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。

## 神经网络和深度学习

深度学习是机器学习的一支，是神经网络重命名的，以可微分的参数化函数为特征。深度学习是来源于多层可微分函数的连接。

虽然所有的机器学习方法都是以基于过去的观察为特征来进行预测，但是深度学习不仅是要预测也是要正确的表述数据。给定要求的输入输出映射，深度学习可以通过将数据通过一个网络产生一系列转换，直到最终可以预测输出。这种转换通过给定输入输出的映射来实现，以便于每一个转换能够使数据与相应的标签相联系。

人们设计网络的架构，训练机制，提供合适的输入输出集合，编码输入数据以合适的方式，网络学习正确的表示，通过网络架构自动运行。

## 深度学习在NLP

神经网络在自然语言问题的处理上很有吸引力。神经网络对于语言的使用的一个中心部分使编码层，将离散的符号映射到连续的向量上在一个相对低维的空间。当对单词进行嵌入时将孤立的符号转化为数学可以操作的对象。尤其是，向量之间的距离可以等同于单词之间的距离，是将一些行为可以泛化到各个单词上。把单词映射为向量的表示是作为网络训练过程大的一部分。沿着网络的层次结构往上，网络学会了将词向量组合起来以有利于预测任务的方式。这种能力一定程度上缓解了离散性和数据稀疏性的问题。

有两种主要的神经网络架构可以以一些方式进行组合，前馈神经网络和循环神经网络。

前向网络，特别是多层感知机，允许以固定大小的输入输入，或是可变的长度输入但是不考虑元素的顺序。当给网络一组输入组件的时侯，它会学会将它们以有意义的形式组合起来，MLPs曾经被用作一种线性模型。非线性的网络架构带有可以轻易的结合预训练的词嵌入的形式，通常会导致优秀的分类准确率。

卷积神经网络是特殊设计的结构，它在提取数据局部特征上表现优秀，可以输入任意大小的输入，对于特定的单词顺序敏感，而不考虑这种模式出现在哪里，因此他们工作在表示固定长度的短语和成语的指示上表现良好。

循环网络是专门为顺序数据而设计的模型。它们是一些网络可以一些序列化的事物作为输入，产生一个固定大小的向量去总结这个序列.这种总结在不同任务中表现不同（例如回答关于句子情感的信息和关于语法的信息是不同的）。循环网络很少独立的进行任务，通常作为一个可训练组件将将数据输入到其他的网络结构中，与他们协同工作。举例可以将循环网络的输出传递给前馈神经网络中，以尝试预测一些值。循环网络通常被用作输入转换器被训练去产生有信息的表示然后传递给前馈网络在顶层对其进行操作。循环网络对于序列是十分有意义的，可能是神经网络对于语言处理给出的最令人激动的贡献。他们允许放弃过去几十年一直流行的马尔科夫假设，并且设计完全基于整个句子的模型，在需要的时候把单词顺序考虑进去，而且不必为数据的稀疏性而苦恼。他的可以进行语言建模的能力令人印象深刻，而且可以承担预测下一个词的任务，这时许多NLP应用的基石。递归网络将循环扩展为树。

自然语言处理的很多问题是结构化的，要求产生复杂的输出结构例如序列或是树。神经网络适合这种需要，既可以通过对已知的线性模型的结构化预测算法进行调整或是通过一些新奇的结构例如S2S(encoder-decoder)模型（我们在这本书中提到了条件生成模型）来达到这一目标。这些模型是现今的机器翻译方法的核心。

最后许多语言任务是彼此相关的，在某种意义上可以通过学习一种任务来进行另一种任务。当我们缺少有标记的训练数据的时候，我们可以简单地使用原始的数据作为支持。我们能够学习未标记的样本吗？神经网络可以通过学习相关领域的问题和半监督学习来解决这一问题。

### 成功的案例

全连接网络在某种意义上可以是大多数情况下线性学习器使用场景下代替线性学习器。这包括二分类和多分类问题和更加复杂的结构化预测问题。网络的非线性以及能够很容易地整合预先训练的单词嵌入的能力通常导致卓越的分类准确率。一系列的工作成功地获得了改进的语法解析结果，通过将线性模型代替为全连接的前馈前向神经网络。将前馈网络作为分类器替换的简单应用（通常是使用预先训练的单词向量）为许多语言任务提供了好处，包括语言建模的非常基本的任务，CCG超级标签，对话状态跟踪，以及统计机器翻译的预序。Iyyer等人2015年演示了多层结构前馈网络可以在情绪分类和问答方面提供有竞争力的结果回答问题。周等人2015年和Andor等人将他们整合到一个定向搜索中的结构预测系统，在句法分析、序列标注上达到卓越的精度和等等。具有卷积和池化层的网络对于分类任务非常有用，我们希望找到针对分类对象有关的强有力的局部线索，但这些线索可以出现在输入的不同位置。例如，在一个文档分类任务中，一个关键短语（或ngram）可以帮助确定文档的主题。卷积和池化架构在许多任务上显示了有希望的结果，包括文档分类、短文本分类，情绪分类，实体之间的相对类型分类，事件检测，解释识别，语义角色标签，问题回答，预测电影票房收入（基于评论家的评论），对文本趣味度的建模，以及对字符序列和词性标注之间的关系进行建模。

