“中文信息处理平台(CIPP)”工程

李东　董振东　黄昌宁

1993.9.2

1. 引言

“CIPP”工程的目标在于开发中文信息处理的平台，以便支持各种中文信息处理应用系统的研究和开发。

该工程是国家“八五”计划(1991－1995)的科技攻关项目。由电子部计算机与微电子发展研究中心(CCID)组织国内该领域著名的专家、学者和单位共同实施。参加单位包括：清华大学、北京大学、北京语言学院、东北大学、河南财经学院等。

CCID 高级顾问，中文信息学会理事长陈力为教授是该工程的高级顾问。

关键字：

复杂特征集(CFS)：属性/值对的集合。

短语结构树(PST)：短语结构语法生成的一棵句法树，该树表示句子中词从左到右的顺序以及词组成短语和句子的层次结构。

语义网络(NET)：表示句子中概念之间以及事件之间的语义关系的一种超网。

谓词框架(VFR)：描述概念的语义组合“潜力”，包括概念在语义组合中所选择的语义角色及其语义上的选择性限制(又称为语义约束)。

规则描述语言(RDL)：一种描述自然语言句法规则的形式化体系。

语料库：支持语言信息处理的汉语文本数据库。

生语料库：磁盘形式的未经加工的真实文本库。

熟语料库：对生语料库进行不同深度的加工所得到的带有一定的语言知识的语料库。

句法词典(SYN)：记录每个词条的词性和句法组合信息。

语义词典(SEM)：记录每个词条的语义分类和语义特征信息。

规则(RUL)：用规则形式描述短语及句子的结构类型及关系类型。

语料库管理系统：支持语料检索统计的软件。

自动分词系统：在给定文本中把汉字串自动切分为词串的程序系统。

自动词性标注系统：在给定文本中自动判定每个词的词类及其次范畴的程序系统。

分析器(PARSER)：根据词典和规则运用一定的分析算法来得到表示输入句意义的语义网络。

后编辑系统：解决分析器的遗留问题，如歧义判别等。

词典维护系统：句法词典和语义词典的代码化及合成，以及支持词条信息查询和增、删、改等功能的管理软件。

规则管理系统：完成规则的编辑、检索及编译等。

2. 总体设计

2.1 任务概述

- 目的：研究中文信息处理的基础性关键技术，为我国计算机系统建立一个中文信息处理的平台，以便支持各类实用化汉语信息处理系统的开发。如：智能化全文检索、文电自动标引、自动文摘系统、自然语言接口系统、机器翻译系统等。
- 年限：1991——1995 年
- 目标：

a. 开发一个通用的大型的信息处理用汉语词汇信息库(Lexicon)

b. 开发具有广泛覆盖面的汉语句法语义规则体系(Rule Base)

