第1章 绪 论

自然语言作为人类思想情感最基本、最直接、最方便的表达工具，无时无刻不充斥在人类社会的各个角落。人们从出生后的第一声啼哭开始，就企图用语言（声音）来表达自己的情感和意图。随着信息时代的到来，人们使用自然语言进行通信和交流的形式也越来越多地体现出它的多样性、灵活性和广泛性。然而，人脑是如何实现自然语言理解这一思维过程的？我们应该如何建立语言、知识与客观世界之间可计算的逻辑关系，并实现具有较高区分能力的语义计算？为什么世界上不同人种在拥有几乎相同的大脑结构和语声工作机理的情况下，却无法实现不同语言之间的相互理解？众多科学问题至今仍困扰着我们，目前计算机处理自然语言的能力在大多数情况下都不能满足人类社会信息化时代的要求。有关专家已经指出，语言障碍已经成为制约21世纪社会全球化发展的一个重要因素。因此，如何尽早实现自然语言的有效理解，打破不同语言之间的固有壁垒，为人际之间和人机之间的信息交流提供更便捷、自然、有效和人性化的帮助与服务，已经成为备受人们关注的极具挑战性的国际前沿研究课题，也是全球社会共同追求的目标和梦想。

从研究内容上来看，自然语言处理研究集认知科学、计算机科学、语言学、数学与逻辑学、心理学等多种学科于一身，其研究范畴不仅涉及对人脑语言认知机理、语言习得与生成能力的探索，而且，包括对语言知识的表达方式及其与现实世界之间的关系，语言自身的结构、现象、运用规律和演变过程，以及不同语言之间的语义关系等。因此，自然语言处理是现代信息科学研究不可或缺的重要内容，从事这项研究不仅具有重要的科学意义，而且具有巨大的实用价值。

1.1 基本概念

1.1.1 语言学与语音学

我们知道，语言作为人类特有的用来表达情感、交流思想的工具，是一种特殊的社会现象，由语音、词汇和语法构成。语音和文字是构成语言的两个基本属性，语音是语言的物质外壳，文字则是记录语言的书写符号系统[黄伯荣等，1991]。

根据《现代语言学词典》[克里斯特尔，2002]的定义，语言学（linguistics）是指对语言的科学研究。作为一门纯理论的学科，语言学在近期获得了快速发展，尤其从20世纪60

年代起，已经成为一门知晓度很高的广泛教授的学科。

根据语言学家的注意中心和兴趣范围，语言学可以区分为一些不同的分支，例如，历时语言学（diachronic linguistics）或称历史语言学（historical linguistics）、共时语言学（synchronic linguistics）、一般语言学（general linguistics）、理论语言学（theoretical linguistics）、描述语言学（descriptive linguistics）、对比语言学（contrastive linguistics）或类型语言学（typological linguistics）、结构语言学（structural linguistics）等。

语音学（phonetics）是研究人类发音特点，特别是语音发音特点，并提出各种语音描述、分类和转写方法的科学。语音学一般有三个分支：①发音语音学（articulatory phonetics），研究发音器官是如何产生语音的；②声学语音学（acoustic phonetics），研究口耳之间传递语音的物理属性；③听觉语音学（auditory phonetics），研究人通过耳、听觉神经和大脑对语音的知觉反应。仪器语音学（instrumental phonetics）则是利用各种物理设备，如测量气流或分析声波的仪器等，来研究上述三个问题的任一方面[克里斯特尔，2002]。

由于语音学家研究的目标通常是发现支配语音性质和使用的普世原则，因此，语音学又常称作一般语音学或通用语音学（general phonetics）。实验语音学（experimental phonetics）具有相同的含义。

从研究方法上来看，如果研究者关心的只是语音发音、声学或知觉的一般性规律和特点，那么，语音学研究与语言学的关系不大。但是，如果研究者关注的重点是具体语言或方言（或语言、方言群）的语音特点时，我们往往很难说清楚语音学到底是一门独立的学科还是应看作语言学的一个分支。而在有些大学里，有的相关的系称为“语言学系”，有的则称为“语言学和语音学系”，但实际上“语言学系”也同样教授语音学。因此，为了避免这种名称上的差异可能给人们造成的错觉，一些聪明的外国人采用一种折中的办法，用复数的“语言科学（linguistic sciences）”来作为整个学科的统称，既包括语言学，也包括语音学。在本书中，我们愿意沿用这种复数的语言科学名称。

1.1.2 自然语言处理

自然语言处理（natural language processing，NLP）也称自然语言理解（natural language understanding，NLU），从人工智能研究的一开始，它就作为这一学科的重要研究内容探索人类理解自然语言这一智能行为的基本方法。在最近二三十年中，随着计算机技术，特别是网络技术的迅速发展和普及，自然语言处理研究得到了前所未有的重视和长足的进展，并逐渐发展成为一门相对独立的学科，备受关注。

语言学家刘涌泉在《大百科全书》（2002）中对自然语言处理的定义为：“自然语言处理是人工智能领域的主要内容，即利用电子计算机等工具对人类所特有的语言信息（包括口语信息和文字信息）进行各种加工，并建立各种类型的人-机-人系统。自然语言理解是其核心，其中包括语音和语符的自动识别以及语音的自动合成。”

冯志伟对“自然语言处理”的解释为：自然语言处理就是利用计算机为工具对人类特有的书面形式和口头形式的自然语言的信息进行各种类型处理和加工的技术[冯志伟，1996]。

美国计算机科学家马纳瑞斯(Bill Manaris)给自然语言处理的定义为:“自然语言处理是研究人与人交际中以及人与计算机交际中的语言问题的一门学科。自然语言处理要研制表示语言能力(linguistic competence)和语言应用(linguistic performance)的模型,建立计算框架来实现这样的语言模型,提出相应的方法来不断地完善这样的语言模型,根据这样的语言模型设计各种实用系统,并探讨这些实用系统的评测技术。”[Manaris,1999]

“计算语言学(computational linguistics)”这一术语是在美国国家科学院成立于 20 世纪 60 年代的自动语言处理咨询委员会(Automatic Language Processing Advisory Committee,ALPAC)于 1966 年宣布的对机器翻译译文质量的评估报告中首次提出来的,目前仍没有统一的严格定义,我们能够看到的定义基本上都是解释性的,但这并不影响我们对它的理解。根据英国《大不列颠百科全书》的解释,计算语言学是利用电子数字计算机进行的语言分析。虽然许多其他类型的语言分析也可以运用计算机,计算机分析最常用于处理基本的语言数据——例如建立语音、词、词元素的搭配以及统计它们的频率[翁富良,1998]。当然,从目前情况来看,这种解释似乎有点过时,因为它仅仅强调的是计算机作为辅助工具用于对自然语言进行一些相关的分析和统计,而没有把计算机作为一种可以提供主动服务,能够帮助人类达到对话、翻译、检索等若干目的的智能工具。

《现代语言学词典》[克里斯特尔,2002]中对计算语言学的定义为:语言学的一个分支,用计算技术和概念来阐述语言学和语音学问题。已开发的领域包括自然语言处理,言语合成,言语识别,自动翻译,编制语词索引,语法的检测,以及许多需要统计分析的领域(如文本考释)。

