PROCEEDINGS OF
1983 INTERNATIONAL CONFERENCE ON CHINESE INFORMATION PROCESSING

中文信息处理国际研讨会

论 文 集

(2)

中国中文信息研究会和联合国教科文组织联合举办

1983年10月12日至14日于北京

CO-SPONSORED BY CIPSC & UNESCO
BEIJING, PEOPLE'S REPUBLIC OF CHINA
OCTOBER 12-14, 1983

数据加载失败，请稍后重试！

第 二 集

目 录

汉语语言的理解系统

吉林大学计算机科学系　**王锡龙**

摘要：1.从汉语语句中词的基本要素：词形(X)、词音(Y)、词义(I)和词性(N)出发，给出了理解语言的其他三个信息：词域(U)、前位属性(Q)、和后位属性(H)。从而构成了理解语言的全部信息量(记为W)：

$$W=\{X,Y,N,U,Q,H\}。$$

于是，语句的理解式L_x可表示为：

$$L_x=\{V\mid \varphi(W_1,W_2,\cdots,Wn)\}。$$

其中，V为语句x中使得条件$\varphi(W_1,W_2,\cdots,Wn)$成立的所有词V的集合。

2. 进一步给出了分析语句结构关系(语法和语义)的树型理解图。并给出构造理解图的算法。

3. 最后，给出了限定型语言理解系统的结构框图，并举例说明系统的功能。

由于计算机的广泛应用，迫切需要解决用汉字使用计算机和用计算机处理汉语语言的问题。本着这个目的，我大胆地提出一些看法并给出一些结果，借此与语言学家和计算机科学事业的同行们进行探讨，以便促进我国的语言自动化和计算机的普及与应用。

一、语言的表示和信息量

任何一种语言（自然语言和计算机语言）都是定义在某一个符号集上的无穷多个有限的符号行(串)集。若满足某种语言规范条件的符号行（称为语句）时，则语言就是定义在符号集上的无穷多个语句集合。一般地，可用如下形式表示[1]：

$$L=\{x\mid \varphi(x)\}。$$

其中，x是代表“使得条件$\varphi(x)$成立的所有语句x的集合”。

事实上，任何有穷多个符号集上的语句x的集合，都是可数无穷多集。语言只是这个集上的子集。而且，并非是可数无穷集[2]。由于语言的复杂性（语法的和语义的），我们只能给出某些有穷表示，用来辅助计算机理解（而不是识别）自然语言。

1. 符号集

汉语书面语言（社会科学的与自然科学的）材料，一般使用如下符号：

〈字〉：指方块字（目前，计算机采用汉字编码输入）。字是形和音结合的最小单位。计算机只能采用有限多个字。如字典字。所有字典中的字记为Z。

〈字符〉：字母（如拼音字母与英文字母）和有限个符号（如星号“*”，求和号“Σ”等），记为M。

〈标点〉：语言的停顿符号，即汉语标点符号，记为B。

若汉语语言的符号集用字母“V”表示时，则有：

$$V=\{Z,M,B\}。$$

由于语言的最小单位是词，词才具有形、音、义、性四个基本要素[3]，才是理解语言的最小单位。若所有的词(系指某词典词)用符号“C”表示，则语言的符号集V*,可表示如下：

$$V^* = \{C, M, B\}。$$

2. 语言的信息量

词构成语句，句构成段（章、节），然后组成文章。由于词是语言的最小单位，句是语言的基本单位，所以，我们还是从词、句的信息讨论入手,从而了解语言的信息量。词的本身要素是词形、词音、词义。词的基本要素是词形、词音、词义和词性。以后分别用X,Y,I,N表示。词性是词在语句中，组词搭配的特性分类，所以，词性对于理解语句尤为重要[8,9]。

由于汉语缺少形态变化，在许多语法、词法分析上，都出现了“中间状态”，很难截然分开。这就给计算机的处理带来了很大困难。所以，计算机对汉语语言的处理，至目前为止，只能是限定性的和辅助性的[4,5,6]。然而，几千年来的实践表明：汉语语言是最灵活、最丰富、表述最简明的语言之一。比如，同样一篇文章，用汉语的书面语言表达，一般来说，总比用其它语言文字表达，要明了简短。所以，与其说它不发达，不如说它更高级，理解起来更具有智能性。请看下例语句中，标记词的基本要素。

