第 1 章　自然语言处理领域探索

本书旨在帮助专业人士将自然语言处理（natural language processing，NLP）技术应用到他们的工作中，无论他们是在从事 NLP 项目还是在其他领域（例如数据科学）中使用 NLP。本书的目的是向你介绍自然语言处理领域及其底层技术，包括机器学习（machine learning，ML）和深度学习（deep learning，DL）。

本书强调数学基础（例如线性代数、统计和概率）以及优化理论的重要性，这些对于理解自然语言处理中使用的算法是必不可少的。本书还附有 Python 代码示例，可让你预先练习、实验并生成书中介绍的一些开发成果。

本书将讨论自然语言处理面临的挑战，例如理解单词的上下文和含义、单词之间的关系以及对标记数据的需求。

本书还将介绍自然语言处理的最新进展，包括 BERT 和 GPT 等预训练语言模型，以及大量文本数据的可用性，这些都提高了自然语言处理任务的性能。

本书将讨论语言模型对自然语言处理领域的影响，包括提高自然语言处理任务的准确率和有效性、开发更先进的自然语言处理系统以及让更广泛人群能够使用等。

本章包含以下主题：

- 自然语言处理的定义
- NLP 的历史和演变
- 自然语言机器处理的初步策略
- 成功的协同效应——自然语言处理与机器学习的结合
- 自然语言处理中的数学和统计学简介
- 理解语言模型——以 ChatGPT 为例

1.1　本书目标读者

本书的目标读者是那些需要在项目中处理文本的专业人士，这可能包括自然语言处理从业者（初学者也在此列）以及那些不以通常方式处理文本者。

1.2 自然语言处理的定义

NLP 是人工智能（artificial intelligence，AI）的一个领域，专注于计算机与人类语言之间的交互。它涉及使用计算技术来理解、解释和生成人类语言，使计算机能够自然而有意义地理解和响应人类的输入。

1.3 NLP 的历史和演变

对 NLP 历史的探索将让我们进入一段迷人的时光之旅，它最早可以追溯到 20 世纪 50 年代，因为正是在那个年代，艾伦·图灵（Alan Turing）等先驱为此做出了重大贡献。图灵的开创性论文 *Computing Machinery and Intelligence*（计算机器与智能）引入了图灵测试（Turing test）的概念，为未来在 AI 和 NLP 领域的探索奠定了基础。这一时期标志着符号 NLP 的诞生，其特点是使用基于规则的系统（rule-based system），例如 1954 年著名的乔治城实验（Georgetown experiment），该实验雄心勃勃地试图通过将俄语内容翻译成英语来解决机器翻译问题。有关该实验的详细信息，可访问：

```
https://en.wikipedia.org/wiki/Georgetown%E2%80%93IBM_experiment
```

乔治城实验引起了广泛的关注，成为机器翻译历史上最具影响力的实例之一，在当时引发了一股兴奋和乐观的狂潮，但事实证明，该项目的进展非常缓慢，直至最后不了了之。这也揭示了人类语言理解和生成的复杂性。

20 世纪 60 年代和 70 年代见证了早期自然语言处理系统的发展，该系统展示了机器使用有限的词汇和知识库进行类似人类交互的潜力。这个时代还见证了概念本体的创建，这对于以计算机可理解的格式构建现实世界的信息至关重要。

基于规则的系统的局限性导致了 20 世纪 80 年代后期科学家们的转向，他们开始转向统计 NLP 范式，这也得益于机器学习的进步和计算能力的提高。

这种转变使得机器从大型语料库中更有效地学习成为可能，大大推进了机器翻译和其他自然语言处理任务的发展。这种范式转变不仅代表了技术和方法的进步，而且还强调了自然语言处理中语言学方法的概念演变。

在摆脱预定义语法规则的僵化机制之后，这种转变采用了语料库语言学（corpus linguistics），这种方法允许机器通过大量接触文本来“感知”和理解语言。这种方法反映了

