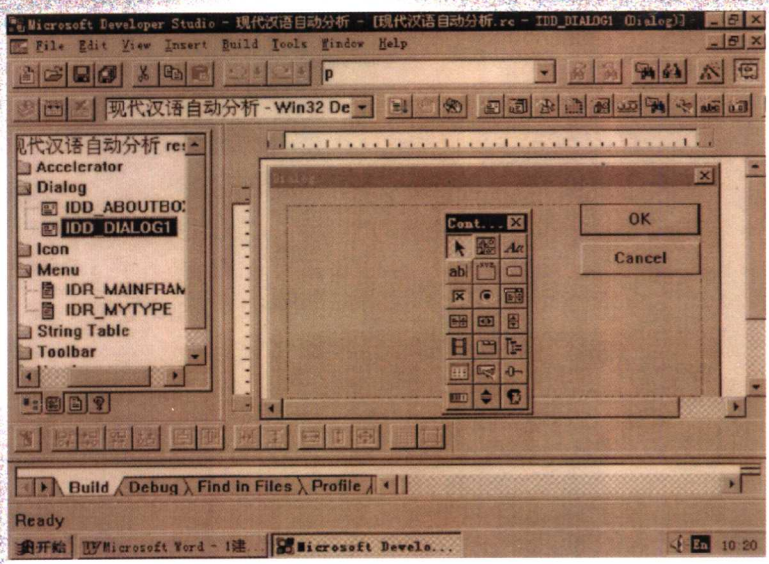
现代汉语自动分析

Visual C++实现

Automatic Analysis Of Contemporary Chinese Using Visual C++

陈小荷 著



北京语言文化大学出版社

现代汉语自动分析

——Visual C++实现

Automatic Analysis of Contemporary Chinese Using Visual C++

北京语言文化大学出版社

（京）新登字 157 号

图书在版编目（CIP）数据

现代汉语自动分析：Visual C＋＋实现/陈小荷著．
－北京：北京语言文化大学出版社，1999
ISBN 7－5619－0800－8
Ⅰ．现…
Ⅱ．陈…
Ⅲ．汉语－言语分析－现代
Ⅳ．H085.2
中国版本图书馆 CIP 数据核字（1999）第 73290 号

责任印制：乔学军
出版发行：北京语言文化大学出版社
（北京海淀区学院路 15 号　邮政编码 100083）
印　　刷：北京北林印刷厂
经　　销：全国新华书店
版　　次：2000 年 3 月第 1 版　2000 年 3 月第 1 次印刷
开　　本：787 毫米×1092 毫米　1/16　印张：12
字　　数：230 千字　印数：0001－3000
书　　号：ISBN 7－5619－0800－8/H・9105
定　　价：19.00 元

致　　谢

首先感谢我的已故导师朱德熙先生。他的严谨的治学态度和不倦追求新知的精神使我终生受益。正是在他的指导下，我得到了语言学研究的基本训练，并且产生了对计算语言学的最初兴趣。

张普先生是汉语信息处理研究的先驱者之一，我有幸在他的带领下从事现代汉语自动分析的研究，经常得到他的鼓励和鞭策，并蒙他慨允为这本小书作序。

当我刚刚步入计算语言学这一新领域时，一些热心的朋友给过我许多启蒙知识。自动分词的最大匹配法是黄建平先生教我的；白拴虎先生慷慨地把他自己编制的源程序拷贝给我，词性标注的统计方法就是从这些源程序中学到的。

感谢本书的责任编辑张旺熹先生，感谢他的辛勤劳动和认真负责的态度。

陈小荷

chenxh@blcn.edu.cn

序

张　普

语言是人类最重要的交际工具。语言交际的模式主要是表达和理解，一方表达，另一方理解。语言交际就是在一定的场合、依据一定的规则、表达或理解一定的意思，句法、语义和语用在交际（或者在表达和理解）中是三位一体的。