语言学家刘涌泉在我国的《大百科全书》(2002)中给出的解释是:计算语言学是语言学的一个分支,专指利用电子计算机进行语言研究。

根据这些定义和上述(复数)语言科学的概念可以看出,计算语言学实际上包括以语音为主要研究对象的语音学基础及其语音处理技术研究和以词汇、句子、话语或语篇(discourse)及其词法、句法、语义和语用等相关信息为主要研究对象的处理技术研究。在前面介绍的《现代语言学词典》给出的定义中,自然语言处理属于计算语言学研究的范畴。但实际上,近几年来由于自然语言处理技术的迅速发展,相关技术不断与语音识别(speech recognition)、语音合成(speech synthesis)等技术相互渗透和结合已经形成了若干新的研究分支,例如,基于语音输入输出的人机对话系统,语音翻译(speech-to-speech translation),语音文档摘要(speech document summarization)和语音文档检索(speech document retrieval)等。因此,从目前情况来看,自然语言处理一般不再被看作是计算语言学范畴内的一个研究分支,而两者基本上是处于同一层次上的概念。

从术语的字面上来看,似乎“计算语言学”更侧重于计算方法和语言学理论等方面的研究,而“自然语言理解(或处理)”更偏向于“理解”过程和“处理”方法等方面的研究,语言工程和应用系统实现方面的含义似乎更多一些,但是,在很多情况下我们很难绝对地区分开“计算语言学”与“自然语言处理”之间到底存在怎样的包含或重叠关系以及各自不同的内涵和外延。因此,很多人在谈到“计算语言学”、“自然语言处理”或“自然语言理解”这些术语时,往往默认为它们是同一个概念,至少在其外延上不再细究其差异。甚至有些专著

中于脆直接这样解释:计算语言学也称自然语言处理或自然语言理解[刘颖,2002]。

在本书中,我们主要介绍以词汇、语句、篇章和对话等为主要处理对象的自然语言处理技术的基本理论和实现方法,不涉及语音技术的细节。

1.1.3 关于"理解"的标准

当人们提到关于"理解"的标准时,总是不会忘记著名的英国数学家图灵(Turing)1950年提出的测试标准。当时图灵提出这个测试的目的是用来判断计算机是否可以被认为"能思考"。后来这个测试被称为图灵测试(Turing test),现已被多数人承认。图灵试图解决长久以来关于如何定义思考的哲学争论,他提出了一个虽然主观但可以操作的标准:如果一个计算机系统的表现(act)、反应(react)和互相作用(interact)都和有意识的个体一样,那么,这个计算机系统就应该被认为是有意识的。为此,图灵设计了一种"模仿游戏",即现在所说的图灵测试:测试人在一段规定的时间内,在无法看到反应来源的情况下,根据两个实体(被测试的计算机系统和另外一个人)对他提出的各种问题的反应来判断做出反应的是人还是计算机。通过一系列这样的测试,从计算机被误判为人的几率就可以测出计算机系统所具有的智能程度。

在自然语言处理领域中,人们采用图灵实验来判断计算机系统是否"理解"了某种自然语言的具体准则可以有很多,例如:通过问答(question-answering)系统测试计算机系统是否能够正确地回答输入文本中的有关问题;通过文摘生成(summarizing)系统测试计算机系统是否有能力自动产生输入文本的摘要;通过机器翻译(machine translation,MT)系统测试计算机系统是否具有把一种语言翻译成另一种语言的能力;通过文本释义(paraphrase)系统测试计算机系统是否能够用不同的词汇和句型来复述其输入文本,等等[石纯一,1993]。

实际上,人们在自然语言处理领域研究的任何一个应用系统都可以拿来做图灵测试。按照人的标准对这些系统的输出结果进行评价,从而判断计算机系统是否达到了"理解"的效果。显然,被测试系统所表现出来的性能反映了计算机系统的"理解"能力。因此,我们从事自然语言理解研究的任务也就是研究和探索针对具体应用目的的新方法和新技术,使实现系统的性能表现尽量符合人类理解的标准和要求。

1.2 自然语言处理研究的内容和面临的困难

1.2.1 自然语言处理研究的内容

自然语言处理研究的内容十分广泛,根据其应用目的不同,我们可以大致列举如下一些研究方向:

(1) 机器翻译(machine translation,MT):实现一种语言到另一种语言的自动翻译。

(2) 自动文摘(automatic summarizing 或 automatic abstracting):将原文档的主要内容和含义自动归纳、提炼出来,形成摘要或缩写。

(3) 信息检索(information retrieval):信息检索也称情报检索,就是利用计算机系统

从海量文档中找到符合用户需要的相关文档。面向多语言的信息检索叫作跨语言信息检索(cross-language/trans-lingual information retrieval)。

(4) 文档分类(document categorization/classification):文档分类也称文本分类(text categorization/classification)或信息分类(information categorization/classification),其目的就是利用计算机系统对大量的文档按照一定的分类标准(例如,根据主题或内容划分等)实现自动归类。

(5) 问答系统(question-answering system):通过计算机系统对人提出的问题的理解,利用自动推理等手段,在有关知识资源中自动求解答案并做出相应的回答。问答技术有时与语音技术和多模态输入、输出技术,以及人机交互技术等相结合,构成人机对话系统(human-computer dialogue system)。

(6) 文字编辑和自动校对(automatic proofreading):对文字拼写、用词,甚至语法、文档格式等进行自动检查、校对和编排。

(7) 信息过滤(information filtering):通过计算机系统自动识别和过滤那些满足特定条件的文档信息。通常指网络有害信息的自动识别和过滤,主要用于信息安全和防护等。

(8) 语言教学(language teaching):借助计算机辅助教学工具,进行语言教学、操练和辅导等。

(9) 文字识别(optical character recognition,OCR):通过计算机系统对印刷体或手写体等文字进行自动识别,将其转换成计算机可以处理的电子文本。相对而言,文字识别研究的主要内容属于字符(汉字)图像识别问题,但对于高性能文字识别系统而言,相关的语言理解技术一般是不可缺少的。

(10) 语音识别(speech recognition):将输入计算机的语音信号识别转换成书面语表示。语音识别也称自动语音识别(automatic speech recognition,ASR)。

(11) 文语转换(text-to-speech conversion):将书面文本自动转换成对应的语音表征,又称语音合成(speech synthesis)。

(12) 说话人识别/认证/验证(speaker recognition/identification/verification):对一说话人的言语样本做声学分析,依此推断(确定或验证)说话人的身份。

实际上,我们所能够想到的涉及人类语言的任何研究几乎都隐含着计算语言学的问题,这里不再一一列举了。

1.2.2　自然语言处理涉及的几个层次

如果暂时撇开语音学研究的层面,自然语言处理研究的问题一般会涉及自然语言的形态学、语法学、语义学和语用学等几个层次。

形态学(morphology):形态学(又称"词汇形态学"或"词法")是语言学的一个分支,研究词的内部结构,包括屈折变化和构词法两个部分。由于词具有语音特征、句法特征和语义特征,形态学处于音位学、句法学和语义学的结合部位,所以形态学是每个语言学家都要关注的一门学科 [Matthews,2000]。

语法学(syntax):研究句子结构成分之间的相互关系和组成句子序列的规则。其关注的中心是,为什么一句话可以这么说,也可以那么说?

