第1章 导 论

Dave Bowman：HAL，请你打开太空舱的分离舱门。
HAL：对不起，Dave，我不能这样做。①

Stanley Kubrick 和 Arthur C. Clarke
2001 年电影剧本：《太空漫游》

使计算机获得处理人类语言的能力的想法就像计算机本身的想法一样古老。本书就是一本论述和实现这种令人激动的想法的专著。在本书中，我们将介绍的内容涉及到很多不同的方面，它们形成了一个独具风格的、动人心魄的交叉学科领域（interdisciplinary field），这个交叉学科领域由于侧重点的差异而对应于不同的学科名称，诸如**语音和语言处理**（language and speech processing）、**人类语言技术**（human language technology）、**计算语言学**（computational linguistics），以及**语音识别与合成**（speech recognition and synthesis）等。这个新兴的交叉学科的目标在于让计算机实现与人类语言有关的各种任务，例如，使人与计算机之间的通信成为可能，改进人与人之间的通信，或者简单地让计算机进行文本或语音的自动处理，等等。

这些有用的任务的一个实例是**会话代理**（conversational agent）。在 Stanley Kubrick 的 2001 年的电影《太空漫游》中有一台称为 HAL 的 9000 计算机，这台计算机具有 20 世纪最受人们认可的一些特征。影片中的 HAL 是一个具有高级的语言处理能力并且能够说英语和理解英语的智能机器人（artificial agent），在影片情节的关键时刻，HAL 甚至能够进行唇读（reading lip），上面就是电影中的角色 Dave 先生请求智能机器人 HAL 打开宇宙飞船的分离舱门（pod bay doors），与 HAL 之间进行的一段对话。HAL 的作者 Arthur C. Clarke 曾经乐观地预言，到一定的时候，我们就可以制造出如 HAL 这样的智能机器人。但是，现在我们离这样的预言还有多远呢？为了让 HAL 具有与语言相关的能力，我们还应该做些什么呢？我们认为，如 HAL 这样的机器人至少应该通过语言与人类进行交流。我们把 HAL 这样的能够使用自然语言与人类会话的程序称为**会话代理**（conversational agents）或者**对话系统**（dialogue systems）。在本书中，我们要研究设计这样的现代会话代理的各个组成部分，其中包括语言输入[**自动语音识别**（automatic speech recognition）和**自然语言理解**（natural language processing）]和语言输出[对话与回答的规划（dialogue and response planning ）以及**语音合成**（speech synthesis）]。

让我们转到另一个与语言密切关联的问题，这就是怎样使不会讲英语的读者能够懂得英语网页上的数量可观的科学信息，怎样为讲英语的读者把用其他语言写的数以亿计的网页翻译成英语以便他们阅读。**机器翻译**（machine translation）的目标就是自动地把文献从一种语言翻译成另一种语言。我们将介绍一些算法和工具，使读者理解现代的机器翻译系统是如何工作的。机器翻译迄今还是一个远远没有解决的问题，我们将介绍目前在这个领域中所使用的各种算法以及一些重要的局部性的研究工作。

---

① 这是本章的开场白，为了便于读者理解，我们把英文原文写出来。
Dave Bowman：Open the pod bay doors，HAL..
HAL：I'm sorry Dave，I'm afraid I can't do that.
Stanley Kubrick and Arthur C. Clarke，
screenplay of 2001：*A Space Odyssey* ——译者注

与网络有关的自然语言处理的问题还有很多。除了机器翻译之外,还有**基于网络的问答系统**(Web-based question answering)。这种基于网络的问答系统是简单的网络搜索的进一步发展,在基于网络的问答系统中,用户不只是仅仅键入关键词进行提问,而是可以用自然语言提出一系列完整的问题,从容易的问题到困难的问题都可以提。例如,下面的问题,

- What does"divergent" mean?(divergent 的意思是什么?)
- What year was Abraham Lincoln born?(亚伯拉罕·林肯生于哪一年?)
- How many states were in the United States that year?(那一年在美国有多少个州?)
- How much Chinese silk was exported to England by the end of the 18th century?(18 世纪末有多少中国的丝绸出口到英国?)
- What do scientists think about the ethics of human cloning?(关于克隆人的论理学问题科学家们是如何考虑的?)

在这些问题中,有的问题只要求回答**定义**(definition),有的问题只要求回答诸如日期、地点等简单的**新闻要素**(factoid),对于这样的问题,使用搜索引擎就可以回答了。但是对于需要抽取嵌入在网页的其他文本中的信息才能回答的那些更加复杂的问题,就要进行**推理**(inference),也就是根据已经知道的事实推出结论,或者从多重的信息源或网页中对信息进行综合或摘取。在本书中,我们将研究建造这种现代的自然语言理解系统的各个组成部分,包括**信息抽取**(information extraction)、**词义排歧**(word sense disambiguation),等等。

尽管这些问题现在还远远没有完全解决,有的研究领域仍然非常活跃,很多的技术已经商品化。在本书的其他部分,我们还将简短地总结为了完成上述的这些任务[以及其他的诸如**拼写校正**(spelling correction)、**语法检查**(grammar checking)等任务]所必需的各种知识,并且介绍数学模型,各种数学模型的介绍是贯穿全书的。

1.1 语音与语言处理中的知识

自然语言处理的这些应用与其他的应用系统的区别在于,自然语言处理要使用语言知识。例如,UNIX 的 wc 程序可以用来计算文本文件中的字节数、词数、行数。当我们用它来计算字节数和行数的时候,wc 只是用于进行一般的数据处理。但是,当我们用它来计算一个文件中的词的数目的时候,我们就需要关于"什么是一个词"的语言知识,这样,这个 wc 也就成为了一个自然语言处理系统。

当然,wc 只是一个非常简单的系统,它只具有极为有限的语言知识。如 HAL 这样有更复杂的语言能力的智能机器人、机器翻译系统、鲁棒的问答系统将要求更加广泛和更加深刻的语言知识。我们只要读一读本章开头 HAL 和 Dave 进行的对话,或者看一看问答系统中如何回答上面所列的问题,我们就可以了解到这些更加复杂的应用所需要的语言知识的范围和种类。

HAL 必须能够分析它所接收的声音信号,并且从单词序列生成声音信号。要完成这两方面的任务,需要**语音学**(phonetics)和**音系学**(phonology)的知识:单词是怎样发出音来而成为声音序列的,而每一个声音又是怎样在语音学上实现的。

值得注意的是,与 Star Trek 的指令数据不同,HAL 还能够说出如 I'm 和 can't 的缩约形式。产生并且识别单词的各种变体(例如,识别 Doors 是复数)要求**形态学**(morphology)方面的知识,说明单词是怎样分解成它的组成成分的,而这些成分又是怎样负荷如单数和复数这样的意义的。

除了处理一个一个的单词之外，HAL 还应该知道怎样使用结构的知识恰当地把这些单词组织成单词串并构成回答。例如，HAL 必须知道，下面的单词序列对于 Dave 是没有意义的，尽管这个单词系列所包含的单词与它原来的回答中所包含的单词完全一样：

(1.1) I'm I do, sorry that afraid Dave I'm can't.

