第 1 章　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，是一门以计算为手段对自然语言进行研究和处理的学科。例如，用计算机对自然语言的音、词汇、句法、语义和语用等信息进行处理。

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 5 个过程。

(1) 提出计算机要处理的语言学问题。

(2) 根据语言学理论把需要研究的语言学问题形式化，使之能严谨规范并采用一定的数学描述方式描述出来(对应于自然语言处理的规则方法)。或把需要研究的语言学问题抽象成数学模型(对应于自然语言处理的统计方法)。

(3) 设计计算机算法，使得计算机能自动地处理形式化的语言学问题，或者按着数学模型进行相应的统计和处理。

(4) 根据算法编写计算机程序，运行计算机程序来实现计算机的自然语言处理。

(5) 对计算机处理的结果进行实验分析，得出结论。

因此，为了处理自然语言，不仅要有语言学方面的知识，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交叉学科(冯志伟 1996)。

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和可统计的概率模型。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，进行检索、统计、计算、推导、分析、转换、生成等，从实现角度对模型进行检验。因此，计算语言学家一方面需要研究现有的所有语言学理论，另一方面需要研究现有的数学统计模

型、计算机算法和高级程序设计语言。了解哪些问题可以用现有语言学理论解决,哪些是不可以解决的。哪些可以用抽象的统计模型来解决,哪些不能。还必须了解处理语言学问题适合的算法和编程语言(侯敏 1999;姚亚平 1999)。

1.1.3 计算语言学与语言学

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并对其进行自动处理。传统语言学和计算语言学的区别主要在于以下方面。

(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,一种方法是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。另一种方法是根据大规模语料库中语言单位出现的概率来计算所要处理问题的概率 (冯志伟 1996)。

(4) 计算语言学的理论必须要通过计算机实践来检验,从实验结果中检验计算语言学的理论是否可行。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和处理后理解,理解是分析和处理的结果。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

计算语言学就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数学、计算机科学以及人工智能等发生了密切的联系,使得语言的研究逐渐走上了现代化的道路。句法分析、语义消歧、机器翻译、自动问答等语言自动处理技术的出现,要求精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:(1)代数语言学;(2)统计语言学;(3)应用数理语言学。

代数语言学:采用集合论、数理逻辑、图论、形式文法、自动机等离散的、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、人机对话以及信息检索的技巧与方法,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，既有理性主义方法，也有经验主义研究方法，还有把二者结合起来的研究方法。

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、Java 语言、Perl 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，适合于计算机使用和处理。

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）自然语言在语音、词汇、句法、语义和语用层面都存在歧义。而人工语言中的歧义则可以由人来进行限定。

（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、语篇和语用。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）。

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。计算机形态分析是研究如何将一个词分析为词素的组合，从而导出该词的组成结构和意义。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,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句法。根据语法学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"parsing",目前,parsing 技术比较成熟,有 Earley 分析算法、Tomita 分析算法、Chart 分析算法和 CYK 算法等。用于计算机自动处理的语言学理论有广义短语结构语法,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、依存语法、链语法等。

计算语义学:研究如何利用计算机来分析自然语言的语义。目前,语义分析集中于利用大规模语料库中的上下文来确定一个词在所在句子中的确定含义、对同义词进行辨析或利用词典和上下文来确定词与词之间的语义关系等。计算机处理的语义学理论有 R. Wilks的优选语义学、Fillmore 的格语法、R. C. Shank 的概念依存理论、R. F. Simmons 的语义网络理论和 R. Montague 的蒙塔格语法等。

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,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。

语料库语言学是利用计算机强大的检索、统计和处理语料的能力,从大规模的语料库中检索符合研究问题的实例,对其进行统计。在大量实例和统计数据的基础上,对研究问题进行定性分析,从功能上对其进行语言学解释。利用计算机和语料库可以对语言各个层面的特征(单个特征或多个特征)进行分析和研究。

