计 算 语 言 学

刘 颖　编著

清 华 大 学 出 版 社

(京)新登字 158 号

内 容 简 介

计算语言学是一门涉及语言学、计算机科学和数学等多门学科交叉的学科,覆盖面很广,本书侧重最经典的工作,阐述计算语言学的基本理论和方法。主要介绍现代句法理论和语义理论,词法、句法和语义阶段重要的分析算法及语料库和统计语言学。本书结构完整,层次分明,条理清楚。既便于教学,又便于自学。

本书可作为中文、外语、计算机等专业高年级本科生和研究生教材,也可供从事自然语言处理或信息处理的研究者参考。

图书在版编目(CIP)数据

计算语言学/ 刘颖编著 .—北京:清华大学出版社,2002
ISBN 7-302-05788-5

Ⅰ .计… Ⅱ .刘… Ⅲ 数理语言学 Ⅳ .H087

中国版本图书馆 CIP 数据核字(2002)第 062478 号

出 版 者: 清华大学出版社(北京清华大学学研大厦,邮编 100084)
http:// www .tup .tsinghua .edu .cn
责任编辑: 马庆洲
印 刷 者: 清华大学印刷厂
发 行 者: 新华书店总店北京发行所
开 本: 787×1092 1/ 16 印张: 12 .5 字数: 281 千字
版 次: 2002 年 10 月第 1 版 2003 年 8 月第 2 次印刷
书 号: ISBN 7-302-05788-5/ H・428
印 数: 3001～ 4000
定 价: 22 .00 元

前　言

计算语言学,也称自然语言处理或自然语言理解,它是研究如何利用计算机来理解和生成自然语言。计算语言学是植根于计算机科学、语言学和数学等多学科沃土而成长起来的一门新兴学科。一般情况下,处理自然语言不仅要有语言学方面的知识,而且还要有数学和计算机科学方面的知识。因此,计算语言学就成为一门介于语言学、数学和计算机科学之间的边缘性交叉学科。

本书第 1 章主要介绍计算语言学与计算机科学、数学和语言学学科之间的关系,并介绍了计算语言学的基本方法、主要内容以及发展过程。第 2 章主要介绍汉语的切词、切词歧义以及如何消歧,并介绍英语的形态分析及主要分析算法。第 3 章主要介绍词性标注的 4 种方法——规则方法、统计方法、规则与统计结合方法,以及基于转换的错误驱动方法。重点介绍用隐马尔可夫模型(Hidden Markov Model,简称 HMM)进行词性标注的统计处理过程。第 4 章主要介绍乔姆斯基(Chomsky)的 4 种类型的文法和 4 种类型的自动机。文法和自动机是刻画语言的有效手段,文法用来生成语言中的句子,自动机用来识别语言的句子,就描述一种语言而言,两者是统一的。前者属于形式语法理论,后者属于自动机理论。第 5 章主要介绍自 20 世纪 50 年代发展起来的、用于自然语言处理的一些重要的句法理论,主要有基于类的语法理论和基于词的语法理论,基于类的语法理论有:转换生成语法、树连接语法、词汇功能语法、功能合一语法、广义短语结构语法和中心词驱动的短语结构语法。基于词的语法理论包括:范畴语法、依存语法和链语法等。乔姆斯基提出的短语结构语法分析能力不高,分析时难以区分大量的不合语法的句子,生成能力过强。后来乔姆斯基提出了转换生成语法来克服短语结构语法的这些局限性,但转换生成语法本身也有局限性,它的生成能力过强,于是,乔姆斯基提出管辖约束理论来限制转换生成语法过强的生成能力。然而,由于转换生成语法通常要涉及到若干个句子之间的关系,在机器翻译和自然语言处理中使用起来很不方便,不如短语结构语法那样,就一个句子来分析一个句子,它的成分结构是单一的,非常便于进行机器翻译的语法分析和自然语言处理。计算语言学的学者们抛弃了转换生成语法,又转向短语结构语法,于是 20 世纪 80 年代以来出现了各种增强的短语结构语法。例如,词汇功能语法、功能合一语法、广义短语结构语法、中心词驱动的短语结构语法等等,这些语法都采用了复杂特征结构来改进短语结构语法单一的特征,采用合一运算来改进传统的集合运算,从而有效地克服了短语结构语法的缺点,保持了短语结构语法的优点。基于词的语法与基于类的语法不同,把语言知识主要都记录在词典中。第 6 章主要介绍了用于自然语言分析的扩充转移网络、厄尔利(J .Earley)分析算法、富田胜(Tomita)分析算法和线图(Chart)分析算法。基于扩充转移网络的句法分析的优点在于所定义的操作接近人在理解语言时所采用的操作,缺点

