CHAPTER 1

第 1 章 NLP 基础

在本章中，你将学到与 NLP（自然语言处理）相关的基础知识。

本章的要点包括：

- ▼ NLP 基础概念
- ▼ NLP 的发展与应用
- ▼ NLP 常用术语以及扩展介绍

1.1 什么是 NLP

1.1.1 NLP 的概念

NLP（Natural Language Processing，自然语言处理）是计算机科学领域以及人工智能领域的一个重要的研究方向，它研究用计算机来处理、理解以及运用人类语言（如中文、英文等），达到人与计算机之间进行有效通讯。所谓“自然”乃是寓意自然进化形成，是为了区分一些人造语言，类似 C++、Java 等人为设计的语言。在人类社会中，语言扮演着重要的角色，语言是人类区别于其他动物的根本标志，没有语言，人类的思维无从谈起，沟通交流更是无源之水。在一般情况下，用户可能不熟悉机器语言，所以自然语言处理技术可以帮助这样的用户使用自然语言和机器交流。从建模的角度看，为了方便计算机处理，自然语言可以被定义为一组规则或符号的集合，我们组合集合中的符号来

传递各种信息。自然语言处理研究表示语言能力、语言应用的模型，通过建立计算机框架来实现这样的语言模型，并且不断完善这样的语言模型，还需要根据该语言模型来设计各种实用的系统，并且探讨这些实用技术的评测技术。这一定义有点宽泛，但是语言本身就是人类最为复杂的概念之一。这些年，NLP 研究取得了长足的进步，逐渐发展成为一门独立的学科，从自然语言的角度出发，NLP 基本可以分为两个部分：自然语言处理以及自然语言生成，演化为理解和生成文本的任务，如图 1-1 所示。

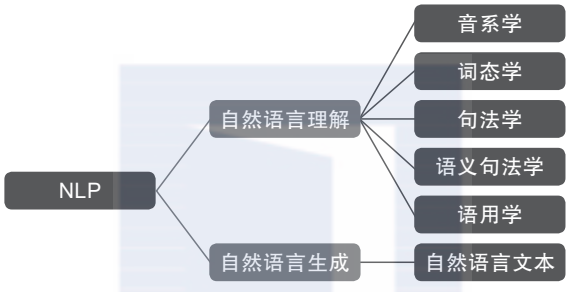

图 1-1　NLP 的基本分类

自然语言的理解是个综合的系统工程，它又包含了很多细分学科，有代表声音的音系学，代表构词法的词态学 , 代表语句结构的句法学，代表理解的语义句法学和语用学。

- ▼ 音系学：指代语言中发音的系统化组织。
- ▼ 词态学：研究单词构成以及相互之间的关系。
- ▼ 句法学：给定文本的哪部分是语法正确的。
- ▼ 语义学：给定文本的含义是什么？
- ▼ 语用学：文本的目的是什么？

语言理解涉及语言、语境和各种语言形式的学科。而自然语言生成（Natural Language Generation，NLG）恰恰相反，从结构化数据中以读取的方式自动生成文本。该过程主要包含三个阶段：文本规划（完成结构化数据中的基础内容规划）、语句规划（从结构化数据中组合语句来表达信息流）、实现（产生语法通顺的语句来表达文本）。

1.1.2 NLP 的研究任务

NLP 可以被应用于很多领域，这里大概总结出以下几种通用的应用：

- ▼ 机器翻译：计算机具备将一种语言翻译成另一种语言的能力。
- ▼ 情感分析：计算机能够判断用户评论是否积极。
- ▼ 智能问答：计算机能够正确回答输入的问题。
- ▼ 文摘生成：计算机能够准确归纳、总结并产生文本摘要。
- ▼ 文本分类：计算机能够采集各种文章，进行主题分析，从而进行自动分类。
- ▼ 舆论分析：计算机能够判断目前舆论的导向。
- ▼ 知识图谱：知识点相互连接而成的语义网络。