从第一台电子计算机诞生到现在半个多世纪过去了，人类已经由发明工具延伸自己的四肢进化到延伸大脑。计算机一开始只是科学家、工程师的计算工具，现在已经逐步进入大众的日常生活，因特网将世界各地的电脑连成一片，人们可以在网上查询或发布信息，可以在网上聊天、漫游、购物甚至在网上生存。“电脑＋网络”正在从延伸人的大脑发展为延伸人类的生存空间，网络社会使人类的交际行为和方式发生重大变化，语言已经从单纯人类最重要的交际工具发展到可以进行人和机器之间的交流。

我在《关于大规模真实文本语料库的几点理论思考》（见《语言文字应用》1999 年第一期）一文的“关于交际”一节中曾经谈到：从表达方和理解方来看，现在交际行为至少有以下四种模式：

A　人表达←——→人理解
B　机器表达←——→人理解
C　人表达←——→机器理解
D　机器表达←——→机器理解

以电脑为发送信息的“一方”或接收信息的“另一方”，研究电脑如何表达人的语言（模式 B）是“自然语言生成”，研究电脑如何理解人的语言（模式 C）就是“自然语言理解”。机器翻译（MT）需要电脑理解一种自然语言，然后转换生成为另一种语言，所以是既包括自然语言理解研究，也包括自然语言生成研究（属于模式 D），还包括语言之间的转换研究。因此，研究“自然语言处理”（包括生成与理解），不可以不研究语言交际，不可以不研究人脑的语言机制和模拟人脑的语言机制。

陈小荷博士的《现代汉语自动分析——Visual C++ 实现》，主要讲的是自动分析，所以是属于对现代汉语这一自然语言的理解（主要是书面语言的理解）的研究，他没有多讲生成，因为他认为“到目前为止，自然语言理解方面研究得比较多，生成方面研究得少。”当然，这只是

相比较而言的，实际上理解和生成是密切相关的，自然语言理解方面的许多技术和方法也可以运用到自然语言生成方面。

陈小荷博士在书中主要介绍了自动分词、词性标注和句法分析等基础的自动分析技术，作为自动分析的入门知识，为了使读者掌握自动分析的方法并能独立地研究问题，他还介绍了一些有关语言统计模型的知识、建立实验平台的步骤和字词检索等实用技术。他为什么不讨论语义、语用方面的自动分析问题？是不是他认为这两个问题不那么重要？不是的。他清楚地知道“自然语言理解，说到底是对句子、篇章的语义和语用的理解”，只不过他希望向读者介绍一些比较成熟的技术和方法。本书副标题“Visual C++ 实现”就是强调所介绍的方法和技术都是可以用程序设计语言来实现的，自然语言的自动分析是看得见、摸得着的。他深入浅出地介绍了汉语理解方面的一些比较成熟的理论、方法和技术，并且用详细的程序代码展示了各个具体的实现步骤。“模型—算法—程序”是他这本书叙述问题的基本线索。

汉语信息处理首先要过汉字这一关，这是书面汉语的特殊性所决定的。70 年代末 80 年代初武汉大学建立现代汉语文学语言资料库时，电脑上还不能直接输入汉字。现在汉字输入已经不成为主要问题，但是理解书面汉语还需要进一步解决自动分词问题。电脑理解口语也需要进行自动分词，口语的自动分词需要有语音识别技术的支持，还可以利用一些韵律特征，当然口语和书面语的自动分词也有一定的共性。书面汉语自动分词看上去似乎是一个文字问题，实际上已经涉及汉语理解。要取得比较理想的分词结果，必须解决歧义切分和未登录词问题，为此电脑需要汉语词汇、构词法、句法和语义等许多方面知识的支持。按照语料库深加工的一般步骤，自动分词之后还需要进一步标注文本中每个词的词性、词义，然后再进行句法分析、语义分析和语用分析，最终达到对句子、段落和篇章的理解。汉语的自动分词、词性标注已经有一定程度的突破，能满足一般性实用的要求，词义标注困难大一些，至于语用的自动分析则至今尚未针对大规模真实文本展开。目前的热点是词汇语义分析和句法语义分析，理论上和方法上都正酝酿着重大突破，实验内容从少量的短句向大规模真实文本扩展。关于汉语信息处理的概况和前沿成果，读者可以从已经发表的专著和论文中读到，陈小荷博士的这本书不以“新”、“全”取胜，而着意求“实”。所谓“实”包含三方面的意思：一是用程序代码去实现，这一点前面已经谈到；二是所介绍的计算模型和算法都是最基本的，没有太深奥难懂的内容；三是实实在在地为他心目中的读者着想，介绍了建立应用程序、文件处理、建立词库、字频统计和字词检索等相关技术。因此本书很值得一读，特别是值得语言学专业一些关心计算语言学和语言信息处理的本科生、研究生以及语言学界的青年语言工作者一读。