是随着结点的增多,计算的复杂性就会急剧地增长,修改时非常困难。富田胜分析算法、线图分析算法等都可以运用复杂特征集和合一运算机制对短语结构语法进行分析。富田胜分析算法改进了 LR 分析算法,是一种高效的自然语言分析方法。线图分析算法采用了线图来记录分析结果,线图可以表示互不相连的树,可以表示歧义。第 7 章主要介绍了用于自然语言处理的一些语义理论以及如何运用这些理论。第 8 章介绍了语料库语言学的一些基本概念和基本模型,同时介绍了语料库以及语料库对齐技术。第 9 章系统地介绍了机器翻译的原理、方法、困难以及评价。

本书可作为中文、外语、计算机等专业高年级的本科教材,教授学时可为 32 至 64 学时。教师可根据学时,安排上机。比如:词法分析、词性标注和句法分析等。如果学生掌握了基本理论和算法,同时上机实现了一些重要算法,则对学生掌握本门课程和掌握计算机处理自然语言打下坚实基础。

本书在写作时尽量做到通俗易懂,所有的算法都举例进行了详细说明,并列出了计算机处理自然语言的详细过程。本书的读者如果具有一定的计算机科学方面的知识(如离散数学、数据结构等),则能更好地理解本书的所有内容。

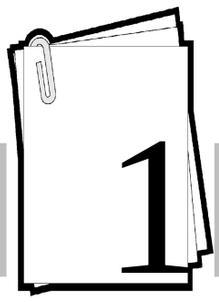
本书的写作参考了许多学者的论文和著作,本书能够出版与他们所作的工作紧密相关,谨向他们表示衷心感谢。

由于本人水平和时间限制,本书难免存在疏漏和不足之处。欢迎各位读者批评指正。

刘 颖

2002 年 2 月 20 日

目　　录



1 计算语言学简介

1.1 计算语言学

1.1.1 计算语言学概念

计算语言学,也称自然语言处理或自然语言理解,它是研究如何利用计算机来理解和生成自然语言。例如,用计算机对自然语言的形、音、义等信息进行处理,即对字、词、句、篇章的输入、输出、识别、分析、理解、生成等的操作和加工。

自然语言处理这个术语主要用于说明方法,计算语言学这个术语主要用于说明理论。

计算机对自然语言的研究和处理,一般应经过如下 4 个过程:

1.从语言学角度提出自然语言处理的问题和理论(linguistic problem)。

2.把需要研究的问题在语言学上加以形式化(linguistic formalism),使之能以一定的数学形式,严密而规整地表示出来。

3.把这种严密而规整的数学形式表示为算法,使之在计算机上形式化(computational formalism)。

4.根据算法编写计算机程序,使之在计算机上加以实现(computer implementation)。

因此,为了处理自然语言,不仅要有语言学方面的知识,而且,还要有数学和计算机科学方面的知识,这样计算语言学就成为了一门介于语言学、数学和计算机科学之间的边缘性交叉学科,它同时涉及到文科、理科和工科三大领域(冯志伟 1996)。

第一、第二个过程属于计算语言学的理论部分,第三和第四个过程属于计算语言学的方法部分,也称为自然语言的计算机处理。

1.1.2 计算语言学与计算机科学

计算语言学一方面要求把计算机科学处理问题的一些基本思想、基本方法引到语言学研究中来,从新的角度观察语言学,建立和传统语言学不同的语言学理论,这些语言学理论要精确地描述和解释语言的结构、现象和规律,建立语言的严谨的可计算的形式化模型。另一方面,计算机科学提供相应的算法,在这些模型的基础上,进行计算、推导、分析、转换、生成等,从实现角度来对模型进行检验。因此,计算语言学家必须了解哪些问题是

计算机可以解决的,哪些是不可以解决的;还必须了解如何使计算机按照他所设计的算法去解决问题。因此,计算语言学的理论研究与成果应用要以计算机科学和技术为基础;计算语言学也应该和必然推动计算机科学的深入与普及(侯敏 1999) (姚亚平 1999)。

