1 计算语言学简介

1.1 计算语方学

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乎于语言学、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计算机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：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。计算机形态分析指计算语言学研究如何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法结构（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。

机器翻译经过萧条以后也逐渐复苏，机器翻译的研究者们从失败中汲取教训并且认识到，原语和译语两种语言的差异，不仅表现在词汇上，还表现在句法结构的不同上。因此，这一时期的机器翻译系统几乎都把句法分析放在第一位，把语法与算法分开，而且语义分析在机器翻译中越来越受到重视。这一时期机器翻译系统的典型代表有：1976 年加拿大蒙特利尔大学与加拿大联邦政府翻译局开发的实用性的翻译系统 TAUM-METEO，提供天气预报方面的翻译。美国在乔治敦大学机器翻译系统的基础上，进一步开发了大型翻译系统 SYSTRAN，可进行俄英、英俄、德英、汉法、汉英机器翻译，是目前应用最为广泛、所开发的语种最为丰富的一个实用化的机器翻译系统。日本富士通公司开发了 ATLAS-I 和 ATLAS-II。ATLAS-I 以句法分析为中心，ATLAS-II 以语义分析为中心，用于日英翻译。法国诺布尔理科医科大学应用数学研究所自动翻译中心的俄法机器翻译系统 ARIANE-78，采用"独立分析—独立生成一相关转换"的方法，即原语词法分析一原语句法分析—原语译语词汇转换一原语译语结构转换一译语句法生成—译语词法生成。此外，还有一些大规模的机器翻译系统正在研制中，如 EUROTRA 计划、Mu 系统、ODA 计划、DLT 系统等。从实用化商品化的角度来看，机器翻译的研究者们对语法和词典都下了不少工夫，研究的规模也扩充了，因而翻译时未登录的词减少，句子分析的成功率提高，多义词选择的准确性和歧义判别能力也进一步得到了改进。

随着互连网的广泛使用，为了克服互连网络上的语言障碍，最近日本的一些公司开发出了一大批网络上的英语日语互译的自动翻译系统。网上翻译将是机器系统进入实用领域的一个新的突破口。近年来，国内外还开始了自动翻译电话的研究和口语翻译系统的研制。

在计算语言学发展期，各种计算语言学的理论逐渐成熟，出现了一大批理论成果。乔姆斯基的形式语言理论是影响最大的早期计算语言学的句法理论。乔姆斯基定义了 0 型文法、上下文无关文法、上下文有关文法和有限状态文法。其中上下文无关文法又叫做短语结构语法，广泛应用于自然语言的自动句法分析和生成中。但由于短语结构语法的分析能力不高，分析时难以区分大量的不合语法的句子，生成能力过强，20 世纪 50 年代末期，乔姆斯基指出了短语结构语法在描述自然语言方面的种种局限性，并提出了转换生成语法来克服短语结构语法的这些局限性。70 年代以来，乔姆斯基发现，就是转换生成语法本身也有局限性，它的生成能力过强，它不仅可以生成一切人类的语言，还可以生成许多人类语言之外的符号串。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，一个句子只有一个成分结构，句子与句子之间在成分结构上没有联系，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，于是出现了各种增强的短语结构语法，如扩充转移网络、词汇功能语法、功能合一语法、广义短语结构语法、中心词驱动的短语结构语法等。这些语法都采用了复杂特征结构来改进短语结构语法，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（俞如珍，金顺德 1994）。

1969 年，厄尔利提出了厄尔利算法，把自底向上分析与自顶向下分析结合起来，提高了分析效率。1980 年，马丁·凯（Martin Kay）提出了线图分析法（chart parsing），为短语结构语法的自动分析提供了一种较好的控制方法。1985 年富田胜提出了富田胜算法，这是一种基于上下文无关文法的高效的自然语言剖析算法。这些都为自然语言自动句法分析提供了理论基础。