例一

①今天大家都到校园里种花。 zhǒng huā	“种花”为栽培花草之义。这是由于语句环境，特别是“校园里”给定的。
②医生给小朋友种花。 zhǒng huā	这句的“种花”与①中是同形、同音、异义词。“种花”是取义“种牛痘”方言。
③这是一株种花。 Zhǒng huā	这里与①②中的种花已经是同形、异音、异义、异性词，指留籽儿的花。
④这种花真好看。 Zhǒng huā	此句中，“种”已是量词，与这字结合成指示代词。同时，读音也与①②不同,虽与③同音而义不同。请注意!这里“好”还有hǎo与hào之分。
⑤“种花还没嫁人吧？” Chóng huā 种大嫂关心地问。 Zhǒng	种字的有两个异体字，即繁体字。它们为异形、异音、同义、同性字。此句都为姓。种花是位女人，因为有“嫁人”一词。当然，如不作标记，种花肯定是词典外词，计算机不会识别出来。种chóng与种zhǒng音也不易分清。
⑥种花两家都富起来了。 Chóng huā (Zhòng)	因为有后边的“两家”一词得知是姓种的和姓花的两家富起来了。而不是种花草的两家富起来了。若“种花”与“两家”中有“的”字，那就取后一种含义了。
⑦这钱中花。 Zhōng huā	这句中，因为与“钱”字搭配，立即读Zhōng huā，而不会读Zhòng huā。“花”字在此句是动词，指耗费。与①～⑥中的花含义都不同。

通过上例①～⑦可见，词的基本要素都具有多值性。然而，给定一个语句后，一般来说，人是可以很快理解的。汉语中的词具有如下性质：

性质1.有词必有义，有义必可用词来表达[7]。

性质2.词在语句中的基本要素是客观确定的。

性质1说明了词汇的丰富性、能产性和完备性。性质2说明了语言的可理解性。但是，对于计算机来说，就不那么容易判定和理解了。我们必须限定某些条件，给出计算机理解语言的某种算法，让计算机一步一步的去判定，去理解。通过例一可看出，只从词的基本要素理解语句是不充分的。还需具备一些其它必要信息。下面引进一些新的概念。

词域：词所属的科类别，称为词域，记为U。

词域体现了词在语句中的客观环境。比如：数学，物理，化学，天文，地理，医学，美工，经济，……。词典中，应对词加以适当的域标记，以助理解词义、句义。

词位和属性：设语句x为如下形式：

$$\underbrace{\cdots\cdots V_1}_{x\text{的前句}}\underbrace{V_2\cdots\cdots V_i\cdots\cdots V_{n-1}}_{x}\underbrace{V_n\cdots\cdots}_{x\text{的后句}}$$

其中，$V\in V^*$，i称为词V_i的位号，词在语句中的位置称为词位。用i表示。

若V_i与V_j，当$i<j$时，称V_i为V_j的前位词符；V_j为V_i的后位词符。每个词都具有一定的性质，称为属性。V_i的属性称为V_j的前位属性；而V_j的属性，称为V_i的后位属性[10]。前位属性和后位属性分别用Q和H表示。位属性提供了参照前后词的意义来判断某一词的条件，表现了词的动态关系。

综上所述，词在语句中的信息（记为W）可表示为：

$$W=\{X,Y,I,N,V,Q,H\}$$

共为七个信息量。

3． 语句理解式

现在，我们可以给出语句x的一般理解表达式为：

$$L_x=\{V\mid\varphi(W_1,W_2,\cdots,W_n)\}。$$

其中，V是使得条件$\varphi(W_1,W_2\cdots,W_n,)$成立的语句x中的全部词符的集合。

同时，我们对一般语言表达式：$L=\{x\mid\varphi(x)\}$的理解也就加深了一步。

自然语言是复杂的，不能简单地靠数学运算求解。我们需要提供给计算机处理语言的一个算法，使得计算机按此算法，对任意给定的一个语句x，找出基本符合语句x中全部词符V的信息量W的统一值。这样，就说计算机理解了该语句。并且，可以回答某些需求结果。下面就来讨论这些问题。