1.1.3 计算语言学与语言学的区别

语言学是研究语言现象及其规律的科学。计算语言学是语言学的一个分支,是运用计算机的手段研究语言现象和规律的。传统语言学和计算语言学的区别主要在于:

1.传统语言学是一门经验学科,而计算语言学既是一门理论学科,又是一门实验科学(侯敏 1999)。

2.计算语言学要面对整个自然语言现象,因此,它必须研究计算机处理语言的带有普遍性和总体性的一般问题;而传统语言学家喜欢深入研究某一特殊的语言现象,更加重视研究语言中的某个特殊问题(冯志伟 2001)。

3.传统语言学主要是描述性的,而计算语言学要求的语言学理论必须具有可操作性,要想操作,就首先要把一个句子中所有的信息,包括词法的、句法的、语义的都形式化,变成机器可以识别的规则,这样它才能一步步操作,最后达到理解这个句子的目的。计算语言学最根本、最关键的方法就是要指出各种语言形式出现和变换的条件。只有指出了条件,计算机才可能根据有关的条件,执行相应的动作,从而使整个系统成为一个动态地执行的过程。不论哪一种计算机,在执行有关程序时,总免不了给它指出条件,有了条件,并且让计算机知道究竟是什么样的条件,计算机才能执行相应的动作,这就是可操作性,而计算机的任何操作都可以归结为一个公式:“条件→动作”偶对。要使自然语言的语法规则成为可供计算机执行的形式,就必须指出各种语法现象出现的条件。比如,$N+N$,传统语言学是这样描述的:在汉语中可以构成定中关系、主谓关系、并列关系、复指关系。如:学校图书馆、工人农民、鲁迅先生、今天星期六分别构成定中关系、并列关系、复指关系、主谓关系。可是对计算机,这条规则就不行了,它形式化和具体化的程度都不够,必须指出在什么条件下“$N+N$”是定中关系,什么条件下“$N+N$”是并列关系,什么条件下“$N+N$”是复指关系,什么条件下“$N+N$”是主谓关系。如:$N+N$,当前面的 N 是专有名词,后面的名词是称呼名词时,是复指关系(冯志伟 1996)。

4.计算语言学的研究成果必须要通过自然语言处理来检验,计算语言学的理论要说得通,更要重视理论的实用性。而传统语言学则要求讲道理,重视逻辑的完美性(冯志伟 2001)。

5.计算语言学研究语言时必须先分析后理解,理解是分析的结果。而传统语言学是先理解后分析,理解是分析的必要前提(冯志伟 2001)。

1.1.4 计算语言学与数理语言学

计算语言学相当于应用数理语言学,是数理语言学的一个分支。数理语言学是运用数学思想和数学方法来研究语言现象的一门新兴的语言学科。数理语言学的出现,使得作为一门人文科学的语言学与现代数学、计算机科学、信息论以及人工智能等发生了密切

的联系,逐渐走上了现代化的道路。机器翻译、情报检索、自然语言理解等语言自动处理技术的出现,要求　确地描述和解释语言的结构,建立语言的数学模型,并用数学方法来研究语言的语法和语义结构(冯志伟 1985)。

数理语言学主要研究: 代数语言学、统计语言学、应用数理语言学。

代数语言学: 采用集合论、数理逻辑、算法理论、模糊数学、图论、格论等离散的、代数的方法来研究语言。

统计语言学: 采用概率论、数理统计和信息论等统计数学的方法来研究交际过程中语言成分使用的统计规律。

应用数理语言学: 把代数语言学和统计语言学应用于机器翻译、人机对话以及情报自动检索的技巧与方法,就是应用数理语言学的研究内容。

代数语言学是基于规则的,它代表着数理语言学中的理性主义方法;统计语言学是基于统计的,它代表着数理语言学中的经验主义研究方法;而在数理语言学的实际应用中,则应该把理性主义方法和经验主义研究方法结合起来。

1.1.5　计算语言学与自然语言

计算语言学研究和处理的对象是自然语言,而不是人工语言或其他的形式语言。

世界上的语言,绝大多数是自然语言。自然语言是人类发展过程当中自然产生、约定俗成的用于人类社会交际的语言,如英语、汉语、日语等。自然语言中有少数是通过人为的力量创造或规定下来的语言,比如世界语。