这里所说的关于单词的排列顺序以及组词成句的知识，称为**句法**(syntax)。

现在我们来讨论在问答系统中是如何处理下面的问题的：

(1.2) How much Chinese silk was exported to Western Europe by the end of the 18th century? (18 世纪末有多少中国的丝绸出口到西欧?)

为了回答这个问题，我们需要关于**词汇语义学**(lexical semantics)的知识以便了解问句所有单词(export 或 silk)的意义，我们还需要**组合语义学**(compositional semantics)的知识：相对于 Eastern Europe 或 Southern Europe 这样的组合，Western Europe 的语义是怎样组合而成的；当 end 与 the 18th century 结合在一起的时候，它的含义是什么。我们还需要知道，by the end of the 18th century 中的 by 是表示时间终点的，而不是描述施事(agent)的，而在下面句子中的 by 则是描述施事的：

(1.3) How much Chinese silk was exported to Western Europe by southern merchants? (南方的商人出口了多少中国丝绸到西欧去?)

我们还需要知识能够使 HAL 确定，Dave 说的话是关于要 HAL 采取某种行动的一个请求，这样的请求不同于下面关于陈述客观世界的简单命题，也不同于下面关于 door 的问话，它们是 Dave 请求的不同变体：

请求：HAL, open the pod bay door. (HAL, 请打开分离舱的门。)
陈述：HAL, the pod bay door is open. (HAL, 分离舱的门是开着的。)
问话：HAL, is the pod bay door open? (HAL, 分离舱的门是开着的吗?)

另外，尽管智能机器人 HAL 的行为还不十分熟练，它也应该充分地懂得如何对 Dave 表示礼貌。例如，它不要简单地回答 No 或者 No, I won't open the door。HAL 首先用表示客气的话回答 I'm sorry 和 I'm afraid，然后委婉地说 I can't，而不是直截了当地(并且老老实实地)说 I won't①。这种关于说话人使用句子来表达意图的行为的知识就是**语用学**(pragmatic)或**对话**(dialogue)的知识。

在回答下面的问题的时候，需要另一种关于语用学或**话语**(discourse)的知识：

(1.4) How many states were in the United States *that year*? (那一年在美国有多少个州?)

在这个问题中，that year 究竟是哪一年？为了解释如 that year 这样的单词的含义，问答系统需要检查前面已经回答过的问题；在这个问题的情况下，前面的问题谈的是关于 Lincoln 诞生的年份，因此，that year 就是 Lincoln 诞生的那一年。这种**同指消解**(coreference resolution)的任务需要确认诸如 that 或 it 或 she 这样的代词究竟涉及前面话语中的哪一个部分的知识。

总而言之，在复杂的语言行为中需要的语言知识可以分为 6 个方面：

- 语音学与音系学——关于语言语音的知识。

① "我不愿意关门"。这样的回答显得非常生硬、呆板和偏执。

- 形态学——关于词的有意义的组成成分的知识。
- 句法学——关于词与词之间结构关系的知识。
- 语义学——关于意义的知识。
- 语用学——关于意义与说话人的目的和意图之间关系的知识。
- 话语学——关于比一个单独的话段更大的语言单位的知识。

1.2 歧义

上述6个方面的语言知识存在着一个令人吃惊的事实,这就是:语音和语言计算机处理的绝大多数或者全部的研究都可以看成是在其中的某个层面上消解歧义。如果我们想把某个意思输入计算机,而存在着若干个不同的结构来表示这个意思,那么,我们就说这样的输入是有歧义的。我们来考虑口语中的一个句子 I made her duck。这个句子可能有5个不同的意思(还会更多),以下是歧义的若干实例:

(1.5) I cooked waterfowl for her. (我给她烹饪鸭子。)

(1.6) I cooked waterfowl belonging to her. (我烹饪属于她的鸭子。)

(1.7) I created the (plaster?) duck she owns. [我把她的(石膏?)鸭子做了创新。]

(1.8) I caused her to quickly lower her head or body. (我使她很快地把她的头或者身体放低一些。)

(1.9) I waved my magic wand and turned her into undifferentiated waterfowl. (我挥动魔杖把她变成了一只人们一点儿也看不出破绽的鸭子。)

这些不同的意思都是由于歧义引起的。首先,duck 和 her 的词类在形态或句法上是有歧义的。duck 可以是动词或名词,而 her 可以是表示给予格的代词或表示所属格的代词。其次,make 在语义上是有歧义的,它的意思可以是 create(创造),也可以是 cook(烹饪)。最后,动词 make 还可以有不同的句法歧义。make 可以作及物动词,带直接宾语(1.6);make 也可以作双及物动词,带两个宾语(1.9),表示把第一宾语(her)变成了第二宾语(duck);make 还可以带一个直接宾语和一个动词(1.8),表示使直接宾语(her)去进行某个动作(duck)。此外。在口语的句子中,还可以有一种更为深刻的歧义,第一个词可以被理解为 eye,或者第二个词可以被理解为 maid。

这样,歧义就更加复杂了。在本书中,我们会经常介绍**消解**(resolve)这些歧义,或者**排歧**(disambiguation)的模型和算法。例如,使用**词类标注**(part-of-speech tagging)的办法来确定 duck 是名词还是动词。使用**词义排歧**(word sense disambiguation)的办法来确定 make 的意思是 create(创造)还是 cook(烹饪)。词类排歧和词义排歧是**词汇排歧**(lexical disambiguation)的两个主要内容。很多研究都可以纳入到词汇排歧的框架之内。例如,在文本-语音合成系统中,当读到单词 lead 的时候,必须判断这个 lead 是按照 lead pipe 中的 lead 读音呢,还是按照 lead me on 中的 lead 读音。此外还有**句法排歧**(syntactic disambiguation)。例如,当我们判断 her 和 duck 是属于不同的实体,如例句(1.5)或例句(1.8),还是属于同一个实体,如例句(1.6),这样的问题就属于句法排歧的问题了,可以通过**概率剖析**(probabilistic parsing)的方法来解决。在上述例子中没有出现的一些歧义(例如,判断一个句子是陈述句还是疑问句),可以通过**言语行为解释**(speech act interpretation)的办法来解决。

