进行估计。

约束留一估计

- 单调性约束：$p_{r-1} <= p_r$；折扣约束：$p <= r/N$。
- 约束留一估计：让计数估计 $r^* = p_r \cdot N$ 处于距其最近的绝对频数之间：

$$\frac{r-1}{N} \le p_r \le \frac{r}{N}, 1 \le r \le R$$

在这个约束下，单调性约束自然满足。

- 计算方法：计算μ时检查每个 p_r 是否满足约束，不然就用约束的上下界进行裁剪，然后重新计算μ，一直迭代下去直到所有 p_r 满足约束。

折扣模型

- Katz 指出 Turing-Good 公式实质是对模型中观察到的事件进行折扣，将折扣得来的概率摊到所 n_0 个未现事件中。在这个思想的指导下，估计公式可以下成如下形式：

$$p_r = \begin{cases} \dfrac{r-d_r}{N}, 0 < r \le R \\ \dfrac{1}{Nn_0}\sum_{s>0} n_s d_s, r = 0 \end{cases}$$

其中， dr 是对计数为 r 的事件的计数的一个折扣函数。

绝对折扣模型

- 若折扣函数定义为：$d_r=b$，其中 b 为一个大于 0 的常数。那么未现事件的总概率为：

$$n_0 \cdot p_0 = \frac{1}{N}\sum_{r=1}^{R} n_r d_r = b \cdot \frac{K - n_0}{N}, \text{其中} K = \sum_{r=0}^{R} n_r, 0 < b \leq 1$$

对应绝对折扣模型的估计公式为：

$$p_r = \begin{cases} \dfrac{r-b}{N}, 0 < r \leq R \\ b \cdot \dfrac{K - n_0}{N \cdot n_0}, r = 0 \end{cases}$$

线性折扣模型

- 若折扣函数定义为： $d_r = \alpha \cdot r$ ，其中α为一个大于 0 的常数。那么未现事件的总概率为：

$$n_0 \cdot p_0 = \frac{1}{N}\sum_{r=1}^{R} n_r d_r = \alpha, \text{其中} 0 < \alpha < 1$$

对应线性折扣模型的估计公式为：

$$p_r = \begin{cases} (1-\alpha)\dfrac{r}{N}, 0 < r \le R \\ \dfrac{\alpha}{n_0}, r = 0 \end{cases}$$

若$\alpha = n1/N$，则 $n_0 p_0 = n_1/N$，与 Turing-Good 估计相同。

删除插值法（Deleted Interpolation）

- 其基本思想是，由于 N-Gram 比 N+1-Gram 出现的可能性大的多，所以使用 N-Gram 估计 N+1-Gram 的概率，例如 trigram 的计算公式如下：

$$p(w_3 \mid w_1 w_2) = \lambda_3 f(w_3 \mid w_1 w_2) + \lambda_2 f(w_3 \mid w_2) + \lambda_1 f(w_3)$$

其中，

$$\lambda_3 + \lambda_2 + \lambda_1 = 1$$

- 参数λ的确定：将训练数据分为两部分，一部分用于估计 $f(w_i \mid w_1 w_2 \ldots w_{i-1})$，一部分用于计算参数λ，求使语言模型的困惑度最小的λ。