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,以及语料库方法在词汇、语法、语义、语篇结构、语域变异、语言习得、作家作品风格分析等领域中的应用。

语料库语言学的优势在于:(1)可以利用计算机的强大功能,进行快速、准确的分析;(2)语料库规模大,所包括的语域全面,文本量大,语言信息范围广;(3)既有定量分析,又有定性的功能解释,对语言的描写全面;(4)语料库方法与以往的方法相比能做出更概括和更全面的调查。因此,基于语料库的方法可以扩大以往调查的范围和调查语言的新应用。语料库语言学已经成为语言研究的主流,它正对许多语言研究领域产生越来越大的影响。

1.3 计算语言学理论的主要用途

1.3.1 机器翻译

机器翻译:利用计算机将一种自然语言自动翻译成另外一种自然语言。中英翻译就是利用计算机自动地把汉语翻译成英语。很多大学、科研院所和公司展开了对两种语言或多种语言之间的两两互译。Google、Microsoft 和百度公司都开发了在线多语言机器翻译系统。MSN、社交网络 Facebook 和通信工具 GoogleTalk 都提供了即时翻译任务。在欧美和日本已经开发出 SYSTRAN、TAUM-METEO、METAL 等多个实用的机器翻译系统。

比较有名的在线机器翻译系统是 Systran：http://www.systransoft.com，ReadWorld：http://www.readworld.com。

1.3.2　语音自动识别和自动生成

语音自动识别：用计算机将人的语音自动转换成文本。语音识别是与声学、语音学、语言学、计算机科学、数字信号处理理论、信息论等学科紧密相关的一个多学科交叉的领域。20 世纪 90 年代后，语音识别的研究向实用化的方向发展。IBM 公司推出的 ViaVoice 可以针对大词汇量和非特定人的连续语音进行识别。微软开发的 Speech Application SDK、SUN 公司开发的 JavaSpeechAPI 和 IBM 的 Dutty++等都能识别多个不同国家的语言，比如英语、日语和中文等，语音识别技术比较成熟。语音识别在旅游、铁路、宾馆预订、民用航空可用来建立人机对话的无人管理问讯处。在侦查部门用来作“声纹”刑事侦破系统，还可以用于口语翻译。

语音自动合成：就是用计算机技术或数字信号处理技术把文本转换成人类的语音。语音合成与语音识别经常结合形成人机对话系统。IBM 公司开发的智能词典 2000，能对单词、短语、句子以及段落等准确发音。AT&T 公司开发的真人语音合成系统，它发出的语音让人无法辨出真假。微软公司开发的 SAPI SDK 语音应用工具包，支持多种语言的识别和朗读，包括：汉语、日语和英语。

1.3.3　自动文摘

自动文摘：用计算机将反映原电子文档核心内容的文本自动地抽取出来，生成一篇语意连贯的文本。自动文摘应具有概括性、客观性、可理解性和可读性（俞士汶 1996）。它应以提供文献内容梗概为目的，简明、确切地记述文献重要内容的短文。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。比较著名的自动文摘系统有 MITRE 公司开发的 Mani 和 Bloedorn 系统、南加州大学开发的 Marcu 系统、卡内基-梅隆大学开发的 Goldstein 和 Carbonell 系统、哥伦比亚大学开发的 Mckeown 系统等。目前，关于自动文摘最有影响的会议是文本分析会议（Text Analysis Conference，简称 TAC），包含文档理解会议（Document Understanding Conference）和文本检索会议（Text Retrieval Conference，简称 TREC）。

1.3.4　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。如果校对的方式还停留在人工校对上，则与出版业其他环节的逐步自动化不相匹配。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，因而提出了自动校对。文字处理软件 Word 和 Wordpefect 都嵌入了英文拼写检查功能。ExpertEase 公司推出的 DealProof，Newton 公司推出的 Proofread 是互联网上见到的英文单词拼写检查系统。黑马校对系统、金山校对系统等是已经商品化的中文文本校对系统。

