统计语言模型

语言统计模型的马尔可夫假设

- 问题：如何在没有神经网络的情况下，仅用概率统计来计算一个句子的概率？
- 核心思想：马尔可夫假设（Markov Assumption），一种务实的简化
  - 假设当前词只依赖于最近的 $N-1$ 个词，而非整个历史

$$P(x_t \mid x_{<t}) \approx P(x_t \mid x_{t-n+1}, \ldots, x_{t-1})$$

- 参数估计：最大似然估计 (MLE)。其本质就是“数数”。例如，对于Bigram模型

$$P(\text{word} \mid \text{context}) = \frac{\text{Count}(\text{context} + \text{word})}{\text{Count}(\text{context})}$$

- 优点：简单、计算快、可解释性强

N-gram 语言模型

- 将马尔可夫假设应用于语言，我们得到 N-gram 模型
  - 一元语法 (Unigram, n=1): 词与词之间相互独立

$$P(x_1, \dots, x_T) \approx \prod_{t=1}^{T} P(x_t)$$

  - 二元语法 (Bigram, n=2): 每个词只依赖于前一个词

$$P(x_1, \dots, x_T) \approx P(x_1) \prod_{t=2}^{T} P(x_t | x_{t-1})$$

  - 三元语法 (Trigram, n=3): 每个词依赖于前两个词

$$P(x_1, \dots, x_T) \approx P(x_1, x_2) \prod_{t=3}^{T} P(x_t | x_{t-2}, x_{t-1})$$

- 参数估计： 通过最大似然估计（即“数数”）来计算概率

$$P(w_t | w_{t-n+1}, \dots, w_{t-1}) = \frac{\text{Count}(w_{t-n+1}, \dots, w_{t-1}, w_t)}{\text{Count}(w_{t-n+1}, \dots, w_{t-1})}$$

N-gram 的根本缺陷 (1): 数据稀疏性与零概率问题

- 数据稀疏 (Data Sparsity)
  - 如果一个N-gram在训练语料中未出现过，其计数为0，导致概率估计为0
    - 如，训练集中没见过 "to wreck a nice beach"，模型会认为这句话不可能出现
  - 经典对策：平滑 (Smoothing)
    - 拉普拉斯平滑：给所有可能的 N-gram 计数加一个小的平滑值 α

$$P(w_t|\dots) = \frac{\text{Count}(\dots, w_t) + \alpha}{\text{Count}(\dots) + \alpha|V|}$$

    - 问题： 拉普拉斯平滑过于粗暴，它会从高频事件中“偷走”太多概率，分配给海量的未见事件。对于语言的长尾分布（Zipf's Law），效果很差

N-gram 的根本缺陷 (2): 维度灾难

- 模型缺陷的根源： N-gram 模型的参数空间会随着 n 的增加而指数级增长
- 一个大小为 $|V|$ 的词汇表，N-gram 模型的参数数量级为 $O(|V|^n)$
  - 假设 $|V| = 20000$ (一个中等大小的词汇表)
    - Bigram ($n = 2$): $20000^2 = 4 \times 10^8$
    - Trigram ($n = 3$): $20000^3 = 8 \times 10^{12}$
    - 4-gram ($n = 4$): $20000^4 = 1.6 \times 10^{17}$
- 后果：
  - 存储灾难： 无法存储如此庞大的参数表
  - 统计灾难： 即使能存储，也绝无可能在有限数据上可靠地估计这些参数
- 被迫的妥协： 实践中，只能使用非常小的 n (通常 $n \leq 5$)。意味着模型只能看到非常短的上下文，无法捕捉句子中的长程依赖 (Long-term Dependencies)
  - 例： "The boy who lives in that big house on the hill ... is happy."
  - 主语 "boy" 和动词 "is" 之间的单复数一致性，需要跨越长距离，N-gram 模型无法捕捉

N-gram 的根本缺陷 (3): 语义的缺失与泛化无能

- N-gram 模型将每个单词视为一个独立的、离散的符号（one-hot vector）
- 它无法理解单词之间的语义相似性
  - 假设模型在语料中见过 "the cat is walking"
  - 计算 "the dog is walking" 概率时，无法利用 "cat" 和 "dog" 都是动物、都可作为主语语义信息
  - 在模型眼中，历史 (the, cat) 和 (the, dog) 两个不相关的、正交的实体。它们的知识无法共享
- 根本原因： 基于计数的离散表示法，无法学习到一个平滑、泛化的语义空间。模型只能“记忆”见过的模式，而不能“理解”和“泛化”到未见的、但语义相似的模式

- 结论
  - 统计方法的内在缺陷已经走到了尽头
  - 我们需要一种全新的表示方法，它必须是稠密的、低维的、并且能编码语义