1.3 模型和算法

50年来的自然语言处理研究说明,前一节中所描述的那些知识可以使用数量有限的形式模型或理论来获得。值得庆幸的是,这些模型和理论都来自计算机科学、数学和语言学的工具,在

这些领域受过训练的人，对这样的工具一般都不会感到生疏。其中最重要的部分是**状态机器**（state machine）、**形式规则系统**（formal rule system）、**逻辑**（logic）、**概率模型**（probabilistic models）和**向量空间模型**（vector-space models）。这样的模型本身又可以给出为数不多的算法。其中最重要的算法是如**动态规划**（dynamic programming）算法的**状态空间搜索**（state space search）算法、分类器（classifiers）和期望最大算法（Expectation-Maximization，EM）的机器学习算法以及其他的学习算法。

简单地说，状态机器就是形式模型，形式模型应该包括状态、状态之间的转移以及输入表示等。这种基本模型的变体有**确定的有限状态自动机**（deterministic finite-state automata）、**非确定的有限状态自动机**（non-deterministic finite-state automata）和**有限状态转录机**（finite-state transducers）。

同这些过程性模型紧密联系的模型是陈述性模型，也就是形式规则系统。这些陈述性模型中，如果既考虑概率模型，也考虑非概率模型，我们认为最重要的有**正则语法**（regular grammars）、**正则关系**（regular relations）、**上下文无关语法**（context-free grammars）、**特征增益语法**（feature-augmented grammars）以及与这些语法相应的概率语法变体。状态机器和形式规则系统是用于处理音系学、形态学和句法学的主要工具。

对于获取语言知识起着关键性作用的第三种模型是基于逻辑的模型。我们将讨论**一阶逻辑**（first order logic），即**谓词演算**（predicate），以及诸如λ运算（lambda-calculus）、特征结构（feature structures）、语义基元（semantic primitives）等有关的形式化方法。在传统上，这些逻辑表达方法用于建立语义学、语用学的形式模型，尽管最近的工作倾向于集中力量研究那些从非逻辑的词汇语义学中借鉴来的潜在地更加具有鲁棒性的技术。

概率论是我们获取语言知识的技术中的最为关键的一个部分。其他的各种模型（状态机器、形式系统和逻辑）都可以使用概率得到进一步的提高。例如，状态机器可以使用概率论来提升，成为**加权自动机**（weighted automaton ），或马尔可夫模型（Markov model）。我们将用很多的时间来讨论隐马尔可夫模型（Hidden Markov Models，HMM），在自然语言处理的领域内到处都在使用 HMM，在词性标注、语音识别、对话理解、文本-语音转换和机器翻译中，HMM 都发挥了作用。概率论的一个关键性的优点是能解决前面我们讨论过的各种歧义问题；几乎所有的语音处理和语言处理问题都可以这样来表述："对于某个歧义的输入给出 N 个可能性，选择其中概率最高的一个。"

基于线性代数的向量空间模型是信息检索和词义处理的基础。

典型地说，使用这些模型来处理语言就是通过表示输入假定的状态的空间来进行搜索。在语音识别中，我们通过音子序列的空间来搜索它们所对应的正确的单词。在句法剖析中，我们通过树的空间，对于输入的句子来搜索它们所对应的句法剖析树。在机器翻译中，我们通过翻译假设的空间，对于一个句子来搜索它在其他语言中所对应的正确翻译。对于那些涉及状态机器的非概率的任务，我们使用诸如**深度优先搜索**（depth-first search）之类的众所周知的图算法。对于那些具有概率的任务，我们使用**最佳优先搜索算法**（best-first）和 **A*** **搜索算法**（A* search）等试探性算法的变体，并依靠动态规划算法来提高计算的可循性。

分类器（classifiers）和**序列模型**（sequence models）之类的机器学习工具在自然语言处理的很多工作中起着重要作用。根据所描述客体的属性，分类器把一个单独的客体指派到一个单独的类别中去，而序列模型则对于一个客体序列进行分类，把它指派到一个类别序列中去。

例如，在判定一个单词的拼写是否正确的时候，就可以使用诸如**决策树**（decision trees）、**支持向量机**（support vector machines）、**高斯混合矩阵**（Gaussian mixture models）和**逻辑回归**（logistic regression）等分类器在某一时刻对于某一单词进行二分判定，从而确定这个单词的拼写是正确的还是不正确的。

最后,自然语言处理的研究者们还使用很多机器学习研究中在方法论上相同的工具,使用独特的训练集和测试集,使用诸如**交叉验证**(cross-validation)这样的统计技术,细心地对训练系统进行评测。

1.4 语言、思维和理解

如果计算机能够像人类一样熟练地处理语言,那么,这就意味着计算机已经达到了真正的智能机器的水平。这种信念是基于这样的事实:语言总是与我们的认知能力纠缠在一起。Alan Turing(1950)是第一个认识到计算机与认知能力之间有着如此密切关系的科学家。在他的一篇著名的论文中,Turing 提出了**图灵测试**(Turing test)的想法。Turing 在他的论文的开头就指出,关于什么是机器思维的问题是不能回答的,因为"机器"(machine)与"思维"(think)这两个术语本身就是含糊不清的。因此,他建议做一个游戏来进行测试,在游戏中,计算机对于语言的使用情况就可以用来作为判断计算机是否能进行思维的根据。如果计算机在游戏中获胜,那么就可以判断计算机具有智能。

在 Turing 的游戏中有三个参加者:两个人和一台计算机。其中的一个人充当提问者的角色,他要使用电传打字机向另外两个参加者提出一系列问题,根据这两个参加者的回答判断哪一个回答是计算机做出的。计算机的任务是尽量设法来愚弄提问者,对于提问者的问题,尽量做出如人一样的回答,设法使提问者相信它真的是一个人。而第二个参加游戏的人则尽量设法使提问者相信第三个参加者是计算机,只有他和提问者才是人。

下面是 Turing 在他的论文中所描述的一个交互过程。显而易见,计算机要模拟人,并不能要求它成为一个无所不知的专家。

Q: Please write me a sonnet on the topic of the Forth Bridge.
A: Count me out on this one. I never could write poetry.
Q: Add 34957 to 70764.
A: (30 秒钟之后给出回答)105621.

