这本书主要包括自然语言处理，计算语言学，信息检索，口语理解，重点在于面向新技术，新应用，人类语言学的两个或多个子领域。

书的前半段涉及监督机器学习和前馈神经网络的基础，通过语言数据进行机器学习的基本知识，以及使用矢量的方式而不是符号对单词进行表示，同时涵盖了计算图抽象，这种方式允许我们简单地定义和训练任意神经网络，并且式现代神经网络软件库设计的基础。

书的第二部分包括一些特定的神经网络架构，包括一维卷积神经网络，循环卷积神经网络，条件生成模型以及注意力模型。这些架构和技术是机器翻译，句法分析和其他许多应用程序的推动力量。最后，我们还讨论了树形网络，结构化预测和多任务学习前景。

# 引语

自然语言处理指的是人类语言自动计算处理。包括采用人类产生的文本作为输入的算法，和产生类似自然的文本作为输出的算法。这其中的困难在于二义性，变化性以及没有明确的定义。

自然语言本质上是符号，在开始进行自然语言处理的尝试中也是符号型的，基于逻辑，规则和知识。然而自然语言具有高度的模糊性和可变性，这些特性需要统计的方法来解决。实际上当前主流的语言处理是基于统计机器学习。过去的十年自然语言理解基于线性模型的方法，主要包括感知机，线性SVM，和逻辑回归，通过非常高维但非常稀疏的特征向量进行学习。

在2014年左右，在密集输入的非线性神经网络取得了更大的成功。一些神经网络是线性模型的简单推广，并且可以近似看作是线性分类器的直接替代。一些新的创造性的改变，例如使用RNN减轻对马尔科夫性质的依赖在序列模型上表现良好。允许适应任意长的序列和产生高效的特征提取，这些导致了语言模型，自动翻译，和其他应用上的突破。

作者在这本书里传输了一些基本概念，术语以及工具。

这本书是从Goldberg，2016上发展而来的。目标那些希望采用现有技术解决实际问题的读者。如果要进行进一步的。讨论，Bengio的书被高度推荐。随后推荐了机器学习上的一些著作以及其他的相关著作。

这里主要涉及一些已经很成熟的技术和一些有希望的技术。

# 1章

## 1.1 NLP的挑战

NLP是设计（输入或输出非结构化的，自然语言数据的模型和算法）。自然语言是模糊的，随语境的改变而改变，也是变化和发展的。人类是语言很好的使用者，但在理解和解释语言规则方面很无力。

使用计算机理解和创造语言是十分困难的。事实上，已知处理语言数据的最好方法是使用监督学习的方法，这些方法从一些预先标注的输入和输出对中获得一些模式和规则。机器学习的算法适合在一些很难定义规则但可以通过标注给定输入得到输出很容易的领域。

除了语义模糊和规则难以定义外，自然语言往往是离散的，组合的和稀疏的。

语言是符号性的和离散的，书写语言的基础元素是字符。字符形成单词，反过来表示目标，概念，事物，动作和概念。字符和单词均是离散的符号，单词例如汉堡知识的是某种确定的心理感知的表示，汉堡包和披萨之间从符号本身和单独的字母之间没有内在的联系。这里作者也对图像与语言之间进行了比较，例如颜色来源于机器视觉，它是连续的，可以从彩色图像映射到灰度图像通过数学操作，或是通过内在的属性可以简单的比较粉红色和红色，而在语言上并不容易做到这一点。

语言也是组合性的，字母形成了单词，单词形成了短语和句子，短语的意义可能比单独的单词的意义更大，并且遵循一些复杂的规则。为了理解文本，除了要在字母和单词上工作之外，还要看词语的序列例如句子或者完整的文件。

上述特性的组合造成了数据的缺少。不同方式组合起来单词形成的意义是实际上是无限的。有效的句子量极为庞大，不能期望对其进行枚举，就像本文中出现的许多四字的意义就是很新的。十年前的一些人，品牌，公司，网络用语，技术名词和未来十年你会碰到的都是新颖的。因此没有明确的形式从一句话推断出另一句话或是定义出他们的相似性。这并不依赖于他们我们未知的含义。这对我们要从一些例子中进行学习进而观察到从未出现在样本集中的的实例和在数据集中非常不同的实例很有挑战性。