陈小荷博士在绪言中提到，本书主要是面向语言学工作者的，这正是他心目中的读者。我和陈小荷博士一样都是中文系出身，本身就是语言学工作者。我们自己以及我们的研究生都有体会——文科背景的人从事计算语言学研究，理解计算模型和算法存在一定的困难。

怎样解决这一困难呢？陈小荷博士的书采取的做法是结合具体的实例作尽可能通俗的讲解。这些实例不是信手拈来随意点缀的，而是按照汉语分析的基本步骤渐次展开并且结合成一个完整的应用程序。我们的研究生都学过 C 语言，有了 C 语言的基础知识，就有可能通过阅读程序代码而达到对书中所涉及的计算模型和算法的彻底理解。这种做法好不好，读者最有权评判。理科背景的读者可能会觉得不够简洁、精练，没学过程序设计的读者需要先过语言关(倒不一定非得是 Visual C++)，恐怕没有两全其美的办法。任何技术性较强的著作都不可能适合各层面的所有读者，特别是那些交叉性学科或综合性边缘学科的技术著作。不同学科背景的读者可以从本书中各取所需，对汉语分析有强烈兴趣和实际需要的读者还可以根据自己的具体情况补充某些其他方面的知识，那就是入门之后的事情了。

计算语言学主要是计算机科学与技术和语言学交叉结合的一门学科。1990 年 2 月发布的国家标准《汉语信息处理词汇 01 部分：基本术语》(GB12200·1 — 90)中对计算语言学的解释是："语言学的一个分支学科。它应用计算机技术来研究和处理语言文字，内容包括：字频和词频统计、语音的识别与合成、机器词典的编纂、机器翻译、自然语言理解、计算机的自然语言接口等。"这是列举式的定义或说明，大致指出了计算语言学的主要应用领域。陈小荷博士也认为："计算语言学是一门应用科学。跟一般语言学相比，计算语言学的特点之一是实践性强。"过去，《辞海》没有收计算语言学条目，机器翻译等内容就是列举在应用语言学条目之下的。

随着语言信息处理的深入，计算语言学研究的广度和深度不断拓展，从学科体系的角度来审视，作为语言学的一个新兴的分支学科，计算语言学的下位学科体系正在逐步清晰化。我个人的看法是，我们现在已经大体上可以从计算语音学、计算词汇学、计算句法学、计算语义学、计算语用学等几个层面描述计算语言学，汉语的计算研究还应包括计算汉字学，人们已经用树形图、拓扑结构和四元组等描述汉字的分解与合成。计算风格学和计算词典学可以分别作为计算语用学和计算词汇学的下位。从学科体系看，陈小荷博士的书涉及到计算词汇学的一些内容，但主要讲的是计算句法学，是对现代汉语句法自动分析的介绍。

计算语言学是在 20 世纪后半叶产生的，它还在发展之中，还有许多的空白领域。随着信息社会的发展，计算语言学的地位会越来越重要，关注计算语言学的人也会越来越多。陈小荷博士虽然是语言学专业出身，但是他已经具备了较好的计算机科学和数学等领域的交叉知识结构，且积年从事语言信息处理研究，这为他研究计算语言学奠定了坚实的基础。他年富力强，勤于钻研，善于思考，正是跨世纪的人才，我相信他在计算语言学领域会有不断的推进，并期待着他的下一部力作。