第一个提问要求写出有关"第四号桥"主题的一首商籁体诗(十四行诗),这显然是高智能的工作,因此,回答是拒绝写这样的诗歌:"不要问我这个问题。我从来不写诗"。第二个提问要求做 5 位数的加法:34957 加 70764 是多少,需要 30 秒钟的思索之后才能够给出正确的回答:105621。这样的问题不论对于人来说,还是对于计算机来说,都并非容易之事。

Turing 预言,大约到 20 世纪末,具有 10GB 存贮量的计算机经过 5 分钟的提问之后,将会有 30% 的机会使提问者感到困惑,辨不清回答问题的是人还是计算机。如果你也总是像计算机那样地去愚弄别人,你也可能说不清楚究竟采用什么样的严格标准去愚弄,才可以使别人感到困惑。然而,对于 Turing 来说,关键的问题是,只要使用人的语言就足以作为测定智能的可操作标准了。

令人惊讶的是,在 1966 年,就有人设计了一个称为 ELIZA 的程序来验证 Turing 的思想(Weizenbaum, 1966)。ELIZA 是一个早期的自然语言处理系统,这个系统能与用户进行有一定限制的对话。下面是 ELIZA 系统模拟心理医生的一段对话:

用户$_1$: You are like my father in some way.(你在某些方面很像我的父亲。)
ELIZA$_1$: WHAT RESEMBLANCE DO YOU SEE(你看是哪些方面像呢?)
用户$_2$: You are not very aggressive but I think you don't want me to notice that.(你不是一个非常坚持自见的人,可是你不愿意我提起这一点。)

ELIZA$_2$：WHAT MAKES YOU THINK I AM NOT AGGRESSIVE（你凭什么说我不是一个非常坚持自见的人?）

用户$_3$：You don't argue with me.（因为你不想和我争论。）

ELIZA$_3$：WHY DO YOU THINK I DON'T ARGUE WITH YOU（为什么你认为我不想和你争论呢?）

用户$_4$：You are afraid of me.（因为你怕我。）

ELIZA$_4$：DOES IT PLEASE YOU TO BELIEVE I AM AFRAID OF YOU（凭什么使得你认为我怕你?）

显而易见，ELIZA 只是使用模式匹配的方法来处理输入并且把它转换成适当的输出，这是一个非常简单的系统，我们将在第 2 章中更加详细地讨论这个问题。事实上 ELIZA 并没有必要懂得如何去模拟心理医生，它只是使用简单的模式匹配就取得了成功。正如系统的设计人 Weizenbaum 所说的，在 ELIZA 系统中，听话者的所作所为就好像他们对于周围的世界一无所知。

ELIZA 与 Turing 思想的深刻联系在于，很多与 ELIZA 进行过交互的人都相信，ELIZA 确实理解了他们所说的话以及他们所提出的问题。Weizenbaum（1976）指出，甚至在把程序的操作过程向人们作了解释之后，仍然有不少的人继续相信 ELIZA 的能力。近年来，人们又以不同的形式重复着 Weizenbaum 的工作。自 1991 年以来，在 Loebner 奖的比赛中，人们试图设计各种计算机程序来做 Turing 测试。尽管这些比赛的科学意义不是很大，不过，这些比赛的成绩说明，哪怕是很粗糙的程序有时也会愚弄人们的判断力（Shieber，1994a）。哲学家和人工智能研究者对于 Turing 测试究竟是否适合用来测试智能的争论已经持续很多年了，但是，上述比赛的结果，并没有平息这样的争论（Searle，1980）。

就本书的目的而言，这样的比赛结果与计算机究竟能否思维，或者计算机究竟能否理解自然语言的问题是风马牛不相及的。更为重要的是，在社会科学中的有关研究证实了 Turing 在同一篇文章中的预见：

> 然而，我相信，在本世纪的末叶，词语的使用和教育的舆论将大大地改变，使我们有可能谈论机器思维而不致遭到别人的反驳。

现在已经清楚，不管人们相信什么，不管人们是否已经知道了计算机的内部工作情况，他们都在谈论计算机，并且都在与计算机进行着交互，把计算机当成一个社会实体。人们把计算机当成人一样地对待，他们要对它讲礼貌，他们把它当成团队中的成员，并且期望计算机能够理解人们的需求，能够非常自然地与人们进行交互。例如，Reeves and Nass（1996）发现，当计算机要求人们来评价计算机的所作所为好不好的时候，人们要针对不同计算机提出的同样的问题做出更多的正面的回答。人们似乎担心他们给计算机的回答不够礼貌。Reeves 和 Nass 在另外的实验中还发现，如果计算机对人们说一些奉承的话，人们给计算机的评价也就会高一些。给出这样的一些预设，使用语音和语言的系统就能够给众多的用户在很多应用方面提供更加自然的交互界面。这些导致了一个称为**会话代理**（conversational agents）的研究焦点，所谓会话代理就是通过会话进行交际的计算机人造实体，会话代理的研究将会持续很长的时间。

1.5　学科现状与近期发展

> 尽管我们只能往前看很短的距离，但是我们能看清楚什么是我们需要做的事情。
>
> Alan Turing

现在语音和语言处理正处于激动人心的时刻。普通计算机用户可以使用的计算资源正以惊人的速度迅速增长,互联网的兴起成为了无比丰富的信息资源,无线移动通信日益普及并且日益增长起来,这些都使得语音和语言处理的应用成为了当前科学技术的热门话题。这里我们想列举该学科一些当前的应用项目,并提出该学科近期发展的一些可能的方面。

- Amtrak 旅行社、美国联合航空公司以及其他的一些旅行社可以与智能会话代理进行交互,在智能会话代理的指导下,他们能够自动地处理关于旅行中的订票、到达、离开等方面的信息。
- 汽车制造公司可以给汽车驾驶员提供语音识别和文本-语音转换系统,使得他们可以通过语音来控制他们的环境、娱乐以及导航系统。在国际空间站的宇航员也可以使用简单的口语对话系统来帮助他们的工作。
- 一些视频搜索公司使用语音识别技术,可以在网络上提供多达数百万小时的视频资料的搜索服务,并且在语音资料中搜索到与之相应的单词。
- Google("谷歌")在网上提供跨语言信息检索和自动翻译服务,用户可以使用他们自己的母语来提问,以便搜索其他语言中的有关信息。Google 还可以对用户提出的问题进行自动翻译,找出与所提出的问题最相关的网页,然后自动地把它们翻译成用户的母语。
- 如 Pearson("培生")的大型出版集团和如 ETS 的测试服务公司使用自动系统来分析数千篇学生的作文,对于这些作文进行自动打分、自动排序和自动评价,而且计算机的打分结果与人的打分结果几乎毫无二致,难以分辨。
- 具有生动活泼的动画特征的交互式虚拟智能代理可以充当教员来教儿童学习如何阅读(Wise et al., 2007)。
- 文本分析公司根据用户在互联网论坛和用户群体组织中表现出来的意见、偏好、态度的自动测试结果,对用户提供智能化的服务,帮助用户在市场上购买到符合他们要求的商品。