机器翻译是自然语言处理中最为人所熟知的场景，国内外有很多比较成熟的机器翻译产品，比如百度翻译、Google 翻译等，还有提供支持语音输入的多国语言互译的产品（比如科大讯飞就出了一款翻译机）。

体验下，百度在线翻译：

http://fanyi.baidu.com/?aldtype=16047#auto/zh

情感分析在一些评论网站比较有用，比如某餐饮网站的评论中会有非常多拔草的客人的评价，如果一眼扫过去满眼都是又贵又难吃，那谁还想去呢？另外有些商家为了获取大量的客户不惜雇佣水军灌水，那就可以通过自然语言处理来做水军识别，情感分析来分析总体用户评价是积极还是消极。

智能问答在一些电商网站有非常实际的价值，比如代替人工充当客服角色，有很多基本而且重复的问题，其实并不需要人工客服来解决，通过智能问答系统可以筛选掉大量重复的问题，使得人工座席能更好地服务客户。

体验下，图灵机器人：

http://www.tuling123.com/experience/exp_virtual_robot.jhtml? nav=exp

文摘生成利用计算机自动地从原始文献中摘取文摘，全面准确地反映某一文献的中心内容。这个技术可以帮助人们节省大量的时间成本，而且效率更高。

文本分类是机器对文本按照一定的分类体系自动标注类别的过程。举一个例子，垃圾邮件是一种令人头痛的顽症，困扰着非常多的互联网用户。2002 年，Paul Graham 提出使用“贝叶斯推断”来过滤垃圾邮件，1000 封垃圾邮件中可以过滤掉 995 封并且没有一个是误判，另外这种过滤器还具有自我学习功能，会根据新收到的邮件，不断调整。也就是说收到的垃圾邮件越多，相对应的判断垃圾邮件的准确率就越高。

舆论分析可以帮助分析哪些话题是目前的热点，分析传播路径以及发展趋势，对于不好的舆论导向可以进行有效的控制。

知识图谱（Knowledge Graph/Vault）又称科学知识图谱，在图书情报界称为知识域可视化或知识领域映射地图，是显示知识发展进程与结构关系的一系列各种不同的图形，用可视化技术描述知识资源及其载体，挖掘、分析、构建、绘制和显示知识及它们之间的相互联系。知识图谱的一般表现形式如图 1-2 所示。