目　录

绪　言

科幻作品中，机器人常常是神通广大的主角。从外表看，它们有的几乎跟人类一模一样，会说话，会走路，有各种感知能力，甚至还有七情六欲。此外，它们还有高速处理各种信息的能力。在它们面前，自然语言不是障碍，有词典和语法书就够了：不管是自然科学还是人文科学，把书本飞快地翻一遍就能掌握其中的所有知识。人们常常忧虑，是否有那么一天，机器人会替代人类而成为世界的主宰？我没有这种远虑。我感兴趣的是，是否真有一天，我们能够造出一大批这样的机器人来为人类服务？

对于智能机器人，我们当然希望它们感觉敏锐，动作敏捷，力量强大，但是更希望它们具有丰富的知识和聪明的头脑。自然语言是通向人类知识宝库的一把钥匙。作为智能机器人，首先需要掌握这把钥匙，具有人类的语言能力，会听说读写，能方便自如地跟人类交换信息。人工智能就是以建立智能化的、自主的计算机为目标的一门学科，它的主要课题包括：专家系统、问题求解、逻辑与不确定性问题、自然语言处理、机器人学、学习机、视觉与模式识别、与真实世界交互等等。其中，自然语言处理是人工智能的核心课题，包括自然语言的分析理解和综合生成。起初，自然语言处理强调理性主义，用人工智能方法（如状态空间搜索、知识表达和机器学习等）来处理自然语言。在这一阶段，人们集中力量建立各种规则系统，试图通过规则的演算来解决语言的分析和生成问题。但不久人们便发现，自然语言远不是一个精确定义的符号体系，精心构造的规则只能在严格受限的领域内起作用，无法处理大规模真实文本中的种种复杂的语言现象。的确，语言学中的规则几乎都有例外。例如，词类问题上，语法书说汉语的名词可以做主语、宾语、定语、名词性短语里的中心语，不能做状语，等等。但是“咱们电话联系吧”、“现在广播找人”就是名词做状语（孙德金，1996）；“会见时”、“晚饭时”的“时”一般认为是名词，但它从来不做主语、宾语、定语。短语层面上，通常认为组合式述宾结构必须加上“的”才能做定语，但我们常说“落实知识分子政策问题”、“建设有中国特色社会主义理论”，显然就是这一规则的例外。[①]

计算语言学是计算机科学和语言学相结合的一门学科。它也是研究自然语言的分析和生成，跟自然语言处理有相似的任务。自然语言处理属于人工智能科学，在总目标上是为建立智能计算机服务的，其学科性质更靠近计算机科学。顾名思义，计算语言学属于语言学，在总目标上是为研究人类语言的一般规律服务的，其学科性质更靠近语言学。诚然，计算语言学是要用计算机来研究语言的，但是不能说凡是用计算机来研究语言就是计算语言学。例如，仅仅用计算机检索例句写了一篇语言学论文，就不一定属于计算语言学。计算语言学是通过建立形式化的计算模型来处理自然语言的，例如，隐马尔科夫模型、概率上下文无关

语法就是形式化的计算模型。在计算语言学中，计算模型占有中心的地位，它是用计算机处理语言问题的基本思路。有了计算模型，人们才能研究实现模型的具体算法，编制出实现算法的程序代码。“模型—算法—程序”是本书叙述问题的一条基本线索。

到了80年代后期，计算语言学开始注重对大规模真实文本的处理，出现了语料库语言学，其特点是以经验主义为旗帜，以对大规模真实文本的统计为主要方法。语料库语言学认为，人类的一切语言知识都蕴涵在大规模的语料之中，可以通过统计来发掘这些知识。以前人们凭主观内省而得到的语言知识是不完整的，甚至可能是错误的。语料库语言学在真实文本的词性标注方面打了一场胜仗。以前人们制定了各种语言学规则来标注词性，正确率并不高，而运用统计方法和训练语料（已标注过的语料）来标注词性，正确率高达95%左右。各种语料库如雨后春笋般地涌现，规模也越来越大（黄昌宁，1992，1993；McEnery，et al，1996）。然而，语料统计方法也不是万能的，随着语言信息处理层次的提高，简单的统计模型越来越暴露出其弱点：难以处理长距离依赖的语言现象，难以获取高度概括的知识以顺利地处理训练语料之外的语料。目前，在对大规模真实文本进行自动句法分析这一关键问题上，还没有取得突破性进展。