1.6 语音和语言处理简史

在历史上,语音和语言处理曾经在计算机科学、电子工程、语言学和心理认知语言学等不同的领域分别进行研究。之所以出现这种情况,是由于语音和语言处理包括了一系列性质不同而又彼此交叉的学科,它们是:语言学中的**计算语言学**(computational linguistics)、计算机科学中的**自然语言处理**(natural language processing)、电子工程中的**语音识别**(speech recognition)、心理学中的**计算心理语言学**(computational psycholinguistics)。本节中,我们将把在语音和语言处理中这些不同的历史线索做总结性的说明。不过,本节只是提供一个梗概,相应领域的更详细的介绍请参阅本书相关的章节。

1.6.1 基础研究:20 世纪 40 年代和 20 世纪 50 年代

这个领域的研究最早可以追溯到第二次世界大战刚结束时的那个充满了理智的时代,那个时代刚发明了计算机。从 20 世纪 40 年代到 20 世纪 50 年代末的时期有两项基础性的研究值得注意:一项是**自动机**(automaton)的研究,另一项是**概率模型**(probabilistic models)或**信息论模型**(information-theoretic models)的研究。

20 世纪 50 年代提出的自动机理论来源于 Turing 的算法计算模型(1936),这种模型被认为是现代计算机科学的基础。Turing 的工作首先导致了 McCulloch-Pitts 的**神经元**(neuron)理论(McCulloch-Pitts, 1943)。一个简单的神经元模型就是一个计算的单元,它可以用命题逻辑来描述。

接着，Turing 的工作导致了 Kleene（1951，1956）关于有限自动机和正则表达式的研究。Shannon（1948）把离散马尔可夫过程的概率模型应用于描述语言的自动机。Chomsky（1956）从 Shannon 的工作中吸取了有限状态马尔可夫过程的思想，首先把有限状态自动机作为一种工具来刻画语言的语法，并且把有限状态语言定义为由有限状态语法生成的语言。这些早期的研究工作产生了**形式语言理论**（formal language theory）这样的研究领域，采用代数和集合论把形式语言定义为符号的序列。Chomsky 在研究自然语言的时候首先提出了上下文无关语法（1956），但是，Backus（1959）和 Naur et al.（1960）在描述 ALGOL 程序语言的工作中也独立地发现了这种上下文无关语法。

这个时期的另外一项基础研究工作是用于语音和语言处理的概率算法的研制，这是 Shannon 的另一个贡献。Shannon 把通过诸如通信信道或声学语音这样的媒介传输语言的行为比喻为**噪声信道**（noisy channel）或者**解码**（decoding）。Shannon 还借用热力学（thermodynamics）的术语“**熵**”（entropy）来作为测量信道的信息能力或者语言的信息量的一种方法，并且他用概率技术首次测定了英语的熵。

在这个时期，还研究了声谱（Koenig et al ., 1946），声谱和实验语音学的基础研究为之后语音识别的研究奠定了基础。这导致了 20 世纪 50 年代第一个机器语音识别器的研制成功。1952 年，贝尔实验室的研究人员建立了一个统计系统来识别由一个单独的说话人说出的 10 个任意的数字（Davis et al ., 1952）。该系统存储了 10 个依赖于说话人的模型，它们粗略地代表了英语数字的头两个元音的共振峰。贝尔实验室的研究人员采用选择与输入具有最高相关系数模式的方法，达到了 97% ~99% 的准确率。

1.6.2 两个阵营：1957 年至 1970 年

在 20 世纪 50 年代末期到 20 世纪 60 年代初期，语音和语言处理明显地分成两个阵营：一个阵营是符号派（symbolic），一个阵营是随机派（stochastic）。

符号派的工作可分为两个方面。一方面是 20 世纪 50 年代后期以及 20 世纪 60 年代初期和中期 Chomsky 等的形式语言理论和生成句法的研究，很多语言学家和计算机科学家的剖析算法研究，早期的自顶向下和自底向上算法的研究，后期的动态规划的研究。最早的完整的剖析系统是 Zelig Harris 的“转换与话语分析课题”（Transformation and Discourse Analysis Project，TDAP）。这个剖析系统于 1958 年 6 月至 1959 年 7 月在宾夕法尼亚大学研制成功（Harris，1962）[①]。另一方面是人工智能的研究。在 1956 年夏天，John McCarthy，Marvin Minsky，Claude Shannon 和 Nathaniel Rochester 等学者汇聚到一起组成了一个为期两个月的研究组，讨论关于他们称之为“人工智能”（Artificial Intelligence，AI）的问题。尽管有少数的 AI 研究者着重于研究随机算法和统计算法（包括概率模型和神经网络），但是大多数的 AI 研究者着重研究推理和逻辑问题。典型的例子是 Newell 和 Simon 关于“逻辑理论家”（logic theorist）和“通用问题解答器”（general problem solver）的研究工作。早期的自然语言理解系统都是按照这样的观点建立起来的。这些简单的系统把模式匹配和关键词搜索与简单试探的方法结合起来进行推理和自动问答，它们都只能在某一个领域内使用。在 20 世纪 60 年代末期，学者们又研制了更多的形式逻辑系统。

随机派主要是一些来自统计学专业和电子学专业的研究人员。在 20 世纪 50 年代后期，贝叶斯方法（Bayesian method）开始被应用于解决最优字符识别的问题。Bledsoe and Browning（1959）建立

① 这个系统最近又重新建立起来，Joshi and Hopely（1999）以及 Karttunen（1999）对这个系统作了描述，他们指出，该系统的剖析本质上是用层叠式的有限状态转录机实现的。

了用于文本识别的贝叶斯系统,该系统使用了一部大词典,计算词典的单词中所观察的字母系列的似然度,把单词中每一个字母的似然度相乘,就可以求出字母系列的似然度来。Mosteller and Wallace(1964)用贝叶斯方法来解决在《联邦主义者》(The Federalist)文章中的原作者的分布问题。