二、理解图和算法

语句结构关系的分析，语言学家已经给出了多种分析办法[11,12]。我们将根据计算机理解语言的需要和可能，选择和确定一种比较合适的办法。

1． 结构关系的表示

下面，我们通过例句，列出几种分析方法，以便比较，决定取舍。

例句：我们厂最近试制出新产品。

方法一：（我们）厂‖〔最近〕试制＜出＞|（新）产品。

符号说明：‖：主语部分与谓语部分的分界符号；

|：谓语部分与宾语部分的分界符号；

=：主语；——：谓语；ʷʷ：宾语；

（ ）：定语；〔 〕：状语；〈 〉：补语[11]。

符号标记简单，但结构关系不清晰。

方法二（见图1）：

树型结构，层次清楚。但是，没有结构关系。方法一与方法二都是语句的语法成分分析，属层次结构型。

方法三（见图2）：

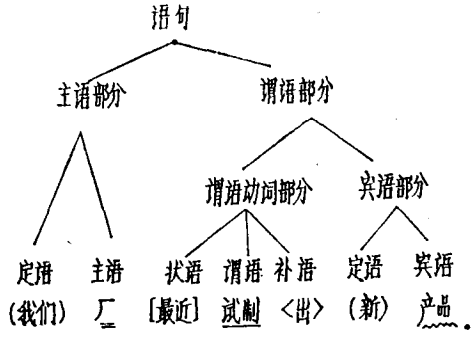

图 1

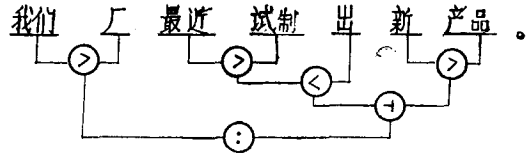

图 2

符号说明：＞表示修饰成分在前；＜表示修饰成分在后；⊣符号表示补足成分在后；⊢符号表示补足成分在前；：是表述关系（即主谓关系）。另外，还有：＋表示并列关系等[12]。

这种方法层次清楚，强调结构关系，即语义关系。

方法四：

〈语句〉::=〈主语部分〉〈谓语部分〉

〈主语部分〉::=〈定语〉〈主语〉

〈谓语部分〉::=〈谓语动词部分〉〈宾语部分〉

〈谓语动词部分〉::=〈状语〉〈谓语〉〈补语〉

〈宾语部分〉::=〈定语〉〈宾语〉

〈主语〉::=厂

〈谓语〉::=试制

〈宾语〉::=产品

〈定语〉::=我们

〈定语〉::=新

〈状语〉::=最近

〈补语〉::=出

这种表示为 BNF 表示法（即Backus-Normal Form 或 Backus-Naur Form 的缩写符）。::=表示“定义为”。这是计算机语言中常用的一种递归定义。显然，对自然语言是不合适的。比如，按上述定义就会有“新厂最近试制出我们产品”的二义性语句。可以说“我

们的产品”，“我们厂”，而不能说“我们产品”。

本系统将基本上采用方法三。重点放在结构关系（即语义理解）上，兼顾语法成份。对于该例句将产生如下的树型结构形式，称为语句k的理解图。

语句 k：

我们	厂	最近	试制	出	新	产品	。
1	2	3	4	5	6	7	8

理解图见图3。

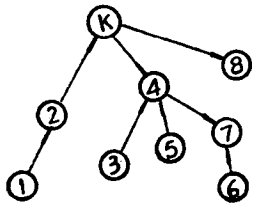

图3　语句k的理解图

其中，k 为语句序号。(k) 为根节点。根节点至多有一条接收线。体现了语句若有主语，则只有一个，否则为无主句。根节点的发射线至少有一个。根节点的直接子女中，主语节点只有一个向根节点的发射线。可以有多个接收线，做为他的附加成分。谓语节点发射线中，若为一个，则是一个宾语节点；若为两个，则为双宾语结构；它可以有多个接收线，做为附加成分节点。当任一子女节点中，接收线节点词的词序号，大于该节点的词序号时，如⑤→④，则为补足成分，其余皆为修饰、限定成份。其余根节点的直接子女，有且只有一个接收线。宾语节点类似谓语节点，附加成分节点的发射线和接收线各为一条。