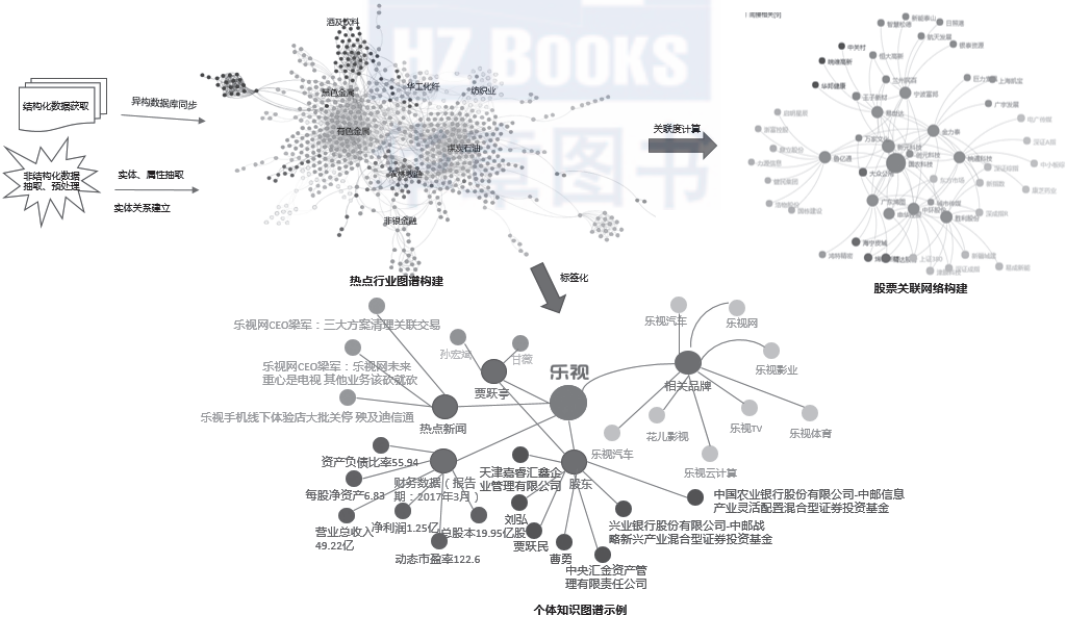

图 1-2　知识图谱图示

1.2 NLP 的发展历程

NLP 的发展大致经历了 3 个阶段：1956 年以前的萌芽期，1980 年～ 1999 年的快速发展期和 21 世纪的突飞猛进期。

萌芽期（1956 年以前）

早期的自然语言处理具有鲜明的经验主义色彩。如 1913 年马尔科夫提出马尔可夫随机过程与马尔可夫模型的基础就是“手工查频”，具体说就是统计了《欧根·奥涅金》长诗中元音与辅音出现频度；1948 年香农把离散马尔可夫的概率模型应用于语言的自动机，同时采用手工方法统计英语字母的频率。

然而这种经验主义到了乔姆斯基时期出现了转变。1956 年乔姆斯基借鉴香农的工作，把有限状态机作为刻画语法的工具，建立了自然语言的有限状态模型，具体来说就是用“代数”和“集合”将语言转化为符号序列，建立了一大堆有关语法的数学模型。这些工作非常伟大，为自然语言和形式语言找到了一种统一的数学描述理论，一个叫做“形式语言理论”的新领域诞生了。但乔姆斯基否定了有限状态模型在自然语言中的适用性，然后主张采用有限的、严格的规则去描述无限的语言现象，提出了风靡一时的转换生成语法。这一时期，虽然诸如贝叶斯方法、隐马尔可夫、最大熵、支持向量机等经典理论和算法也有提出，但自然语言处理领域的主流仍然是基于规则的理性主义方法。

快速发展期（1980 年 ～ 1999 年）

这种情况一直持续到 20 世纪 80 年代初期才发生变化，很多学者开始反思有限状态模型以及经验主义方法的合理性。20 世纪 80 年代初，话语分析（Discourse Analysis）也取得了重大进展。之后，由于自然语言处理研究者对于过去的研究进行了反思，有限状态模型和经验主义研究方法也开始复苏。

90 年代后，基于统计的自然语言处理开始大放异彩。首先是在机器翻译领域取得了突破，因为引入了许多基于语料库的方法。1990 年在芬兰赫尔辛基举办的第 13 届国际

计算语言学会议确定的主题是“处理大规模真实文本的理论、方法与工具”，研究的重心开始转向大规模真实文本了，传统的基于规则的自然语言处理显然力不从心了。学者们认为，大规模语料至少是对基于规则方法有效的补充。在 1994 年～ 1999 年间，经验主义空前繁荣，如句法剖析、词类标注、参照消解、话语处理的算法几乎把“概率”与“数据”作为标准方法，成为自然语言处理的主流。

20 世纪 90 年代中期，有两件事从根本上促进了自然语言处理研究的复苏与发展。一件事是 20 世纪 90 年代中期以来，计算机的运行速度和存储量大幅增加，为自然语言处理改善了物质基础，使得语音和语言处理的商品化开发成为可能；另一件事是 1994 年 Internet 商业化和同期网络技术的发展使得基于自然语言的信息检索和信息抽取的需求变得更加突出。这样，自然语言处理的社会需求更加迫切，自然语言处理的应用面也更加宽广，自然语言处理不再局限于机器翻译、语音控制等早期研究领域了。

从 20 世纪 90 年代末到 21 世纪初，人们逐渐认识到，仅用基于规则或统计的方法是无法成功进行自然语言处理的。基于统计、基于实例和基于规则的语料库技术在这一时期开始蓬勃发展，各种处理技术开始融合，自然语言处理的研究再次繁荣。