在自然语言中，我们经常使用任意大小的结构化数据，比如序列和树。我们希望能够在这种结构中捕捉规律，或者模拟这种结构之间相似性。循环网络是被设计模拟序列的，递归网络是被设计模拟树的.循环网络在语言建模上表现强大，以及序列标注，机器翻译、解析、处理和许多其他任务，包括噪声的文本规范化、对话状态跟踪，响应生成，并建模字符序列之间的关系和词性标记。

递归模型被证明可以产生最先进的或接近最先进的结果对于选区的依赖性解析重新排序，话语解析，语义关系分类，基于解析树的政治意识形态检测，情绪分类，目标独立的情感分类，问答。

## 书籍组织结构

e书由四个部分组成。第一部分介绍了我们将在书中使用的基本学习机制：监督学习、mlp、基于梯度的训练，以及用于实现和训练神经网络的计算图抽象。第二部分将第一部分中介绍的机器与语言数据连接起来。它介绍了在处理语言数据时可用的主要信息来源，并解释了如何将它们与神经网络机制集成在一起。它还讨论了文字嵌入算法和分布假设，以及语言建模的前馈方法。第三部分处理专门的体系结构及其对语言数据的应用：用于处理ng的一维卷积网络，以及用于建模序列和堆栈的RNNs。RNNs是将神经网络应用于语言数据的主要创新，而第三部分的大部分内容都是针对它们的，包括它们所促进的强大的条件生成框架和基于注意力的模型。第四部分是各种高级主题的集合：用于建模树的递归网络、结构化预测模型和多任务学习。

第一部分，涵盖了神经网络的基础，包括四章。第2章介绍了监督机器学习的基本概念，参数化函数，线性和对数线性模型，正则化和损失函数，训练为最优化，和基于梯度的培训方法。它从底层开始，为以下内容提供所需的材料章。熟悉基础学习理论和基于梯度的学习的读者可以考虑跳过这一章。第三章阐明了线性模型的主要局限性，激发了对非线性模型的需要，为多层神经网络奠定了基础和动力。第四章介绍了前馈神经网络和mlp。它讨论了多层网络，它们的理论力量，以及普通的子组件，比如非线性和损失函数。第五章处理神经网络训练。介绍了计算图允许对任意网络进行自动梯度计算的抽象（反向传播算法），并为有效的网络提供了几个重的技巧和技巧培训。

第二部分介绍语言数据，由七章组成。第6章介绍了常见语言处理问题的类型，并讨论了使用语言数据时可用的信息来源（功能）。第7章提供了具体的案例研究，展示了前一章描述的特征如何用于各种自然语言任务。熟悉语言处理的读者可以跳过这两章。第8章连系第6章和第7章的材料与神经网络，并讨论了各种编码方式基于语言的特征作为神经网络的输入。第9章介绍语言建模任务和前馈神经语言模型体系结构。这也为后面章节中讨论预训练的词嵌入。第10章讨论分布式和词义表征的分布方法。它引入了单词上下文矩阵进行分布式语义表示，以及神经语言建模通过词嵌入算法，比如GLOVE和Word2Vec，并讨论它们和分布式方法之间的联系。第11章介绍了如何使用外部的词嵌入神经网络的背景。最后，第12章介绍了一个案例研究

第三部分介绍了专门的卷积和重复架构，由五章组成。第13章介绍了卷积神经网络，它专门学习信息性的ngram模式。还讨论了另一种hash-kernel技术。这部分的其余部分，第14章17节，是关于rnn的。第14章描述了用于建模序列和堆栈的RNN抽象。第15章描述了RNNs的具体实例化，包括简单的RNN（也称为艾尔曼RNNs）和封闭式架构，如长短期内存（LSTM）和门控的重复单元（GRU）。第16章提供了使用RNN抽象建模的示例，展示了它们在具体应用程序中的使用。最后，第17章介绍了条件生成框架，它是最先进的机器翻译背后的主要建模技术，以及无监督的句子建模和许多其他的创新应用。第四部分是高级和非核心主题的混合，由三章组成。第18章介绍了树木结构的递归网络。虽然这一模式非常吸引人，但它仍处于研究阶段，而且还没有展现出令人信服的成功故事。尽管如此，对于那些致力于将建模技术推广到最先进水平的研究人员来说，这是一个重要的模型家族。那些对成熟和健壮的技术感兴趣的读者可以安全地跳过这一章。第19章涉及结构化预测。这是一个相当技术性的章节。那些对结构化预测特别感兴趣的读者，或者已经熟悉线性模型或语言处理的结构化预测技术的读者，很可能会欣赏这种材料。其他人可能会相当安全地跳过它。最后，第20章介绍了多任务和半监督学习。神经网络为多任务和半监督学习提供了充足的机会。是重要的技术，仍处于研究阶段。然而，现有的技术相对容易实现，并且确实提供了真正的收益。章在技术上并不具有挑战性，并被推荐给所有的读者。

## 什么没有介绍

没有关于语音方面的介绍，对于对话系统，文件总结或是知识问答这类open的问题没有涉及。这里描述的技术可以在这些领域使用，但没有具体讨论。语义分析也超出了范围。关于语言和其他模态（视觉）。数据的关系只是简单的提及，最后是以英语为中心的。