c. 建立支持信息处理的汉语语料库(Chinese Corpus)并对语料库进行加工。

d. 开发中文信息处理的基本软件

- 汉语自动分词系统
- 汉语词性自动标注系统
- 通用汉语分析器

e. 开发中文信息处理的支撑环境

- 词典维护系统
- 规则管理系统
- 语料库管理系统

- 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

- SUN 工作站
- 微机

2.2.2 软件环境

- UNIX
- XENIX

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 词性兼类判别

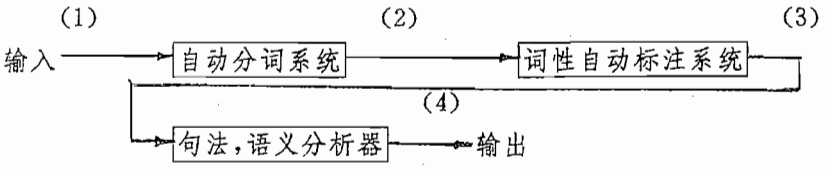
采用基于语料库面向统计的技术(运用"二元语法"模型)。

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。

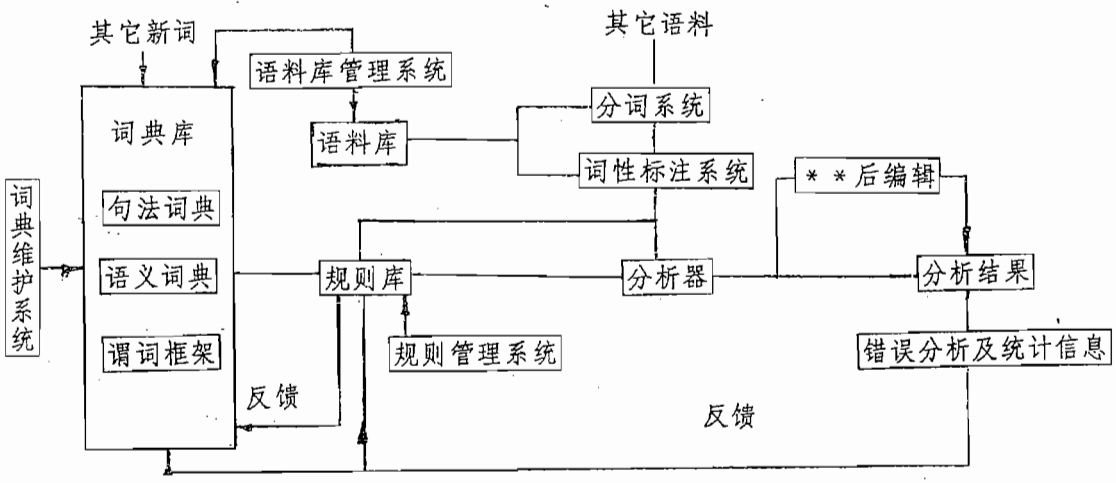
2.3.2 处理流程



（图 1:处理流程图）

[附录 1]为一个句子的处理流程示例。

2.4 结构



（图 2:系统结构图）

[注] “* *”表示待开发的内容

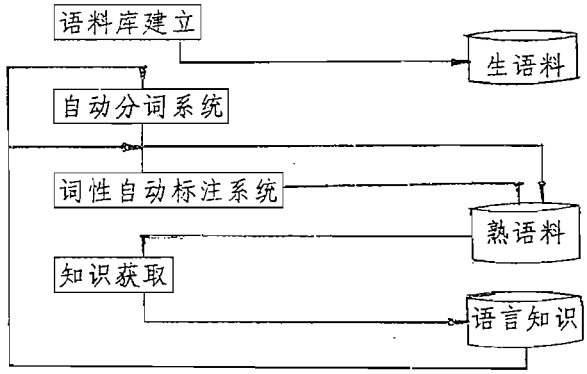
2.4.1 语料库

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。

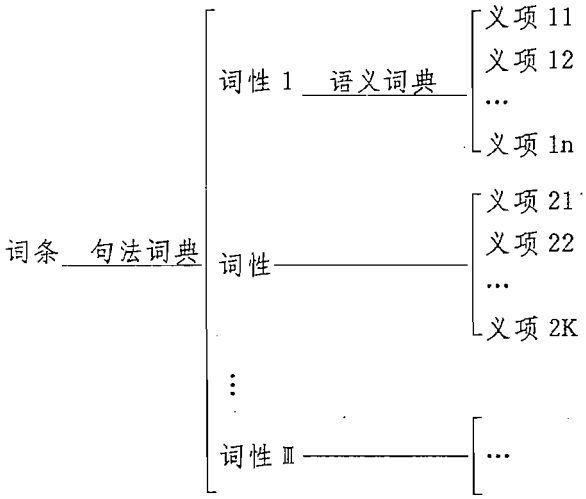
b.结构



（图 3:语料库子专题结构图）

2.4.2 词典

词典总体结构



（图 4:词典结构图）

2.4.2.1 句法词典

a. 功能:充分描述词条的词法和句法信息。

b. 规模与结构:

规模 5 万条,结构见图 5。

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。

[附录 2]为词条描述示例 2.4.2.2 语义词典:

a. 功能

- 语义词典:建立现代汉语事物类和性状类语词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。
- 谓词框架:对每个谓词给出详尽的谓词框架描写。

b. 规模与结构

句法词典中词语的语义描述

结构见图 6

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)。

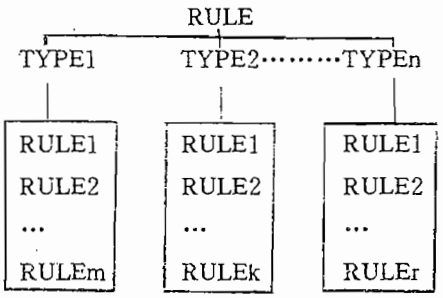
- 对于事物类的描述:基本原则是在研究共同语义特征的基础上,建立义类体系,对于不宜归类的语义特征采用复杂特征集描述,即"分类+特征描述"的方法。
- 对于性状类的描述:基本原则是按其所描述的属性分类,同时用语义指向指出其可描述的主体。
- 对于运动类的描述:则采用谓词框架来描述运动类概念的语义搭配框架。