形式语言是人们有意识地通过形式化的定义所规定的语言,典型的形式语言包括程序设计语言(比如 C 语言)和符号逻辑语言(比如一阶逻辑语言)。形式语言是具有严格结构的符号系统,适合于计算机等具有符号化信息处理能力的计算和通信装置使用。

在计算机软件中,早已设计了许多人工语言,如 Basic、Pascal、Cobol、lisp、C、Java 等程序设计语言,这些人工语言都遵循着形式语言的规律和法则。对这些人工语言的词法、句法、语义的分析和生成,技术已比较成熟,发展成为一门新的学科“编译原理”,但自然语言比人工语言要复杂得多,因而用计算机处理起来也就困难得多。

自然语言与人工语言的区别,主要表现在下面 4 个方面(冯志伟 2001):

1. 自然语言中充满着歧义,而人工语言中的歧义则是可以控制的。

2. 自然语言的结构复杂多样,而人工语言的结构则相对简单。

3. 自然语言的语义表达千变万化,迄今还没有一种简单而通用的途径来描述它,而人工语言的语义则可以由人来直接定义。

4. 自然语言的结构和语义之间有着错综复杂的联系,一般不存在一一对应的同构关系;而人工语言则常常可以把结构和语义分别进行处理,人工语言的结构和语义之间有着整齐的一一对应的同构关系。

由于自然语言的这些独特性质,使得自然语言处理成为人工智能的一大难题。

1.2 计算语言学主要研究的内容

按照语言学上一般的分析,语言可分为如下的一些层次:语音、词汇、语法、语义、语用。计算机在语言学上各个层次的应用便形成了计算语音学、计算词汇学、计算语法学、计算语义学、计算语用学等,它们都是计算语言学的分支学科(冯志伟 1999),也是计算语言学主要研究的内容。

计算语音学:研究如何利用计算机对语音信息进行处理,实现语言的自动合成与识别。

计算词汇学:研究如何用计算机处理自然语言的词汇、建立语言词汇库、术语数据库等机器可读词典。对于印欧语言主要研究形态分析。计算机形态分析指如何用计算机将一个词分析为词素的组合,从而导出该词的意义。例如,将词 friendly 分析为名词 friend 和后缀 ly 的组合,计算机可以得知 friendly 是由 friend 导出的形容词。一个自动词法分析方案可包括一部词干词典和一套描述词形变化和构词的规则系统,这样,在分析时,给出词干,计算机就可以自动地列举出它的所有的变化形态,而给出一个变化形式,计算机就可以自动地把它切分为词干、词缀和词尾。对于汉语,主要研究汉语的自动分词。因为汉语中单词与单词之间没有空格,因此必须首先进行分词(罗振声,袁毓林 1996)。

计算语法学:研究如何用计算机来分析自然语言的语法。根据语言学理论所提供的关于语法结构的规则,推导出一个语句的所有可能的语法结构。这种研究在计算机中叫做"剖析(parsing)"。目前,剖析技术比较成熟,有自顶向下分析法、厄尔利(Earley)分析算法、富田胜分析算法、线图分析算法、马库斯(M.Marcus)提出的确定性分析算法等。语言学理论有乔姆斯基(Chomsky)提出的短语结构语法理论、转换生成语法和管辖与约束理论,广义短语结构语法,词汇功能语法,功能合一语法,基于中心词驱动的短语结构语法等。

计算语义学:如何利用计算机来分析自然语言的语义,如威尔克斯(Y.Wilks)的优选语义学,菲尔摩(C.J.Fillmore)的格语法,商克(R.Schank)的概念依存理论,西蒙斯(R.F.Simmons)的语义网络理论,蒙塔格(R.Montague)的蒙塔格语法等,都是计算语义学的重要研究成果。另外计算语言学还研究计算机语言学习和语料库语言学等。

计算机语言学习:以上每个问题,都需要应用大量的语言知识。解决某一问题需要哪些知识,如果都需要由人工决定,并形式化地表达这些知识的话,则需要大量的人工及专家知识。计算机语言学习的目的就是通过机器学习,自动地获得语言处理所需要的专门知识,并将这些知识形式化地表达出来。