语义学(semantics):是一门研究意义,特别是语言意义的学科[毛茂臣,1988]。语义

学的研究对象是语言的各级单位(词素、词、词组、句子、句子群、整段整篇的话语和文章,乃至整个著作)的意义,以及语义与语音、语法、修辞、文字、语境、哲学思想、社会环境、个人修养的关系,等等[陆善采,1993]。其重点在探明符号与符号所指的对象之间的关系,从而指导人们的言语活动。它所关注的重点是:这个语言单位到底说了什么?

语用学(pragmatics):是现代语言学用来指从使用者的角度研究语言,特别是使用者所作的选择、他们在社会互动中所受的制约、他们的语言使用对信递活动中其他参与者的影响。目前还缺乏一种连贯的语用学理论,主要是因为它必须说明的问题是多方面的,包括直指、会话隐含、预设、言语行为、话语结构等。部分原因是由于这一学科的范围太宽泛,因此出现多种不一致的定义。从狭隘的语言学观点看,语用学处理的是语言结构中有形式体现的那些语境。相反,语用学最宽泛的定义是研究语义学未能涵盖的那些意义[克里斯特尔,2002]。因此,语用学可以是集中在句子层次上的语用研究,也可以是超出句子,对语言的实际使用情况的调查研究,甚至与会话分析、语篇分析相结合,研究在不同上下文中的语句应用,以及上下文对语句理解所产生的影响。其关注的重点在于:为什么在特定的上下文中要说这句话?

在实际问题的研究中,上述几方面的问题,尤其是语义学和语用学的问题往往是相互交织在一起的。语法结构的研究离不开对词汇形态的分析,句子语义的分析也离不开对词汇语义的分析、语法结构和语用的分析,它们之间往往互为前提。

1.2.3 自然语言处理面临的困难

根据上面的介绍,自然语言处理涉及形态学、语法学、语义学和语用学等几个层面的问题,其最终应用目标包括机器翻译、信息检索、问答系统等非常广泛的应用领域。其实,如果进一步归结,实现所有这些应用目标最终需要解决的关键问题就是歧义消解(disambiguation)问题和未知语言现象的处理问题。一方面,自然语言中大量存在的歧义现象,无论在词法层次、句法层次,还是在语义层次和语用层次,无论哪类语言单位,其歧义性始终都是困扰人们实现应用目标的根本问题。因此,如何面向不同的应用目标,针对不同语言单位的特点,研究歧义消解和未知语言现象的处理策略和实现方法,就成了自然语言处理面临的核心问题。请看如下这些典型的歧义例句:

例句 1 Put the block in the box on the table.

在例句 1 中,"on the table"既可以修饰"box",也可以限定"block"。于是,我们可以得到两种不同的句法结构:

(1) Put the block [in the box on the table].

(2) Put [the block in the box] on the table.

如果在这个句子中再增加一个介词短语(...in the kitchen),我们可以得到 5 种可能的分析结果,另外再增加一个的话,就可以得到 14 种可能的分析结构[Samuelsson *et al.*,2000]。

类似地,见例句 2:

例句 2　I saw a man in the park with a telescope.

可以得到 5 种不同的分析结构[冯志伟,1996],而 W. A. Martin 曾报道他们的系统对于以下句子可以给出 455 个不同的句法分析结果[Martin *et al.*,1987]:

例句 3　List the sales of the products produced in 1973 with the products produced in 1972.

实际上,这种歧义结构分析结果的数量是随介词短语数目的增加呈指数上升的,其歧义组合的复杂程度随着介词短语个数的增加而不断加深,这个歧义结构的组合数称为开塔兰数(Catalan numbers,记作 C_n),即如果句子中存在这样 n(n 为自然数)个介词短语,C_n 可以由下式获得[Samuelsson *et al.*, 2000]:

$$C_n = \binom{2n}{n}\frac{1}{n+1}$$

由此,歧义结构数目的急剧增加,使得句法分析算法面临的困难迅速增大,句法分析算法不得不消耗大量的时间在这样一个组合爆炸的候选结构中搜索可能的路径,以实现局部歧义和全局歧义的有效消解。

在现代汉语中,尽管一般不会出现像上述英语例句那样由于多个介词结构的挂靠成分不同而引起句子歧义结构数目大量存在的现象,但是,汉语中的各类歧义现象却也是普遍存在的。请看如下例句:

例句 4　喜欢乡下的孩子。

这个句子可以理解为"[喜欢/乡下]的孩子。"也可以理解为"喜欢[乡下/的/孩子]。"而句子:

例句 5　关于鲁迅的著作。

可以解析为"关于[鲁迅/的/著作]",也可以解析为"[关于/鲁迅]的著作"。

句法结构歧义固然是自然语言处理中典型的问题,而词汇的词类(part-of-speech)歧义、词义歧义和句子的语义歧义等,也同样是自然语言处理中普遍存在的问题。例如,英语动词"swallow"通常需要有生命的动物作为主语,客观存在的有形的东西(被吞咽的对象)作为宾语,但在实际运用中,当用于隐喻时就出现了例外。例如[Manning *et al.*, 1999]:

例句 6　I swallowed his story, hook, line, and sinker.

例句 7　The supernova swallowed the planet.

在汉语中,似是而非、模棱两可的句子更是司空见惯。句子"咬死猎人的狗"既可以指"那只狗是咬死了猎人的狗",也可以指"把那只猎人的狗咬死";我们说"今天中午吃食堂"绝不意味着今天中午要把食堂吃下去,而是要在食堂吃午饭;我们夸奖一个人说"这个人真牛"时,并不是说这个人是真正的牛,而是夸奖他真能;"火烧圆明园"与"火烧驴肉"也绝非同一种结构和含义。在《现代汉语词典》(1999,商务印书馆)里"打"字做实词使用时就

有25种含义，在“打鼓、打架、打球、打酒、打电话、打毛衣”等用法中，“打”字的含义各有不同。除此之外，“打”字还可以用作介词（如：自打今天起）和量词（如：一打铅笔）。如何根据特定的上下文让计算机自动断定“打”字的确切含义恐怕不是一件容易的事情。

作为一个例子，请看如下这段幽默小片断：

他说：“她这个人真有意思（funny）。”她说：“他这个人怪有意思的（funny）。”于是人们以为他们有了意思（wish），并让他向她意思意思（express）。他火了：“我根本没有那个意思（thought）！”她也生气了：“你们这么说是什么意思（intention）？”事后有人说：“真有意思（funny）。”也有人说：“真没意思（nonsense）”。（原文见《生活报》1994.11.13.第六版）[吴尉天，1999]

在整个片断中，“意思”一词在不同的语境里共有6个不同的含义。如果实现这个词义的自动理解，恐怕不是目前的自然语言处理系统所能够胜任的。当然，这个片断可能是人为编造出来的，实际运用中一般不会出现如此复杂的用词方法。我们使用这个例子的意思也绝不是说一个自然语言处理系统必须具备如此复杂的歧义消解能力才算得上是真正实用的系统，而只是想说明，歧义是自然语言中普遍存在的语言现象，它们广泛地存在于词法、句法、语义、语用和语音等每一个层面。任何一个自然语言处理系统，都无法回避歧义的消解问题。

