“八五”国家重点科研成果论文集

中文信息处理应用平台工程

THE CHINESE LANGUAGE INFORMATION PROCESSING INFRASTRUCTURE PROJECT '91～95

陈力为　袁琦　主编

電子工業出版社

“八五”国家重点科研成果论文集

中文信息处理应用平台工程

THE CHINESE LANGUAGE
INFORMATION PROCESSING
INFRASTRUCTURE PROJECT ' 91～95

陈力为　袁琦　主编

电子工业出版社

内 容 简 介

本书收录了“八五”(1991～1995年)期间，电子工业部计算机与微电子发展研究中心(CCID)组织实施“905工程”中，参加项目的各单位围绕各子专题的有关研究内容所发表的论文。这些论文展现了我国计算语言学研究与应用的部分成果，其中包括电子词典的设计、汉语的句法和语义分析、汉语分析器设计、语料库建设与语料加工技术等三十余篇论文。

本书的出版对今后我国中文信息处理的基础研究和产品开发具有重要的参考价值，特别是对我国建设与国际接轨的国家信息基础结构(NII)，推动信息高速公路上有中国特色的语言工程的建设有重要意义。

本书可供从事中文信息处理的研究开发人员、计算机、语言学等专业的科研人员、工程技术人员、大学教师和研究生阅读。

“八五”国家重点科研成果论文集

中文信息处理应用平台工程

THE CHINESE LANGUAGE INFORMATION
PROCESSING INFRASTRUCTURE PROJECT ’91～95

陈力为 袁琦 主编

责任编辑：竞力

*

电子工业出版社出版

北京市海淀区万寿路173信箱(100036)

电子工业出版社发行 各地新华书店经销

三河市兴华印刷厂印刷

开本：787×1092毫米 1/16 印张：17.75 字数：410千字

1995年12月第一版 1995年12月北京第一次印刷

印数：300册 定价：100.00元

ISBN 7-5030-3477-8/TP.1380

前　言

二十年前，周恩来总理亲自批准的“748工程”开始在我国建立的汉字信息处理工程，对计算机的推广应用和计算机产业的发展具有战略意义。五年前，国防科工委和电子工业部下达的国家科研项目“905工程”开始在我国实施的中文信息处理应用平台工程，具有同样的战略意义。

在军队指挥自动化系统中，从情报获取、情报处理，直到作战方案制定和命令下达的各个环节，以语言文字方式传递的军事情报占有重要地位。“905工程”旨在研究计算机汉语理解技术中的基础性、关键性技术，建立应用平台，以便支持各种实用化汉语信息处理系统的开发。“905工程”的实施，不仅在汉语语言模型构造、电子词典和规则知识库建造、汉语分析系统设计的理论和技术等方面，取得了一系列中间成果，而且为我国尽快建立与“905工程”相衔接的，更具战略意义的“汉语信息处理工程”打下了基础。

现将“八五”（1991～1995年）期间，电子工业部计算机与微电子发展研究中心（CCID）组织实施“905工程”中，参加项目的各单位围绕各子专题的研究内容所发表的论文汇总成册，并且作为“905工程”预研成果的一部分，出版发行。

目　录

汉语书面语的分词问题 — 一个有关全民的信息化问题

陈力为

电子部计算机与微电子发展研究中心

汉语的书面语是按句连写的，词间无间隙。因此在汉语书面语的处理中，例如，统计、分析、理解等，我们首先遇到的问题是词的切分。把按句连写转换为按词连写。所以，词的正确切分是进行汉语书面语处理的必要条件；它的任何错误都将使处理结果受到或大或小的影响，有时是严重的影响。

从80年代初起，很多学者、专家致力于汉语书面语的自动分词[4]，取得了不少可用的分词系统，但在实用的过程中，又遇到不少新问题，困扰着我们[6].，例如，人名、地名、企业名、新词等未登录词[5,6,7]。对于这些问题，经过业界的努力，近两年来，又取得若干可喜的突破。但随着国民经济信息化的不断发展，中文信息处理的广泛、深入地开展，对分词系统的要求将越来越高。难度越来越大。现在，汉语书面语的分词技术已经悄悄地形成了一门新兴的富有挑战性的学问。

过去经验告诉我们，我们中文信息处理技术是在不断克服困难中前进的。书面语的分词也不会例外。我们相信业界将根据客观需要，继续研究分词中的难点，推动分词技术的前进。但是，现在我们需要冷静地想一想，汉语书面语的切分是汉语固有的属性呢，还是人们强加给它的呢?