1.3.5 自然语言理解

自然语言处理包括自然语言理解和自然语言生成。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。自然语言理解是人工智能研究中的热点和难点之一。日本研制第 5 代计算机的主要目标之一,就是要使计算机具有理解和运用自然语言的功能。

1.3.6 信息自动检索

信息自动检索:又称信息检索,是从大规模非结构化数据(通常是文本)的集合(一般保存在计算机上)中找出满足用户信息需求的资料的过程(王斌 2010)。随着 Internet 的迅速发展,网络上的信息越来越多,面对浩瀚的信息许多用户手足无措,无法准确地获取自己所需要的信息。针对这种情况有些组织和个人开发出用以查找网络信息的检索工具——搜索引擎。目前世界上最大的搜索引擎是 Google、MSN 和雅虎,MSN 主要是美国商业目录搜索引擎,主要为用户提供教育、新闻、媒体及娱乐信息。中文综合性搜索引擎有:百度、Google、中国搜索联盟、新浪、搜狐、网易、雅虎等,其中百度是目前最具影响力的中文搜索引擎。

广泛用于科学研究和论文检索的著名三大文献检索工具是 SCI、SSCI 和 EI 检索,它们已成为评价科研工作人员、科研机构乃至一个国家的学术研究水平的重要指标,是科学研究领域研究人员查阅学术资源的重要工具。

科学引文索引(Science Citation Index,SCI),创刊于 1963 年,是美国科学信息研究所(http://www.isinet.com)出版的世界性学术期刊文献检索工具。SCI 收录世界上的重要期刊约 3500 种,而网络 SCI 扩展版(SCI Expanded)收录 5900 种。内容涵盖数、理、化、农、林、医、生命科学、天文、地理等自然科学各学科领域。

社会科学引文索引(Social Science Citation Index,简称 SSCI)是美国科学信息研究所建立的综合性社科文献数据库。SSCI 收录全球 50 多个语种的 1700 多种重要社会科学期刊论文,内容涉及社会科学、人类学、考古学、商业、财政、经济、教育、地理、历史、法律、语言、政治等 50 多个学科领域。SSCI 是评价人文及社会科学领域学者学术水平的权威的参考指标之一。

工程索引(Engineering Index,简称 EI)创刊于 1884 年,是美国工程信息公司出版的工程技术类综合性检索工具。EI 收录生物工程、交通运输、化学和工艺工程、照明和光学技术、农业工程和食品技术、计算机和数据处理、应用物理、电子和通信、控制工程等各学科领域 5100 种工程类期刊、会议论文集和技术报告。EI 具有综合性强、资料来源广、地理覆盖面广、数据资源丰富、信息质量高、权威性强等特点。文献是否被 EI 收录是衡量工程技术研究领域学者学术水平的重要指标之一。

国内比较重要的检索是中国科学引文数据库 CSCD(Chinese Science Citation Database)、南京大学中国社会科学研究评价中心研制的中文社会科学引文索引 CSSCI(Chinese Social

Sciences Citation Index)、中国科技期刊引证报告 CJCR(Chinese Journals Citation Reports)、北京大学研制的中文核心期刊要目总览 PKU 以及 MEDLINE(MED)等。

1.3.7　自动问答

自动问答:针对计算机用自然语言句子提问,并且能为用户直接返回所需的答案。对于问答系统,用户不需要把自己的问题分解成关键词,可以把整个问题直接交给问答系统。问答系统结合自然语言处理技术,通过对问题理解,能够直接提交给用户想要的答案。问答系统就是新一代的搜索引擎。问答系统一般包括三个主要部分:问题分析、信息检索和答案抽取。国外已经开发了一些相对成熟的问答系统。麻省理工开发出了一个问答系统 Start,1993 年开始发布在 Internet 上(http://www.ai.mit.edu/projects/infolab/)。可以回答有关地理、历史、文化、科技、娱乐等方面的问题。另外还有一个比较成熟的问答系统 AnswerBus(http://misshoover.si.umich.edu/zzheng/qa-new/)。AnswerBus 是一个多语种的自动问答系统,它不仅可以回答英语的问题,还可以回答法语、西班牙语、德语、意大利语和葡萄牙语的问题。在每年一度的文本信息检索(TREC)会议上,自动问答是最受关注的主题之一。越来越多的大学和科研机构参与了 TREC 会议的自动问答(郑实福 2002)。