计算语言学是一门应用科学。跟一般语言学相比，计算语言学的特点之一是实践性强，表现在以下几个方面：第一，它研究的许多问题来自社会生产、生活的实际需要。例如，人与计算机交换语言信息不能限于键盘输入和屏幕显示，于是要研究印刷体和手写体文字的自动识别问题，研究语音识别和语音合成问题；要从庞大的资料库中快速、准确地检索出有用的信息，就需要研究文献检索、自动文摘，自然也就涉及句法、语义分析和话语理解等。第二，它要求每一个研究目标都有现实的可行性。一般语言学所研究的许多问题，在计算语言学看来是很有价值的，但是不一定都能够在目前的计算机上实现。对于那些暂时不能实现的目标，就只能先放一放。第三，它更重视一般语言现象。如前所说，自然语言不是一个精确定义的符号体系，对于任何“规律”，几乎都能找出不少例外。一般语言学重视这些例外，力图从中探索出隐蔽更深的规律。而计算语言学以计算机为工具，在计算机尚不具有人类智慧的情形下，它有时需要暂时“忽视”这些例外，先处理那些大量出现的、反映一般规律的语言现象。实践性特点使得计算语言学的发展获得了强大的社会推动力。但是这绝不意味着计算语言学不重视研究理论或永远忽视例外。如何自动学习，自动获取知识特别是概括程度较高的知识，有没有适合自然语言处理的计算模型和更高效率的算法等等，这些都是计算语言学非常关注的问题。

计算语言学是一门文理交叉的边缘学科，除了语言学和计算机科学之外，还涉及哲学、心理学、逻辑学、数学、信息论等多种学科。当然，语言学和计算机科学是其中的主角，计算语言学的发展，主要有赖于语言学和计算机科学两个领域的专家通力合作。合作需要有共识，共识来源于对对方的基本了解。本书主要面向语言学工作者，介绍计算语言学的一些基本问题和处理这些问题的基本方法和技术，说明计算机目前在自然语言处理方面能做些什么，做得怎样。语言学工作者掌握了这些基本方法和技术之后，就可以自己动手利用计算机来做一些研究。

到目前为止，自然语言理解方面研究得比较多，生成方面研究得少。这里至少有两个原

因:第一,关于理解的研究,难度相对小一些;第二,这方面研究的迫切性也更大一些。本书书名为《现代汉语自动分析》,顾名思义,是关于现代汉语这一自然语言的理解问题的介绍。当然,理解和生成是密切相关的,自然语言理解方面的许多技术和方法可以运用到自然语言生成方面。

现代汉语自动分析,是用计算机来分析现代汉语语料,这里主要是指分析现代汉语书面语料,而且特别强调把大规模的真实文本作为研究对象。词汇分析和句法分析是其基本内容。作为入门知识,本书将介绍自动分词、词性标注和句法分析等基础的自动分析技术。为了使读者掌握自动分析的方法并能独立地研究问题,我们还将介绍有关语言统计模型的知识,建立实验平台的步骤和字词检索等实用技术。熟悉语言学的读者可能会奇怪,为什么不讨论语义、语用方面的自动分析问题?是不是在计算语言学中这两个问题不那么重要?不是的。恰恰相反,这两个问题实在是太重要了。自然语言理解,说到底是对句子、篇章的语义和语用的理解;计算语言学只有把话语的意思和用意计算出来才算真正达到了目的。遗憾的是,在这两个问题上,理论基础都太薄弱,虽然有过一些蓝图式的勾勒和"积木世界"之类的实验,但从总体上来说还很不成熟。前沿的理论探索可以给读者开拓广阔、深邃的思考空间,但如果处理得不好,也可能留下一个"众说纷纭,莫衷一是,姑妄言之,姑妄听之"的印象。本书不打算这样做,而只是介绍一些比较成熟的技术和方法。本书的副标题"Visual C++ 实现",就是强调本书所介绍的方法和技术都是可以用程序设计语言来实现的,自然语言的自动分析是看得见、摸得着的。