在汉语中什么是词，到现在并无公认的定义。今天也并非讨论什么是词的时候。但人的思维是以词为基本单位进行的。人们表达自己的思想有两种途径：语言，文字；前者叫做口语，后者叫书面语。口语中，词间有'顿挫'(按词说出)，而书面语中词间无间隙。很明显，口语忠实地表达了人们(说话人)的思想(表情、手势等人体动作除外)，而书面语则把人们思想的非常关键的信息，词间间隙，给滤掉了。因此书面语的读者首要的任务是：使用自己的全部知识，进行词的切分，边分词边理解，把书面语滤掉的信息给补上。实际上，这对读者是十分沉重的负担；只是习惯了，误认为这是自己应该干的事。

上述书面语和口语的鲜明对照，使我们清醒地认识到，汉语书面语的词的切分问题，并非汉语所固有的，而是人们强加给它的，是人为的。若要恢复汉语原来的面貌，其办法是显而易见的：这就是，由书面语(文章)的作者按词连写(词间加间隙)就是。只是所需空间增加了四分之一。在这样的书面语面前，词的切分歧义问题不见了，象'乒乓球拍卖完了'这类的拦路虎也自动解体了(这句话指的是'乒乓球'还是'球拍'，难道还会难倒使用这句话的人吗?)；未登录词不见了。把一件易如反掌的事情变为一座难以攻破的堡垒，这是我们现行的汉语书面语书写规范(按句连写)造成的后果。必须引起我们的深思。

大约在50年代，语言学界有一次辩论：是否把按句连写改为按词连写[8]，未能通过。在1987年中文信息处理国际会议上，本文作者也提到同样的问题[2]. 最近在香山科学会议第42次会议[9]上有多位学者在发言时，提到这个问题。周锡令教授在计算机世界上又从软件的中译本方面出发，指出这个问题的迫切性[3]。

看来，汉语书面语的书写规范已经到了必须修改的时候了。

回顾一下汉语书面语书写规范的沿革是有帮助的。在古代，汉语书面语中不要任何标点，于是标注文章成了一门高深的学问。从汉代起，读书人才注意断句(句读，'读'音dou)问题。只是在大约70年前，'五四运动'以后，人们才开始使用现行的全套标点符号。可以看出，每次改革都使原始书写者通过书面语，传递更多的信息；虽然书写者得要多费些力气，也增加了费用，但由于信息含量的增多，含糊和歧义减少了，不仅为读者带来了好处，社会效益也增加了。这样的大好事情当然

只能留给书面语的写作者去做了。

必须指出，汉语书面语书写规范的修改是一桩有关全民、全社会的工作和生活的大事。它的拟定和实施将遇到一系列的问题，这些问题都要一个一个的予以解决。同时，它也是一个复杂的系统工程，需要有组织有计划地进行。其中最复杂的是习惯势力（例如：看不惯，写不惯等）；它必然有形、无形地发生着制约的作用。当然，在技术上，也存在一些问题，例如要分清什么是词。从时间上讲，它不是三年五年的事情，可能是跨世纪的大工程。但是，只要我们有决心，这些问题都是可以解决的。

国民经济信息化的迅速发展将迎来我国社会生活的美好前景，并将推动信息高速公路的创建。量大惊人的信息在公路上飞驰。为了抽取其中有用的信息资源，人们对信息处理的速度和精度将提出极为严格的要求。面对这样严峻的挑战，难道我们的信息处理仍然容忍被人们强加给汉语的词的切分问题继续困扰下去吗？否！我们还有其他更重要、更迫切的课题要去解决。