在语义自动分析方面，50 年代，美国人类语言学家在分析亲属词时提出了义素分析法。1966 年，菲尔摩提出了格语法，建立了句法和语义之间的关系。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语义网络（semantic network）。1972 年，美国人工智能专家西蒙斯和斯乐康（J. Slocum）首先将语义网络用于自然语言理解系统中。威尔克斯于 1974 年提出了优选语义学，提高了英法机器翻译的译文质量。商克提出了概念依存理论，用于英语的自动理解。20 世纪 70 年代初，美国数理逻辑学家蒙塔格（Richard Montague）提出的蒙塔格语法用数理逻辑来研究自然语言的句法结构和语义关系，开辟了一条新途径。

这些基础理论的研究，为计算语言学的进一步发展奠定了坚实的理论基础。计算语言学的发展表明，这一学科的进步不仅有利于机器翻译技术的进步，而且在当今世界上，它有着重大的理论意义和现实意义。语言能力是人类的智能行为之一，长期以来是语言学、认知科学、心理学和人工智能等学科关注的焦点之一。计算语言学从另外的角度促进了这些学科的发展，有助于人类早日搞清楚语言发生、运作的机理。同时，计算语言学在机器翻译、信息检索、人机接口等信息处理领域有着广泛的应用前景，意义非凡。

3. 繁荣期

从 1989 年，计算语言学进入了大规模真实文本处理的新时期。这个新时期的重要标志是在基于规则的技术中引入了语料库方法，其中包括统计方法、基于实例的方法、通过语料加工手段使语料库转化为语言知识库的方法等。

基于实例的机器翻译最早是日本机器翻译专家长尾真（Makoto Nagao）于 1984 年提出的。基本思想是，人们在翻译一个简单句时并没有作深层的语言分析，而是首先将句子拆分为适当的片段，然后将这些片段翻译成目标语言片段，最后将这些目标语言片段组合为一个完整的句子。目前，基于实例的机器翻译系统主要有日本京都大学长尾真和佐藤的 MBT1 和 MBT2 系统。美国卡内基一梅隆大学的多引擎机器翻译系统 PAGLOSS，这个系统的主要引擎是基于知识的机器翻译，基于实例的机器翻译系统是它的一个引擎。日本口语翻译通信研究实验室的 ETOC 和 EBMT 系统。

语料库语言学试图从大规模真实文本的语料库中获取语言知识，以求得对于自然语言规律的更为客观、准确的认识。随着人们对大规模真实文本处理的日益关注，越来越多的学者认识到，基于语料库的分析方法（经验主义的方法）至少是对基于规则的分析方法（理性主义的方法）的一个重要补充。但是，一个语料库不管规模多大，如果未经加工，就只是一些文本的简单累积，它的研究价值和使用价值都是极其有限的。为了从语料库中获取有关的语言学知识，就必须对语料进行必要的加工，将生语料加工成熟语料。20 世纪 80 年代初，马莎（Mashall）设计了第一个用统计方法的词性标注系统 CLAWS，对 LOB 语料库进行自动标注，使标注正确率提高到 97%。如果把基于规则方法与基于统计方法结合，必定会推动计算语言学的进一步发展。

2 词法分析

传统语言学根据词的形态结构把语言分为三大类（冯志伟 1996）：

分析型语言：词基本上没有专门表示语法意义的附加成分，形态变化很少，语法关系靠词序和虚词来表示。如汉语、藏语等。

黏着型语言：词内有专门表示语法意义的附加成分，一个附加成分表达一种语法意义，一种语法意义也基本上由一个附加成分来表达，词根或词干跟附加成分的结合不紧密。如芬兰语、日语等。

屈折型语言：用词的形态变化表示语法关系，一个形态成分可以表示若干种不同的语法意义，词根或词干跟词的附加成分结合得很紧密，往往不易截然分开。如：英语、德语和法语等。