突飞猛进期（2000 年至今）

进入 21 世纪以后，自然语言处理又有了突飞猛进的变化。2006 年，以 Hinton 为首的几位科学家历经近 20 年的努力，终于成功设计出第一个多层神经网络算法——深度学习。这是一种将原始数据通过一些简单但是非线性的模型转变成更高层次、更加抽象表达的特征学习方法，一定程度上解决了人类处理“抽象概念”这个亘古难题。目前，深度学习在机器翻译、问答系统等多个自然语言处理任务中均取得了不错的成果，相关技术也被成功应用于商业化平台中。

未来，深度学习作为人工智能皇冠上的明珠，将会在自然语言处理领域发挥着越来越重要的作用。

1.3 NLP 相关知识的构成

1.3.1 基本术语

为了帮助读者更好地学习 NLP，这里会一一介绍 NLP 领域的一些基础专业词汇。

（1）分词（segment）

词是最小的能够独立活动的有意义的语言成分，英文单词之间是以空格作为自然分界符的，而汉语是以字为基本的书写单位，词语之间没有明显的区分标记，因此，中文词语分析是中文分词的基础与关键。中文和英文都存在分词的需求，不过相较而言，英文单词本来就有空格进行分割，所以处理起来相对方便。但是，由于中文是没有分隔符的，所以分词的问题就比较重要。分词常用的手段是基于字典的最长串匹配，据说可以解决 85% 的问题，但是歧义分词很难。举个例子，“美国会通过对台售武法案”，我们既可以切分为“美国 / 会 / 通过对台售武法案”，又可以切分成“美 / 国会 / 通过对台售武法案”。

（2）词性标注（part-of-speech tagging）

基于机器学习的方法里，往往需要对词的词性进行标注。词性一般是指动词、名词、形容词等。标注的目的是表征词的一种隐藏状态，隐藏状态构成的转移就构成了状态转移序列。例如：我 /r 爱 /v 北京 /ns 天安门 /ns。其中，ns 代表名词，v 代表动词，ns、v 都是标注，以此类推。

（3）命名实体识别（NER，Named Entity Recognition）

命名实体是指从文本中识别具有特定类别的实体（通常是名词），例如人名、地名、机构名、专有名词等。

（4）句法分析（syntax parsing）

句法分析往往是一种基于规则的专家系统。当然也不是说它不能用统计学的方法进

行构建，不过最初的时候，还是利用语言学专家的知识来构建的。句法分析的目的是解析句子中各个成分的依赖关系。所以，往往最终生成的结果是一棵句法分析树。句法分析可以解决传统词袋模型不考虑上下文的问题。比如，“小李是小杨的班长”和“小杨是小李的班长”，这两句话，用词袋模型是完全相同的，但是句法分析可以分析出其中的主从关系，真正理清句子的关系。

（5）指代消解（anaphora resolution）

中文中代词出现的频率很高，它的作用的是用来表征前文出现过的人名、地名等。例如，清华大学坐落于北京，这家大学是目前中国最好的大学之一。在这句话中，其实“清华大学”这个词出现了两次，“这家大学”指代的就是清华大学。但是出于中文的习惯，我们不会把“清华大学”再重复一遍。

（6）情感识别（emotion recognition）

所谓情感识别，本质上是分类问题，经常被应用在舆情分析等领域。情感一般可以分为两类，即正面、负面，也可以是三类，在前面的基础上，再加上中性类别。一般来说，在电商企业，情感识别可以分析商品评价的好坏，以此作为下一个环节的评判依据。通常可以基于词袋模型 + 分类器，或者现在流行的词向量模型 +RNN。经过测试发现，后者比前者准确率略有提升。

（7）纠错（correction）

自动纠错在搜索技术以及输入法中利用得很多。由于用户的输入出错的可能性比较大，出错的场景也比较多。所以，我们需要一个纠错系统。具体做法有很多，可以基于 N-Gram 进行纠错，也可以通过字典树、有限状态机等方法进行纠错。

（8）问答系统（QA system）

这是一种类似机器人的人工智能系统。比较著名的有：苹果 Siri、IBM Watson、微软小冰等。问答系统往往需要语音识别、合成，自然语言理解、知识图谱等多项技术的