请看看英语吧。英语书面语，除了词间有间隔外，专用名词的首字母还要大写。书面语带来的信息超过了口语，为信息处理提供了有利的条件。那么，要求书面汉语恢复汉语的本来面目，词间增加间隙，也是理所当然的了。若是在专用名词上再增加下划线，那就喜出望外了。但这并不稀奇，从'五四'前后有语体文到本世纪50年代，一直就是这样的。现在，少数古籍的整理仍然使用。

很多键盘输入系统是按词输入的，但在完成输入任务以后，又把分词信息抹掉了。十分可惜。

结束语

近几年来虽然多次提到书面汉语的改革问题，但都未取得共识，更未见诸行动。其原因不外乎：

1. 未有充分的实践经验使我们认识到它的严重危害性
2. 未感受到国民经济信息化的进程对信息处理的猛烈冲击

今天不同了。我们认识到：书面汉语的改革已经刻不容缓了。而且，语言学界和信息处理界的结合也为书面汉语的改革创造了有利条件。

这样一个重大改革，必须分阶段进行。第一步，可考虑在自然科学和技术科学领域中试行，摸索经验。

第二步，从小学语文教育开始，逐步推广到全社会。

参考文献

1. 陈力为，' Some Key Issues in Chinese Language Information Processing and Their Prospective Developments' 1987 ICCIP Conference, Beijing
2. 陈力为，'当前中文信息处理中的几个问题及其发展前景'计算机世界，1987年，第21期，第34版。
3. 周锡令，'软件书籍中译本的可读性和几点建议' 计算机世界，1995年，第41期，第15版
4. 梁南元，'再论汉语自动分词和切分知识' 1987 ICCIP Conference, Beijing。
5. 郑家恒，刘开瑛，'自动分词系统中姓氏人名处理策略探讨'计算语言学研究与应用，1993。
6. 宋柔 等，'基于语料库和规则库的人名识别法' 计算语言学研究与应用，1993。
7. 沈达阳 等，'中国地名的自动辨识' 计算语言学进展与应用，1995。
8. 许嘉璐，在香山科学会议第42次会议的发言

建立汉语信息处理工程 推动国民经济信息化和国防现代化进程

袁 琦

电子部计算机与微电子发展研究中心

世界正进入信息化时代。人类的社会进步和经济发展对信息资源、信息技术和信息产业的依赖程度越来越大。特别是随着计算机性能的提高，成本下降，其应用日益广泛。从为国民经济信息化和国防现代化服务的高度出发，抓住机遇，建立与发展具有我国特色的计算机中文信息处理产业是我们义不容辞的历史责任。

一、语言信息处理技术是国际上竞争激烈的高新技术

八十年代以来，国际上在“语言产业”、“语言工程”方面，发表了一系列论述。在研究开发体制上，也相应成立了政府主管和投资的各种形式的开发实体，组织和实施了一系列国家工程。它们代表了各国政府倾向性的意见和行动。例如，欧洲共同体委员会发表的文件中指出：“在今后向信息化社会发展的进程中，必须同时实现三个领域技术的并行进展，即信息技术(IT)，电信技术(TT)和自然语言处理技术(NLP)”。日本在通产省的支持下，投入巨额资金启动的 EDR 工程(机读电子词典工程)、ATR 工程(自动语音电话翻译工程)和 MMT 工程(多国语机器翻工程)有力地推动了语言处理技术在国家经济和产业部门中的应用。以长尾真教授为代表的一批学者1994年提出了面向21世纪信息化社会的高新技术--语言处理技术的中长期发展规划。

九十年代初，美国在国防部高级计划局 DARPA 的资助下，相继成立了计算机用词汇研究联合机构(CLR)和计算机用语言学数据研究联合机构(LDC)。在美国军方的支持下，还开展了一系列语言处理技术的关键性课题研究。有代表性的成果是 FASTUS 信息提取系统和 MARIE 智能信息检索系统。

