题目 an empirical study of smoothing techniques for language modeling

主要内容

1 N 元语言模型

  1.1 语言模型的性能评价：利用交叉熵和困惑度（测试集中每个词汇的概率的几何平均值的倒数和交叉熵的关系）

2 主流平滑模型

- 加法平滑：普遍加上一个常数，即使得默认他比实际上多发生一个常数次
- Good-Turing 估计法
  - 假设一个出现 r 次的内容应该出现 r*次 $r^* = (r+1)\frac{n_{r+1}}{n_r}$，这样调整之后，多出来的部分可以分配给其他所有未分配的内容。本方法是后面很多方法的基础。
- Jelinek-Mercer 平滑方法
  - 在二元模型中加入一元模型，当二元模型的值相等时，加入一元模型考虑，更加优先的产生频率较高的词对
  - 采用插值模型，在二元模型中加入一个考虑一元模型的量，参数采用最大似然
- Katz 平滑方法
  - 利用分段的方法，采用两个模型分别针对高频和低频进行不同的模型处理。

$$c_{\mathrm{katz}}(w_{i-1}^{i}) = \begin{cases} d_r r & \text{if } r > 0 \\ \alpha(w_{i-1})\, p_{\mathrm{ML}}(w_i) & \text{if } r = 0 \end{cases}$$

  - 高频的用一个折减率扣减，低频的就采用一阶语言模型的最大似然法
- Witten-Bell 平滑方法
  - 进一步提升的 JM 插值法，采用 n 元模型和 n-1 元模型结合的方式处理
- 绝对减值方法
  - 插值模型中，对插入值的系数公式的设计 $D = \frac{n_1}{n_1 + 2n_2}$，n1，n2 分别是出现的一元二元语法模型的总数。
- kNeser-Ney 平滑方法
  - 认为使用一元文法的概率不应该与单词出现的次数成正比，而是与他前面出现的不同单词数目成正比
  - Chen and Goodman 曾提出不同的推导方法
- 算法小节：

algorithm	$\alpha(w_i \mid w_{i-n+1}^{i-1})$	$\gamma(w_{i-n+1}^{i-1})$	$p_{\mathrm{smooth}}(w_i \mid w_{i-n+2}^{i-1})$
additive	$\frac{c(w_{i-n+1}^{i})+\delta}{\sum_{w_i} c(w_{i-n+1}^{i})+\delta\lvert V\rvert}$	0	n.a.
Jelinek-Mercer	$\lambda_{w_{i-n+1}^{i-1}} \, p_{\mathrm{ML}}(w_i \mid w_{i-n+1}^{i-1}) + \ldots$	$(1-\lambda_{w_{i-n+1}^{i-1}})$	$p_{\mathrm{interp}}(w_i \mid w_{i-n+2}^{i-1})$
Katz	$\frac{d_r r}{\sum_{w_i} c(w_{i-n+1}^{i})}$	$\frac{1-\sum_{w_i:c(w_{i-n+1}^{i})>0} p_{\mathrm{katz}}(w_i \mid w_{i-n+1}^{i-1})}{\sum_{w_i:c(w_{i-n+1}^{i})=0} p_{\mathrm{katz}}(w_i \mid w_{i-n+2}^{i-1})}$	$p_{\mathrm{katz}}(w_i \mid w_{i-n+2}^{i-1})$
Witten-Bell	$(1-\gamma(w_{i-n+1}^{i-1}))p_{\mathrm{ML}}(w_i \mid w_{i-n+1}^{i-1}) + \ldots$	$\frac{N_{1+}(w_{i-n+1}^{i-1}\bullet)}{N_{1+}(w_{i-n+1}^{i-1}\bullet)+\sum_{w_i} c(w_{i-n+1}^{i})}$	$p_{\mathrm{WB}}(w_i \mid w_{i-n+2}^{i-1})$
absolute disc.	$\frac{\max\{c(w_{i-n+1}^{i})-D,0\}}{\sum_{w_i} c(w_{i-n+1}^{i})} + \ldots$	$\frac{D}{\sum_{w_i} c(w_{i-n+1}^{i})} N_{1+}(w_{i-n+1}^{i-1}\bullet)$	$p_{\mathrm{abs}}(w_i \mid w_{i-n+2}^{i-1})$
Kneser-Ney (interpolated)	$\frac{\max\{c(w_{i-n+1}^{i})-D,0\}}{\sum_{w_i} c(w_{i-n+1}^{i})} + \ldots$	$\frac{D}{\sum_{w_i} c(w_{i-n+1}^{i})} N_{1+}(w_{i-n+1}^{i-1}\bullet)$	$\frac{N_{1+}(\bullet w_{i-n+2}^{i})}{N_{1+}(\bullet w_{i-n+2}^{i-1}\bullet)}$

- ■

3　其他平滑模型

- Church-Gale P 平滑方法：把对应概率分布分段，类似数据挖掘中数据预处理中使用的分箱，效果只在二元有体现，三元以上就很差了
- 贝叶斯平滑方法　效果较差，否了
- 修正的 Kneser-Ney 平滑方法
  - 复杂化的 KN 模型，针对出现 1 次 2 次 3 次以上的分别用不同模型

4　文中做的提升工作

- JM 算法采用类似 bagging 的思想，训练集分开，然后用来找最好的参数
- Katz 算法　减值率的改进
- WB 算法　不采用插值而是采用分段方法实现
- 绝对减值算法　不采用插值，而是采用分段方法实现

最终在不同实验集应用还是各有优劣，所以不做赘述。