语料库语言学:语料库语言学用概率统计来研究语言,它的手段是语料库。语料库语言学研究的基础是机器可读的大容量语料库和一种易于实现的统计处理模型,两者相辅相成,缺一不可。语料库语言学的基本任务是研究机器可读的自然语言文本的采集、存储、检索、统计等,以及语料库方法在语言定量分析、词典编纂、作品风格分析、自然语言理解和机器翻译等领域中的应用。

1.3 计算语言学理论的主要用途

1.机器翻译 (machine translation)

机器翻译: 将一种语言自动翻译成另外一种语言。

2.语音自动识别、语音自动生成

语音自动识别: 用计算机对语音作出明确无误的辨认。语音识别在铁路、民用航空可用来建立人机对话的无人管理问讯处,在民航部门用来作“声纹”刑事侦破系统,还用于口语翻译的语音识别。

语音自动合成: 就是用计算机技术或数字信号处理技术来重新产生人类的语言。目前一些系统已达到实用化。

3.自动文摘

自动文摘: 用计算机将反映原文档中心的某方面内容自动地抽取出来,并用同于或不同于原文的句子表示出来。目前,网上文本信息大量涌现,人们越来越关心如何能快捷、准确、全面地获取这些信息,而浏览全文的摘要是一条有效途径。

4.自动校对

自动校对: 目前出版业(尤其是电子出版)发展非常迅速,其中校对环节的工作量也大大增加了。而校对的方式还停留在人工校对的方式上,这与出版业其它环节的逐步自动化形成了鲜明的对照。因而对于自动校对提出了要求。如果能由计算机来完成其全部或部分工作,则会减轻繁重的校对工作,减少大量的劳力。

5.自然语言理解(understanding natural language)

自然语言理解: 又叫人机对话(man-machine dialogue),研究如何让计算机理解和运用人类的自然语言,使得计算机懂得自然语言的含义,并对人给计算机提出的问题,通过对话的方式,用自然语言进行回答。自然语言理解系统可以用作专家系统、知识工程、情报检索、办公室自动化的自然语言人机接口,有很大的实用价值。

6.情报自动检索(automatic information retrieval)

情报自动检索: 利用计算机从众多的文献资料中找出符合特定需要的文献或情报的过程。又称信息自动检索或信息检索。目前已经成为科技情报工作现代化的核心内容。

7.术语数据库

术语数据库: 存储在计算机中的记录概念和术语的自动化电子词典。术语数据库研制的目的是可以满足翻译人员的需要,为了标准化的需要,满足图书出版商的需要等。

8.计算机辅助教学

计算机辅助教学: 计算机要按着人们事先安排好的语言教学计划进行课堂教学和辅助课外操练。最早开始计算机辅助教学研究的是美国 IBM 公司的沃斯顿研究中心,在 1958 年,利用一台 IBM650 计算机连接一台电传打字机来教小学生学习二进制算术。目前,已开发了数学、工程、医学、商业、外语、哲学、音乐、计算机等课程的辅助教学系统。

9.电子词典

电子词典: 机器可读的,形式化的知识库,而不是数据库,应该是大容量的、高水平的。

电子词典是语言知识的重要资源。电子词典的构造是一个民族语言现代化的基本建设。

10 .汉字自动识别

汉字自动识别: 分为印刷体汉字识别和手写体汉字识别。如果汉字自动识别准确率高,必然会大大提高中文信息计算机处理的效率。

11 .文献自动分类(information classification) 等。

1.4 计算语言学研究的基本方法

1.4.1 理性主义和经验主义

理性主义研究方法认为,人的很大一部分的语言知识是生来俱有,由遗传决定的。理性主义研究方法从 20 世纪 60 年代到 80 年代中期主宰了计算语言学。与理性主义相反的是经验主义的研究方法。它认为人并不是生来俱有一套有关语言的原则和处理方法,人的知识是通过感官输入,经过一些简单的联想(association)与通用化(generalization)的操作而得到的。经验主义研究方法从 20 世纪 20 年代到 50 年代主宰了计算语言学,并在 80 年代中期后重新受到了重视(翁富良,王野翊 1998)。二者具体区别:

1 .理性主义主要研究人的语言知识结构(语言能力 language competence)。实际的语言数据只提供了这种内在知识的间接证据。而经验主义的研究对象直接是这些实际的语言数据。