在建设全球性信息基础结构(GII)的趋势下，语言资源的共享、高效的语言处理技术和访问技术的实现是国际上共同关心的问题。全球性信息化的发展也推动了语言处理技术的国际交流与合作。1993年美日两国联合召开了电子词典和语言技术国际会议，展望21世纪语言产业的技术发展及其研究方向。1994年9月，美、日、英、法、德等十六个国家参加的“信息高速公路上的语言工程” 国际会议上，很多代表阐述了国家信息基础结构(NII)与语言信息处理技术之间的依存关系；各国间交流了个自的研究开发课题和政府的实施政策。为研究信息高速公路上的语言处理技术，欧洲正在与美国 TEI(Text Encoding Initiative)合作，开发共享的语料库系统。亚洲也准备通过国际合作，研究和开发可为亚洲语言共享的语料库系统。

二、中文信息处理的发展已得到国家的重视

中文信息处理技术是我国重要的计算机应用技术。在计算机产业中，唯有中文信息处理技术是我国的专长，在国际上享有得天独厚的优势。这是任何国家所不能比拟的。中文信息处理产品也是最具有中国特色的我国自主版权的软件产品。在国际市场上，也是产业界竞争最激烈的高新

技术领域。

我国的计算机应用，离不开对中文的处理，无论是基层机关、企事业单位的局部办公自动化系统，还是推动国民经济信息化的“金”系列工程都是如此。十几年来，我国计算机的普及应用证明，凡有中文信息处理特点的技术和产品都得到了发展，而且成为国际市场竞争中过得硬的技术和产品。

国务院1992年批准下达的国家中长期科学技术发展纲领中指出：中文信息处理技术是高新技术发展的重点。1995年1月下发的国家电子工业“九五”规划(草案)中，以及1995年1月胡启立部长在全国电子行业工作会议上的讲话对发展中文信息处理软件提出了战略性、前瞻性和指导性的意见：国家应重点支持有自主开发能力，有市场前景的科研成果产业化，我国软件产业发展的重点是中文信息处理软件。

“八五”期间，在国防科工委和电子部的支持下，在中国工程院院士陈力为教授、软件工程专家周锡令教授的指导下，电子部计算机与微电子发展研究中心(CCID)聘请董振东、黄昌宁教授为顾问，组织国内从事中文信息处理研究和开发的主要单位－－清华大学、北京大学、东北大学、北京语言学院、河南财经学院，从应用研究(计算机如何处理和理解汉语)着手，以中文信息处理技术的智能化为目标，组织实施了中文信息处理应用平台工程。通过五年的研究开发，在语料库及其管理系统、电子词典系统、规则库系统、汉语自动分词和词性标注系统、基于复杂特征集和合一运算的汉语自动分析系统，以及商品化全文信息检索系统和智能型全文检索实验模型等方面，都取得了预期的成果。从长期目标看，这是一项旨在全面提高我国中文信息处理技术及产品水平的语言基础结构工程(language infrastructure project)，其成果可直接用于“九五”智能化的中文信息处理技术和产品的开发。

我们应当在“八五”中文信息处理应用平台工程成果的基础上，根据国家“九五”期间重点支持中文信息处理软件发展的建议，总结工程实施经验，从国民经济信息化和国防现代化需求出发，尽快制定出“九五”计划和2010年远景目标。

三、建立我国的“汉语信息处理工程”是当务之急

为贯彻和落实中共中央关于制定国民经济和社会发展“九五”计划和2010年远景目标的建议，探讨跨世纪的中文信息处理技术的发展趋势，全国22个单位的专家代表出席了1995年10月30日至11月2日召开的香山科学会议第42次会议。与会专家一致认为，20年前，周恩来总理亲自批准的“748工程”开始在我国建立的汉字信息处理工程具有战略意义，现在建立与“748工程”相衔接的，更具战略意义的“汉语信息处理工程”正是当务之急。