另一方面，对于一个特定系统来说，总是有可能遇到未知词汇、未知结构等各种意想不到的情况，而且每一种语言又都随着社会的发展而动态变化着，新的词汇（尤其是一些新的专用词汇）、新的词义、新的词汇用法（新词类），甚至新的句子结构都在不断出现，尤其在口语对话或计算机网络对话（通过MSN、QQ等形式）中，稀奇古怪的词语和话语结构更是司空见惯。因此，一个实用的自然语言处理系统必须具有较好的未知语言现象的处理能力，而且有足够的对各种可能输入形式的容错能力，即我们通常所说的系统的鲁棒性（robustness）问题。当然，还对于机器翻译、信息检索、文本分类等特定的自然语言处理任务来说，还存在若干与任务相关的其他问题，诸如如何处理不同语言的差异、如何提取文本特征等。

总而言之，目前的自然语言处理研究面临着若干问题的困扰，既有数学模型不够奏效、有些算法的复杂度过高、鲁棒性太差等理论问题，也有数据资源匮乏、覆盖率低、知识表示困难等知识资源方面的问题，当然，还有实现技术和系统集成方法不够先进等方面的问题。正是这些问题和困难，才使得计算语言学研究更加充满挑战性，更需要我们去创造和探索。

1.3 自然语言处理的基本方法及其发展

1.3.1 自然语言处理的基本方法

一般认为，自然语言处理中存在着两种不同的研究方法，一种是理性主义（rationalist）方法，另一种是经验主义（empiricist）方法[Church *et al.*，1993]。

理性主义方法认为，人的很大一部分语言知识是与生俱来的，由遗传决定的。持这种观点的代表人物是美国语言学家乔姆斯基（Noam Chomsky），他的内在语言官能（innate

language faculty)理论被广泛地接受。乔姆斯基认为,很难知道小孩在接收到极为有限的信息量的情况下,在那么小的年龄如何学会了如此之多复杂的语言理解的能力。因此,理性主义的方法试图通过假定人的语言能力是与生俱来的、固有的一种本能来回避这些困难的问题。

在具体的自然语言问题研究中,理性主义方法主张建立符号处理系统,由人工整理和编写初始的语言知识表示体系(通常为规则),构造相应的推理程序,系统根据规则和程序,将自然语言理解为符号结构——该结构的意义可以从结构中的符号的意义推导出来。按照这种思路,在自然语言处理系统中,一般首先由词法分析器按照人编写的词法规则对输入句子的单词进行词法分析,然后,语法分析器根据人设计的语法规则对输入句子进行语法结构分析,最后再根据一套变换规则将语法结构映射到语义符号(如逻辑表达式、语义网络、中间语言等)。

而经验主义的研究方法也是从假定人脑所具有的一些认知能力开始的。因此,从这种意义上讲,两种方法并不是绝对对立的。但是,经验主义的方法认为人脑并不是从一开始就具有一些具体的处理原则和对具体语言成分的处理方法,而是假定孩子的大脑一开始具有处理联想(association)、模式识别(pattern recognition)和通用化(generalization)处理的能力,这些能力能够使孩子充分利用感官输入来掌握具体的自然语言结构。在系统实现方法上,经验主义方法主张通过建立特定的数学模型来学习复杂的、广泛的语言结构,然后利用统计学、模式识别和机器学习等方法来训练模型的参数,以扩大语言使用的规模。因此,经验主义的自然语言处理方法是建立在统计方法基础之上的,因此,我们又称其为统计自然语言处理(statistical natural language processing)方法。

在统计自然语言处理方法中,一般需要收集一些文本作为统计模型建立的基础,这些文本称为语料(corpus)。经过筛选、加工和标注等处理的大批量语料构成的数据库叫作语料库(corpus base)。由于统计方法通常以大规模语料库为基础,因此,又称为基于语料(corpus-based)的自然语言处理方法。

实际上,理性主义和经验主义试图刻画的是两种不同的东西。Chomsky 的生成语言学理论试图刻画的是人类思维(I-language)的模式或方法。对于这种方法而言,某种语言的真实文本数据(E-language)只是提供间接的证据,这种证据可以由以这种语言为母语的人提供。而经验主义方法则直接关心如何刻画这些真实的语言本身(E-language)。Chomsky 把语言的能力(linguistic competence)和语言的表现(linguistic performance)区分开来了。他认为,语言的能力反映的是语言结构知识,这种知识是说话人头脑中固有的,而语言表现则受到外界环境诸多因素的影响,如记忆的限制、对环境噪声的抗干扰能力等。

1.3.2　自然语言处理的发展

理性主义和经验主义在基本出发点上的差异导致了在很多领域中都存在着两种不同的研究方法和系统实现策略,这些领域在不同的时期被不同的方法主宰着。

在 20 世纪 20 年代到 60 年代的近 40 年时间里,经验主义方法在语言学、心理学、人工智能等领域中处于主宰的地位,人们在研究语言运用的规律、言语习得、认知过程等问

题时，都是从客观记录的语言、语音数据出发，进行统计、分析和归纳，并以此为依据建立相应的分析或处理系统。

大约从20世纪60年代中期到20世纪80年代中后期，语言学、心理学、人工智能和自然语言处理等领域的研究几乎完全被理性主义研究方法控制着，人们似乎更关心关于人类思维的科学，人们通过建立很多小的系统来模拟智能行为，这种研究方法一直到今天还仍然有人在使用。但是，这种做法常常受到批评，因为这种做法只能处理一些小的问题，而不能对研究方法的有效性给出一个总的客观的评估，因此，这些小系统有时也被轻蔑地称为玩具[Manning *et al.*，1999]。

无论如何，我们必须承认，这一时期的计算语言学理论得到了长足的发展并逐渐成熟，出现了一系列重要的理论研究成果，其中，乔姆斯基的形式语言理论[Chomsky，1956]是影响最大的早期计算语言学句法理论。后来乔姆斯基又分别在20世纪50年代和70年代提出了转换生成语法和约束管辖理论。随后，很多学者又提出了扩充转移网络、词汇功能语法、功能合一语法、广义短语结构语法和中心驱动的短语结构语法等。1969年厄尔利(J. Earley)提出了Earley句法分析算法[Earley，1970]；1980年马丁·凯(Martin Kay)提出了线图句法分析算法(chart parsing)[Allen，1995]；1985年富田胜(M. Tomita)提出了Tomita句法分析算法[Tomita，1985，1987]。这些研究成果为自然语言自动句法分析奠定了良好的理论基础。在语义分析方面，1966年菲尔摩(C. J. Filmore)提出了格语法；1968年美国心理学家奎廉(M. R. Quilian)在研究人类联想记忆时提出了语义网络(semantic network)的概念；1972年美国人工智能专家西蒙斯(Simmous)等人首先将语义网络用于自然语言理解系统中；1974年威尔克斯(Y. Wilks)提出了优选语义学；20世纪70年代初，美国数理逻辑学家蒙塔格(Richard Montague)提出的蒙塔格语法，首次提出了利用数理逻辑来研究自然语言的句法结构和语义关系的设想，为自然语言处理研究开辟了一条新的途径。

总之，这一时期的理论成果不仅为计算语言学的进一步发展奠定了坚实的理论基础，而且对我们今天研究人类语言能力这一智能行为，促进认知科学、语言学、心理学和人工智能等相关学科的发展，具有重要的理论意义和现实意义。