[附录 2]为词条描述示例

2.4.3 规则

a. 功能:提供汉语的句法体系及句法组合中的句法、语义约束信息。

b. 结构:



(图 7:规则库结构)

其中 RULEi 的一般格式为:

规则号〈上下文无关文法〉:

〈属性约束〉,

〈属性传递〉。

c. 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

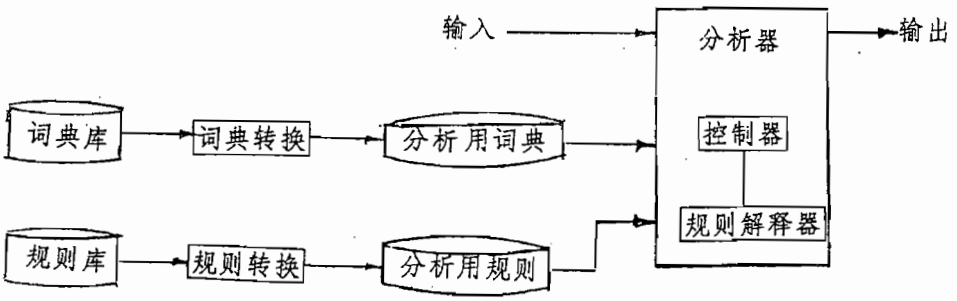
运用单一标记(词性标记和短语标记,如 N、V、NP、VP 等)来描述短语和句子归并的层次结构,并且引入复杂特征集的合一概念,充分利用句法属性和语义属性对单一标记过强的生成能力加以约束,来解决汉语短语和句子的歧义及其内部句法和语义关系问题,采用规则描述语言来体现规则的形式化描述。

[附录 3]为规则描述示例

2.4.4 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. 结构:



（图 8:分析器结构）

c. 原则和方法

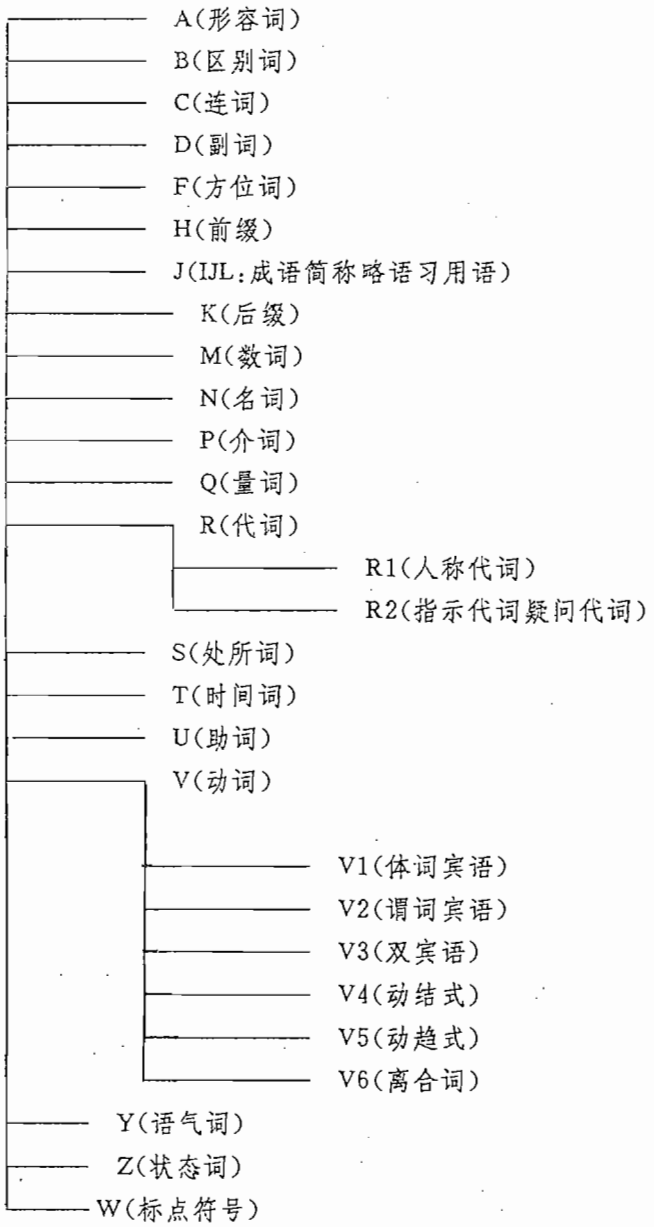
继承形式语法已有的成熟算法，针对单一标记的不足，引入复杂特征集的合一运算。

3. 结束语

“CIPP”工程从 1990 年 5 月正式立项。经过 1990 年一年的预研工作，于 1991 年起已进入工程实施阶段。该项目采取技术方案总体设计，按功能划分子专题进行分散开发的模式(参见 2.4 系统结构图)。目前各子专题都在按预订计划和进度顺利实施，并已有一些可喜的阶段成果。如现代汉语语词的静态语义分析体系和一定规模谓词框架的建立，在汉语语义形式化方面有一定突破。汉语短语和句子的句法和语义规则的形式化体系描述已有数千条，而句法词典的词条已有数万条。这些都为汉语的计算机处理奠定了坚实的基础。

在汉语分析的理论和技术上我们进行了一定的探讨，将目前国际上比较活跃的两种方法，经验主义(基于大规模语料库分析)和理性主义(基于语言学理论)相结合，以寻求汉语分析的有效途径。目前比较有代表性的成果有汉语词性自动标注系统，它采用基于语料库面向统计的方法。具体讲采用“二元语法”模型，113 个标注集较成功得解决了汉语的词性兼类判别问题，标注的准确率达 97%。从而减轻了分析器的负担(规则量得以减少)。然而，由于该项工程是基础性的平台研究和开发，加之汉语计算机处理的基础比较薄弱，在开发过程中有许多遗留的和崭新的问题有待解决。如汉语语义结构的形式化描述以及语义关系的自动分析，汉语歧义的自动判别以及分析器效率的提高等等。都是我们要进一步研究和解决的关键问题。

图 5:句法词典的库结构


G(总库)
A(形容词)
B(区别词)
C(连词)
D(副词)
F(方位词)
H(前缀)
J(IJL:成语简称略语习用语)
K(后缀)
M(数词)
N(名词)
P(介词)
Q(量词)
R(代词)
R1(人称代词)
R2(指示代词疑问代词)
S(处所词)
T(时间词)
U(助词)
V(动词)
V1(体词宾语)
V2(谓词宾语)
V3(双宾语)
V4(动结式)
V5(动趋式)
V6(离合词)
Y(语气词)
Z(状态词)
W(标点符号)

[附录 1]

[例]:

(1)会议正在进行。

(2)会议正在进行。

(3)会议(N)正在(D)进行(V)。(W)

(4)

Tree:

```
Sp(会议正在进行)
|
|—(Hed)——→Vp(进行)
|              |
|Hed)——→Vm(进行)
|          |
|          |—(Hed)——→V(进行)
|          |
|          |—(Adm)——→D(正在)
|          |
|          |—(Sbj)——→Np(会议)
|                        |
|                        —(Hed)——→N(会议)
```

Net:

```
会议正在进行(,jin4xing2,321,)
|
|—(Exp)——→会议(,hui4yi4,112,)
(,)
```

[附录 2]

[例]

SYN

词语=会议
粘着=自
子类=a
个集=个
个体量词=次
前代的=你们

SEM

词语=会议
词类=n
语义类别=事情
代码=112
……

VFR

词事=进行
词类=V
义类=促进
施事=+(人类)
当事=+(事情|活动)
……

[附录 3]

[例]

Np←——Ap　　Np1：	<=[短语不生式]==
⟨ApCtrCset>=[Hed,Adm], <NpiCtrHedSyn 词类>=N, <ApCtrHedSyn 定语>=～'否', <ApCtrAdmSyn 子类>=1, <ApCtrHedSem 语义类别>;	<=[约束条件部分]== ——→短语结构约束 ——→词汇句法属性约束 ——→子成分句法属性约束 ——→词汇语义属性约束
<NpCtr>:=<Np1Ctr>, <NpCtrAtt>:=ApCtr>, <NpNetKer>:NpiNerKer>, <NpNet 属性>:=<ApNetKer>	<=[结果传递部分]== ——→短语结构关系获取 ——→短语语义网络获取

图 6:语义词典的体系结构