“汉语信息处理工程”是国民经济和国防信息化建设的重要基础。在这些领域中，80%以上的信息是以语言文字为载体的。这些文本型信息的自动输入和勘校、分类和提取、检索和翻译等等都必须得到“汉语信息处理工程”的技术支撑，为信息的及时有效利用，进一步提高政府及国防安全等部门决策支持系统的智能化程度创造良好环境，它关涉到国家科技进步、国民经济发展、国家的安全、社会主义制度的巩固和两岸的和平统一。对建设有中国特色并与国际接轨的国家信息化基础结构也有重要意义。

“汉语信息处理工程”是一项大型的计算机中文信息处理应用集成系统，它以汉语为主要处理对象，利用现有中文平台，建立包括大规模电子词典、语料库及语言知识库的语言信息处理基础结构(LIPI)，研究自然语言分析与生成技术，开发新一代中文输入输出系统、文本勘校系统、文本分类和信息提取系统，以及信息检索系统和语言翻译系统等等。

四、智能型语词、语句处理技术及产品是语言信息处理产业发展必然趋势

综观我国中文信息处理技术的发展，可大致划分为字、词、句三个阶段。前者属于表层课题(surface layer topic)研发，后者属于深层课题(deep layer topic)研发。中文信息处理技术的发展，必然与人工智能和知识工程相结合朝向深层课题－即自然语言处理方向发展。

当前，语言信息处理的发展趋势是，充分利用迅速发展的计算机技术和计算语言学的最新成果，不断提出新的语言模型。通过采用新的处理策略，建造各种语言知识库，不断提高系统的顽健性(robustness)和对大规模真实文本的处理能力。

然而，从语言工程角度，我们应该看到一种语言理论和计算机模型的提出，直到实用化和商品化系统的开发，中间要经过小规模模型、实验模型、实证模型、和实用模型等数个研究阶段。在每一个研究阶段都要付出相当大的投入。例如，没有大量的语言知识数据作为依托的语言理论是站不住脚的。只有在对大量真实文本数据进行分析和统计的基础上，才能提出可供实用的语言理论和模型。又如，日本通产省发起的国际合作项目多国语言机器翻译工程经过五个国家八年的共同努力，才通过3000语句翻译的实证演示阶段，其成果尚与实用化系统研究、商品化系统开发还有相当距离。因此，我们认为在计算机应用技术当中，语言信息处理属于大型技术（Large-scale technologies），需要国家主管部门作出长期规划，并分阶段实施。

基于计算机汉语理解的智能型语词、语句处理技术涉及到用可计算的方法描述人类语言的思维机制和认知规律，以及语言知识获取和知识表示等等。它是一项长期的研究开发课题，在短期内不可能取得突破性进展。但是，在开展其基础理论和关键技术研究的同时，可以利用不断取得的中间成果逐步提高现有系统的智能程度，改善系统性能。通过推向市场，扩大使用面，经济效益和社会效益的提高，验证计算机中文信息处理技术智能化的必要性，争取国家给以更多的支持。

参考文献

1、建立中国语言工程 推动国民经济信息化进程
——香山科学会议第42次会议全体专家紧急呼吁

2、H. Nomura : Information Extraction and Generation on Information Highway International Symposium on Natural Language Processing, 1995

3、H. Nomura : Language Engineering on the Information Highway International Symposium on Language Engineering, 1994

4、A. Nijholt : Linguistic Engineering: Tools and Products Proceedings of the second Twente Workshop on Language Technology, 1991

5、H. Suematsu : Current Status of the EDR Electronic Dictionary Project and Its Evaluation Research New Generation Computing, No.4, 1994

