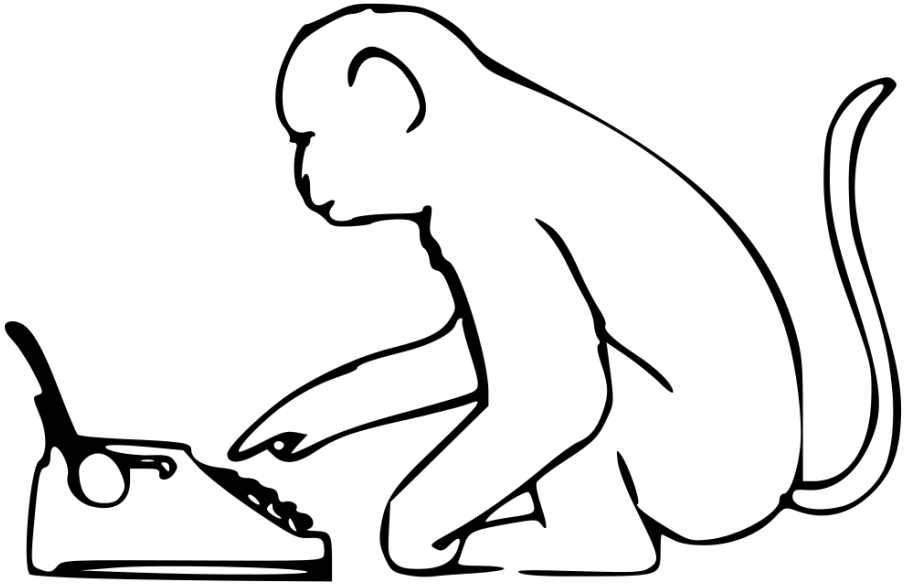
语言模型



语言模型

- 给定文本序列 $x_1, \ldots, x_T$，语言模型的目标是估计联合概率 $p(x_1, \ldots, x_T)$
- 它的应用包括
  - 做预训练模型（eg BERT，GPT-3）
  - 生成本文，给定前面几个词，不断的使用$x_t \sim p(x_t \mid x_1, \ldots, x_{t-1})$ 来生成后续文本
  - 判断多个序列中哪个更常见，e.g.“to recognize speech” vs“to wreck a nice beach

使用计数来建模

- 假设序列长度为2，我们预测

$$p(x, x') = p(x)p(x'|x) = \frac{n(x)}{n}\frac{n(x, x')}{n(x)}$$

  - 这里 n 是总词数，$n(x), n(x, x')$ 是单个单词和连续单词对的出现次数

- 很容易拓展到长为3的情况

$$p(x, x', x'') = p(x)p(x'|x)p(x''|x, x') = \frac{n(x)}{n}\frac{n(x, x')}{n(x)}\frac{n(x, x', x'')}{n(x, x')}$$

N元语法

- 当序列很长时，因为文本量不够大，很可能 $n(x_1, \ldots, x_T) \leq 1$
- 使用马尔科夫假设可以缓解这个问题
  - 一元语法：
$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2)p(x_3)p(x_4)$$
$$= \frac{n(x_1)}{n}\frac{n(x_2)}{n}\frac{n(x_3)}{n}\frac{n(x_4)}{n}$$
  - 二元语法：
$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_2)p(x_4 \mid x_3)$$
$$= \frac{n(x_1)}{n}\frac{n(x_1, x_2)}{n(x_1)}\frac{n(x_2, x_3)}{n(x_2)}\frac{n(x_3, x_4)}{n(x_3)}$$
  - 三元语法：$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2 \mid x_1)p(x_3 \mid x_1, x_2)p(x_4 \mid x_2, x_3)$

总结

- 语言模型估计文本序列的联合概率
- 使用统计方法时常采用n元语法