对语言的更加经验化和数据驱动的理解，其中的模式和含义来自实际的语言使用而不是理论构造，从而实现了更加细致入微和灵活的语言处理能力。

进入 21 世纪，网络的出现提供了大量数据，促进了无监督和半监督学习算法的研究。

2010 年，神经网络自然语言处理的出现带来了突破，深度学习技术开始占据主导地位，在语言建模和解析方面提供了前所未有的准确性。这个时代的特点是 Word2Vec 等复杂模型的发展和深度神经网络的普及，推动 NLP 朝着更自然以及更有效的人机交互的方向发展。随着我们继续推进这些进步，NLP 站在了人工智能研究的最前沿，其历史反映了人们对理解和复制人类语言细微差别的不懈追求。

近年来，NLP 还被广泛应用于医疗保健、金融和社交媒体等众多行业，用于自动化决策和增强人机之间的沟通。例如，NLP 已用于分析客户反馈、从医疗文档中提取信息、在不同语言之间翻译文档以及搜索大量帖子等。

1.4　自然语言机器处理的初步策略

自然语言机器处理的传统方法通常是先进行文本预处理（text preprocessing）——文本准备（text preparation），然后应用机器学习方法。

文本预处理是自然语言处理和机器学习应用中必不可少的步骤，其操作主要是清洗和转换原始文本数据，使其成为机器学习算法可以轻松理解和分析的形式。

预处理的目标是消除噪声和不一致之处，并使数据标准化，使其更适合高级自然语言处理和机器学习方法。

预处理的一个主要好处是它可以显著提高机器学习算法的性能。例如，删除停用词（即没有太多含义的常用词，如英语的"the"和"is"，中文的“哎呀”和“哦”等）可以帮助降低数据的维度，使算法更容易识别模式。

以下面的句子为例：

```
I am going to the store to buy some milk and bread.
```

删除停用词后，得到以下内容：

```
going store buy milk bread.
```

在上述例句中，停用词"I" "am" "to" "the" "some"和"and"不会给句子增加任何额外的含义，可以删除而不会改变句子的整体含义。

需要强调的是，停用词的删除需要根据具体目标进行量身定制，因为删除某个词在某

种情况下可能微不足道，但在另一种情况下却可能有很大影响。

此外，词干提取（stemming）和词形还原（lemmatization，指将单词简化为其基本形式）可以帮助减少数据中唯一单词的数量，从而使算法更容易识别它们之间的关系。下文将更详细地解释这些操作。

以下面的句子为例：

```
The boys ran, jumped, and swam quickly.
```

在应用词干提取之后，将每个单词简化为其词根或词干形式，忽略单词时态或派生词缀，可得到以下内容：

```
The boy ran, jump, and swam quick.
```

词干提取可将文本简化为其基本形式。在上述示例中，"ran" "jumped"和"swam"分别简化为"ran" "jump"和"swam"。

可以看到，"ran"和"swam"没有变化，这是因为词干提取通常会产生接近其词根形式但不完全是词典基本形式的单词。此过程有助于降低文本数据的复杂性，使机器学习算法更容易匹配和分析模式，而不会因同一单词的变体而陷入困境。

以下面的句子为例：

```
The boys ran, jumped, and swam quickly.
```

词形还原将考虑单词的形态分析，旨在返回单词的基本形式或词典形式（称为词根），在应用词形还原之后，这样得到的是以下内容：

```
The boy run, jump, and swim quickly.
```

可以看到，词形还原可将"ran" "jumped"和"swam"准确地转换为"run" "jump"和"swim"。也就是说，此过程考虑了每个单词的词性，确保在语法和语境意义上还原为恰当的基本形式。

与词干提取不同，词形还原可以更精确地还原基本形式，确保处理后的文本仍然有意义且上下文是准确的。这提高了自然语言处理模型的性能，使它们能够更有效地理解和处理语言，降低数据集的复杂性，同时保持原始文本的完整性。

预处理的另外两个重要方面是数据规范化（data normalization）和数据清洗（data cleaning）。数据规范化包括将所有文本转换为小写、删除标点符号以及标准化数据格式。这有助于确保算法不会将同一单词的不同变体视为单独的实体，从而导致结果不准确。

