6.1 语言模型

语言模型（language model）是自然语言处理的重要技术。自然语言处理中最常见的数据是文本数据。我们可以把一段自然语言文本看作一段离散的时间序列。假设一段长度为T的文本中的词依次为$w_1, w_2, \ldots, w_T$，那么在离散的时间序列中，w_t（$1 \leq t \leq T$）可看作在时间步（time step）t的输出或标签。给定一个长度为T的词的序列$w_1, w_2, \ldots, w_T$，语言模型将计算该序列的概率：

$$P(w_1, w_2, \ldots, w_T).$$

语言模型可用于提升语音识别和机器翻译的性能。例如，在语音识别中，给定一段“厨房里食油用完了”的语音，有可能会输出“厨房里食油用完了”和“厨房里石油用完了”这两个读音完全一样的文本序列。如果语言模型判断出前者的概率大于后者的概率，我们就可以根据相同读音的语音输出“厨房里食油用完了”的文本序列。在机器翻译中，如果对英文“you go first”逐词翻译成中文的话，可能得到“你走先”“你先走”等排列方式的文本序列。如果语言模型判断出“你先走”的概率大于其他排列方式的文本序列的概率，我们就可以把“you go first”翻译成“你先走”。

6.1.1 语言模型的计算

既然语言模型很有用，那该如何计算它呢？假设序列$w_1, w_2, \ldots, w_T$中的每个词是依次生成的，我们有

$$P(w_1, w_2, \ldots, w_T) = \prod_{t=1}^{T} P(w_t \mid w_1, \ldots, w_{t-1}).$$

例如，一段含有4个词的文本序列的概率

$$P(w_1, w_2, w_3, w_4) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1, w_2)P(w_4 \mid w_1, w_2, w_3).$$

为了计算语言模型，我们需要计算词的概率，以及一个词在给定前几个词的情况下的条件概率，即语言模型参数。设训练数据集为一个大型文本语料库，如维基百科的所有条目。词的概率可以通过该词在训练数据集中的相对词频来计算。例如，$P(w_1)$可以计算为w_1在训练数据集中的词频（词出现的次数）与训练数据集的总词数之比。因此，根据条件概率定义，一个词在给定前几个词的情况下的条件概率也可以通过训练数据集中的相对词频计算。例如，$P(w_2 \mid w_1)$可以计算为w_1, w_2两词相邻的频率与w_1词频的比值，因为该比值即$P(w_1, w_2)$与$P(w_1)$之比；而$P(w_3 \mid w_1, w_2)$同理可以计算为w_1、w_2和w_3三词相邻的频率与w_1和w_2两词相邻的频率的比值。以此类推。

6.1.2 n元语法

当序列长度增加时，计算和存储多个词共同出现的概率的复杂度会呈指数级增加。n元语法通过马尔可夫假设（虽然并不一定成立）简化了语言模型的计算。这里的马尔可夫假设是指一个词的出现只与前面n个词相

关，即n阶马尔可夫链（Markov chain of order n）。如果$n = 1$，那么有$P(w_3 \mid w_1, w_2) = P(w_3 \mid w_2)$。如果基于$n - 1$阶马尔可夫链，我们可以将语言模型改写为

$$P(w_1, w_2, \ldots, w_T) \approx \prod_{t=1}^{T} P(w_t \mid w_{t-(n-1)}, \ldots, w_{t-1}).$$

以上也叫n元语法（n-grams）。它是基于$n - 1$阶马尔可夫链的概率语言模型。当n分别为1、2和3时，我们将其分别称作一元语法（unigram）、二元语法（bigram）和三元语法（trigram）。例如，长度为4的序列w_1, w_2, w_3, w_4在一元语法、二元语法和三元语法中的概率分别为

$$\begin{aligned}
P(w_1, w_2, w_3, w_4) &= P(w_1)P(w_2)P(w_3)P(w_4),\\
P(w_1, w_2, w_3, w_4) &= P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_2)P(w_4 \mid w_3),\\
P(w_1, w_2, w_3, w_4) &= P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1, w_2)P(w_4 \mid w_2, w_3).
\end{aligned}$$

当n较小时，n元语法往往并不准确。例如，在一元语法中，由三个词组成的句子“你走先”和“你先走”的概率是一样的。然而，当n较大时，n元语法需要计算并存储大量的词频和多词相邻频率。

那么，有没有方法在语言模型中更好地平衡以上这两点呢？我们将在本章探究这样的方法。

小结

- 语言模型是自然语言处理的重要技术。
- N元语法是基于$n - 1$阶马尔可夫链的概率语言模型，其中n权衡了计算复杂度和模型准确性。