配合才会实现得比较好。

1.3.2 知识结构

作为一门综合学科，NLP 是研究人与机器之间用自然语言进行有效通信的理论和方法。这需要很多跨学科的知识，需要语言学、统计学、最优化理论、机器学习、深度学习以及自然语言处理相关理论模型知识做基础。作为一门杂学，NLP 可谓是包罗万象，体系化与特殊化并存，这里简单罗列其知识体系：

- 句法语义分析：针对目标句子，进行各种句法分析，如分词、词性标记、命名实体识别及链接、句法分析、语义角色识别和多义词消歧等。
- 关键词抽取：抽取目标文本中的主要信息，比如从一条新闻中抽取关键信息。主要是了解是谁、于何时、为何、对谁、做了何事、产生了有什么结果。涉及实体识别、时间抽取、因果关系抽取等多项关键技术。
- 文本挖掘：主要包含了对文本的聚类、分类、信息抽取、摘要、情感分析以及对挖掘的信息和知识的可视化、交互式的呈现界面。
- 机器翻译：将输入的源语言文本通过自动翻译转化为另一种语言的文本。根据输入数据类型的不同，可细分为文本翻译、语音翻译、手语翻译、图形翻译等。机器翻译从最早的基于规则到二十年前的基于统计的方法，再到今天的基于深度学习（编解码）的方法，逐渐形成了一套比较严谨的方法体系。
- 信息检索：对大规模的文档进行索引。可简单对文档中的词汇，赋以不同的权重来建立索引，也可使用算法模型来建立更加深层的索引。查询时，首先对输入比进行分析，然后在索引里面查找匹配的候选文档，再根据一个排序机制把候选文档排序，最后输出排序得分最高的文档。
- 问答系统：针对某个自然语言表达的问题，由问答系统给出一个精准的答案。需要对自然语言查询语句进行语义分析，包括实体链接、关系识别，形成逻辑表达式，然后到知识库中查找可能的候选答案并通过一个排序机制找出最佳的答案。
- 对话系统：系统通过多回合对话，跟用户进行聊天、回答、完成某项任务。主要

涉及用户意图理解、通用聊天引擎、问答引擎、对话管理等技术。此外，为了体现上下文相关，要具备多轮对话能力。同时，为了体现个性化，对话系统还需要基于用户画像做个性化回复。知识结构结构图如图 1-3 所示。

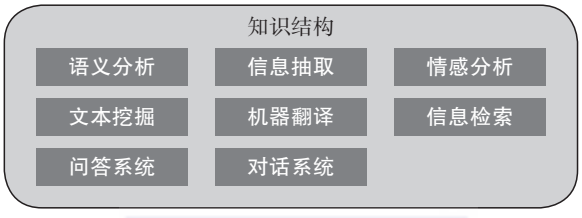

图 1-3　知识结构图示

扩展阅读

自然语言的学习，需要有以下几个前置知识体系：

▼ 目前主流的自然语言处理技术使用 python 来编写。
▼ 统计学以及线性代数入门。

1.4　语料库

巧妇难为无米之炊，语料库就是 NLP 的“米”，本书用到的语料库主要有：

（1）中文维基百科[⊖]

维基百科是最常用且权威的开放网络数据集之一，作为极少数的人工编辑、内容丰富、格式规范的文本语料，各类语言的维基百科在 NLP 等诸多领域应用广泛。维基百科提供了开放的词条文本整合下载，可以找到你需要的指定时间、指定语言、指定类型、指定内容的维基百科数据，中文维基百科数据是维基提供的语料库。

⊖ https://dumps.wikimedia.org/zhwiki/。

（2）搜狗新闻语料库[⊖]

来自若干新闻站点 2012 年 6 月～ 7 月期间国内、国际、体育、社会、娱乐等 18 个频道的新闻数据，提供 URL 和正文信息。

（3）IMDB 情感分析语料库[⊜]

互联网电影资料库（Internet Movie Database，简称 IMDb）是一个关于电影演员、电影、电视节目、电视明星和电影制作的在线数据库。IMDb 的资料中包括了影片的众多信息、演员、片长、内容介绍、分级、评论等。对于电影的评分目前使用最多的就是 IMDb 评分。