数据清洗包括删除重复或不相关的数据，以及纠正数据中的错误或不一致之处。这在

大型数据集中尤其重要，因为手动清洗既耗时又容易出错。自动预处理工具可以帮助快速识别和消除错误，使数据更可靠，便于分析。

图 1.1 描绘了一个全面的预处理流程。我们将在第 4 章中介绍此代码示例。

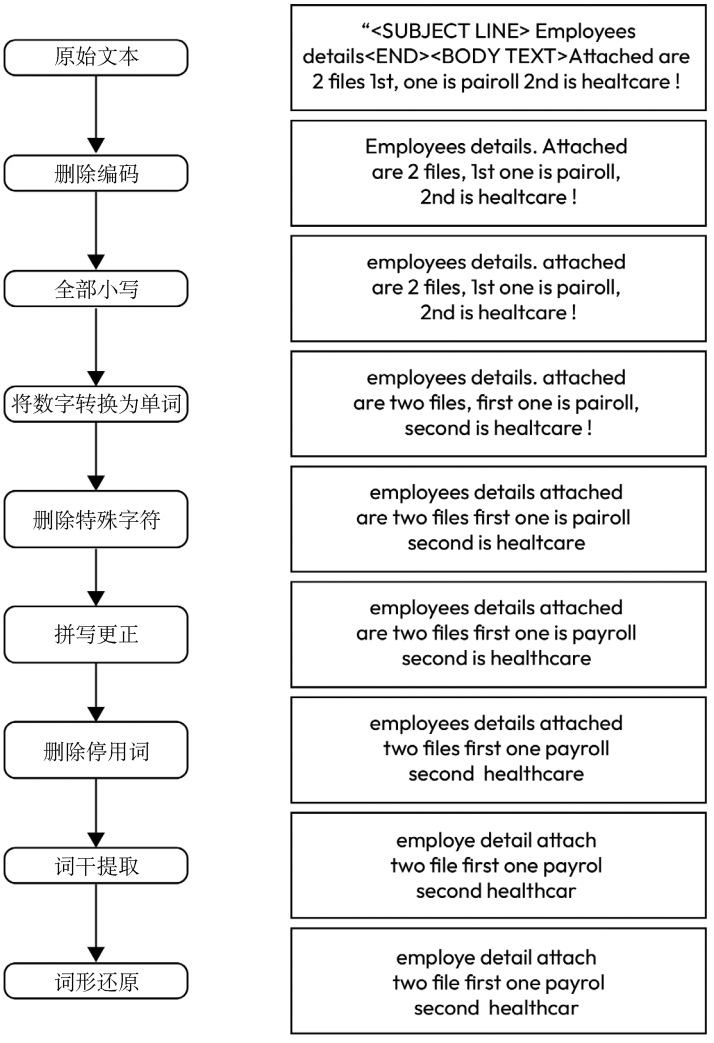

图 1.1　全面的预处理流程

总之，文本预处理是自然语言处理和机器学习应用中的重要步骤，它可以通过消除噪声和不一致之处并标准化数据来提高机器学习算法的性能。

在自然语言处理任务中，数据准备和数据清洗起着至关重要的作用。通过在预处理上投入时间和资源，可以确保数据质量高，并为高级自然语言处理和机器学习方法做好准备，从而获得更准确、更可靠的结果。

在准备好文本数据之后，接下来要做的就是为其拟合机器学习模型。

1.5　成功的协同效应——自然语言处理与机器学习的结合

机器学习（ML）是人工智能的一个分支，它指的是训练算法从数据中学习，让算法无须明确编程即可做出预测或决策。机器学习正在推动许多不同领域的进步，例如计算机视觉、语音识别，当然还有 NLP。

如果你深入研究机器学习的具体技术，则会发现 NLP 中使用的一种特殊技术是统计语言建模（statistical language modeling），它涉及在大型文本语料库上训练算法以预测给定单词序列的可能性。这有广泛的应用，例如语音识别、机器翻译和文本生成。