大约在1989年以后，人们越来越多地关注工程化、实用化的解决问题方法，经验主义方法被人们重新认识并得到迅速发展。在自然语言处理研究中，重要的标志是基于语料库的统计方法被引用到自然语言处理中，并发挥着重要作用，很多人开始研究和关注基于大规模语料的统计机器学习方法及其在自然语言处理中的应用，并客观地比较和评价各种方法的性能。这种处理思路和重心的转移经常反映在我们使用的一些新术语上，比如说，“语言技术”或“语言工程”。在这一时期，基于语料库的机器翻译(corpus-based machine translation)方法得到了充分发展，尤其是IBM的研究人员提出的基于噪声信道模型(noisy channel model)的统计机器翻译(statistical machine translation)模型[Brown *et al.*，1990，1993]及其实现的Candide翻译系统[Berger *et al.*，1994]，为经验主义方法的复苏和兴起吹响了号角，并成为机器翻译领域的里程碑。

IBM的研究人员基于统计翻译模型实现的法语到英语的机器翻译实验系统Candide，

以加拿大议会辩论记录的英法双语语料 Hansard[①] 作为训练数据，取得了较好的实验效果，几乎半数的短语得到了完全正确或基本正确的翻译结果。根据 ARPA 的测试结果，Candide 系统译文的流利程度(fluency)甚至超过了著名的商品化机器翻译系统 SYSTRAN[②]，在当时引起了轰动。与此同时，日本著名学者长尾真(Makoto Nagao)提出的基于实例的机器翻译(example-based machine translation)方法也得到长足发展，并建立了实验系统[Sato，1990；Sumita *et al.*，1991；McLean，1992]。这些标志性成果的诞生结束了基于规则的机器翻译系统一统天下的单一局面[Hutchins，2001a；US NSF Report，1999]。

另外，值得指出的是，隐马尔可夫模型(Hidden Markov Model，HMM)等统计方法在语音识别中的成功运用对自然语言处理的发展也起到了推波助澜的作用，甚至是关键的作用，统计机器翻译中的许多思想都来源于语音识别中统计模型成功运用的经验，或在某种程度上受到了统计语音识别研究思路的启发[Hutchins，2001a；Abney，2002a；Och *et al.*，2001a；Church，1993；Church *et al.*，2003]。实践证明，除了语音识别和机器翻译以外，很多自然语言处理的研究任务，包括汉语自动分词和词性标注、文字识别、拼音法汉字输入等，都可以用噪声信道模型来描述和实现。

同时，随着统计方法在自然语言处理中的广泛应用和快速发展，以语料库为研究对象和基础的语料库语言学(corpus linguistics)迅速崛起。由于语料库语言学从大规模真实语料中获取语言知识，以求得对于自然语言规律更客观、准确的认识，因此，越来越多地得到广大学者的认识。尤其是随着计算机网络的迅速发展和广泛使用，语料的获取更加便捷，语料库规模更大、质量更高，因此，语料库语言学的崛起又反过来进一步推动了计算语言学其他相关技术的快速发展，一系列基于统计模型的自然语言处理系统相继被开发，并获得了一定的成功，例如，基于统计方法的汉语自动分词与词性标注系统、句法解析器、信息检索系统和自动文摘系统等。

作者认为，经验主义方法的复苏与快速发展一方面得益于计算机硬件技术的快速发展，计算机存储容量的迅速扩大和运算速度的迅速提高，使得很多复杂的原来无法实现的统计方法能够容易地实现；另一方面，统计机器学习等新理论方法的不断涌现，也进一步推动了自然语言处理技术的快速发展。

在 20 世纪 80 年代末期和 90 年代初期，曾经引发了关于理性主义和经验主义两种不同观点的激烈争论。但随着时间的推移，当人们从那些空泛的辩论中冷静下来以后，逐渐地认识到，无论是理性主义也好，还是经验主义也罢，任何一种方法都不可能完全解决自然语言处理这一复杂问题，只有将两种方法很好地结合起来，寻找一种融合的解决问题办法，甚至建立一种新的理论方法才是自然语言处理研究的真正出路。理性主义方法和经验主义方法从对立状态的结束到相互结合、共同发展，使得目前的计算语言学研究正处于一个前所未有的繁荣发展时期。

如果从 1946 年世界上第一台计算机诞生、英国人 A. D. Booth 和美国人 W. Weaver 提出利用计算机进行机器翻译研究开始算起，自然语言处理技术经过了 60 多年的发展历

① http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T20

② http://www.systransoft.com/index.html

程，期间潮起潮落，几经曲折。冯志伟曾将整个发展历程归纳为“萌芽期、发展期和繁荣期”三个历史阶段[冯志伟，2001b]。

回顾自然语言处理技术半个多世纪的发展历程，黄昌宁等(2002b)认为这一领域的研究取得了两点重要认识，即：①对于句法分析，基于单一标记的短语结构规则是不充分的；②短语结构规则在真实文本中的分布呈现严重的扭曲。换言之，有限数目的短语结构规则不能覆盖大规模真实语料中的语法现象，这与原先的预期大相径庭。NLP 技术的发展在很大程度上受到这两个事实的影响。从这个意义上说，本领域中称得上里程碑式的成果有三个：①复杂特征集和合一语法的提出；②语言学研究中词汇主义的建立；③语料库方法和统计语言模型的广泛运用。大规模语言知识的开发和自动获取成为目前 NLP 技术的瓶颈问题。因此，语料库建设和统计学理论将成为该领域中研究的关键课题。实际上，近几年来在众多词汇资源的开发过程中，语料库和统计学方法发挥了很大的作用，这也是经验主义方法和理性主义方法相互融合的可喜开端。

1.4 自然语言处理的研究现状

关于自然语言处理技术的研究现状不是一个容易回答的问题，因为自然语言处理涉及太多的领域和分支，而且各个领域和分支都有一定的相对独立性，发展起点和速度也不一样。但是，如果我们不考虑具体的技术分支，从自然语言处理研究的总体状况来看，可以简单地用以下三点来粗略地反映自然语言处理技术所处的现状：

(1) 很多技术已经达到或基本达到实用程度，并在实际应用中发挥着巨大作用。例如，文字输入、编辑、排版，文字识别，电子词典，语音合成等。

(2) 许多新的研究方向不断出现。正如我们前面指出的，受实际应用的驱动，自然语言处理技术不断与新的相关技术相结合，用于研究和开发越来越多的实用技术。例如，网络内容管理、网络信息监控和有害信息过滤等，这些研究不仅与自然语言处理技术密切相关，而且涉及图像理解、情感计算和网络技术等多种相关技术。而语音自动翻译则是涉及语音识别、机器翻译、语音合成和通信等多种技术的综合集成技术。语音自动文摘、语音检索和基于图像内容及文字说明的图像理解技术研究等，都是集自然语言处理技术和语音技术、图像技术等于一体的综合应用技术。对于这些新的任务，研究尚刚刚开始或者仅处于非常初步的探索阶段，离问题的最终解决和达到实用化目标还十分遥远。

(3) 许多理论问题尚未得到根本性的解决。尽管许多理论模型在自然语言处理研究中发挥着重要作用，并且很多方法已经得到实际应用，如上下文无关文法、HMM、噪声信道模型等，但是，许多重要的问题仍未得到彻底、有效的解决，如语义的形式化与计算问题、句法分析问题、指代歧义消解问题、汉语自动分词中的未登录词(unknown word)识别问题等。纵观整个自然语言处理领域，尚未建立起一套完整、系统的理论框架体系。许多理论研究仍处于盲目的探索阶段，如尝试一些新的机器学习方法或未曾使用的数学模型，这些尝试和实验带有很强的主观性和盲目性。在技术实现上，许多改进往往仅限于对一些边角问题的修修补补，或者只是针对特定条件下一些具体问题的处理，未能从根本上建立一套广泛适用的、鲁棒的处理策略。总之，面对自然语言问题的复杂性和多变性，现有