20 世纪 60 年代还出现了基于转换语法的第一个人类语言计算机处理的可严格测定的心理模型;并且还出现了第一个联机语料库:Brown 美国英语语料库,该语料库包含一百万单词的语料,样本来自不同文体的 500 多篇书面文本,涉及的文体有新闻、中篇小说、写实小说、科技文章等。这些语料是布朗大学(Brown University)在 1963 年到 1964 年收集的(Kučera and Francis, 1967; Francis, 1979; Francis and Kučera, 1982)。王士元(William S. Y. Wang)在 1976 年建立了 DOC(Dictionary on Computer),这是一部联机的汉语方言词典。

1.6.3 四个范型:1970 年至 1983 年

在这个时期,语音和语言的计算机处理中出现了四个研究范型:它们至今还在语音和语言的计算机处理中起着支配的作用。

随机范型(stochastic paradigm)在语音识别算法的研制中起着重要的作用。其中特别重要的是隐马尔可夫模型和比喻为噪声信道与解码的模型。这些模型是分别独立地由两支队伍研制的。一支是 Jelinek, Bahl, Mercer 和 IBM 的 Thomas J. Watson 研究中心的研究人员,另一支是卡内基梅隆大学的 Baker 等人,Baker 受到普林斯顿国防分析研究所的 Baum 和他的同事们的工作的影响。AT&T 的贝尔实验室也是语音识别和语音合成的中心之一,详情可参阅 Rabiner and Juang (1993)对这方面工作的全面描述。

基于逻辑的范型(logic-based paradigm)肇始于 Colmerauer 和他的同事们(Colmerauer, 1970, 1975)关于 Q 系统(Q-system)和变形语法(metamorphosis grammar)的工作,Colmerauer 是 Prolog 语言的先驱者。定子句语法(definite clause grammar, Pereira and Warren, 1980)也是在基于逻辑的范型方面的早期工作之一。Kay 对于功能语法的研究(1979),稍后 Bresnan 和 Kaplan 在词汇功能语法(Lexical Function Grammar, LFG, 1982)方面的工作,都是特征结构合一(feature structure unification)研究方面的重要成果。

这个时期的**自然语言理解**(natural language understanding)肇始于 Terry Winograd 的 SHRDLU 系统,这个系统能够模拟一个嵌入玩具积木世界的机器人的行为(Winograd, 1972a)。该系统的程序能够接受自然语言的书面指令[例如,"Move the red block on top of the smaller green one"(请把绿色的小积木块移动到红色积木块的上端)],从而指挥机器人摆弄玩具积木块。迄今为止我们还没有看到如此复杂和精妙的系统。这个系统还首次尝试建立基于 Halliday 系统语法的全面的(在当时看来是全面的)英语语法。Winograd 的模型还清楚地说明,句法剖析也应该重视语义和话语的模型。Roger Schank 和他在耶鲁大学的同事和学生们(经常被称为耶鲁学派)建立了一些语言理解程序,这些程序构成一个系列,他们重点研究诸如脚本、计划和目的这样的人类的概念知识以及人类的记忆机制(Schank and Abelson, 1977; Schank and Riesbeck, 1981; Cullingford, 1981; Wilensky, 1983; Lehnert, 1977)。他们的工作经常使用基于网络的语义学理论(Quillian, 1968; Norman and Rumelhart, 1975; Schank, 1972; Wilks, 1975c, 1975b; Kintsch, 1974),并且在他们的表达方式中(Simmons, 1973)开始引进 Fillmore 关于格角色的概念(Fillmore, 1968)。

基于逻辑的范型和自然语言理解的范型还可以在系统中融合起来,例如,LUNAR 问答系统(Woods, 1967, 1973)是一个自然语言理解系统,在该系统中,就使用谓词逻辑来进行语义解释。

话语模型范型(discourse model paradigm)集中探讨了话语研究中的四个关键领域。Grosz 和她的同事们研究了话语中的子结构(substructure)和话语焦点(discourse focus)(Grosz, 1977a;

Sidner, 1983)；一些研究者开始研究自动参照消解(automatic reference resolution)(Hobbs, 1972)。在基于逻辑的言语行为研究中，建立了“信念－愿望－意图”的框架，即 **BDI**(Belief-Desire-Intention)的框架(Perrault and Allen, 1980；Cohen and Perrault, 1979)。

1.6.4　经验主义和有限状态模型的复苏：1983 年至 1993 年

在 1983 年至 1993 这 10 年中，语音和语言处理的研究又回到了 20 世纪 50 年代末期到 20 世纪 60 年代初期几乎被否定的有限状态和经验主义这两种模型上去，这两种模型之所以出现这种复苏，其部分原因在于过去 Chomsky 对于 Skinner 的“言语行为”(Verbal Behavior)的很有影响的评论(Chomsky, 1959b)在这时遭到了理论上的反对。第一种模型是有限状态模型，由于 Kaplan and Kay (1981)在有限状态音系学和形态学方面的工作，以及 Church(1980)在句法的有限状态模型方面的工作，这种模型又重新得到注意。本书自始至终都会讨论到与有限状态模型有关的工作。

在这个时期的第二个倾向是所谓的“重新回到经验主义”；这里值得特别注意的是语音和语言处理的概率模型的提出，这样的模型受到 IBM 的 Thomas J. Watson 研究中心的语音识别概率模型的强烈影响。这些概率模型和其他数据驱动的方法还传播到了词类标注、句法剖析、附着歧义的判定以及从语音识别到语义学的联接主义方法的研究中去。

在这个时期，自然语言的生成研究也取得了引人瞩目的成绩。

1.6.5　不同领域的合流：1994 年至 1999 年

在 20 世纪的最后 5 年，语音和语言处理这个领域发生了很大的变化。这主要表现在三个方面。首先，概率和数据驱动的方法几乎成为了自然语言处理的标准方法。句法剖析、词类标注、参照消解和话语处理的算法全都开始引入概率，并且采用从语音识别和信息检索中借过来的评测方法。其次，由于计算机的速度和存储量的增加，使得在语音和语言处理的一些子领域，特别是在语音识别、拼写检查、语法检查这些子领域，有可能进行商品化的开发。语音和语言处理的算法开始被应用于增强交替通信(Augmentative and Alternative Communication, AAC)中。最后，Web 的发展使得进一步加强基于语言的信息检索和信息抽取的需要变得更加突出。

1.6.6　机器学习的兴起：2000 年至 2008 年

在 21 世纪，从 20 世纪 90 年代后期开始的经验主义倾向进一步以惊人的步伐加快了它的发展速度。这样的加速发展在很大的程度上受到下面三种彼此协同的趋势的推动。