程序设计语言是目前计算机所能理解的最高级的语言,要把我们的意图转化为可以执行的机器指令,程序代码中就不能有半点含糊不清的地方。本书提供的代码虽然未必反映出多少程序设计技巧,但都是经过我自己反复调试证明是可以运行的。[②]用自然语言叙述复杂的模型、概念和方法,难免会有一些模糊不清之处,这些代码以及所附的文字注释或许可以在一定程度上弥补这一缺陷。

过去写人工智能程序通常使用 List 或 Prolog,现在看来,这些程序设计语言效率不够高,在人工智能领域之外极少使用,处于编程语言的主流之外。Visual C++ 是一个非常好用的软件开发平台,效率高,功能强大。虽然它比较复杂,但是,我们并不需要把它全部弄懂之后才开始使用。"沧海虽大,我取一瓢饮",这种态度是明智的。当然这并不意味着本书介绍的方法和技术只能用 Visual C++ 来实现。原理是一样的,至于采用何种编程语言,多少还跟各人的习惯和偏好有关。

从科幻作品回到现实世界,我们知道智能机器人离我们还十分遥远。它的动作还十分笨拙,它感知到的信息太具体而琐碎,以致难以抽象出事物的本质特征。虽然有"深蓝"战胜了国际象棋特级大师,但是还没有一台计算机能够真正具有普通儿童的思维能力和语言能力。机器翻译是人们了解得比较多的课题,包括自然语言的理解和生成,虽然并不要求完全的理解,也不是创作意义上的生成,但是计算机目前在这方面的表现仍然难以令人满意。不过,我们也用不着泄气。"千里之行,始于足下",计算语言学的发展,也有赖于读者诸君的努力探索。

注释：

①关于“组合式述宾结构”，参见朱德熙(1982)。组合式述宾结构是跟粘合式述宾结构相对而言的，后者专指单个动词跟单个名词组成的述宾结构。因此，“落实知识分子政策”是组合式述宾结构，但修饰“政策”时可以不加“的”。后一个例子更能说明问题，“有中国特色”修饰“社会主义”不加“的”，“建设……社会主义”修饰“理论”仍然不加“的”。组合式述宾结构的这种用法，看来有逐渐增加的趋势。

②这些代码并未形成商品软件，所以读者可以随意使用，但同时我也不能对读者在机器上使用这些代码所可能产生的问题承担责任。

第1章　自动分析中的统计方法

“基于规则”(rule based)和“基于语料库”(corpus based)是计算语言学论著中经常遇到的两个术语。基于规则的方法,其核心就是根据语言学原理和知识制定一系列共性规则和个性规则,以处理自动分析中所遇到的各种语言现象。另一种意见认为,从大规模真实文本中可以观察到,自然语言远不是一个经过事先精心规划的系统,我们难以用一套规则去准确地预测真实文本中所出现的各种变异,也就是说,这些变异有相当的随机性。因此应当用基于语料库的统计方法来研究自然语言。[①]两种方法之争反映了语言研究中的理性主义和经验主义的对立。有的学者认为应该把两种方法结合起来,并且已经产生了一些有代表性的研究成果。[②]语言学工作者对于基于规则的方法容易理解,倒是有必要多了解一些基于语料库的统计方法,作为我们知识的补充。