1.3.8　自动分类

自动分类:用计算机系统代替人工根据文章的内容、按照一定的标准对文献等对象进行分类。一般包含自动聚类与自动归类。自动聚类是由计算机系统从被考察对象中抽取有关特征,根据一定的标准(如类别的数量限制,同类对象的亲近程度等),将相近、相似或相同特征的对象聚合在一起的过程。自动归类是指计算机系统按照一定的分类标准或分类参考,将被考察对象划归到不同类目的过程(自动分类-百度百科 http://baike.baidu.com/)。分类的目的是为了人们更为系统地管理文档,方便、快捷地检索资料。文本分类可用于垃圾网页的自动判定、情感判定或发现、邮件自动分类、搜索结果聚类、文档集聚类、基于聚类的检索等(王斌 2010)。自动分类较为成功的系统有麻省理工学院为白宫开发的邮件分类系统、卡内基集团为路透社开发的 Construe 系统等。

1.3.9　信息抽取

信息抽取:是把文本里包含的信息进行结构化处理,变成表格一样的组织形式。信息抽取系统从各种各样的文档中抽取相关的信息,以统一的形式集成在一起。为了方便检索和比较,抽取出来的信息以表格形式集成在一起。随着互联网近年来的飞速发展,信息抽取获得了空前的发展。因为同一主题的信息在网上通常分散存放在不同网站上,表现的形式也各不相同。信息抽取的目的就是将这些信息收集在一起,用结构化形式储存起来,方便那些把因特网当成是知识来源的人[①]。

① http://baike.baidu.com

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生来具有,是由遗传决定的。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具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和归纳而得到的。经验主义研究方法从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)。

1. 理性主义研究方法

理性主义研究方法实现的自然语言处理系统,通常需要设计规则集或程序,将自然语言语句的结构或意义推导出来。而句子的结构和意义一般是由句子中的词、短语的结构和意义根据规则推导出来。比如:语法分析器按照人所设计的语法规则,将输入语句分析为语法结构(树结构、共享森林等)。自然语言处理系统中的规则通常是先验的,由人设计好输给计算机。

2. 经验主义研究方法

经验主义研究方法目前包含有指导的学习和无指导的学习两种。

有指导的学习首先需要具有标注好的训练语料库和合适的统计模型,然后统计模型的参数在训练语料库的概率,再把参数概率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)和训练语料库(带有正确词性标注的语料库),然后统计训练语料库中每个词用作不同词类的概率和两两词类共现的概率(模型中的参数值),最后把这些参数值应用到新的语料中确定出每个词的词性。有指导的词性标注方法依赖于大规模的带有正确词性标注的语料库,从该语料库中统计出一个句子中每个词的词性概率,通常我们认为概率高的是正确的标注。

无指导的统计学习方法不需要具有标注正确的语料库,计算机需要从未带标注的语料库中自动学习知识,进而来处理相应的问题。比如:统计金庸和古龙多部小说中的词类频率,由所有考虑的词类组成一个向量空间,根据词类向量空间对金庸和古龙的每部小说进行自动聚类就是无指导的统计学习方法。

经验主义研究方法广泛应用于词性标注、语法分析、语义消歧、机器翻译、语音识别、自动分类、信息检索等语言处理领域。

3. 理性主义与经验主义的结合

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。

1.4.2　理性主义和经验主义的区别

（1）理性主义主要研究人的语言知识结构，人类使用的语言为理性主义提供了间接证据。而经验主义的研究对象是实际使用的语言。

（2）理性主义方法一般根据乔姆斯基的语言原则（principles）来进行研究。它根据语言所必须遵守的一系列原则来描述语言。也就是：当一个句子遵守了语言原则，则这个句子是正确的。而当一个句子违反了语言原则便是错误的。经验主义方法是根据香农（Shannon）的信息论，对语言单位（字、词、短语或句子等）根据其在实际语料库中使用的频率来计算其概率。一个语言单位的概率越大，则认为其越常见，否则是罕见的。