另一项重要技术是深度学习，它是机器学习的一个子领域，涉及在大量数据上训练人工神经网络。深度学习模型——例如卷积神经网络（convolutional neural network，CNN）和循环神经网络（recurrent neural network，RNN）已被证明适用于语言理解、文本摘要和情感分析等自然语言处理任务。

图 1.2 描绘了 AI、ML、DL 和 NLP 之间的关系。

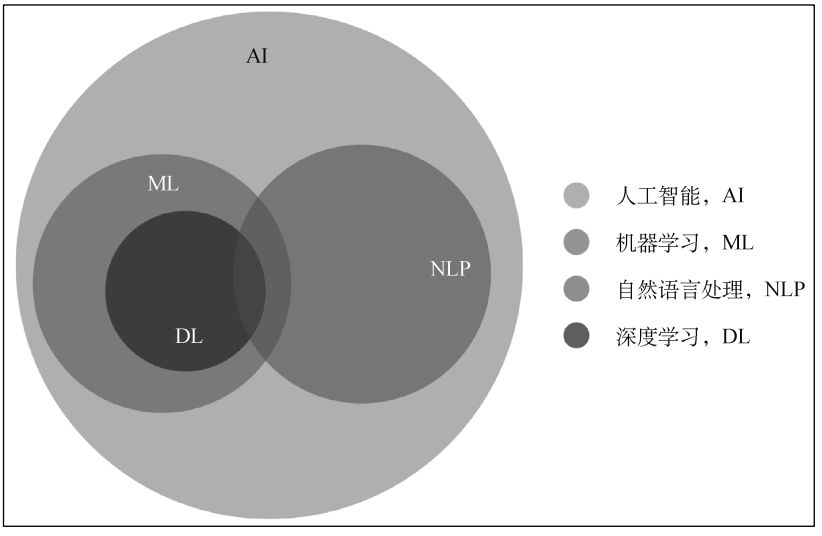

图 1.2　不同学科之间的关系

1.6　自然语言处理中的数学和统计学简介

NLP 和机器学习的坚实基础是算法所基于的数学知识。具体来说，关键基础是线性代数、统计和概率以及优化理论。第 2 章将介绍理解这些主题所需的关键概念。本书将提供各种方法和假设的证明和论证。

NLP 的挑战之一是处理人类语言产生的大量数据。这包括理解上下文、单词的含义以及它们之间的关系。为了应对这一挑战，研究人员开发了各种技术，例如嵌入（embedding）和注意力（attention）机制，它们分别以数字格式表示单词的含义并可以帮助识别文本中最关键的部分。

NLP 的另一个挑战是需要已标记的数据，因为手动注释大型文本语料库既昂贵又耗时。为了解决这个问题，研究人员开发了可以从未标记的数据中学习的无监督和弱监督方法，例如聚类（clustering）、主题建模（topic modeling）和自监督学习（self-supervised learning）。

总体而言，NLP 是一个快速发展的领域，有可能改变我们与计算机和信息进行交互的方式。它的应用非常广泛，从聊天机器人（chatbot）、语言翻译、文本摘要（text summarization）到情感分析（sentiment analysis），都可以看到它的身影。使用统计语言建模和深度学习等机器学习技术对于开发这些系统至关重要。还有一些研究正在进行中，它们将解决剩余的挑战，例如理解上下文和处理缺乏已标记数据的问题。

NLP 领域最重大的进步之一是预训练语言模型的开发，例如 bidirectional encoder representations from transformers（BERT）和 generative pre-trained transformer（GPT）。这些模型都已经在大量文本数据上进行了训练，并且可以针对特定任务（例如情感分析或语言翻译）进行微调。

Transformer 是 BERT 和 GPT 模型背后的技术，它使机器能够更有效地理解句子中单词的上下文，从而彻底改变了 NLP。与以前线性处理文本的方法不同，Transformer 可以并行处理单词，通过注意力机制捕捉语言中的细微差别。这使它们能够辨别每个单词相对于其他单词的重要性，大大增强了模型掌握复杂语言模式和细微差别的能力，并为自然语言处理应用程序的准确性和流畅性树立了新标准。这促进了自然语言处理应用程序的创建，并提高了各种自然语言处理任务的性能。