面对语言现象的随机性,需要有一种研究随机现象的工具。概率论和统计学都是数学的分支:概率论研究随机现象中有关事件的规律性;统计学研究如何以有效的方式收集、整理和分析受到随机性影响的数据,从而对所考察的问题作出统计推断。这种统计推断是以概率论的理论为基础的,因此可以说概率论是统计学的理论基础,而统计学是概率论的一种应用。要了解语言自动分析中的统计学方法,先要了解概率论的一些基本概念。

1.1 概率

“概率”(probability)是概率论中的一个最基本的概念。这一节我们着重用语言学中的例子来说明概率、条件概率、转移概率和在语言统计中经常用到的贝叶斯公式。

1.1.1 概率

我们常常说某件事“有可能”(或“很可能”、“不可能”、“必然”)发生,这是用模糊的自然语言来描述事件发生的可能性。概率则是对事件发生的可能性大小的定量描述,必然发生的事件的概率为1,不可能发生的事件的概率为0,有可能或很可能发生的事件,其概率在0与1之间,但后者的概率大于前者。对于随机事件,通常需要观察它的发生频率,从频率来估计概率。例如,假定语料规模是100万词次(tokens),其中“的”出现了35000次,我们就可以估计“的”的概率是0.035,表示为:

$$P(W=\text{“的”})=0.035$$

这意味着当我们从语料中任意挑选一个词(word token)恰好为“的”的可能性为3.5%。在不发生误解时,可以简单地写成P(“的”)=0.035。

注意，我们刚才是说“从语料中任意挑选一个词”，不是说从词典中任意挑选一个词条(word type)。假如一部词典有5万个词条，每个词条被选中的可能性是相等的，那么，从中任意挑选一个词条，该词条为“的”的概率将是5万分之一。不过，无论对于word type还是对于word token，概率的含义都是一样的：从100万词次的语料中任意选词时，每个token被选中的概率都是100万分之一；但由于“的”出现了35000次(假定)，因此作为一个word type，“的”被选中的概率则为3.5万/100万，即0.035。

另外，从频率来估计概率，随机试验的次数应该足够大。例如，掷一枚硬币，正面朝上和反面朝上的概率各为0.5；假如只做3次试验，其中2次正面朝上，这时，说正面朝上的概率为2/3显然是很不可靠的。拿语料处理来说，就是语料规模应该相当大，这时才能从比较可靠的词频数据来估计每个词的出现概率。

1.1.2 条件概率和转移概率

条件概率(conditional probability)是指在另一些事件发生的条件下某事件发生的概率。由于增加了新的条件，因此一个事件的概率一般不同于该事件的条件概率。例如，“的”的概率大约是0.035，但是如果已知前一个词是“绿油油”，那么当前词是“的”的条件概率几乎为1；如果已知前一个词是“多少”，那么当前词是“的”的条件概率几乎为0。条件概率一般表示为：P(A|B)，其中A是当前事件，B是作为条件的事件。但我们所举的例子其实还施加了更多的条件(先后邻接顺序)，所以用$P(W_2 = \text{“的”} \mid W_1 = \text{“绿油油”})$这样的形式来表示更为清楚。

语言统计中常常用到转移概率(transitive probability)，指的是从一个状态转到另一个状态的概率，这实际上是一种特殊的条件概率，即规定了先后邻接顺序的条件概率。例如，从“绿油油”转移到“的”的概率为$P(W_2 = \text{“的”} \mid W_1 = \text{“绿油油”})$，可估计为“绿油油的”的出现次数除以“绿油油”的出现次数。类似地，从名词转移到动词的概率是$P(T_2 = \text{动词} \mid T_1 = \text{名词})$，可估计为名词和动词相邻出现的次数除以名词的出现次数。“状态转移”这种概念可能不太容易把握，像上面两个例子，就可以分别理解为：