6、田中康仁 : **自然言語処理を支える知識データ**

中文信息处理应用平台工程

吴 升　陈志明

电子部计算机与微电子发展研究中心

一、引言

近年来，随着计算机软硬件、网络、多媒体技术的迅猛发展，各个发达国家政府部门、军事部门以及国际上在信息处理方面处于领先地位的著名商业公司在继续巩固信息化方面取得的已有成绩的基础上，纷纷着手制定和启动能大大提高自身语言信息处理能力的政策和行动计划，斥巨资对有着较好前途的自然语言理解和处理技术以及应用进行大规模地研究和开发工作。当前信息技术的重点开始从过去重视研究和开发对数据表层的存储、传输和处理等技术向当前重视对隐含在数据（尤其是以自然语言形式表达的）中深层意义的挖掘、抽取和综合技术方向转移 。 这样的变化不仅标志着信息处理技术进入了一个新的阶段、业界向着更加成熟和健康的方向发展，而且折射出我们正在经历着的这个社会的时代特征，即知识和信息是这个社会发展的主要动力。

我国在汉语语言理解和处理方面的研究工作主要起步于“七五”，那时是我国刚刚解决了汉字在计算机上面的输入输出问题的时期，与语言理解和处理有关的基础工作十分薄弱。主要的研究人员均集中在各个高校和研究所，以学习消化模仿国外的先进技术和方法为主。在这一时期，我国首次较大规模地对汉语的词频进行了统计。

1990年 5月在国防科工委和电子部的支持下，在中国工程院院士陈力为教授的直接指导下，电子部计算机与微电子发展研究中心 （CCID）中文信息处理开放实验室（CIPOL） 联合清华大学、北京大学、北京语言学院、东北大学等国内多所著名大学和研究所，正式启动并实施了旨在研究汉语自然语言理解和处理技术中的基础性、关键技术，为各种实用化的汉语自然语言理解和处理系统建立公共应用平台的工程项目（简称“９０５”工程）。董振东和黄昌宁教授是工程总体组领导成员，本项目的高级顾问。

通过五年的工程实践，现已取得了多项实用成果，积累了许多宝贵经验，并培养了一支从事汉语信息处理研究与应用的专业技术队伍。

具体成果如下：

1. 在知识库的基本建设方面，建立了从句法、语义和事件框架等多个角度对词汇进行描述，属性多达上百个，规模巨大的电子词典，其中：

    语法词典　51,000 条
    语义词典　43,000 条
    谓词框架　4,300　条

2. 针对现代书面汉语中出现的常用句型，制定了便于语言学家使用的规则描述语言，并以此为

工具建立了汉语分析规则　2,000条

3. 为汉语自动分析服务，经过分词、词性标注处理的大型语料库 5,000,000 字

在上述工作的基础上，获得了对汉语分析有意义的统计数据，开发了对汉语自动分析中较具有普遍意义的一些基础性、关键性技术和软件系统。

4. 精度分别为99%和96%的汉语自动分词和词性标注技术和实用系统
5. 基于复杂特征集和合一运算的汉语自动分析技术和实验系统

和两个和具体应用相关的实用和实验系统：

6. 实用化的 TIR 全文检索系统和智能型全文检索实验模型。

TIR全文检索系统已在国家部委机关和直属机构、部队系统，以及中国计算机行业协会、计算机用户协会等部门应用并受到好评。

7. 通用人机接口和问答实验系统

二、项目介绍

2.1 任务概述

. 目的：研究汉语理解和处理中的基础性、关键技术，为计算机系统建立一个汉语理解和处理的应用平台，以便支持各类实用化的汉语理解和处理系统。如全文检索系统，信息自动抽取系统，汉语文本自动分类系统，自动文摘系统，自然语言接口系统，机器翻译系统等等。

. 年限：1991年至1995年

. 目标：

a. 开发一个大型的汉语理解和处理用通用词典库
b. 开发具有广泛覆盖面的现代汉语语法规则系统
c. 建立支持汉语理解和处理的现代汉语语料库系统
d. 开发汉语理解和处理的基本软件
   - 现代汉语自动分词与词性标注系统
   - 现代汉语分析理解系统
e. 开发汉语理解和处理的支撑环境
   - 词典库管理系统
   - 规则库管理系统
   - 语料库管理系统
f. 智能型全文检索实验模型
g. 通用人机接口和问答实验系统

