Chapter 1

使用浅层模型进行语义嵌入

1.1 词向量

词向量（word vector）是许多应用中非常有用的构建模块。它可以对词间的语义关系进行捕获和编码，并最终将单词转换为数字序列，从而形成非常适合训练深度学习模型的密集向量。

1.1.1 经典方法

构建单词表示的传统方法一般使用词袋模型。在该模型中，词表示将各个单词视为彼此独立的。因此此类表示通常使用独热编码生成句子或文档的向量表示，以显示句子中单词的存在与否。但这种表示在实际应用中鲜有使用，因为单词的含义会根据周围单词而变化。在使用词袋模型（其中，单词以其自身维度进行编码）的经典方法中，实际上无法对这种语义上的相似性进行编码。

传统方法的另一个不足是无法体现单词在句子中出现的顺序。传统的词袋方法统计文档中文本的词汇量，以获得存在单词的表示形式，但这丢了失上下文。与前面讨论的编码类似，它假定文档中的单词彼此独立。这种方法还有一个局限：会导致数据稀疏，使得统计模型的训练变得更加困难。

1.1.2 Word2vec

单词的向量表示可以实现语义相似单词的连续表示，其中相关的单词会被映射到高维空间内彼此靠近的点上。这种单词表示方法基于以下事实：有相似上下文的单词也有相似的语义。Word2vec 就是这样的一种模型，它试图通过使用相邻的单词来直接预测单词并学习小且密集的向量（也称为嵌入）。Word2vec 可从原始文本中学习词嵌入，是一种在计算上很有效率的无监督模型。为了学习这些密集向量，Word2vec 有两种形式：连续词袋（CBOW）模型和跳字（skip-gram）模型（由 Mikilov 等提出）。

Word2vec 是一个浅层的三层神经网络，其中第一层和最后一层构成输入和输出，中间层构建潜在表示以便将输入单词转换为输出向量表示形式。

Word2vec 单词表示法可以探索词向量之间有趣的数学关系，这也是单词的一种直观表达。

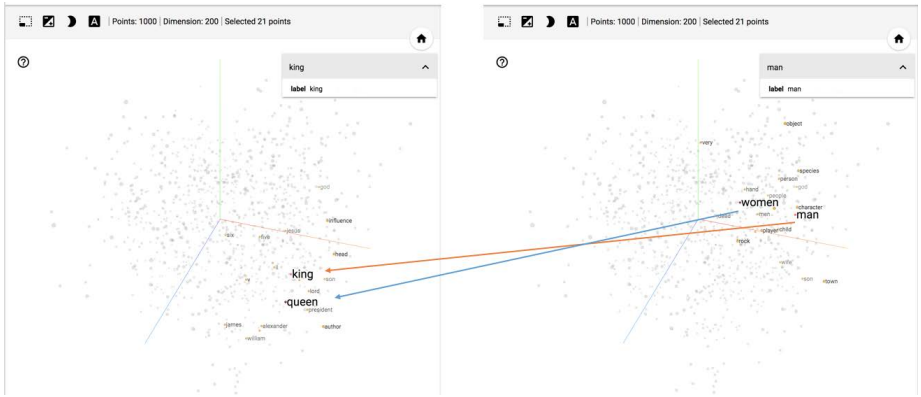

Figure 1.1: 表明词向量是如何从 woman 转换到 queen 的，这与从 man 转换到 king 具有相似之处。使用 Word2vec 可以理解该关系，该模型使用了一个简单的三层神经网络来预测周围的单词（给定输入单词）或预测该单词（给定周围的单词）。这两种方法都是 Word2vec 的变体，其中使用输入单词来预测周围单词的方法是跳字模型，而使用周围单词来预测目标单词是连续词袋模型。

1.1.3 连续词袋模型

Word2vec 的连续词袋模型从一组输入源的上下文单词来预测目标单词。

1.1.4 跳字模型

跳字模型执行连续词袋任务的逆操作，通过使用目标单词来预测上下文中的相邻单词。通常，当数据集更大时，跳字方法倾向于产生更好的单词表示。

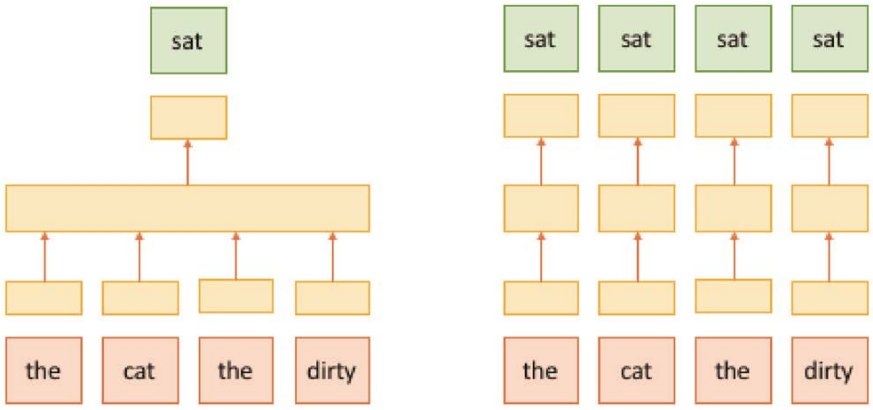

Figure 1.2: 跳字给定目标单词来预测上下文，而连续词袋会根据目标单词周围固定大小窗口的单词（以词袋表示）来学习预测目标单词。（这个图似乎不太好理解）