首先是建立带标记语料库的趋势。在语言数据联盟(Linguistic Data Consortium, LDC)和其他相关机构的帮助下，研究者们可以获得口语和书面语的大规模的语料。重要的是，在这些语料中还包括一些标注过的语料，如宾州树库(Penn Treebank)(Marcus et al ., 1993)、布拉格依存树库(Prague Dependency Treebank)(Hajič, 1998)、宾州命题语料库(PropBank)(Palmer et al ., 2005)、宾州话语树库(Penn Discourse Treebank)(Miltsakaki et al ., 2004b)、修辞结构库(RSTBank)(Carlson et al ., 2001)和 TimeBank(Pustejovsky et al ., 2003b)。这些语料库是带有句法、语义和语用等不同层次的标记的标准文本语言资源。这些语言资源的存在大大地推动了人们使用有监督的机器学习方法来处理那些在传统上非常复杂的自动剖析和自动语义分析等问题。这些语言资源也推动了有竞争性的评测机制的建立，评测的范围涉及剖析(Dejean and Tjong Kim Sang, 2001)、信息抽取(NIST, 2007a；Tjong Kim Sang, 2002；Tjong Kim Sang and De Meulder, 2003)、词义排歧(Palmer et al ., 2001；Kilgarriff and Palmer, 2000)、问答系统(Voorhees and Tice, 1999)、自动文摘(Dang, 2006)等领域。

第二是统计机器学习的趋势。对于机器学习的日益增长的重视，导致了学者们与统计机器

学习的研究者更加频繁地交互,彼此之间互相影响。对于支持向量机技术(Boser et al ., 1992; Vapnik, 1995)、最大熵技术以及与它们在形式上等价的多项逻辑回归(Berger et al ., 1996)、图式贝叶斯模型(Pearl, 1988)等技术的研究,都成为了计算语言学的标准研究实践活动。

第三是高性能计算机系统发展的趋势。高性能计算机系统的广泛应用,为机器学习系统的大规模训练和效能发挥提供了有利的条件,而这些在上一个世纪是难以想象的。

最后应当指出,在这个时期结束时,大规模的无监督统计学习方法得到了重新关注。机器翻译(Brown et al ., 1990; Och and Ney, 2003)和主题模拟(Blei et al ., 2003)等领域中统计方法的进步,说明了也可以只训练完全没有标注过的数据来构建机器学习系统,这样的系统也可以得到有效的应用。由于建造可靠的标注语料库要花费很高的成本,建造的难度很大,在很多问题中,这成为了使用有监督的机器学习方法的一个限制性因素。因此,这个趋势的进一步发展,将使我们更多地使用无监督的机器学习技术。

1.6.7 关于多重发现

尽管我们这里只是简单地回顾语音和语言处理的发展历史,我们已经可以看出,在不少场合下,同样的思想可能会多次地在不同的地方独立地被发现。在本书中,我们将讨论这种"多重发现"。例如,动态规划在序列比较中的应用就被 Viterbi, Vintsyuk, Needleman and Wunsch, Sakoe and Chiba, Sankoff and Reichert 等,以及 Wagner and Fischer 分别独立地提出过(见第 3 章、第 5 章和第 6 章)。语音识别中的 HMM 模型和噪声信道模型就被 Jelinek, Bahl 和 Mercer 分别独立地提出过(见第 6 章、第 9 章和第 10 章)。上下文无关语法就被 Chomsky, Backus 与 Naur 分别独立地提出并研究过(见第 12 章)。瑞士德语中存在着非上下文无关的句法的证明就被 Huybregts 和 Shieber 分别独立地研究过(见第 16 章)。把合一运算应用于语言处理就被 Colmerauer 等人和 Kay 分别独立地提出过(见第 15 章)。

这些多重的发现难道是令人惊讶的巧合吗?科学社会学者 Robert K. Merton(1961)反对巧合的说法,他指出:

> 一切科学发现,包括那些从表面上看来似乎是独一无二的科学发现,原则上都是多重的。

显而易见,历史上确实存在着许多众所周知的多重的科学发现和科学发明的事例。Ogburn and Thomas(1922)曾经列出了一个多重发现表,其中列举了很多的事例。例如,Leibnitz 和 Newton 分别发明了微积分;Wallace 和 Darwin 分别研究了自然选择的理论;Gray 和 Bell 分别发明了电话。[①]然而,Merton 举出了进一步的事例提出的这样的假设:多重发现是一个规律,而不是偶然的例外。很多公认的独一无二的发现原来是过去没有公布过的工作或者是没有被接受的工作的再发现。他根据人类学方法论还提出一个更加有力的论点,这种论点认为,科学家本身总是在把多重发现作为准则的假定下从事研究工作的。因此,科学生活的很多方面都是为了帮助科学家避免被别人"抢先得到"他的发现而设计的。例如,在给杂志提交论文时要注明日期;在科学研究的记录中要仔细地注明日期;及时周转预研报告或技术报告。

1.6.8 心理学的简要注记

本书的许多章节都有关于人类对于语言处理的心理学研究的简要说明。显而易见,理解人

① 一般认为,Ogburn 和 Thomas 注意到了,多重发现的普遍存在意味着文化的环境是科学发现的决定性的原因,而个人的天分并不是科学发现的决定性原因。然而 Merton 半开玩笑地援引 19 世纪以及 19 世纪以前的材料说明,这样的说法本身也是一种多重发现。

类的语言处理不论对于这种研究本身还是对于认知科学的整个领域的一个部分来说，都是极为重要的科学探索的目标。而且，理解人类对于自然语言的处理，经常能够帮助我们建立起更好的语言处理的机器模型。不过，这样的看法似乎与人们通常对此的认识相矛盾，因为人们通常认为，对于自然的算法的直接模仿在工程应用中并没有很大的作用。其论据是：如果我们一点不差地复制自然并不能导致工程技术上的成功，例如，如果飞机也像鸟一样地摆动它的机翼，这样的设计并不适用于飞机制造工程；因为具有固定机翼的飞机在工程技术上是更为成功的解决方法。然而，语言并不是航空。如果说模仿自然只是有时对于航空有所用处(因此，飞机才有了机翼)，但是，如果我们试图解决以人类为中心的问题，模仿自然就特别有用了。飞机的飞行与鸟的飞行有着不同的目的；但是语音识别系统的目的与法院的记录员每天工作的目的却是非常一致的：两者的目的都是要把口语的对话转写下来。由于人类在这一方面已经做得很成功了，我们就可以学习自然原本的解决方法。由于语音和语言处理系统的一个重要应用是人机交互，所以，模拟人类习以为常的解决方法肯定是能奏效的。