要求基础：合适的评价机制和数据标注。这些并没有被讨论。合适的评价机制要求对给定任务可以给出评估标准，和其他人的工作做对比，表现为误差的分析以及统计意义上的评估。

数据标注是NLP重要组成部分，要对数据的来源有了解，考虑到数据形成过程中的影响。数据标注是很大的话题，包括对标注任务选取合适的形式，决定标注的指导方针。决定标注数据的源，它覆盖的范围，类属性，好的训练测试集分割。和使用注释器，合并决策，验证注释器和注释的质量，以及各种各样的类似的主题。

## 术语的提示

“特征”是指一个具体的、语言的输入，例如一个单词，一个后缀，或者一个部分。语义标签。例如，在一阶的语义标记中，特征可能是“当前的单词，前一个词，下一个词，前一段话”。“输入向量”是用来指对于输入到神经网络分类器的实际输入。相对于其他的文献，这里讲输入向量拿了出来，而不是与特征混在一起。

## 数学符号

用大写字符表示矩阵，小写表示向量。一系列相关的矩阵和向量用上标索引。加粗表示代指，不加粗是具体化。向量被认为是行向量，矩阵乘向量是xW+b。Wi，j i是列，j是行

# 学习的基础和线性模型

神经网络是一类监督机器学习算法。

本章提供了一个关于监督机器学习术语的快速介绍和实践，并介绍了二类分类和多类分的线性和对数线性模型。

这一章节为后面章节的学习设置了准备。

## 监督学习和参数化方法

监督学习的本质是创造一种可以观察实例并由此产生泛化应用。具体而言与其说是设计一个执行任务的算法，不如说是设计一个输入是标记好的实例，输出是一个函数，输入一个实例产生期望的标签的函数。预期产生的函数可以对训练集中没有看到的实例产生一个正确的标签。在所有可能的程序（或所有可能的函数）的集合中搜索是非常困难的（而且相当不明确的）问题，我们经常限制自己去搜索特定族的函数，例如，所有有din输入和dout输出的线性函数的空间，或者所有的关于din变量的决策树空间。这样的函数族被称为假设类（就是假设可能出现的函数）。

下面几张简要介绍一些新学习到的东西

## 数据集划分

在某些情况下随机划分数据集不是好的选择，当你以后实际应用的场景与训练集不是同分布的时候。例如对新闻文章进行学习，但是不同时间下新闻的内容都是变化，这时可以特意保持数据集中有一部分新的数据集，也有一部分旧数据集，这里希望算法能学到随时间的变化数据的变化规律。原则上测试集只有最后才使用。实际上为了与他人方法做比较，这里通常使用别人的数据集设置方式。

## 线性模型

这里作者举了一个利用连续两个字符组成的字符组直方图，这里作者形成了二维的的直方图匹配，实际上从这里看出卷积匹配的是二维图像的特征，但实际情况下在自然语言理解中也可以应用，如果三维空间有一些相关的关系，我们可以通过三维卷积的形式来使用，取决于具体问题，图像先天的适合二维的统计，另外这里也体现了直方图统计在自然语言中的应用可以，例如这里作者举例可以使用频率最高的单词的直方图来解决分类英文、德文文档的问题。

## 分类模型

分类模型更多的是要输出概率，多使用sigmoid的形式，分类损失-log。

## 表示

这里作者以文档分类为例，训练一个多分类器分类文档属于哪种语言。其输出为对各个语言的分数，最大值为分类结果，输入为一个文件的表示向量。这表示可能包含更多信息，例如可以找到文件是以方言写的，或是多语言作者。

输入向量应是包含正则项的双词语计数对于文件进行表示，这种表示和最后用于预测的向量包含相似的信息。但是用于预测的向量更紧凑，针对特定的预测目标更有针对性。shuangshuang

训练得到一个W属于784\*6的矩阵，也就是一种表达形式的学习。W的每一列对应一种特定的语言，将获得一个针对这种语言的特定表示根据双词模式。可以获得6种语言的向量根据相似性进行聚类。这里每一列是按照特定双词序列对应，获得了语言对应的6维向量表示。

这种表示是深度学习的核心，深度学习针对特定输入或是针对输出类别的表示，尽管这可能是不可解释的，但却很有用。

## ONE-HOT 向量与紧密的向量表示

在上面输入的向量是文本中双联词的统计包含了正则化项，这个向量可以一个向量的平均，指的是文件中各个位置双联词，每个D表示一个特定位置的双联词向量是一个784维的one-hot向量。



结果获得的是一个双联词的词袋（指的是文档中所有“单词”的身份信息，不考虑他的位置）。One-hot可以视为单个词的词典。

矩阵W的行视图是文档表示的另一个表示形式，矩阵的每一行对应了一个双联词Di，我们获得了y表示是一个双联词变量在文上的平均表示。

## 损失函数

铰链损失：直接用预测正确那类的预测值与预测不正确那类的预测值的差作为损失函数，这里最近在论文看到的属于这类，这类损失函数用max函数直接限定了预测正确那类预测值要大于其他类，软性版本使用log函数获得差值的损失，而不是用概率将预测值之间的距离放大，感觉适合一些目标不够明确的实验。

硬性版本：



软性版本：log 损失



交叉熵损失：-log最大似然，这里获得的是概率分布，将所有反例的概率值压到最低，最大似然的概念。这个函数衡量的真实概率分布与预测概率分布之间的区别，在图像检测中有明确分类的情况下通常是0，1序列，只有一类没有损失，而在自然语言理解中可能会比较模糊。