还有豆瓣读书相关语料（爬虫获取）、邮件相关语料等。

1.5 探讨 NLP 的几个层面

本书所探讨的自然语言处理可以分为以下三个层面：

（1）第一层面：词法分析

词法分析包括汉语的分词和词性标注这两部分。之前有提过，汉语分词与英文不同，汉语书面词语之间没有明显的空格标记，文本中的句子以字符串的方式出现，句子中由逗号分隔，句子和句子之间常以句号分隔。针对汉语这种独特的书面表现形式，汉语的自然语言处理的首要工作就是要将输入的文本切分为单独的词语，然后在此技术上进行其他更高级的分析。

上述这个步骤称为分词。除了分词之外，词性标注也通常被认为是词法分析的一部分，词性标注的目的是为每一个词赋予一个类别，这个类别可以是名词（noun）、动

⊖ http://download.labs.sogou.com/resource/ca.php。
⊜ https://www.kaggle.com/tmdb/tmdb-movie-metadata。

词（verb）、形容词（adjective）等。通常来说，属于相同词性的词，在句法中承担类似的角色。

（2）第二层面：句法分析

句法分析是对输入的文本以句子为单位，进行分析以得到句子的句法结构的处理过程。对句法结构进行分析，一方面是为了帮助理解句子的含义，另一方面也为更高级的自然语言处理任务提供支持（比如机器翻译、情感分析等）。目前业界存在三种比较主流的句法分析方法：短语结构句法体系，作用是识别出句子中的短语结构以及短语之间的层次句法关系；依存结构句法体系，作用是识别句子中词与词之间的相互依赖关系；深层文法句法分析，利用深层文法，例如词汇化树邻接文法，组合范畴文法等对句子进行深层的句法以及语义分析。

上述几种句法分析，依存句法分析属于浅层句法分析，其实现过程相对来说比较简单而且适合在多语言环境下应用，但是其所能提供的信息也相对较少。深层文法句法分析可以提供丰富的句法和语义信息，但是采用的文法相对比较复杂，分析器的运行复杂度也比较高，这使得深层句法分析不太适合处理大规模的数据。短语结构句法分析介于依存句法分析和深层文法句法分析之间。

（3）第三个层面：语义分析

语义分析的最终目的是理解句子表达的真实语义。但是，语义应该采用什么表示形式一直困扰着研究者们，至今这个问题也没有一个统一的答案。语义角色标注（semantic role labeling）是目前比较成熟的浅层语义分析技术。语义角色标注一般都在句法分析的基础上完成，句法结构对于语义角色标注的性能至关重要。基于逻辑表达的语义分析也得到学术界的长期关注。出于机器学习模型复杂度、效率的考虑，自然语言处理系统通常采用级联的方式，即分词、词性标注、句法分析、语义分析分别训练模型。实际使用时，给定输入句子，逐一使用各个模块进行分析，最终得到所有结果。

近年来，随着研究工作的深入，研究者们提出了很多有效的联合模型，将多个任务

联合学习和解码，如分词词性联合、词性句法联合、分词词性句法联合、句法语义联合等。联合模型通常都可以显著提高分析质量，原因在于联合模型可以让相互关联的多个任务互相帮助，同时对于任何单任务而言，人工标注的信息也更多了。然而，联合模型的复杂度更高，速度也更慢。

本书主要介绍第一层面词法分析和第二层面句法分析的内容。

1.6 NLP 与人工智能

NLP 是计算机领域与人工智能领域中的一个重要分支。人工智能（Artificial Intelligence，AI）在 1955 年达特茅斯特会议上被提出，而后人工智能先后经历了三次浪潮，但是在 20 世纪 70 年代第一次 AI 浪潮泡沫破灭之后，这一概念迅速进入沉寂，相关研究者都不愿提起自己是研究人工智能的，转而研究机器学习、数据挖掘、自然语言处理等各个方向。1990 年迎来第二次黄金时代，同期日本意欲打造传说中的“第五代计算机”，日本当时宣称第五代计算机的能力就是能够自主学习，而随着第五代计算机研制的失败，人工智能再次进入沉寂期。2008 年左右，由于数据的大幅增强、计算力的大幅提升、深度学习实现端到端的训练，深度学习引领人工智能进入第三波浪潮。人们也逐渐开始将如日中天的深度学习方法引入到 NLP 领域中，在机器翻译、问答系统、自动摘要等方向取得成功。