（3）理性主义方法一般通过对一些特殊的语句或语言现象的研究来获得对人的语言能力的认识，而这些语句和语言现象在实际语言应用中并不一定常见。而经验主义方法研究语言的普遍现象，越是常见的、在语料库中出现次数多的语言单位，越受重视。

（4）理性主义方法一般把语言学中的知识用规则描述出来，规则一般是用形式化的数学描述方法（如 Backus 范式）把人类关于语音、词汇、句法、语义甚至语用的知识表示出来。这种规则对人来说，清晰、容易理解。而经验主义方法从大规模语言数据中自动或半自动地获取语言统计知识，获取的知识一定带有频率或概率信息。相比而言，理性主义方法的规则数目较少，而经验主义方法统计的语言单位数量较多。

（5）不管是理性主义方法的规则还是经验主义方法获取的统计知识，都作为自然语言处理系统的知识库。利用理性主义方法的规则，计算机通过一系列的分析、逻辑推理和演算实现自然语言的处理与理解。而对经验主义方法，计算机需要把统计知识运用到适合的数学模型中，通过一系列的计算实现自然语言处理。许多自然语言处理系统既可以采用规则方法来实现，也可以采用统计方法来实现，还可以把两者结合起来。比如：汉语切分、词性标注、句法分析、句法消歧、机器翻译等都有分别用两种方法或两种方法结合实现的系统。

（6）利用理性主义方法实现的系统的准确率和召回率与规则的质量相关。而利用经验主义方法实现的系统的效果与语料库的规模大小和加工精度有很大关系。语料库规模越大、对语料库的标注越精、加工深度越深，效果越好。

（7）基于理性主义方法的系统鲁棒性较差，处理不同于以往的新的语言数据效果可能不好。而基于经验主义方法的系统鲁棒性较强，因为经验主义方法可以根据新的实际数据随时进行参数训练，统计语言模型所需要的语言统计知识。

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 J. P. Eckert 和 J. W. Mauchly 设计的第一台计算机 ENIAC 问世,震惊世界。同一年,英国的A. Donald Booth、美国的 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,一时吸引了人们注意,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、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、九州大学等机构,以及意大利、比利时、英国、捷克、匈牙利、德国等国家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,难以付诸实用。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。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。之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(冯志伟 2001)。

2. 发展期

ALPAC 报告出来以后,计算语言学的研究逐渐转向了自然语言理解。自然语言理解系统可以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之上,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,几乎完全抛开了统计技术。

第一代系统主要有:(1)特殊格式系统,早期的自然语言理解系统根据人机对话内容的特点,采用特殊的格式来进行人机对话。1963 年,R. Lindsay 在美国卡内基技术学院用 IPL-V 表处理语言设计的 SAD-SAM 系统,采用特定格式进行亲属关系方面的人机对话。系统建立了一个关于亲属关系的数据库,可接收关于亲属关系方面的英语句子作为问题,并用英语作出回答。1968 年,波布洛(D. Bobrow)在美国麻省理工学院设计了 STUDENT 系统,这个系统把高中代数应用题中的英语句子归纳为一些基本模式,由计算机来理解这些应用题中的英语句子,列出方程求解,并给出答案。(2)以文本为基础的系统。为了改进特殊格式系统中种种格式限制,出现了以文本为基础的系统。1966 年,R. F. Simmons, J. F. Burger 和 R. E. Long 设计的 PROTOSYNTHEX-Ⅰ系统,对文本信息进行存储和检索,不受特殊格式结构限制。(3)有限逻辑系统。自然语言的句子以更形