图 1.3 详细说明了 Transformer 组件的功能设计。

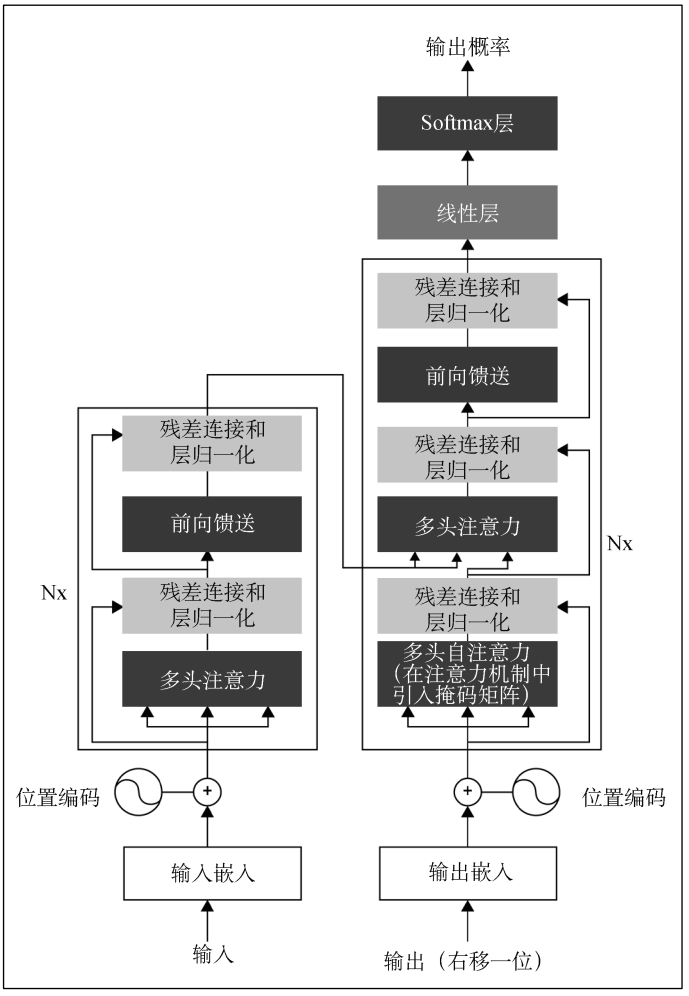

图 1.3　模型架构中的 Transformer

NLP 的另一个重要发展是大量带注释的文本数据的可用性增加，这使得训练更准确的模型成为可能。此外，无监督和半监督学习技术的发展使得在较少量的标记数据上训练模型成为可能，从而有可能将自然语言处理应用于更广泛的场景。

语言模型对自然语言处理领域产生了重大影响。语言模型改变该领域的关键方式之一是提高自然语言处理任务的准确率和有效性。例如，许多语言模型都经过大量文本数据的训练，使它们能够更好地理解人类语言的细微差别和复杂性。这提高了语言翻译、文本摘要和情感分析等任务的性能。

语言模型改变自然语言处理领域的另一种方式是，它能够开发更先进、更复杂的自然语言处理系统。例如，某些语言模型（如 GPT）可以生成类似人类写作风格的文本，这为自然语言生成和对话系统开辟了新的可能性。其他语言模型（如 BERT）则提高了问答、情感分析和命名实体识别等任务的性能。

语言模型也改变了这一领域，让更多人能够接触到它。随着预训练语言模型的出现，开发人员现在可以轻松地针对特定任务微调这些模型，而不需要大量的标记数据或从头开始训练模型的专业知识。这使得开发人员可以更轻松地构建自然语言处理应用程序，并导致基于自然语言处理的新产品和服务激增。

总体而言，语言模型通过提高现有自然语言处理任务的性能、推动更先进的自然语言处理系统的开发，以及使更广泛的人群能够使用 NLP，在推动自然语言处理领域的发展方面发挥了关键作用。

1.7　理解语言模型——以 ChatGPT 为例

大名鼎鼎的 ChatGPT 是 GPT 模型的一个变体，由于其能够生成类似人类写作风格的文本而变得流行，可用于广泛的自然语言生成任务，例如聊天机器人系统、文本摘要和对话系统。