排序损失：在一些情况下，我们没有监督性的标签，而是给定了正确和不正确的项顺序向量，我们的目标是将正确的项向量的得分大于不正确的得分。因此当训练环境只有正样本的时候，我们通过对正样本进行破坏获得负样本。损失函数如下：





可以作为预训练的词嵌入任务的辅助函数，可以将正确的词序列（或是三元组等）打乱，获得不正确的词序列。dos Santos等人给出了排名对数损失的变化，从而为负面和正面类别提供了不同的余量。

### 正则化

在一些标注可能会有些错误，如果网络过度的学习这些特征，就会使网络性能下降。

## 相似性和距离性

这里描述两个向量的的相似性和距离

相似性：



欧式距离：



可训练形式：

相似性：



距离：



## 嵌入层

输入神经网络的包含符号类特征（例如来自一个字典的单词），他们被和每个可能的特征值的特征向量相关联。这些向量被考虑为参数化模型，北河其他参数联合训练。嵌入层表现为R|voc|\*d,每一列代表词典中的一个单词。xE获得词向量，词向量在传输到下一层之前，彼此相连。

## 计算图

计算图是一个有向非循环图。

### 随机开始

使用不同随机初始化的方法多次训练，在验证集上看效果

随机种子选择：使用模型集合的方法，进行多数投票制

### 梯度消失与爆炸

梯度消失：逐层预训练，固定前面的层训练后面的层，使网络窄一些，特殊结构。

梯度爆炸：梯度阶段。

### 饱和死亡的神经元

观察每一层值的范围确定是否发生。饱和可以通过改变初始化的方式，对输入值进行缩放，改变学习速率。死亡神经元：减少学习速率会有帮助。批量归一化比较有用，但在自然语言处理中不太流行。

# 自然语言处理上的工作

## 文本数据的特征

从文本数据抽取特征称为特征提取。虽然深度学习缓解了特征设计的需要，但仍然需要定义一些好的特征集合。这里的符号需要在某种形式上转换为向量。

这里从训练中脱离出来，讨论用于语言数据的特征函数。这里讨论了一些可接受的用于处理文本语言数据的特征。

### 分类问题的类型

自然语言处理的分类问题可以被分为许多类，根据其具体内容，一些问题不能被归于分类问题，例如产生长的文本（文本总结，机器翻译）。

词语：对词语进行分类，需要描述词语的属性（值得使神给，哪种语言，相似词）。这种情况很少见，因为词语通常依赖于上下文背景，很少孤立的去看。

文本:一个文本的片段，包括短语，句子，一段话。（是否是垃圾邮件，是关于什么内容的，表达什么态度，谁写的，是否可信，什么意图）统称为文档分类问题

成对的文本：给定一对较长的文本，描述一些这对文本的特性（A是由B翻译来的？是否是同一作者？句子A的含义与句子B有关？

词在上下文中的含义：给定一段文本，在上下文中对单词进行分类。（获取单词在句子中特定的特性，边界）。这些问题通常产生在更大的目标下，例如标记句子作为文本的一本，将文档切分为句子，发现文档中的命名实体。

词语之间的关系：给定大文档中的两个词语或短语，说出他们之间的关系。（A和B是在一个局子里吗，两个词语之间有什么关系）。

这些分类可以扩展为结构化的问题，在表现一些相关的决定的时候，一个问题的解答可能影响到其他的问题。

什么是一个单词：单词的概念争论很大，在语言学界不甚清楚。

这里的定义是单词是一些空白字符之间的序列。这是过分简单化的。首先，英语中的标点不是用空格分割的。纠正单词是由空格和标点分割的。英语中分词是十分简单的，尽管需要考虑一些缩写词。在其他语言中，事情变的很麻烦。

另一个公共的含义是含义的最小单元。根据这个定义可以看出基于空格的定义时有问题的。这样分割后依然会有实际上是两个词未被分割的情况。还有一些很难界定的情况，例如New York怎么算。

通常我们要区分词和符号，分割器输出是token，不是单词。一个token也许是由多个词组成的，或是多个token是一个单词，有时不同token指的是同一个词。

有时事情是更复杂的，word这个词比较松散，可与token交换。

## NLP问题的特征

通常单词和字母作为离散项，这里通常采用指示器或是计数器的形式。指示器采用0或1作值，表示条件是否满足。计数器通常指一些项出现的次数。

### 直接观察到的属性

单个词的特征：当我们关注的是脱离上下文的单词时，主要的信息是组成单词的字母和他们的顺序，以及有此派生出的单词的长度，，拼写形式，单词的前缀、后缀。

我们也看到一些与外部关系相关的词。这个词出现了几次，这是人名吗：

原型和主干：通常可以看到字典中将booking，booked，books映射到共同的lemma book。这种映射通常用引理词汇或是形态分析。一个单词的引理是模糊的，对其进行引理化通常应该在上下文中处理。词性还原是语言学定义的过程。一个粗糙的过程是对单词的主干化（stem），是一个针对语言启发式的方法，对于一些短序列进行映射，例如不同的词形变化映射到同一个短序列中。在变化后的词形是同一个形式，不一定有意义，如picture 和 pictures 映射到 pictur。又不同的主干化方法，对词形有不同的破坏性。