分析型语言的形态变化很少。比如，书面汉语的单词基本上没有形态变化，但书面汉语不像英语、德语和法语等印欧语言那样，词与词之间用空格分开。一个汉语句子由一串前后连续的汉字组成，词与词之间没有明显的分界标志。因此，书面汉语词法分析的主要任务不是分析单词的形态变化，而是进行单词的自动切分。

这一章我们主要介绍汉语的自动分词和英语的形态还原。

2.1 汉语的自动分词

2.1.1 词与自动分词

词是语言中最小的能独立运用的单位，是信息处理的基本单位。“词”这个概念一直是汉语语言学界纠缠不清而又挥之不去的问题。主要困难在两方面，一方面是单字词与语素之间的划界；另一方面是词与短语（词组）的划界。到目前为止没有一个公认的、具有权威的词表。因此，汉语自动分词的首要任务是确定分词规范。刘源等在《信息处理用现代汉语分词规范》(刘源等 1994)中规定了现代汉语的分词的原则、方法及一系列规则，1992 年由国家技术监督局批准为国家标准（GB—13715）。目前已经被一些系统所采纳。

汉语自动分词：把没有明显分界标志的字串自动切分为词串。包括标点符号、数字、数学符号、各种标记、人名、地名、机构名等未登录词的识别。

因此，汉语自动分词主要包括：

1. 根据分词规范，建立机器词典。
2. 根据分词算法和机器词典，把字串切分为词串。

2.1.2 汉语自动分词的重要性

汉语的词也是汉语语言中最小的独立运用单位。自动分词是现代汉语进行句法分析的第一步，是后续语法和语义分析的基础。因为计算机从事句法分析所凭借的语法知识是机器词典和句法规则库。机器词典收录了每个词条的词法、句法和语义知识。而句法规则库是以词、词类、语义等知识为基础构造的。因此一连串的汉字组成的句子必须先进行分词，才能利用机器词典和规则库，也才有可能进一步进行句法分析。词频统计、作家作品风格学研究、自动标引、自动分类、机器翻译等方面的研究，也必须首先分词，在这些应用和研究领域，没有准确高效的分词策略，汉语的进一步分析必将受到严重影响。

汉语分词的关键在于，好的分词算法和好的分词词库。

2.1.3 汉语自动分词方法

自 20 世纪 80 年代初起，已经提出了许多分词方法。目前，根据所使用的知识资源不同分为基于规则的方法，基于统计的方法，以及两者结合的方法。根据有无分词词典分为有词典分词和无词典分词。有词典分词是主流分词方法。

基于规则的方法一般都需要事先有人工建立好的分词词典和分词规则库。主要是基于字符串匹配的原理进行分词，往往以足够大的词表为依据，采用一定的处理策略将汉语文本的字符串与词表中的词逐一匹配，如若成功，就认为该字串为词。主要有正向最大匹配法、逆向最大匹配法、双向匹配法、逐词遍历匹配法、设立切分标志法、正向最佳匹配法和逆向最佳匹配法等。如果分词词典规模小，覆盖程度有限，则会影响分词的正确率。

基于统计的方法是利用字与字间、词与词间的同现频率作为分词的依据，可以没有建立好的分词词典。这种方法需要大规模的训练文本，用来训练模型参数。这种方法的优点在于它不受应用领域的限制。但训练文本的选择将影响分词结果。下面主要介绍正向最大匹配法、逆向最大匹配法、双向最大匹配法以及联想一回溯法。

1. 正向最大匹配法(Maximum Matching Method,简称 MM 方法)

MM算法的具体算法可以描述如下：

设 MaxLen 表示最大词长，D 为分词词典；

(1) 从待切分语料中按正向取长度为 MaxLen 的字串 str,令 LEN=MAXLEN;

(2) 把 str 与 D 中的词相匹配；

(3) 若匹配成功，则认为该字串为词，指向待切分语料的指针向前移 LEN 个汉字,返回到(1);