“前一词是‘绿油油’，后一词是‘的’”这一事件的概率；③

“前一词是名词，后一词是动词”这一事件的概率。

条件概率不限于先后发生的事件。例如，$P(W_i = \text{“编辑”} \mid T_i = \text{名词})$表示在某词为名词的条件下，其词形是“编辑”的概率，当前事件和作为条件的事件是同时发生的。这种情况下就不应看作是转移概率。求这个条件概率，我们可以用语料库中名词“编辑”的出现次数除以任意名词的出现次数。而$P(T_i = \text{名词} \mid W_i = \text{“编辑”})$表示在某词词形为“编辑”的条件下，其词性为名词的概率，这时可以用语料库中名词“编辑”的出现次数除以任何词性的“编辑”的出现次数。显然，由于名词的出现次数不等于“编辑”的出现次数，因此这两个条件概率是不同的。

概括地说，如果直接从频率来估计概率，那么条件概率可以用下面的公式来求：

$$P(A \mid B) = \frac{N(AB)}{N(B)}$$

即:用事件 AB 的发生的次数,除以事件 B 发生的次数。

如果已知事件 AB、B 的概率,可以用下面的公式来求条件概率:

$$P(A \mid B) = \frac{P(AB)}{P(B)}$$

从条件概率也可得到 AB 的概率,即乘法公式:

$$P(AB) = P(A \mid B)P(B)$$

1.1.3 贝叶斯公式

先看一个具体的例子:④

无线电通讯中,由于随机干扰,当发出信号为“.”时,收到信号为“.”、“不清”和“-”的条件概率分别为 0.7、0.2、0.1。当发出信号为“-”时,收到信号为“-”、“不清”和“.”的条件概率分别为 0.9、0.1、0。如果发报过程中“.”和“-”出现的概率分别是 0.6 和 0.4,当收到信号“不清”时,原发出的信号是什么?试加以推测。

我们用 A_1 表示原发出信号为“.”,A_2 表示原发出信号为“-”,B 表示收到信号“不清”。根据题意,需要分别求出 $P(A_1|B)$ 和 $P(A_2|B)$ 并加以比较。

根据条件概率的定义,有 $P(A_1 \mid B) = \frac{P(A_1B)}{P(B)}$ 和 $P(A_2 \mid B) = \frac{P(A_2B)}{P(B)}$,题目没有直接告诉我们 P(B)。但是可以看出,$A_1B$(原发出信号为“.”并且收到信号为“不清”)和 A_2B(原发出信号为“-”并且收到信号为“不清”)这两个事件是互不相容的,而且加起来就是 B(收到信号“不清”),即:$A_1B \cup A_2B = B$,$A_1B \cap A_2B = \phi$。因此,根据概率的可加性和乘法公式,我们有:

$$\begin{aligned} P(B) &= P(A_1B) + P(A_2B) = P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2) \\ &= 0.6 \times 0.2 + 0.04 \times 0.1 = 0.12 + 0.04 = 0.16 \end{aligned}$$

因此,$P(A_1 \mid B) = \frac{P(A_1B)}{P(B)} = \frac{P(A_1)P(B \mid A_1)}{P(B)} = \frac{0.12}{0.16} = 0.75$,

$P(A_2 \mid B) = \frac{P(A_2B)}{P(B)} = \frac{P(A_2)P(B \mid A_2)}{P(B)} = \frac{0.04}{0.16} = 0.25$

经比较可知,原发出信号为“.”的概率,是原发出信号为“-”的概率的三倍。如果只需知道哪个更可能,并不需要知道条件概率的具体的值,就可以只比较 $P(A_1)P(B|A_1)$ 和 $P(A_2)P(B|A_2)$。

从这个例子加以推广,可以得到全概率公式:如果事件 $A_1, A_2, \ldots, A_n$ 两两互不相容,$P(A_i) > 0 (i = 1,2,\ldots,n)$,并且 $B \subseteq \bigcup_{i=1}^{n} A_i$,那么,$P(B) = \sum_{i=1}^{n} P(A_i)P(B \mid A_i)$。

由乘法公式和全概率公式,可以得到著名的贝叶斯公式(Bayes' law):如果事件 $A_1, A_2, \ldots, A_n$ 两两互不相容,$P(A_i) > 0 (i = 1,2,\ldots,n)$,并且 $B \subseteq \bigcup_{i=1}^{n} A_i$,$P(B) > 0$,那么,