语句k的理解图按着自顶向下，自左至右的顺序读。从主语节点词开始，读取语句的主要成分。然后，按照语句的词序号读取附加成分。

理解图的每个节点词，由计算机按着需求标记信息量的值。所以，它是一个“果树”型。

2．基本规则

理解图是计算机遵循某些基本规则，执行理解算法产生的。下面给出一部分规则，以便理解算法。

规则1　语句中，几个相邻词，若具有同节点性质时，则称为同节点词。同节点词视为一个词结构。

例如，以下相邻词具有同节点词性质。

(1) x＜方位词＞，其中，x为一词符，方位词结构。

(2) x＜助词＞，其中助词中还包括助动词和语气助词。

(3) x＜量词＞，量词结构。

(4) xx　　，同形重叠词结构，

(5) ＜介词＞x，介词短语。

(6) x（词缀），（　）表示可为词头，词尾，词嵌。

(7) ＜否定词＞＜动＞。

(8) ＜肯定词＞＜动＞。

规则2　语句中，凡起到修饰、限定、补足成分的词，称为附加成分。附加词，按照附加结构式标记。

例如，以下结构式为附加结构式：

＜形＞→＜名＞，＜副＞→＜动＞，＜形$_1$＞→＜形$_2$＞，＜量$_1$＞→＜量$_2$＞等。

规则3　凡对理解图根节点，具有直接子女条件的符号，称为语句的直接子女。

例如：＜叹词＞，＜连词＞，＜标点＞等。＜名词＞，＜动＞，＜代词＞具有直接子女性

词。由于代词的指代作用，一般要特殊处理。对人称代词，一般先按直接子女词对待，行再修剪。

规则4 对树型理解图，实施剪枝、嫁接处理的规则，称为剪嫁规则。剪嫁规则分为语法剪嫁，语义剪嫁和综合剪嫁。

(1) “语法剪嫁”例

格式： $\left\{\begin{matrix}名_1\\代_1\end{matrix}\right\}<动>\left\{\begin{matrix}名_2\\代_2\end{matrix}\right\}$

例二 $\underset{1}{今天}\ \underset{2}{学校}\ \underset{3}{开}\ \underset{4}{大}\ \underset{5}{会}$。 (参看图4)

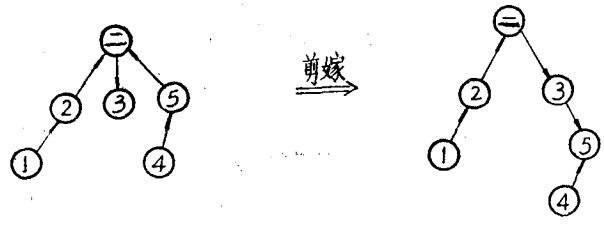

图 4

格式： $\left\{\begin{matrix}名_1\\代_1\end{matrix}\right\}<动>\left\{\begin{matrix}名_2\\代_2\end{matrix}\right\}\left\{\begin{matrix}名_3\\代_3\end{matrix}\right\}$

例三 $\underset{1}{我}\ \underset{2}{问}\ \underset{3}{你}\ \underset{4}{一个}\ \underset{5}{问题}$。 (参看图5)