规格说明

系统输入：汉语句子.
系统输出：句子的短语结构和句法成分树以及广义语义网络.
分析算法：句法制导，合一约束的不确定性算法
功能　：分析现代书面汉语句子的句法结构和语义结构

2.2 运行环境

2.2.1 硬件环境

. IBM-PC 及其兼容微机

2.2.2 软件环境

. 中文版 Windows

2.3 基本设计概念和处理流程

2.3.1 基本设计概念

该系统的设计充分吸取国际上语言信息处理领域比较活跃的理论和技术，继承国内"七五"期间的研究成果，考虑到汉语自身的特点，将理性主义(基于语言规则)和经验主义(基于语料库)的处理技术结合起来，试图寻求解决汉语语言信息处理的较好途径.

技术特点概括如下:

a. 采用复杂特征集对汉语词汇的句法和语义以及语用信息给以充分描述
b. 采用规则描述语言来建立汉语的规则体系. 突出通用性、陈述性和形式化
c. 经验主义和理性主义相结合，解决汉语的歧义判别问题.

c.1 分词和词性标注时歧义的自动判别
采用基于语料库面向统计的技术。其统计模型为"二元语法"模型。

c.2 结构歧义判别
采用基于规则的方法. 引入合一算法，充分运用句法和语义信息来判别歧义结构。

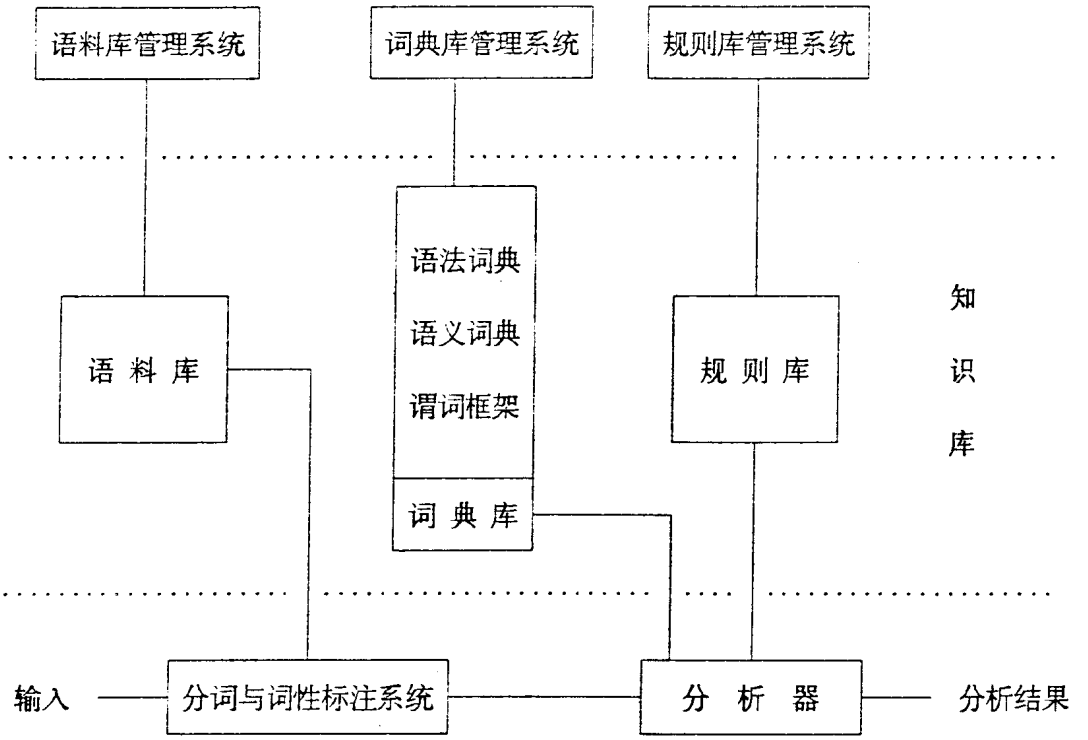
d. 系统按功能划分模块,各个模块自成体系,又互相联系。

2.3.2 处理流程

(1) (2) (3) (4)
输入 --> 自动分词和词性标注系统 --> 句法，语义分析器 --> 输出

2.4 系统结构

汉语理解和处理的支撑环境



汉语理解和处理的基本软件

（系统结构图）

2.4.1 现代汉语语料库系统

a. 功能

语料库建立

收集汉语各种文本，分类建立汉语文本数据库

. 语料库管理系统: 提供库存语料的字表、词表索引和上下文索引(KWIC)以及可以进行型/标比、字频、词频和句长等内容的统计。
. 自动分词系统　　: 对文本实行词切分。
. 词性自动标注系统: 为分词后的文本标注上词性.
. 知识获取系统　　: 从语料库中获取各种语言知识

2.4.2 汉语理解和处理用通用词典库

2.4.2.1 现代汉语语法词典

a. 功能: 充分描述词条的词法和句法信息。
b. 规模: 5万义项.
c. 原则和方法
封闭性词类收全,开放性词类优先选收使用频度高或有代表性的词语,收录的每一个词条,尽可能详尽地描述其词法和句法属性。采用关系数据库的技术分别按词性建库。