词汇资源：对于单词形式的附加信息的来源。本质上是指由机器以编程方式访问的字典，而不是被人类阅读。词汇资源通常包含关于单词的信息，将它们与其他词联系起来，或提供额外的信息。

例如，对于许多语言，有一些词汇表可以将后缀变化的单词形式映射到它们可能的形态分析上（告诉你一个特定的词可能是复数的名词或一个动词）。这样的词汇字典通常也包括引理信息。

WordNet是一个非常出名的词典资源。它是一个人工制作的数据集，试图捕捉一些单词的一些概念性的语义知识。每个单词都属于一个或一些同义词集，每个同义词集描述的是一些认知型的概念。例如名词star属于的第一个同义词集是天体，出名的演员，有能力的人，第二个词集是冠军，王牌，轰动，演奏家等等。词集通过上位词或是下位词（具有更具体含义的词）获得词语间的语义关系。举例，对于第一个词集包含的是一些下为此如太阳，新星和星体等。在WordNet中包含了其他的语义关系，如反义词，全称，简称等。这里包含有名词，动词，形容词，副词的信息。

FrameNet,VerbNet是以动词为核心的词典，包含了关于动词的种种观点。包括动词的施动者，承受者，动词的主题。还有例如一些非核心的概念，例如时间，目标，地点，行为等等。

PPDB是一个巨大的，自动创建的释义数据库，给出了一系列单词和短语以及具有相同释义的单词和短语。这些词汇资源包含大量的信息但在实际工程应用中需要大量的努力才能被使用，通常不被用于神经网络模型，但这种情况也可能会改变。

信息分布：另一个重要的信息关于词语的是分布，在一个文本中一个词语的意思有其他词与之相似吗，后面将讨论如何将词汇资源知识注入到神经网络派生出的分布式词向量中去。与其说是符号，不如说是社会学问题。

文本特征：我们考虑一个句子，段落通常可以观察到的是一个字母或是单词的计数。

步可以使用一些基本元素（例如字母的组合），前面的语言类型判断的方法。

我们可以使用直接用单词或字母驱动的量，如句子有多少个字母或是单词。考虑到孤立的单词，可以使用上面基于单词的特征，如获得具有特定前缀或后缀，或是段单词，长单词的比例。，

权重：同时可以聚焦一些整合外部信息的数据，例如聚焦在给定文档中多次出现的单词，但是在其他文档集中出现的次数较少（与主题相关词）。当用词袋方法的时候，通常使用的是TF-IDF向量，当用词袋的方法的时，这里将其文档作为一个大型文本库的一部分D的一部分，每个词出现的频率，被这个单词在所有文档中出现的频数正则化。，第二项是逆文档概率，是出现该单词的文档的数量。表示的是当前文本单词的重要性。

除了单词，我们也可以考虑连续的单词或是三元组。称之为ngrams。

上下文单词特征：当在一个句子或一个文档中考虑一个词时，这个词的直接可观察特征是它在句子中的位置，以及它周围的单词或字母。和目标更近的单词往往会更具有信息量。（这是一个粗略的概括，在许多情况下，语言表现出了长期的依赖关系，文档开头的单词会影响文档末尾的单词）。

窗口。处于上面的原因，通常会集中注意力与当前单词的周围的窗口位置的直接上下文信息。通常会用表示窗口大小。例如2窗口取目标单词周围的4个单词，前面两个单词，后面两个单词。这是词袋模型的一个版本。

这里可以不考虑词袋假设的顺序，而将单词的相对位置考虑进来。形成相对位置特征。

后面介绍biRNN通过对于窗口特征进行了泛化提供了一个灵活的，可调整的训练窗口。

位置。考虑单词的绝对位置。

词关系特征。考虑单词之间的相对位置，单词之间的距离和同一性。

### 6.2.2 推断语言属性

自然语言中的句子的结构超出了词语之间线性关系的描述。这种结构错综复杂，不可能被直接观察到。这些规则称为语法，这些规则的自然的研究这些自然语言中的规则和性质是语言学研究的目标。而语言的确切的结构一直是个谜，管理许多更复杂的模式要么未被探索，要么仍有待讨论在语言学上，这里仅有一个子集可以被很好的研究和理解。这些概念包括词类单词类（词性标注）、形态学、语法，甚至部分的语义。

而一个文本的语言学属性，不可能被直接观测到通过句子中的词语和他们的顺序，他们可以以不同精度通过句子字符串推断出来。

存在专门的系统来预测词类，句法树，语义规则，语篇关系和其他语言属性，并且具有不同程度的准确性，并且这些预测常常用作进一步分类问题的良好特征。

语言学标注：这探索一些形式的语言学标注。考虑句子可以标注每个的词类。进一步可以考虑语法块的边界，如名词短语，介词，名词短语，动词。。。这里包含一个块包含相应的核心词以及相关的形式。如will open或don’t open

更进一步的是语法树。



上面的数按区域进行嵌套，按层次结构直到单词为止。

另一种语法注释是依赖树。在依赖树，句子中的每个词都是另一个词的修饰词，也就是它的头。句子中的每一个词都是由另一个词引导，除了主词，这通常是一个动词，它是句子的根，由一个特殊的“根”节点领导。