(4) 若匹配不成功；

如果 LEN>1 ,则把 LEN 减 1，从待切分语料中取长度为 LEN 的字串 str,返回到(2)。否则，得到长度为 1 的单字词，指向待切分语料的指针向前前移 1 个汉字，返回到(1)。

说明：在步骤(1)中，如果待切分语料的字串长度小于 MaxLen ,则取字串 str为待切分语料。在步骤(4)中,如果得到的单字不是词,是语素字的话,则需要进行未登录词的识别。

MM 方法优点：(1) MM 扫描方向是从左到右，从长到短的顺序进行匹配，(2) MM 法的原理简单，易于在计算机上实现，时间复杂度也比较低。

MM 方法缺点：(1) 必然会忽视“词中有词”的现象，导致切分错误。例如对字符串“幼儿园地节目”进行切分时,MM 方法的切分结果是“幼儿园／地／节目”，而正确的切分结果应该是“幼儿／园地／节目”。(2) 最大词长的长度比较难于确定，如果定得太长，则匹配时花的时间多，算法的时间复杂度明显提高。如果定得太短，则不能切分长度超过它的词，导致切分正确率降低。

2. 逆向最大匹配法(Reverse Maximum Matching Method ,简称 RMM方法)

这种方法原理与 MM方法相同，但扫描方向由右到左，提出 RMM 方法的意义更在于同 MM方法进行结合运用，即双向匹配法对字符串进行更准确地切分。

3. 双向匹配法

对同一个字符串分别采用 MM 法、RMM 法两种方法进行切分处理，如果能够得到相同的切分结果,则认为切分成功,否则认为有疑点,这时或者采用上下文信息,根据切分歧义规则库进行排歧。或者进行人工干预，选取一种切分为正确的切分。

这个方法克服了 MM 方法里忽视‘词中有词’现象的弊端。例如,使用双向匹配法对“幼儿园地节目”做切分处理时分别使用 MM 法和 RMM 法得到的两个切分结果是‘幼儿园／地／节目’和‘幼儿／园地／节目’,切分系统将报告错误,而不至于将错就错,影响其他的语言处理。

双向匹配法的缺陷是算法复杂度的提高，而且为了使切词词典同时支持正向和逆向两种顺序的匹配和搜索，词典的结构比一般的切词词典要复杂得多。

正向最大匹配法和逆向最大匹配法结合，可以用来查找交叉歧义，如果使用正向最大匹配法得到的切分结果与使用逆向最大匹配法得到的切分结果不相同，则存在交叉歧义。但并不是所有的交叉歧义都可以通过双向最大匹配法找到。

如 原子／结合／成分／子时

MM 和 RMM 得到相同的分词结果，但存在切分歧义。

上面介绍的三种方法都过分依赖分词词典，如果词典的规模受限，则会影响切分正确率。下面介绍两种不是过分依赖词典的分词算法—— 基于联想回溯算法。

4. 基于联想-回溯算法 (Association-Backtracking Word Segmentation ，简称 AB算法)

山西大学采用 AB算法实现了一个分词系统（刘开瑛 2000)。这个系统利用的汉语本身的知识(如构词法、构形法、句法等)比较多,提出了一些歧义结构的实用分词规则,并且采用切分标志法和有穷多次列举的方法来提高分词精度。该系统由知识库和选词控制机制两大部分组成。

知识库包括三个层次：(1)特征词词库；(2) 实词词库；(3)规则库。

(1) 特征词词库。所谓特征词，泛指那些具有可作为分割标识的某种特征的词或词

素，主要包括词缀、虚词、重叠词、联绵词等。

（2）实词词库：主要包括名词、动词、形容词、副词等实词。

（3）规则库包含有专用和通用两类规则。专用规则是通过反复实验从所产生的错误切分结构中抽取整理的。而通用规则主要基于汉语语言本身的词汇知识和句法知识。

选词控制机制由五大功能模块组成。包括：预处理模块；分割模块；细分模块；规则调用模块；人工干预模块。

预处理模块：将源语言（一篇短文或段落）依各种形态标志（主要是标点符号）分解成独立的、可被切分程序直接处理的字串序列。

分割模块：对语料的第一次扫描，它以特征词库中的词作为词切分标志，依靠联想规则将一个字串分割为多个更小的子字串。

细分模块：依据实词库内容将从分割模块得到的子字串切分为语词。采用改进的MM算法，并采用回溯推理机制。当遇到歧义组合结构或产生拒分现象时。便分别转向规则调用模块和人工干预模块。