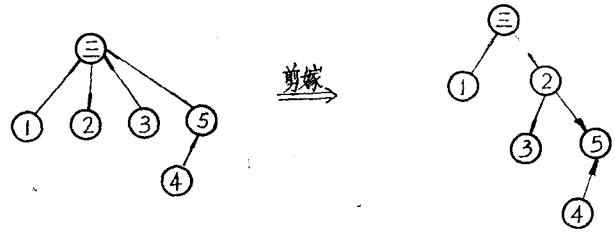

图 5

格式： $\left\{\begin{matrix}名_1\\代_1\end{matrix}\right\}<动_1>\left\{\begin{matrix}名_2\\代_2\end{matrix}\right\}<动_2>$

例四 $\underset{1}{孩子们}\ \underset{2}{听了}\ \underset{3}{故事}\ \underset{4}{很}\ \underset{5}{高兴}$。 (参看图6)

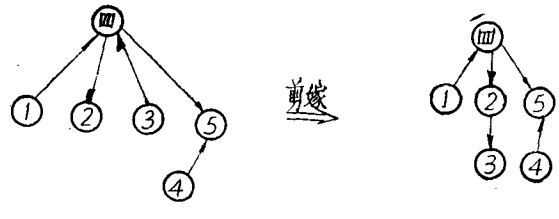

图 6

(2) “语义剪嫁”例

格式：$\begin{Bmatrix}名_1\\代_1\end{Bmatrix}$ $\begin{Bmatrix}名_2\\代_2\end{Bmatrix}$ ‹动›

例五 $\underset{1}{他}$ $\underset{2}{一口}$ $\underset{3}{水}$ $\underset{4}{都}$ $\underset{5}{不喝}$。 （参看图7）

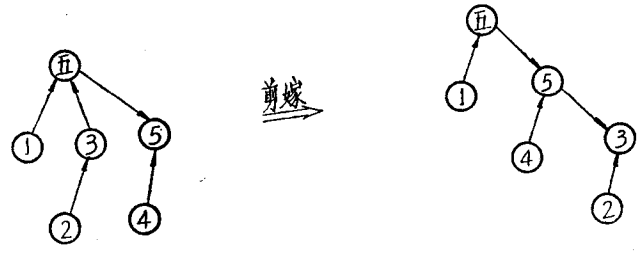

图 7

格式：$\begin{Bmatrix}名_1\\代_1\end{Bmatrix}$ ‹动$_1$› $\begin{Bmatrix}名_2\\代_2\end{Bmatrix}$ ‹动$_2$› $\begin{Bmatrix}名_3\\代_3\end{Bmatrix}$

例六 $\underset{1}{红红}$ $\underset{2}{下}$ $\underset{3}{河}$ $\underset{4}{洗}$ $\underset{5}{衣裳}$。 （参看图8）

图 8

例七 $\underset{1}{大家}$ $\underset{2}{选}$ $\underset{3}{他}$ $\underset{4}{当}$ $\underset{5}{代表}$。 （参看图9）

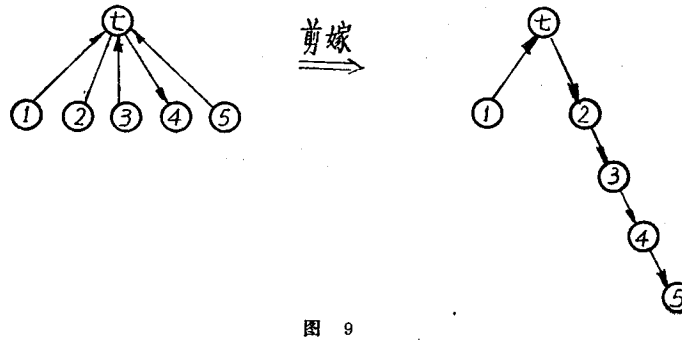

图 9

(3) “综合剪嫁”例

格式：$\begin{Bmatrix}名_1\\代_1\end{Bmatrix}$ （把字式）$\begin{Bmatrix}名_2\\代_1\end{Bmatrix}$ ‹动›