它受欢迎的主要原因是其高质量的输出以及生成与人类书写的文本难以区分的文本的能力。这使得它非常适合需要自然发音文本的应用程序，例如聊天机器人系统、虚拟助手和文本摘要。

此外，ChatGPT 已在大量文本数据上进行预训练，因此能够理解人类语言的细微差别和复杂性。这使其非常适合需要深入理解语言的应用程序，例如问答和情感分析。

此外，通过提供少量与特定任务相关的数据，ChatGPT 可以针对特定用例进行微调，这使其用途更广泛，适用于各种场景。如今，ChatGPT 及其类似产品已广泛应用于行业、研究和个人项目，包括客户服务聊天机器人、虚拟助手、自动内容创建、文本摘要、对话系统、问答和情感分析等。

总体而言，ChatGPT 生成高质量、类似人类写作风格的文本的能力及其针对特定任务进行微调的能力使其成为众多自然语言生成应用的热门选择。

1.8　小　　结

本章引导你探索了自然语言处理领域，它是人工智能的一个子领域。

我们强调了数学基础（例如线性代数、统计和概率以及优化理论）的重要性，这些知识对于理解自然语言处理中使用的算法是必不可少的。

本章还讨论了自然语言处理面临的挑战，例如理解单词的上下文和含义、它们之间的关系以及对已标记数据的需求。

我们介绍了自然语言处理的最新进展，包括预训练语言模型（例如 BERT 和 GPT）以及大量文本数据的可用性，这提高了自然语言处理任务的性能。

我们谈到了文本预处理的重要性，因为你需要了解数据清洗、数据规范化、词干提取和词形还原在文本预处理中的重要性。

此外，我们还讨论了自然语言处理和机器学习的结合如何推动该领域的进步，并成为自动化任务和改善人机交互的越来越重要的工具。

学习完本章之后，你将能够理解 NLP、ML 和 DL 技术的重要性，了解自然语言处理的最新进展，包括预训练语言模型。你还将了解文本预处理的重要性以及它如何在自然语言处理任务的数据准备和数据清洗中发挥关键作用。

在下一章中，我们将介绍机器学习的数学基础。这些基础知识将贯穿整本书。

1.9 问　　答

（1）什么是自然语言处理（NLP）？

- 问：人工智能领域中 NLP 的定义是什么？
- 答：NLP 是 AI 的一个子领域，专注于使计算机能够以对人类用户自然且有意义的方式理解、解释和生成人类语言。

（2）自然语言机器处理的初步策略。

- 问：预处理在 NLP 中有何重要性？
- 答：预处理（包括删除停用词、应用词干提取或词形还原等任务）对于清洗和准备文本数据从而提高机器学习算法在自然语言处理任务上的性能至关重要。

（3）NLP 与机器学习（ML）的协同作用。

- 问：机器学习如何促进自然语言处理的进步？
- 答：机器学习，尤其是统计语言建模和深度学习等技术，通过使算法能够从数据中学习、预测词序列以及更有效地执行语言理解和情感分析等任务，推动了自然语言处理的发展。

（4）自然语言处理中的数学和统计学简介。

- 问：为什么数学基础在自然语言处理中很重要？

- 答：线性代数、统计学和概率等数学基础对于理解和开发自然语言处理技术所依赖的算法（从基本的预处理到复杂的模型训练）至关重要。

（5）自然语言处理的进步——预训练语言模型的作用。

- 问：BERT、GPT 等预训练模型对自然语言处理有何影响？
- 答：预训练模型是在海量文本数据上进行训练的，可以针对情感分析或语言翻译等特定任务进行微调，从而显著简化自然语言处理应用程序的开发并提高任务性能。

（6）理解语言模型中的 Transformer。

- 问：为什么 Transformer 被认为是自然语言处理领域的突破？
- 答：Transformer 可并行处理单词，并使用注意力机制来理解句子中的单词上下文，从而显著提高模型处理人类语言复杂性的能力。