规则调用模块：利用细分模块提示的信息，调用相应的规则处理歧义组合结构。或调用通用规则切分类型词(如,数字词)。

人工干预模块：人工干预常常由词典收词不足引起。包括，修改实词库，追加临时词库，修改规则库，修改特征词词库和利用推理机制自动选词。

5. 统计方法进行汉语切分

令 $S=C_1C_2\cdots C_{n-1}C_n$,其中 $C_i(1\leqslant i\leqslant n)$ 是一个汉字字符。把一个汉语句子切分成词序列就是把这些汉字字符结合成词，比如：

$$\begin{aligned} S &= C_1C_2\cdots C_{n-1}C_n = (C_1\cdots C_{x_1})(C_{x_1+1}\cdots C_{x_2})\cdots(C_{x_{m-1}+1}\cdots C_{x_m}) \\ &= W_1W_2\cdots W_m \end{aligned} \tag{2.1}$$

其中 x_k 是第 k 个词 W_k 的最后字符的下标，$x_0=0$. $x_m=n$,根据信道模型,分词的过程就是求在给定输入字串 C 的条件下所产生的输出词串 W 的概率 $P(W|C)$。根据贝叶斯公式，下面的公式成立：

$$P(W\mid C) = (P(W)P(C\mid W))/P(C) \tag{2.2}$$

因为 C 是给定的字串，$P(C)$ 是一个确定的值，在计算中不起作用。 $P(C|W)$ 是在给定词串的情况下字串出现的概率，可以认为是 1。

$$P(W\mid C)\approx P(W) \tag{2.3}$$

因此，基于统计的词切分过程，可以认为是寻找具有最大概率值的词串过程。

句子 S 的切分可以被惟一地表示成一个整数序列 $x_1,\cdots,x_m$ ，所以可以用相应的整数序列表示一个切分。令 $G(s)=\{(x_1\cdots x_m):1\leqslant x_1\leqslant\cdots\leqslant x_m,\ m\leqslant n\}$ 是句子 S 的所有可能切分。于是对于一个切分 $g(s)=(x_1\cdots x_m)\in G(S)$,由 $L(g(s))$ 对切分 $g(s)$ 进行评分得到：

$$L(g(s)) = \log P_g(w_1\ w_2\cdots w_m) = \sum_{i=1}^{m}\log P_g(w_i\mid h_i) \tag{2.4}$$

其中 $w_i=C_{x_{i-1}+1}\cdots C_{x_i}(i=1,2,\cdots,m)$,h_i 是历史词 $w_1\cdots w_{i-1}$,如果使用二元模型则 $h_i=w_{i-1}$,如果使用三元模型,则 $h_i=w_{i-2}w_{i-1}$.

Veterbi 算法：选择最高的评分作为结果，也即：

$$g^* = \arg\max_{g\in G(S)} L(g(s.)) = \arg\max_{g\in G(S)} \log P_g(w_1\cdots w_m) \tag{2.5}$$

根据动态规划算法：整个句子的最高评分可以通过求解子问题的最佳解得到。

令 $L(k)$为最初 k 个字符的最大评分，则 $L(1)=0$，$L(g^*)=L(n)$，给定$\{L(i);\ 1\leqslant i\leqslant k-1\}$，$L(k)$ 可以如下递归计算：

$$L(k) = \max_{1\leqslant i\leqslant k-1}[L(i) + \log P(C_{i+1}\cdots C_k \mid h_i)] \tag{2.6}$$