的理论模型和方法还远远不够，有待于进一步改进和完善，并期待着新的更有效的理论模型和方法的出现。

综上所述，自然语言处理研究已经取得了丰硕成果，同时也面临着许多新的挑战。无论如何，我们在评价任何一门学科和技术的时候，既不应该因为它所取得的成绩而忽略了问题的存在，也不应该因为问题的存在而全盘否定这门学科的发展。对于评价自然语言处理这门学科更是如此，因为实际上对于自然语言处理的很多问题，具有高度智慧的人类本身解决起来都不能达到非常准确、满意的程度，甚至无法清楚地知道人脑处理这些问题的具体过程，那么，在目前对自然语言处理的一些具体技术提出过高的要求显然没有太多的理由，给予太多的批评和指责也是不公正的。比如说，在现阶段过高地要求机器翻译系统的译文质量和信息抽取系统的准确率等，都是不现实的。相反，这些技术在实际应用中已经在一定程度上为我们提供了很大的帮助和便利。当然，我们并不是不允许人们对某一项技术提出更高的要求和希望，重要的是首先应该建立有效的理论模型和实现方法。这也是自然语言处理这门学科所面临的问题和挑战。

第2章

预备知识

在基于统计方法的自然语言处理研究中，有关统计学和信息论等方面的知识是不可缺少的基础。因此，在本章中我们将对概率论、信息论和支持向量机等有关概念作简要的回顾。我们假设读者已经具备有关这方面知识的基础，因此，对于本章中提到的公式和定理并不作详细的推导和证明。如果读者对其中的某些公式和符号需要进一步了解，可参阅相关的专著或论文。如果读者已经对本章所介绍的内容非常清楚和了解，那么，完全可以越过本章的内容。

2.1 概率论基本概念

2.1.1 概率

概率(probability)是从随机试验中的事件到实数域的映射函数，用以表示事件发生的可能性。如果用 $P(A)$ 作为事件 A 的概率，Ω 是试验的样本空间，则概率函数必须满足如下三条公理：

公理 2-1(非负性)　$P(A)\geqslant 0$

公理 2-2(规范性)　$P(\Omega)=1$

公理 2-3(可列可加性)　对于可列无穷多个事件 $A_1, A_2, \cdots$，如果事件两两互不相容，即对于任意的 i 和 $j(i\neq j)$，事件 A_i 和 A_j 不相交($A_i \cap A_j = \varnothing$)，则有

$$P\left(\bigcup_{i=0}^{\infty} A_i\right) = \sum_{i=0}^{\infty} P(A_i) \tag{2-1}$$

2.1.2 最大似然估计

如果$\{s_1, s_2, \cdots, s_n\}$是一个试验的样本空间，在相同的情况下重复试验 N 次，观察到样本 $s_k(1\leqslant k\leqslant n)$的次数为 $n_N(s_k)$，那么，s_k 在这 N 次试验中的相对频率为

$$q_N(s_k) = \frac{n_N(s_k)}{N} \tag{2-2}$$

由于$\sum_{k=1}^{n} n_N(s_k) = N$，因此，$\sum_{k=1}^{n} q_N(s_k) = 1$。

当 N 越来越大时，相对频率 $q_N(s_k)$ 就越来越接近 s_k 的概率 $P(s_k)$。事实上，

$$\lim_{N\to\infty} q_N(s_k) = P(s_k) \tag{2-3}$$

因此，通常用相对频率作为概率的估计值。这种估计概率值的方法称为最大似然估计(likelihood estimation)。

2.1.3　条件概率

如果 A 和 B 是样本空间 Ω 上的两个事件，$P(B)>0$，那么，在给定 B 时 A 的条件概率(conditional probability) $P(A|B)$ 为

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{2-4}$$

条件概率 $P(A|B)$ 给出了在已知事件 B 发生的情况下，事件 A 的概率。一般地，$P(A|B)\neq P(A)$。

根据公式(2-4)，有

$$P(A \cap B) = P(B)P(A \mid B) = P(A)P(B \mid A) \tag{2-5}$$

这个等式有时称为概率的乘法定理或乘法规则，其一般形式表示为

$$P(A_1 \cap \cdots \cap A_n) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2)\cdots P\Big(A_n \mid \bigcap_{i=1}^{n-1} A_i\Big) \tag{2-6}$$

这一规则在自然语言处理中使用得非常普遍。

条件概率也有三个基本性质：

(1) 非负性：$P(A|B)\geqslant 0$；

(2) 规范性：$P(\Omega|B)=1$；

(3) 可列可加性：如果事件 $A_1, A_2, \cdots$ 两两互不相容，则

$$P\Big(\sum_{i=1}^{\infty} A_i \mid B\Big) = \sum_{i=1}^{\infty} P(A_i \mid B) \tag{2-7}$$

2.1.4　贝叶斯法则

贝叶斯法则，或称贝叶斯理论(Bayesian theorem)，是条件概率计算的重要依据。实际上，根据条件概率的定义公式(2-4)和乘法规则式(2-5)，可得

$$P(B \mid A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A \mid B)P(B)}{P(A)} \tag{2-8}$$

式(2-8)右边的分母可以看作普通常量，因为我们只是关心在给定事件 A 的情况下可能发生事件 B 的概率，$P(A)$ 的值是确定不变的。故有

$$\arg\max_B \frac{P(A \mid B)P(B)}{P(A)} = \arg\max_B P(A \mid B)P(B) \tag{2-9}$$

以下给出事件 A 的概率计算方法。

首先根据乘法规则

$$P(A \cap B) = P(A \mid B)P(B)$$
$$P(A \cap \overline{B}) = P(A \mid \overline{B})P(\overline{B})$$

因此有

$$P(A) = P(A \cap B) + P(A \cap \overline{B})$$
$$= P(A \mid B)P(B) + P(A \mid \overline{B})P(\overline{B})$$

推广到一般形式，假设 B 是样本空间 Ω 的一个划分，即 $\sum_i B_i = \Omega$。如果 $A \subseteq \bigcup_i B_i$，并且 B_i 互不相交，那么 $A = \sum_{i=1} B_i A$，于是 $P(A) = \sum_{i=1} P(B_i A)$。由乘法定理可得

$$P(A) = \sum_i P(A \mid B_i)P(B_i) \tag{2-10}$$

公式(2-10)称为全概率公式。

类似地，我们给出如下贝叶斯法则的精确描述。

假设 A 为样本空间 Ω 的事件，$B_1, B_2, \cdots, B_n$ 为 Ω 的一个划分，如果 $A \subseteq \bigcup_{i=1}^{n} B_i$，$P(A)>0$，并且 $i \neq j, B_i \cap B_j = \varnothing, P(B_i)>0 (i=1,2,\cdots,n)$，则

$$P(B_j \mid A) = \frac{P(A \mid B_j)P(B_j)}{P(A)} = \frac{P(A \mid B_j)P(B_j)}{\sum_{i=1}^{n} P(A \mid B_i)P(B_i)} \tag{2-11}$$

例 2-1 假设一多义词的某一义项很少被使用，平均该词每出现 100 000 次这一义项才有可能被使用一次。我们开发了一个程序来判断该词出现在某个句子中时是否使用了该义项。如果句子中确实使用了该词的这一义项时，程序判断结果为“使用”的概率是0.95。如果句子中实际上没有使用该词的这一义项时，程序错误地判断为“使用”的概率是 0.005。那么，这个程序判断句子使用该词的这一特殊义项的结论是正确的概率有多大?