例八　你(1)　把房间(2)　收拾(3)　干净(4)　。（参看图10）

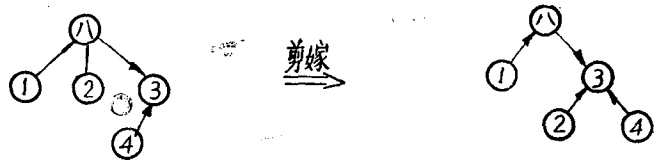

图　10

格式：{名$_1$ / 代$_1$}（被动式）{名$_2$ / 代$_2$}〈动〉

例九　敌人(1)　被我们(2)　消灭了(3)　。　（参看图11）

图　11

3．理解算法

算法：

第一步：取一语句x_k（k为语句号）。若语句全部处理完，则转出口；

第二步：借助词典和知识库进行分词、定词性、编序号[13，14]；

第三步：建立根节点k。依次为名词、代词建立接收指示字；为动词和其它直接子女词符建立发射指示字。

第四步：归并同节点词和处理代词；

第五步：按照附加结构式，建立下一级子女节点；

第六步：根据剪嫁规则，实施剪嫁调整；

第七步：依据需求进行赋值。若取不到值，则转出等待；

第八步：问输出结果吗？是，则输出；否则转第一步。

三、系统框图及应用

1．系统结构

见语言理解的系统结构框图。

共分四部分：

①系统管理部分

控制语言理解系统的管理程序和维护程序，它在操作系统和数据库管理系统支持下[15,16,17]。

②预处理部分

· 语言文本输入程序；

· 分词处理程序；

· 查阅词典和知识库程序；

· 建立静态库程序等。

所谓静态库，是根据需求，将文本的有关信息量，由词典和知识库中取出，并以某种结构形式存入数据库中。以防止多次扫描词典和知识库，用以提高处理速度。

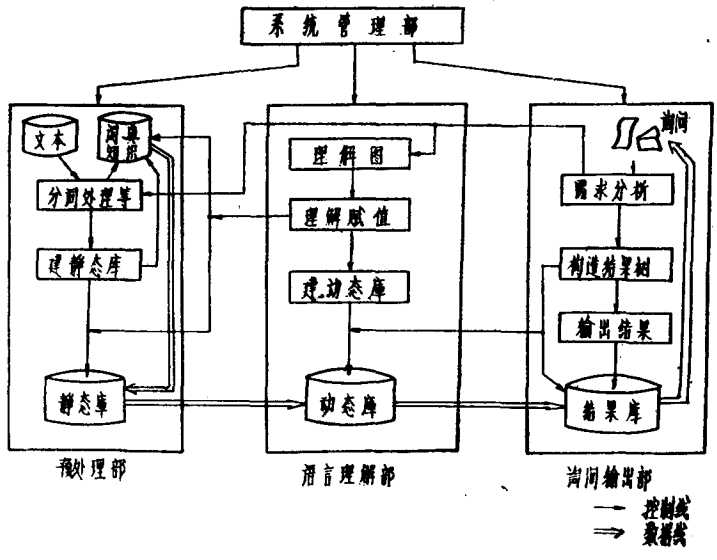

图 12　语言理解的系统结构框图

③语言理解部分

· 语言理解分析程序；

· 建立理解图；

· 赋值程序等。

使赋值后的理解图存于动态库中，以便提供给输出。

④询问输出部分

· 需求分析控制程序；

· 按照需求，从动态库中将结果以输出格式存于结果库中，提供给输出。

· 输出管理程序等。

2. 应用举例

例十　求语句“小毛板着脸不说话”语句中的词音L〔Y〕和第二个词的词性L〔N_2〕。计算机对该语句实施处理后，输出格式如下：

L〔Y〕：xiǎo máo bǎn · zhe liǎn bù shuō huà

L〔N_2〕：板着——动词。

本系统的实现，将对计量语言学、机器翻译、人机对话和计算机辅助教学等有着一定意义。由于篇幅所限，其它应用举例和提问语言省略。

本文曾得到管纪文、孙怀民、张　普和谷新英同志的鼓励与指导，在此表示感谢。

参 考 文 献

〔1〕John E. Hopcroft Jeffrey D. Ullman, FORMAL LANGUAGES AND THEIR RELATION TO AUTOMATA, Addison-Wesley Publishing Company 1969.