1.7　小结

本章介绍语音和语言处理这个领域。下面是本章的要点：

- 理解语音和语言处理研究的一个好办法或者是考察怎样来创造 2001 年电影剧本《太空奥德赛》中的 HAL 这样的智能代理，或者是建立基于 Web 网络的问答系统，或者是设计机器翻译引擎。
- 语音和语言处理技术与音系学、语音学、形态学、句法学、语义学、语用学和话语分析等不同平面上的语言知识的形式模型和形式表示方法有着密切的依赖关系。使用包括状态机器、形式规则系统、逻辑等在内的形式模型以及概率模型就可以获取这样的知识。
- 语音和语言处理的基础是计算机科学、语言学、数学、电子工程和心理学。在语音和语言处理中要使用这些学科的标准框架中为数不多的某些算法。
- 语言和思维之间的密切联系使语音和语言处理技术成为了关于智能机器辩论的中心议题。关于人类怎样与复杂媒体交互的研究表明，语音和语言处理技术在今后智能技术的发展过程中将起着至关紧要的作用。
- 语音和语言处理的革命性的应用目前已经在现实世界的周围呈现出来了。万维网(World-Wide Web)的建设和语音识别与语音合成的最新进展将进一步引导这种技术创造出更加丰富多彩的实际应用前景。

1.8　文献和历史说明

语音和语言处理各个分支领域的研究成果发表在很多会议论文集和杂志上。这些会议和杂志的中心内容都集中在自然语言处理和计算语言学两个方面，主要与美国计算语言学会(Association for Computational Linguistics，ACL)和它的欧洲伙伴欧洲计算语言学会(European Association for Computational Linguistics，EACL)、国际计算语言学会议(International Conference on Computational Linguistics，COLING)有关系。ACL、NAACL 和 EACL 的年会论文集和每两年一次的 COLING 会议是该领域研究工作的首要论坛。相关会议还有诸如自然语言学习会议(Conference on Natural Language Learning，CoNLL)这样的 ACL 特殊兴趣组(Special Interest Group，SIG)，以及自然语言处理中的经验方法会议(Empirical Methods in Natural Language Processing，EMNLP)。

语音识别、理解和合成的研究成果发表于每年一次的 INTERSPEECH 会议上，这个会议又称

为口语处理国际会议(International Conference on Spoken Language Processing, ICSLP),这个会议与欧洲语音通信和技术会议(European Conference on Speech Communication and Technology, EUROSPEECH)每隔一年交替召开。IEEE 国际声学、言语和信号处理会议(IEEE International Conference on Acoustic, Speech and Signal Processing, IEEE ICASSP)每年召开一次。在 IEEE ICASSP 会议或者如 SIGDial 这样的特殊兴趣组专题讨论会上经常发表口语对话研究的成果。

语音和语言的计算机处理的杂志主要有:《计算语言学》(Computational Linguistics)、《自然语言工程》(Natural Language Engineering)、《计算机语音与语言》(Computer Speech and Language)、《语音通信》(Speech Communication)、《IEEE 声频、语音 & 语言处理学报》(IEEE Transaction on Audio, Speech & Language)、《ACM 语音和语言处理学报》(ACM Transaction on Speech and Language)、《语言技术中的语言学问题》(Linguistic Issues in Language Technology)。

《计算语言学》杂志以及 ACL, COLING 会议和其他相关会议的很多论文都可以在 ACL 论文汇编(ACL Anthology)的网页上免费得到。网址:http://www.aclweb.org/anthology-index/。

从人工智能角度研究语言处理的成果发表于美国人工智能学会(American Association for Artificial Intelligence, AAAI)的年会以及两年一次的人工智能国际联合会议(International Joint Conference on Artificial Intelligence, IJCAI)上。下面的人工智能出版物周期性地发表有关语音和语言处理的成果:《机器学习》(Machine Learning)、《机器学习研究杂志》(Journal of Machine Learning)、《人工智能研究杂志》(Journal of Artificial Intelligence Research)。

有关语音和语言处理的各个方面的教科书的数量不少。Manning and Schütze(1999)的《统计语言处理基础》(Foundations of Statistical Language Processing)重点讲述标注、剖析、消歧、搭配等方面的统计模型。Charniak(1993)的《统计语言学习》(Statistical Language Learning)介绍相似的内容,尽管内容有些陈旧,篇幅也比较简短,但是通俗易懂,是一本入门读物。Allen(1995)的《自然语言理解》(Natural Language Understanding)从人工智能的角度,讲述了语言处理的各个方面的材料,覆盖面比较广。Manning et al.(2008)的《信息检索导论》(Introduction to Information Retrieval)着重论述信息检索、文本分类和文本聚类的问题。自然语言处理工具包(Natural Language ToolKit, NLTK)(Bird and Loper, 2004)是一整套的工具,包括程序语言 Python 模块和自然语言处理的数据,还包括在 NLTK 工具报基础上编写的关于自然语言处理的书。Allen(1995)的《自然语言理解》(Natural Language Understanding)从人工智能的角度来论述自然语言处理问题,覆盖面广。Gazdar and Mellish(1989)的《用 Lisp/Prolog/Pop11 进行自然语言处理》(Natural Language Processing in Lisp/Prolog/Pop 11)特别讲述了自动机、剖析、特征和合一等方面的内容,此书可以从网络上免费获得。Pereira and Shieber(1987)的《Prolog 与自然语言分析》(Prolog and Natural Language Analysis)介绍了基于 Prolog 的剖析和解释技术。Russell and Norvig(2002)的《人工智能:现代方法》(Artificial Intelligence: A Modern Approach)是人工智能的导论性读物,其中包括了自然语言处理的章节。Partee et al.(1990)的《语言学中的数学方法》(Mathematical Methods in Linguistics)全面地介绍了数理语言学。Grosz et al.(1986)的《自然语言处理读本》(Readings in Natural Language Processing)搜集了自然语言处理领域很多基础性研究的论文,尽管内容有些陈旧,但是所搜集的文章都是非常优秀的论文。

有很多的地方都可以获得语音语料库和文本语料库。其中最大的一个是语言资源联盟(LDC),LDC 是一个专门从事语料库的建设和分配的非盈利的联盟(http://www.ldc.upenn.edu/)。读者也可以访问如下的网页:CHILDES(http://childes.psy.cmu.edu/),英国国家语料库(the British National Corpus, http://www.natcorp.ox.ac.uk/),国际英语语料库(the International Corpus of English, http://www.ucl.ac.uk/english-usage/ice/index.htm),Gutenberg 项目(Project Gutenberg, http://www.gutenberg.org/)。