其中 h_i为以第 i 个字符 C_i结束的历史词。递归结束时，需要回溯发现切分点。因此需要记录切分点。令 $P(k)$为前一个词最后字符的下标。于是：

$$P(k) = \arg\max_{1\leqslant i\leqslant k-1}[L(i) + \log P(C_{i+1}\cdots C_k \mid h_i)] \tag{2.7}$$

即 $C_{P(k)+1}\cdots C_k$ 为最优切分中直到第 k 个字符的最后一个切分词。

例如：一个句子 $S=C_1C_2\cdots C_5$，根据公式(2.7)得到表 2-1：

表 2-1 $P(k)$ 为前一个词最后字符的下标

字符	C_1	C_2	C_3	C_4	C_5
k	1	2	3	4	5
$P(k)$	0	1	1	3	3

则这个句子的最优切分为：$(C_1)(C_2C_3)(C_4C_5)$

算法的时间复杂度为 $O(n)$(刘颖 2001)。

2.1.4 汉语切分歧义及其处理

对汉语切分会产生切分歧义。切分歧义是影响分词系统切分正确率的重要因素，也是分词阶段最困难的问题。切分歧义包括交集型歧义和组合型歧义（冯志伟 1996，刘开瑛 2000）。

1. 交集型歧义 如果字串 abc 既可切分为 ab/c，又可切分为 a/bc。其中 a，ab，c 和 bc 是词。例如：

(1) 以树型图形式加以描绘。

“图形式”可能切分为“图形／式”，也可能切分为“图／形式”，正确切分为“图／形式”。

(2) 研究生命本质。

“研究生命”可能切分为“研究／生命”，也可能切分为“研究生／命”，正确切分为“研究／生命”。

(3) 白天鹅游过来了。

“白天鹅”可能切分为“白／天鹅”，也可能切分为“白天／鹅”，要根据语境来确定哪一个正确。

(4) 独立自主和平等独立的原则。

“和平等”可能切分为“和／平等”，也可能切分为“和平／等”，正确切分为“和／平等”。

(5) 小说太平淡了。

“太平淡”可能切分为“太／平淡”，也可能切分为“太平／淡”，正确切分为“太／平淡”。

(6) 对这种现象的确切描述。

“的确切”可能切分为“的确/切”,也可能切分为“的/确切”,正确切分为“的/确切”。

2. 组合型歧义 若 ab 为词,而 a 和 b 在句子中又可分别单独成词。例如:

(1)他骑在马上。(切分为:他/骑/在/马/上/。)

马上过来。(切分为:马上/过来/。)

(2)他学会了解数学难题。(切分为:他/学/会/了/解/数学/难/题/。)

我对小华比较了解。(切分为:我/对/小华/比较/了解/。)

(3)请把手抬高一点。(切分为:请/把/手/抬/高/一点/。)

这个把手不好用。(切分为:这/个/把手/不/好用/。)

(4)语言学起来并不十分容易。(切分为:语言/学/起来/并/不/十分/容易/。)

语言学是一门学科。(切分为:语言学/是/一/门/学科/。)

3. 混合型歧义:由交集型歧义和组合型歧义自身嵌套或两者交叉组合而产生的歧义(侯敏,孙建军,陈肇雄 1995)。例如:

(1) 这篇文章写得太平淡了。

这墙抹得太平了!

即使太平时期也不应放松警惕。

“太平淡”是交集型歧义,而“太平”是组合型歧义。

(2)我们学会了解答问题的办法。

他还不了解答题的方法。

他学会了解方程。

我们都了解他。

“了解答”是交集型歧义,而“了解”是组合型歧义。

4. 那么如何采集歧义字串呢?