2.4.2.2 现代汉语语义词典:

a. 功能:

. 语义分析词典　　: 建立现代汉语事物类和性状类词语的静态语义分类体系,对每个词条的义项给出语义类别及语义特征的描写。

. 谓词语义组合词典: 对每个谓词给出详尽的谓词框架描写。

b. 规模: 给出语法词典中词语的语义描述, 约5万条。

c. 信息描述方法

采用分类描写和属性描写相结合的手段.针对不同类型的概念采用不同的信息描述表(复杂特征集的属性表)。

. 对于事物类的描述 : 基本原则是在研究共同语义特征的基础上,建立义类体系,对于不宜归类的语义特征采用复杂特征集描述,即"分类+特征描述"的方法。

. 对于性状类的描述 : 基本原则是按其所描述的属性分类,同时用语义指向指出其可描述的主体。

. 对于运动类的描述 : 则采用谓词框架来描述运动类概念的语义搭配框架。

2.4.3 现代汉语语法规则系统

a. 功能: 提供汉语的语法体系及语法组合中的语法、语义约束信息。

b. 原则和方法:

. 原则
  . 通用性 不依赖于特定的系统，自成体系.
  . 陈述性 规则的描述方式是非过程性的.
  . 形式化 规则的描述具有规范的格式.
. 方法:

运用单一标记(词性标记和短语标记，如N、V、 NP、 VP等)来描述短语和句子归并的层次结构，并且引入复杂特征集的合一概念，充分利用句法属性和语义属性对单一标记过强的生成能力加以约束，来解决汉语短语和句子的歧义及其内部语法和语义关系问题，采用规则描述语言来体现规则的形式化描述。

2.4.4 现代汉语自动分词与词性标注系统

a. 功能: 对原始汉语文本进行分词和词性标注。

b. 原则和方法:

采用基于语料库面向统计的技术。其统计模型为"二元语法"模型。

2.4.5 现代汉语分析理解系统(简称分析器)

a. 功能:

通过对输入句子的分析得到该句子意义的一种形式化表示，即输出的是句子的语义网络。

b. 原则和方法:

继承形式语法已有的成熟算法，针对单一标记的不足，引入复杂特征集的合一运算。

2.5 接口设计

. 分词系统的分词应符合分词规范及通用大词表的规定
. 语法词典的收词也应符合分词规范及大词表
. 词性标注系统所标注词性与语法词典所定词性及代码应完全一致
. 语法词典应依据语料库系统提供的词表描述词条的语法属性