2 .理性主义方法通常是基于乔姆斯基的语言原则(principles)的。它通过语言所必须遵守的一系列原则来描述语言,由此当一个语句遵守了语言原则,则是正确的,而违反了语言原则便是错误的。经验主义方法是基于先农(shannon)的信息论,它将语言事件赋予概率。由此可以说一个语句是常见的还是罕见的。

3 .理性主义方法通常是通过对一些特殊的语句或语言现象的研究来得到对人的语言能力的认识,而这些语句和语言现象在语言的实际应用中并不一定是常见的。而经验主义的方法则偏重于对语料库中人们所实际使用的普通语句的统计表达。

1.4.2 计算语言学研究方法

1 .理性主义研究方法——符号处理系统

符号处理系统是认知心理学家作为人的认知模型而提出的。它在计算语言学得到广泛的应用。在一个符号处理系统中,符号是表示概念和意义的基本单位。

符号必须具有如下的特点: (1)符号具有任意性。一个符号的形状和其所表示的意义没有关系。(2)符号能够按照某些规则递归地构成符号系统,由此产生的组合符号表达式可以表示复杂的意义。

在计算语言学中,用于自然语言处理的符号处理系统通常根据一套规则或程序,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。在一个典型的自然语言处理系统中,由语法分析器按照人所设计的自然语言的语法规则,将输入语句分析为语 结构(比如树结构),再根据一套规则将语法结构映射到语义符号结构(如

逻辑表达、语义网络或中间语言)。自然语言符号处理系统中的规则通常是先验的,也就是由人设计好了以后赋予机器的,这是一种典型的理性主义方法。

2.经验主义研究方法——基于语料库的计算语言学研究

统计学方法:统计学方法首先为要解决的语言处理问题建立统计模型,并且训练语料库来估计统计模型中的参数,然后把参数值应用到模型中处理语言问题。以词性标注为例,首先建立统计模型(比如隐马尔可夫模型),为了能够实现统计,一般在计算时要对隐马尔可夫模型进行简化,然后统计训练语料库中模型中的每个参数值,最后把参数值应用到模型中确定出每个词的词性。统计方法广泛应用于词性标注、语法分析、歧义化解、机器翻译、语音识别等语言处理领域。

基于转换的错误驱动学习:这种方法通过学习得到考虑上下文的规则集,然后计算语料库中应用每个规则时正确和错误的事例个数,再按提高语料库标注的正确率高低来排列规则,从而发现最可能的规则。这个技巧已经用在各个领域,其中包括词性标注、建造短语结构树、文本组块。在每次迭代学习时,把正确的标注语料与已标注语料库进行比较学习,得到一个规则集,统计规则集中每个规则标注这个语料库提高标注的正确率,得到一个按正确率高低排列的有序规则列,选择出正确率最高的规则,用这个规则去标注语料库。再进行迭代学习。直到不能发现新的并能提高语料库标注正确率的规则,学习才停止。这个过程就是基于转换的错误驱动学习过程。

神经网络学习方法:主要应用有动词的形态变换、语法分析、机器翻译等。然而由于自然语言处理中所需要的高层次的知识很难用神经网络中的分布式表达来表示,所以这些系统不具备实用效能。

3.理性主义研究方法与经验主义研究方法的结合

基于规则的理性主义研究方法,其优点是可以不必事先建立一个语料库。研究者只要将语言学家研究的大量现成的语言学知识形式化。这种方法具有较强的概括性,容易推广到一些尚未涉及的领域。但是,基于规则的方法所描述的语言知识颗粒太大,难以处理复杂的、不规则的信息。而且当规则数目增加时,很难保证一致性和健壮性。

基于统计的经验主义研究方法则需事先建立一个语料库,其全部知识都是由计算机通过统计处理大规模真实文本而自动获取的,具有很好的一致性和健壮性。

把基于规则方法和统计方法结合起来,一方面,如果把统计方法作为获取知识的主要途径,依据语言学家的语言学知识对所获取的知识加以取舍,并增加一些统计方法没有得到的、而经过语言学家证明是行之有效的正确的语言规则。另一方面,由于由统计方法获取的语言知识来自大规模真实文本,可以覆盖几乎所有语言现象。这样,便能克服语言学家总结语言规则的片面性和主观性,并使他们集中精力研究那些最常见的、在统计意义上最重要的语言现象。