山西大学在(刘开瑛 2000)中使用双向扫描的方法来采集歧义字串。

(1)正向最大匹配和逆向最大匹配两种方法扫描发现交叉型歧义。

例如:企业要真正具有用工的自主权。

正向最大匹配 企业/要/真正/具有/用工/的/自主/权/。

逆向最大匹配 企业/要/真正/具有/用工/的/自/主权/。

由此发现交集型歧义:“自主权”。

(2)采用正向最大匹配和逆向最小匹配,并且最小匹配从单字词开始的方法发现组合型歧义。

例如:向老人家陈述其中的利害。

逆向最小匹配向/老/人/家/陈述/其/中/的/利/害/。

正向最大匹配 向/老人家/陈述/其中/的/利害/。

这样发现组合型歧义“老人家”“其中”和“利害”。

5. 切分歧义处理方法

目前对于切分歧义消歧主要有三种方法:规则方法,统计方法和规则与统计结合的方法。

(1) 规则方法

主要利用歧义字串、前趋字串和后继字串的句法、语义、语用三个方面的信息来消歧。

句法信息：有些歧义切分字串同其前趋字串和后继字串存在着密切的搭配关系，这时我们就可以利用有关的句法信息得到正确的切分结果。

例如：一阵风吹过来了。

其中“阵”和“风”是由量词和名词组合产生的歧义切分字串，根据汉语的结构，量词之前应该有数词，因此，我们可以建立规则 如果当前歧义字串的前趋字串为数词，则该歧义字串的首段单切，否则，该歧义字串成词。在切词的过程中，遇到该歧义时，就可以调用这条规则，并进行一定的逻辑推理作出正确的切分。

语义信息：当歧义切分字串在句法层次上难以分析时，我们要考虑它的语义信息。

例如：他学会了解数学难题。

歧义字串“了解”是由助词“了”和动词“解”串联组合产生的，可以有两种切分结果：“他／学会／了／解／数学／难题”和“他／学会／了解／数学／难题”，这两种切分结果的词类和句法结构都十分相似，仅仅根据词法和句法知识是难以得到正确的切分结果的，但是根据语义分析可知，动词“解”的义项中，要求宾语应该有“数学公式”或者“扣子”这样的义素，而动词“了解”对宾语则没有这样的要求，由于上述例子里中做宾语的“数学难题”符合动词“解”的义项要求，由此可以判断前一种切分结果是正确的。

语用信息 对于“乒乓球拍卖完了”这个句子，仅根据词法、语义和语义知识是很难判断卖完的东西究竟是“乒乓球”还是“乒乓球拍”，也很难得到正确的切分结果，这个时候，就需要根据语言交际的具体环境和语用方面的知识，才能得到正确的切分。

(2) 统计方法

方法一：孙茂松、黄昌宁等提出了一种利用句内相邻字之间的互信息及 t-测试差这两个统计量解决汉语自动分词中交集型歧义字串的方法（孙茂松、黄昌宁等 1997）。

方法二：刘开瑛提出根据链长和独立成词能力频次库结合的统计方法解决交集型歧义字串的方法(刘开瑛 2000)。

方法三 直接利用 2.1.3 中第 5 部分的统计方法进行切分和歧义消歧一体化处理策略。

(3) 规则与统计结合的方法 把前面两种方法结合。

2.1.5 未登录词的处理

汉语词汇是一个开放集合，无论建立多么庞大的词典，都不可能穷举所有的词。这是因为人们在通过字词组合来创造新词方面有很大的灵活性。而且随着时间的推移，还会不断出现大量的新词。

未登录词：词典中没有登录过的人名、地名、机构名、译名、新词语等（冯志伟 2001）。当采用匹配的方法来切词时，由于词典中没有登录这些词，会引起自动切词的困难。一个开放的系统必须能够识别未登录词，才有可能提高分词系统的正确率。目前，对人名、地名、机构名、译名和新词语的识别，都有人做过研究和实验，并且取得了一定的成果（刘开瑛 2000 宋柔等 1993 孙茂松、张维杰 1993 孙茂松等 1995 张俊盛等 1992 张小衡，王玲