虽然选区树显示了单词到短语中的分组，但依赖树显示了单词之间的修改关系和连接。在句子的表面形式中相距甚远的单词可能会在其依赖树中闭合。例如，男孩和打开在表面形式上有4个词的间隔，但是在依赖树中有一个直接的边连接它们。

依赖关系是句法的：他们关心句子的结构。其他类型的关系是更加具有语义关系的。距离考虑对动词open的修正，也称动词的参数。这种句法树更加清晰的标记了男孩（穿着黑色衬衫），门，以及作为参数的一把钥匙，并且还告诉我们，with a key 是open的论据而不是门的修饰。然而，它并没有告诉我们，关于动词的论据的语义角色是什么，即，男孩是代词执行的动作，而a key是一个工具（相对应the boy opened the door with a smile）在这里，句子将具有相同的句法结构，但除非我们处于魔法世界中，否则微笑是一种行为而不是一种工具。语义角色标注注释揭示了这些结构。



除了可观察的属性（字母，单词，计数，长度，线性距离，频率等）外，我们还可以查看单词，句子和文档的可以推断的语言属性。例如，我们可以查看“一个单词中的单词的词性（POS）（它是一个名词，动词，形容词还是确定词？），单词的句法作用（它是作为动词的主语还是宾语？是否是句子的主要动词？是作为一个状语修饰词吗？），或是它的语义角色（在 ‘the key opened the door’，key表现为一个工具，而在 ‘the boy opened the door中boy是一个代理者）。当在一个句子中给出两个单词时，我们可以考虑句子的句法依赖树，以及连接这棵树内两个单词的子树或路径，以及该路径的属性。 在句子中隔开它们的单词数量相距很远的单词在句法结构中可以彼此接近。

当我们关注句子的时候，我们也许会想要看到将句子联系到一起的论述关系，例如扩展，对立性，因果关系等等。这些关系通常被连接词语表达例如moreover，however，和and，but等，也可能是更间接的联系。

另一个重要的现象是回指的现象 - 考虑男孩用钥匙打开门的句子序列。 它没有锁上，他进入了房间。 他看到一个男人。 他笑了。 回指决议（也称为参与决议）将告诉我们，它指的是门（而不是钥匙或男孩），他指的是男孩，他很可能会指的是男人。

词性标签，句法角色，语篇关系，回指等是基于语言学家长期开发的语言学理论的概念，目的是在非常混乱的系统中捕捉规则和规律 尽管语言规则的许多方面仍然有待辩论，而其他方面可能看起来过于僵化或过于简单化，但这里（和其他方面）探讨的概念的确在语言中捕捉到一系列广泛而重要的概括和规律。

语言概念是否需要？： 一些深度学习的支持者认为，上面提到的手动设计的语言属性是不需要这种推断的，并且神经网络将自己学习这些中间表示（或等价的或更好的表示）。我目前的个人观点是，如果提供足够的数据并且可能推动正确的方向，那么许多这些语言学概念确实可以由网络自行推断。但是，对于许多其他情况，我们没有足够的培训数据可用于我们关心的任务，并且在这些情况下为网络提供更明确的一般概念可能非常有价值。即使我们确实有足够的数据，我们也可能希望将网络集中在文本的某些方面，并通过提供广义的概念来补充或甚至代替词的表面形式来提示它应该忽略其他的方面。最后，即使我们不使用这些语言属性作为输入特征，我们也可能希望通过在多任务学习设置中使用它们作为附加监督来帮助指导网络（参见第20章），或者通过设计网络体系结构或训练范例以更适合学习某些语言现象。总的来说，我们看到足够的证据表明使用语言概念有助于提高语言理解和生产系统。

进一步阅读：在处理自然语言文本时，建议您注意字母和单词以外的语言概念，以及当前可用的计算工具和资源。 这本书几乎没有涉及这个话题。 Bender的书[2013]为针对计算人员的语言概念提供了一个很好和简洁的概述。 有关当前NLP方法，工具和资源的讨论，请参见Jurafsky和Martin [2008]的书以及本系列中的各种专业标题。

### 6.2.3 核心特征和组合特征

在很多情况下，我们对一起出现的特征的连接感兴趣。例如直到词book出现在窗口里和一个动词出现在窗口里比book作为动词出现在窗口里信息量更少。

线性模型不能为事件的联合（X发生和Y发生，...）指定一个分数，这不是它们各自分数的总和，除非联合本身被建模为它自己的特征。例如，当为线性模型设计特征时，我们不仅要定义核心特征，还要定义许多组合特征。这样的·组合数量庞大，而且伴有错误和误差，因此为了构建具有信息量和相对有影响的组合需要专家的知识。

神经网络提供非线性模型，并且不受这个问题的影响。当使用诸如多层感知器（第4章）这样的神经网络时，模型设计者只需指定一组核心特性，并依赖于网络训练程序来获取其自身的重要组合。大大简化了模型设计者的工作。在实践中，神经网络确实能够以核心功能为基础，学习优秀的分类器，有时超越了具有人类设计特征组合的最佳线性分类器。然而，在许多其他情况下，一个具有良好手工制作的功能集的线性分类器是很难被击败的，因为带有核心特征的神经网络模型接近但没有超越线性模型。