解：假设 G 表示事件“句子中确实使用了该词的这一特殊义项”，T 表示事件“程序判断的结论是该句子使用了该词的这一特殊义项”。则有

$$P(G) = \frac{1}{100\,000} = 0.000\,01, \quad P(\overline{G}) = \frac{100\,000-1}{100\,000} = 0.999\,99$$
$$P(T \mid G) = 0.95, \quad P(T \mid \overline{G}) = 0.005$$

于是，可得

$$P(G \mid T) = \frac{P(T \mid G)P(G)}{P(T \mid G)P(G) + P(T \mid \overline{G})P(\overline{G})}$$
$$= \frac{0.95 \times 0.000\,01}{0.95 \times 0.000\,01 + 0.005 \times 0.999\,99} \approx 0.002$$

也就是说，程序判断句子使用该词的这一特殊义项的结论是正确的概率只有 0.002。

2.1.5 随机变量

一个随机试验可能有多种不同的结果，到底会出现哪一种，存在一定的概率，即随机会而定。简单地说，随机变量(random variable)就是试验结果的函数。

设 X 为一离散型随机变量，其全部可能的值为$\{a_1, a_2, \cdots\}$。那么

$$p_i = P(X = a_i), \quad i = 1,2,\cdots \tag{2-12}$$

称为 X 的概率函数。显然，$p_i \geqslant 0$，$\sum_{i=1} p_i = 1$。有时式(2-12)也称随机变量 X 的概率分布，此时，函数

$$P(X \leqslant x) = F(x), \quad -\infty < x < \infty \tag{2-13}$$

称为 X 的分布函数。

2.1.6　二项式分布

假设某一事件 A 在一次试验中发生的概率为 p，现把试验独立地重复进行 n 次。如果用变量 X 来表示 A 在这 n 次试验中发生的次数，那么，X 的取值可能为 $0,1,\cdots,n$。为了确定其分布情况，考虑事件$\{X=i\}$，如果这个事件发生，必须在这 n 次记录中有 i 个 A，$n-i$ 个 $\overline{A}$。那么，每个 A 有概率 p，每个 $\overline{A}$ 有概率 $1-p$。由于在 n 次试验中每次 A 出现与否与其他各次实验的结果无关，因此，根据乘法定理可以得出：每个这样的结果序列 $A\overline{A}AA\cdots\overline{A}$ 发生的概率为 $p^i(1-p)^{n-i}$。又因 A 可能出现在 n 个位置中的任何一处，因此，结果序列有$\binom{n}{i}$种可能。由此可得

$$p_i = \binom{n}{i} p^i (1-p)^{n-i}, \quad i = 0,1,\cdots,n \tag{2-14}$$

X 所遵从的这种概率分布称为二项式分布(binomial distribution)，并记为：$B(n,p)$。如果随机变量 X 服从某种特定的分布 F 时，我们常用 $X\sim F$ 表示。如果 X 服从二项式分布，可记为：$X\sim B(n,p)$。

二项式分布是最重要的离散型概率分布之一。在自然语言处理中，一般以句子为处理单位。为了简化问题的复杂性，通常假设一个句子的出现独立于它前面的其他语句，句子的概率分布近似地被认为符合二项式分布。

2.1.7　联合概率分布和条件概率分布

假设(X_1,X_2)为一个二维的离散型随机向量，X_1 全部可能的取值为 $a_1,a_2,\cdots$；X_2 全部可能的取值为 $b_1,b_2,\cdots$。那么，(X_1,X_2)的联合概率分布(joint distribution)为

$$p_{ij} = P(X_1 = a_i, X_2 = b_j), \quad i = 1,2,\cdots;\ j = 1,2,\cdots$$

一个随机变量或向量 $\boldsymbol{X}$ 的条件概率分布就是在某种给定的条件之下 $\boldsymbol{X}$ 的概率分布。考虑 X_1 在给定 $X_2=b_j$ 条件下的概率分布，实际上就是求条件概率 $P(X_1=a_i \mid X_2=b_j)$。根据条件概率的定义可得

$$P(X_1 = a_i \mid X_2 = b_j) = \frac{P(X_1 = a_i, X_2 = b_j)}{P(X_2 = b_j)} = \frac{p_{ij}}{P(X_2 = b_j)}$$

由于 $P(X_2 = b_j) = \sum_k p_{kj}$，故有

$$P(X_1 = a_i \mid X_2 = b_j) = \frac{p_{ij}}{\sum_k p_{kj}}, \quad i = 1,2,\cdots \tag{2-15}$$

类似地，

$$P(X_2 = b_j \mid X_1 = a_i) = \frac{p_{ij}}{\sum_k p_{ik}}, \quad j = 1,2,\cdots \tag{2-16}$$

2.1.8　贝叶斯决策理论

贝叶斯决策理论(Bayesian decision theory)是统计方法处理模式分类问题的基本理

论之一。假设研究的分类问题有 c 个类别，各类别的状态用 w_i 表示，$i=1,2,\cdots,c$；对应于各个类别 w_i 出现的先验概率为 $P(w_i)$；在特征空间已经观察到某一向量 $\boldsymbol{x}$，$\boldsymbol{x}=[x_1,x_2,\cdots,x_d]^{\mathrm{T}}$ 是 d 维特征空间上的某一点，且条件概率密度函数 $p(\boldsymbol{x}|w_i)$是已知的。那么，利用贝叶斯公式我们可以得到后验概率 $P(w_i|\boldsymbol{x})$如下：

$$P(w_i \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid w_i)P(w_i)}{\sum_{j=1}^{c} p(\boldsymbol{x} \mid w_j)P(w_j)}$$

基于最小错误率的贝叶斯决策规则为：

(1) 如果 $P(w_i|\boldsymbol{x})=\max\limits_{j=1,2,\cdots,c} P(w_j|\boldsymbol{x})$，那么，$\boldsymbol{x}\in w_i$ (2-17)

或者说，

(2) 如果 $p(\boldsymbol{x}|w_i)P(w_i)=\max\limits_{j=1,2,\cdots,c} p(\boldsymbol{x}|w_j)P(w_j)$，则 $\boldsymbol{x}\in w_i$ (2-18)

如果类别只有两类时，即 $c=2$，则有：

(3) 如果 $l(\boldsymbol{x})=\dfrac{p(\boldsymbol{x}|w_1)}{p(\boldsymbol{x}|w_2)}>\dfrac{P(w_2)}{P(w_1)}$，则 $\boldsymbol{x}\in w_1$，否则 $\boldsymbol{x}\in w_2$ (2-19)

其中，$l(\boldsymbol{x})$为似然比(likelihood ratio)，而$\dfrac{P(w_2)}{P(w_1)}$称为似然比阈值(threshold)。

还有一种基于贝叶斯理论的决策方法叫作最小风险的贝叶斯决策，这里我们不再详细介绍了，有兴趣的读者可以参阅文献[边肇祺等，2000]或[杨光正等，2001]等。贝叶斯决策理论在自然语言处理中的词义消歧(word sense disambiguation，WSD)、文本分类等问题的研究中具有重要用途。