那么，为什么深度学习可以在 NLP 中取得这样的成绩呢？现在看来，大概可以归结为两点：

（1）海量的数据。经过之前互联网的发展，很多应用积累了足够多的数据可以用于学习。当数据量增大之后，以 SVM（支持向量机）、CRF（条件随机场）为代表的传统浅层模型，由于模型过浅，无法对海量数据中的高维非线性映射做建模，所以不能带来性能的提升。然而，以 CNN、RNN 为代表的深度模型，可以随着模型复杂度的增大而增强，更好贴近数据的本质映射关系，达到更优的效果。

（2）深度学习算法的革新。一方面，深度学习的 word2vec 的出现，使得我们可以将词表示为更加低维的向量空间，相对于 one-hot 方式，这既缓解了语义鸿沟问题，又降低了输入特征的维度，从而降低了输入层的维度，另一方面，深度学习模型非常灵活，使得之前的很多任务，可以使用端到端的方式进行训练。例如机器翻译，传统的方法需要先进行分词、对齐、翻译，语言模型需要依赖各个模块，每个模块的误差会传递到下个模块，使得整个系统不是一个整体，变得不太可控。而使用端到端的方式，可以直接映射，避免了误差的传递，提升了性能。

深度学习在 NLP 中取得了巨大的成绩，当然随之而来也是诸多挑战。深度学习虽是一把利剑，但由于语音和图像这种属于自然信号，而自然语言是人类知识的抽象浓缩表示，所以意味着深度学习并不能解决 NLP 中的所有问题。人在表达的过程中，由于背景知识的存在会省略非常多的东西，使得自然语言的表达更加简洁，文本所携带的信息也有一定的局限性，在 NLP 处理过程中也会碰到非常多的困难。类似的问题，当样本的数量有限，如何应用深度学习方法和知识信息进行融合提升整个系统的性能，如何能够自动学习知识，达到能够有效应用包括语言学知识、领域知识，如何随着环境的变化而变化，通过强化学习的方式提升系统的性能，以及如何上下文学习，根据上下文增强对当前任务的决策能力。

NLP 过去几十年的发展，从基于简单的规则方法到基于统计学方法，再到现在的基于深度学习神经网络的方法，技术越来越成熟，在很多领域都取得了巨大的成就。展望未来十年，随着数据的积累，云计算，芯片技术发展，人工智能技术的发展，自然语言必将越来越贴近智能。除此之外，随着人工智能各领域的研究细化，每个领域今后将越来越难有大的跨越，所以，跨领域的研究整合将是未来的发展方向。可预见的是 NLP 将会和其他的领域——视觉、听觉、触觉等高度融合，反映在人工智能技术上就是语音识别和图像识别，最后达到“认知智能”，包含语言、知识和推理的真正意义上的智能。当然前途是光明的，路途是坎坷的。还需要各位同仁一起努力，给 NLP 研究添砖加瓦。

1.7 本章小结

本章介绍了 NLP 相关的一些基础知识，主要面向 NLP 刚刚入门的读者。首先介绍了 NLP 的概念、应用场景和发展历程，在学习 NLP 技术之前，有必要了解这些宏观的内容；接着讲解了 NLP 的关键术语、知识结构，以及本书用到的语料库，告诉读者在学习 NLP 的最初，应该做好哪些技术储备；最后宏观地探讨了 NLP 与人工智能的关系，为读者普及相关基本概念，为后面的深入学习打好基础。后续章节我们将介绍通过 Python 处理 NLP 中的一些关键库以及 NLP 日常处理中需要掌握的技术。