### NGRAM 特征

一个特殊的特征组合的例子是ngram，一个指定长度的连续单词序列。我们已经在语言分类案例（第二章）中看到了letter-bigram的特性，，word-bigrams以及字母或单词的三元组（三种序列）也很常见。除此之外，4-grams和5-grams有时用于字母，但由于稀疏的问题，很少使用。从直观上看，为什么word gram比单个单词更能提供信息：它捕捉到了诸如New York，not good，Paris Hilton这样的结构。事实上，bigrams bag的表示比word biagrams更有力量，而且在很多情况下很难被击败。当然，并不是所有的biagram都是一样的信息，像“a”、“boy”之类的“双字母”是非常常见的，对于大多数任务来说，它们的信息并不比它们的单个组件更丰富。然而，很难预先知道哪些n gram对于给定的任务是有用的。一般的解决方案是把所有的n gram都包含到一个给定的长度，让模型正规化，通过分配非常低的权重来丢弃那些不太有趣的。

注意像MLP这样的神经网络架构不能在一般情况下推断n gram特征从一个文档里当输入词袋特征的时候。它可以学会X出现在文档中，Y出现在文档中，而不是X，Y 出现在文档中。因此gram特征在非线性分类的情况下也是很有用的。

多层感知器可以在应用于固定大小的窗口时推断出ngram，位置信息的组合在位置1的位置是X，而在位置2上的单词是Y实际上是bigram XY。更专业的神经网络体系结构，如卷积网络（第13章），旨在根据不同长度的单词序列，为给定的任务寻找信息性的ngram特性。双向rnn（第14章和第16章）进一步概括了ngram的概念，并且对不同长度的信息ng以及在有一定距离的ngram敏感。

### 分布特征

到目前为止，我们对单词的处理是离散的和不相关的符号：就算法而言，“披萨”、“汉堡”和“椅子”等词都是一样的（而且同样是相似的）。

我们通过将它们映射到更粗粒度的类别，例如语言的部分或语法角色（the、a、an都是决定者），从而实现了跨词类型的某种形式的泛化。从受影响的词形成到他们的引理（book，booking，booked都是引理）;看看名单或字典中的成员（约翰、杰克和拉尔夫出现在美国的共同名字中）;或者用诸如WordNet之类的词汇资源来观察它们与其他词的关系。然而，这些解决方案是相当有限的：它们要么提供非常粗粒度的差别，要么依赖于特定的手工编译的字典。除非我们有一份专门的食物清单，否则我们不会知道披萨与汉堡的相似度要比椅子更接近，而且要知道披萨与汉堡更相似，而不是吃冰淇淋，这将更加困难。

# 7 NLP特征的案例学习

这里讨论一些离散NLP分类任务的实例，为他们构建合适的特征。在设计神经网络的时候我们要确保网络能够高效的利用可得到的信号，既可以通过基于合适的设计好的特征或是设计网络架构探索需要的信号，或是加入附加的损失信号。

### 7.1 文档分类：语言判别

这里使用了字母组的形式作为核心特征，获得的是特征在文档中的计数。相似的任务是编码检测。这里使用比特组。

### 7.2 文档分类：主题分类

这里可能词包更有用，也许使用词组包。

如果我们没有很多训练样本，我们可以通过用它的引理来代替每个单词来处理文档。我们也可以用一些分布特征来代替或补充词语，比如词簇或词嵌入向量。

当使用线性分类器时，我们可能还需要考虑词对，也就是。考虑在同一文档中出现的每一对单词（不一定是连续的），作为核心特性。将会产生大量的潜在核心特性，并且需要通过设计一些启发式来减少这个数字，例如只考虑在特定数量的文档中出现的单词对。非线性分类器可以缓解这种需求。

当使用一组词时，有时用比例因子来衡量每个单词的信息量是有用的，例如使用TF-IDF加权（第6.2.1节）。然而，学习算法通常能够自己提出权重。另一种选择是使用word指示器，而不是单词计数：文档中的每个单词（或给定计数之上的每个单词）将被表示一次，而不管文档中出现的次数是多少。

## 7.3 文档分类：作者归因

得到了一份文本，需要推断出作者的身份（来自一组固定的可能作者），或者是文本作者的其他特征，比如他们的性别、年龄或他们的母语。

解决这一问题的信息与主题分类是很不相同的，这些线索是微妙的，涉及到文本的风格属性，而不是内容词。主要集中与语音属性和一些功能词。有些词像on，of，the，and等单词本身不带有自己的含义而是作为一些有内容的单词的连接和赋予他们各个成分的含义。一个近似功能词的列表是语料库出现频率前300的单词或是更多。通过集中于这些特征可以捕捉微妙的风格变化，这对作者来说是独一无二的，而且很难伪造。

一个好的作者属性识别特征包括功能词和代词的词袋，语音标签的词袋，语音标签组的词袋。另外，我们可能想要考虑功能词的密度（例如在文本窗口中，函数词的数量与内容词的数量之比），删除内容词后的功能词的词袋，以及连续函数词之间的距离的分布。

### 7.4 内容词：词性标注

在语音标记任务中，我们被给予一个句子，并且需要在句子中的每个单词中分配正确的词性。这部分的语性来自于一个预定义的集合，在这个例子中假设我们将使用通用Treebank项目的标记集，包括17个标记。