2.1.9 期望和方差

期望值(expectation)是指随机变量所取值的概率平均。假设 X 为一随机变量，其概率分布为 $P(X=x_k)=p_k$，$k=1,2,\cdots$，若级数 $\sum\limits_{k=1}^{\infty} x_k p_k$ 绝对收敛，那么，随机变量 X 的数学期望或概率平均值为

$$E(X) = \sum_{k=1}^{\infty} x_k p_k \tag{2-20}$$

例 2-2 假设某个网页的主菜单栏里共有 6 个关键词，每个关键词被点击的概率一样，经过一段时间后，第 1 到第 6 个关键词被点击的次数分别为 1，2，…，6。那么，平均每个单词被点击次数的期望值 $E(N)$如下：

$$E(N) = \sum_{t=1}^{6} t \times p(w) = \frac{1}{6}\sum_{t=1}^{6} t = \frac{21}{6} = 3\frac{1}{2}$$

其中，变量 t 为关键词被点击的次数，$p(w)$为每个关键词被点击的概率。

一个随机变量的方差(variance)描述的是该随机变量的值偏离其期望值的程度。如果 X 为一随机变量，那么，其方差 var(X)为

$$\begin{aligned} \mathrm{var}(X) &= E((X-E(X))^2) \\ &= E(X^2) - E^2(X) \end{aligned} \tag{2-21}$$

平方根 $\sqrt{\mathrm{var}(X)}$称为 X 的标准差。

这里介绍的概率知识主要是关于离散事件和离散型随机变量方面的，有关其他方面的详细介绍请参阅相关的概率论专著。

2.2　信息论基本概念

2.2.1　熵

香农(Claude Elwood Shannon)于 1940 年获得麻省理工学院数学博士学位和电子工程硕士学位后，于 1941 年加入了贝尔实验室数学部，并在那里工作了 15 年。1948 年 6 月和 10 月，由贝尔实验室出版的《贝尔系统技术》杂志连载了香农博士的文章《通信的数学原理》，该文奠定了信息论的基础。熵(entropy)是信息论的基本概念。

如果 X 是一个离散型随机变量，取值空间为 $\mathbf{R}$，其概率分布为 $p(x)=P(X=x), x\in\mathbf{R}$。那么，$X$ 的熵 $H(X)$ 定义为式(2-22)：

$$H(X)=-\sum_{x\in\mathbf{R}}p(x)\log_2 p(x) \tag{2-22}$$

其中，约定 $0\log 0=0$。$H(X)$ 可以写为 $H(p)$。由于在公式(2-22)中对数以 2 为底，该公式定义的熵的单位为二进制位(比特)。通常将 $\log_2 p(x)$ 简写成 $\log p(x)$。

熵又称为自信息(self-information)，可以视为描述一个随机变量的不确定性的数量。它表示信源 X 每发一个符号(不论发什么符号)所提供的平均信息量[姜丹，2001]。一个随机变量的熵越大，它的不确定性越大，那么，正确估计其值的可能性就越小。越不确定的随机变量越需要大的信息量用以确定其值。

例 2-3　假设 a,b,c,d,e,f 6 个字符在某一简单的语言中随机出现，每个字符出现的概率分别为：1/8，1/4，1/8，1/4，1/8 和 1/8。那么，每个字符的熵为

$$\begin{aligned}H(P)&=-\sum_{x\in\{a,b,c,d,e,f\}}P(x)\log P(x)\\&=-\left[4\times\frac{1}{8}\log\frac{1}{8}+2\times\frac{1}{4}\log\frac{1}{4}\right]=2\frac{1}{2}\text{(比特)}\end{aligned}$$

这个结果表明，我们可以设计一种编码，传输一个字符平均只需要 2.5 个比特：

字符：　a　　b　　c　　d　　e　　f

编码：100　00　101　01　110　111

2.2.2　联合熵和条件熵

如果 X,Y 是一对离散型随机变量 $X,Y\sim p(x,y)$，X,Y 的联合熵(joint entropy) $H(X,Y)$ 定义为

$$H(X,Y)=-\sum_{x\in X}\sum_{y\in Y}p(x,y)\log p(x,y) \tag{2-23}$$

联合熵实际上就是描述一对随机变量平均所需要的信息量。

给定随机变量 X 的情况下，随机变量 Y 的条件熵(conditional entropy)由式(2-24)定义：

$$\begin{aligned}H(Y\mid X)&=\sum_{x\in X}p(x)H(Y\mid X=x)\\&=\sum_{x\in X}p(x)\Big[-\sum_{y\in Y}p(y\mid x)\log p(y\mid x)\Big]\\&=-\sum_{x\in X}\sum_{y\in Y}p(x,y)\log p(y\mid x)\end{aligned}\tag{2-24}$$

将式(2-23)中的联合概率 $\log p(x,y)$ 展开，可得

$$\begin{aligned}H(X,Y)&=-\sum_{x\in X}\sum_{y\in Y}p(x,y)\log[p(x)p(y\mid x)]\\&=-\sum_{x\in X}\sum_{y\in Y}p(x,y)[\log p(x)+\log p(y\mid x)]\\&=-\sum_{x\in X}\sum_{y\in Y}p(x,y)\log p(x)-\sum_{x\in X}\sum_{y\in Y}p(x,y)\log p(y\mid x)\\&=-\sum_{x\in X}p(x)\log p(x)-\sum_{x\in X}\sum_{y\in Y}p(x,y)\log p(y\mid x)\\&=H(X)+H(Y\mid X)\end{aligned}\tag{2-25}$$

我们称式(2-25)为熵的连锁规则(chain rule for entropy)。推广到一般情况，有

$$H(X_1,X_2,\cdots,X_n)=H(X_1)+H(X_2\mid X_1)+\cdots+H(X_n\mid X_1,\cdots,X_{n-1})$$

例 2-4 假设某一种语言的字符有元音和辅音两类，其中，元音随机变量 $V=\{a,i,u\}$，辅音随机变量 $C=\{p,t,k\}$。如果该语言的所有单词都由辅音-元音(consonant-vowel，C-V)音节序列组成，其联合概率分布 $P(C,V)$ 如表 2-1 所示。

表 2-1 概率分布表

元 音	辅 音		
	p	t	k
a	1/16	3/8	1/16
i	1/16	3/16	0
u	0	3/16	1/16

表 2-1 中字符 p,t,k,a,i,u 的概率值分别为：1/16,3/8,1/16,1/4,1/8 和 1/8。

现在来求联合熵。计算联合熵的方法有多种，以下采用的是连锁规则方法。

$$H(C)=-\sum_{c=p,t,k}p(c)\log p(c)=-2\times\frac{1}{8}\times\log\frac{1}{8}-\frac{3}{4}\times\log\frac{3}{4}$$

$$=\frac{9}{4}-\frac{3}{4}\log 3\approx 1.061(\text{比特})$$

$$\begin{aligned}H(V\mid C)&=\sum_{c=p,t,k}p(C=c)H(V\mid C=c)\\&=\frac{1}{8}H\Big(\frac{1}{2},\frac{1}{2},0\Big)+\frac{3}{4}H\Big(\frac{1}{2},\frac{1}{4},\frac{1}{4}\Big)+\frac{1}{8}H\Big(\frac{1}{2},0,\frac{1}{2}\Big)\\&=\frac{11}{8}=1.375(\text{比特})\end{aligned}$$