1.5 计算语言学的发展历程

计算语言学的发展分为萌芽期、发展期和繁荣期(冯志伟 2001)。

1.萌芽期

计算语言学的研究起始于机器翻译。1946 年,美国宾夕法尼亚大学的埃克特

(J .P .Eckert)和莫希莱(J .W .Mauchly)设计的第一台计算机 ENIAC 问世,引起世界震惊。同一年,英国的布斯(A .D .Booth)、美国的韦弗(W .Weaver)就开始了机器翻译的研究。1954 年,美国乔治敦大学在国际商用机器公司(IBM)的协同下,用 IBM-701 计算机进行了世界上第一次机器翻译试验,首次用计算机把俄语译成了英语,并取得初步成功。这是计算机最早的在非数值处理方面的应用,一时引起了人们的注意,许多人认为这是一个大有可为的计算机应用领域。美国的华盛顿大学、麻省理工学院、哈佛大学、密执安大学、宾夕法尼亚大学、美空军国家技术处,苏联语言研究所、苏联科学情报研究所、列宁格勒大学,日本京都大学、九州大学以及意大利、比利时、英国、捷克、匈牙利、德国等国都掀起了一股研究热潮。但是机器翻译的问题很复杂,而早期的机器翻译系统都把机器翻译的过程与解读密码的过程相类比,试图通过查询词典的方法来实现词对词的机器翻译,因而译文的可读性很差,难于付诸实用。1964 年,美国科学院专门成立了一个“自动语言处理咨询委员会”(简称 ALPAC 委员会),调查机器翻译的情况。1966 年,ALPAC 委员会写了一个报告——ALPAC 报告。报告中说:“在目前给机器翻译以大力支持还没有多少理由。”报告出来以后,很多资助都停止了。机器翻译的研究出现了空前萧条的局面。所以造成这样的后果,一方面是机器设备、条件上的原因。另一方面一些有识之士清醒地认识到从计算机处理自然语言的角度研究语言的重要性,在 ALPAC 报告中首次出现了“计算语言学”这个术语,计算语言学就是自然语言计算机处理的基本理论和方法的总称。从此进入了计算语言学的萌芽期。

2 .发展期

ALPAC 报告后,计算语言学研究逐渐转向自然语言理解。自然语言理解系统分为第一代系统和第二代系统两个阶段。第一代系统建立在对词类和词序分析的基础上,分析中经常使用统计方法;第二代系统则开始引进语义甚至语用和语境的因素,几乎完全抛开统计技术。第一代系统主要有:①特殊格式系统,比如,1963 年,林赛(R .Lindsay)设计的 SAD-SAM 系统,采用特定格式进行亲属关系方面的人机对话。②以文本为基础的系统,比如,1966 年,西蒙斯、布格尔(J . F . Burger)和龙格(R . E . Long)设计的 PROTOSYNTHEX-I 系统。③有限逻辑系统,比如,1968 年拉斐尔(B . Raphael)建立的 SIR 系统,采用模式匹配并进行简单的逻辑推理,识别输入句子的结构。④一般演绎系统,如,1968—1969 年,格林(B .Green)和拉斐尔建立的 QA2 和 QA3 系统,采用谓词演算的方式和格式化数据来进行演绎推理,解答问题。

1970 年以来,出现了第二代自然语言理解系统,这些系统绝大多数是程序演绎系统,大量地进行语义、语境甚至语用的分析。其中比较著名的系统是 LUNAR 系统、SHRDLU 系统、MARGIE 系统、SAM 系统和 PAM 系统。LUNAR 系统是伍兹(W .A . Woods)于 1972 年设计的一个自然语言情报检索系统。SHRDLU 是威诺格拉德(T . Winograd)于 1972 年在美国麻省理工学院建立的一个用自然语言指挥机器人动作的系统。MARGIE 是商克于 1975 年在美国斯坦福人工智能实验室研制的一个自然语言理解的直观模型,系统使用概念依存来进行推理。SAM 系统是埃布尔森(Abelson)于 1975 年在美国耶鲁大学建立的采用“脚本”的办法来理解自然语言写的故事。PAM 是威林斯基(R .Wilensky)于 1978 年在美国耶鲁大学建立的另一个理解故事的系统。