词性标注可以通常使用一个结构化的任务来建模，第一个词的标记也许依赖与第三个次的标记，但是可以通过将考虑单词两边形成的窗口的语义标签来给孤立的单词分类。如果我们以固定的顺序标记单词，例如从左到右，我们也可以做出标签估计根据以前的标签。我们的特征函数可以获得句子中的所有单词以及过去的标记决定。在这里，我们讨论一些特性，就好像它们是在一个孤立的分类任务中使用的。在第19章中，我们将使用相同的特性集讨论结构化学习案例。

### 7.5 内容词：命名实体识别

在命名实体识别中，给定一个文档，需要找到其中例如米兰，约翰·斯密斯，巴黎(类似专有名词)等命名实体，并分到预定义的类别中（地点，人名）等。这个任务是依赖于上下文的，例如米兰可能是地名，也可能是一个队名。

虽然NER是一个序列分割任务，它在不重叠的句子之间分配有标记的括号，但它通常被建模为一个序列标记任务，比如pos-tag。使用标记来解决分割任务是使用BIO标记执行的。每个词被分配了一个标签。

**Table 7.1:** BIO tags for named entity recognition

|  |  |
| --- | --- |
| **Tag** | **Meaning** |
| O | Not part of a named entity |
| B-PER I-PER | First word of a person name Continuation of a person name |
| B-LOC I-LOC | First word of a location name Continuation of a location name |
| B-ORG I-ORG | First word of an organization name Continuation of an organization name |
| B-MISC I-MISC | First word of another kind of named entity Continuation of another kind of named entity |

像词性标记一样，NER任务是一个结构化的任务，对于不同的单词的标记决定可能是相互作用的。然而我们再次假设可以近似的用独立的分类决定进行合理的推断。

NER任务与词性标注任务类似，依赖于焦点词两边的一个2个单词的窗口。除了词性标记任务的特征（一些前，后缀）对NER有用外，我们也想要考虑一个实体名词周围的其他相同单词的出现，并且使用指示函数检查单词是否出现在了预编译好的人名，位置名和公司名的列表里。分布式的特征例如单词聚类或词向量对于NER任务也是有用的。

### 7.6 语境中的词，语言特征：介词意思消除歧义

介词，像on，in，with和for起到连接作用用于将谓词与它们的参数和名词与它们的介词修饰符连接起来。介词是非常常见的，并且具有迷惑性。举例，下列句子中的for：

a. We went there for lunch.

b. He paid for me.

c. We ate for two hours.

d. He would have left for home, but it started raining.

这里单词for起到不同的作用，这里会跟随不同的短语代表不同的意思。

为了充分理解句子的意思，你应该会知道它里面介词的正确含义。介词语义消岐会在给定内容的下给介词在语义列表下给定一个正确的含义。在这里给出了一个介词的语义库，涵盖了4250个介词含义，每个都标注了含义。

这是一组很好的特性，用于介词的消除歧义任务？我们关注的是Hovy等人2010年工作所鼓励的特征。显然，介词本身是一个有用的特征（例如in的可能的语义分布与with或是about的分布是非常不同的）。除此之外，我们可以看到单词出现的上下文信息。在信息内容方面，围绕介词的固定窗口可能不太理想。例如，考虑以下句子。

a. He liked the round object from the very first time he saw it.

b. He saved the round object from him the very first time they saw it.

这里的两个from实例有不同的语义（两个from的搭配对象不一样，有的离搭配对象较远），但是在一个窗口下的单词大多数是要不没有信息或者甚至是误导性的。这里需要一种机制更好的机制选择上下文信息。

一个更好的方法是使用启发式的方法，例如左面的第一个动词，右面的第一个名词。从上面两句中会捕捉到saved from him 和liked from time，这样就包含了介词语义的必要成分。在语义上，这种启发式方法帮助捕捉介词的施动者和目标。。通过知道介词的直接管理者和目标就可以大概推论出介词的语义。使用细粒度的语义推理过程即可获得。这种方法要求获得词性的标注器为了识别名词和动词。这种方法可能会很脆弱不过很难假设它失败的情况，一种更鲁棒的方法是使用依赖关系树分析器，这种方法更易读，减少了复杂的启发式方法的需要。



当然，这个分析器用来产生的树可能是错误的。对于鲁棒性，我们可以使用两种方法提取介词的管理词和目标对象作为特征的源。

让学习过程决定哪个源更可靠和如何在他们之间进行平衡。在提取到管理词和目标后，我们可以用他们作为进一步特征提取的基础。这里可以提取下面的信息：

1.词的实际表面形式

2.词的引理词

3.词的词性

4.词的前缀和后缀

5.词聚类或是词的分布式向量

如果允许使用外部的词典资源和不介意大幅度增加特征空间，可以使用基于WordNet的特征将会很有帮助。对于每一个管理者和目标，我们提取许多WordNet指示器，例如：

1.词语是否有一个WordNet条目

2.单词第一个同义词集的上位词

3.单词所有同义词集的上位词

4.单词定义的所有条目

5. 这个词的超感（超感官，也被称为词汇网络术语中的词典-文件，在WordNet层级中是相对较高的层次，它表明了诸如动物、身体部分、情感、食物等概念。

6.不同的其他指示器

这个过程会获得上百个特征针对每个介词，实例。

介词的语义消岐是一种高层次的语义分类问题，因为我们需要一些不能通过表面形式读取的特征，可以收益于语言学上的预处理。（词性标注和语法分析）以及从一些人工词典中获得的语义片段。