1 引言

许多当前的自然语言处理(NLP)系统和技术将词语视为原子单元—词与词之间没有相似性概念，因为它们在词汇表中被表示为索引。这个选择有几个很好的理由—简单性、稳健性，以及简单模型在大量数据上训练优于在较少数据上训练的复杂系统的观察。例如，流行的N元语法模型被用于统计语言建模—如今，几乎可以对所有可用数据(数万亿个词[3])进行N元语法的训练。

然而，简单的技术在许多任务中已经达到极限。例如，自动语音识别领域的相关数据量是有限的—其性能通常受高质量转录语音数据量的限制(通常只有数百万词)。在机器翻译中，许多语言的现有语料库仅包含几十亿个词甚至更少。因此，在某些情况下，单纯扩大基本技术的规模并不会带来显著的进展，我们必须专注于更高级的技术。

随着近年来机器学习技术的进步，在更大规模数据集上训练更复杂的模型变得可能，而这些模型通常优于简单模型。或许最成功的概念是使用词的分布式表示[10]。例如，基于神经网络的语言模型显著优于N元语法模型[1，27，17]。