〔2〕David Gries, COMPILER CONSTRUCTION FOR DIGITAL COMPUTERS, John Wiley & Sons, Inc.

〔3〕“实用词汇新知识”，谭全编著，商务印书馆，1978年3月初版。

〔4〕Schank R. C. & Colby K. M. ed, Computer Models of Thought and Language, W. H. Freeman Co. 1973.

〔5〕Arens, Y., Using Language and Context in the Analysis of Text, IJCAI 7(1981), P. 25.

〔6〕Wilensky, R., A Knowledge-based Approach to Language Processing: A Progress Report, IJCAI 7(1981), p. 52.

〔7〕“汉语词义简析”，朱星，湖北人民出版社。

〔8〕“拼词法和构形法”，张寿康，湖北人民出版社。

〔9〕“词汇知识”，高文达、王立廷编写，山东人民出版社。

〔10〕“用属性文法进行编译”，P. Rechenbrg，浙江大学周炳生译自“Prinzipien des Übersetzerbaus”。

〔11〕高等学校文科教材“现代汉语”，黄伯荣、廖序东主编，甘肃人民出版社。

〔12〕“汉语语法分析问题”，吕叔湘著，商务印书馆，1979年6月第一版。

〔13〕“结合上下文进行辅助分词的学习系统”，吉林大学管纪文、谷新英，会议论文集。

〔14〕“分词词典与知识库”，吉林大学，王锡龙，中国中文信息研究会第二次年会论文。

〔15〕ARTIFICIAL INTELLIGENCE, P. H. Winston 1977.

〔16〕An Introduction to Database Systems, C. J. DATE.

〔17〕Principles of Database Systems, D. Ullman.

结合上下文辅助分词的学习系统

吉林大学　　管纪文　谷新英

摘要：随着自然语言自动化处理中有关工作的进展，解决书面汉语分词自动化的问题日趋迫切。

本文提出一个计算机辅助分词系统，它采用按表核对和参照上下文推导相结合的处理方式，可以在一定程度上独立地解决分词问题。但是，为了保证处理结果的正确性，适当的人工干预是必须的。该系统具有学习功能，故其性能可以在使用过程中不断得到改善。

文章阐述了系统的设计思想和工作原理，并给出了“分词处理程序”的算法。

一、引　　言

本系统的设计立足于现实性、可靠性、通用性、智能性和自完善性。

图 1 是系统结构概图。

系统的核心程序是“分词处理程序”，它以词表和知识库为工具，对语言文本实施分词处理，并在适当的情况下接受人工干预。

下面，我们以分词处理过程为主线展开讨论和叙述，逐步地涉及系统结构的各个部分。最后，我们给出“分词处理程序”的具体算法框图。

二、文本的预处理

内容完整、结构正确的一部书、一篇文章、一组语句或者一个词组都可以是系统的处理对象，这里简称为文本

在把文本送入计算机之前，要人工地进行若干必要的预处理。预处理包括以下内容：

1）　文本线性化

文本应以线性字符串的形式输入计算机，因而对其中的少数例外情形（如底线）要加以特殊处理。

2）　特殊词预标记

特殊词的预标记与系统使用动态词表有关。词表是词的集合，但它还包含一些非词汉字（如：葡、们）和一些非汉字字符（如英文字母、标点符号）。语言的词汇是一个无限集合，任何词表都不可能包容全部词汇。为了使系统具有较好的通用性，同时又不致造成过高的存储和检索代价，我们把词表设计成动态的：一个相对稳定的部分叫作基本词表，保存着一些常见词；另一个比较灵活的部分叫作临时词表，保存着一些用途偏窄的特殊词。

临时词表内容的增生主要依赖于特殊词预标记工作。所谓特殊词预标记，就是由人工事先将文本中那些可能不在词表的词的第一次出现给以特殊标记。系统在处理时若发现果真如此，则把该词加入临时词表。当临时词表增大到一定的限度或对新的处理对象已不适用时，可将其全部或部分地删除，或者转储备用。

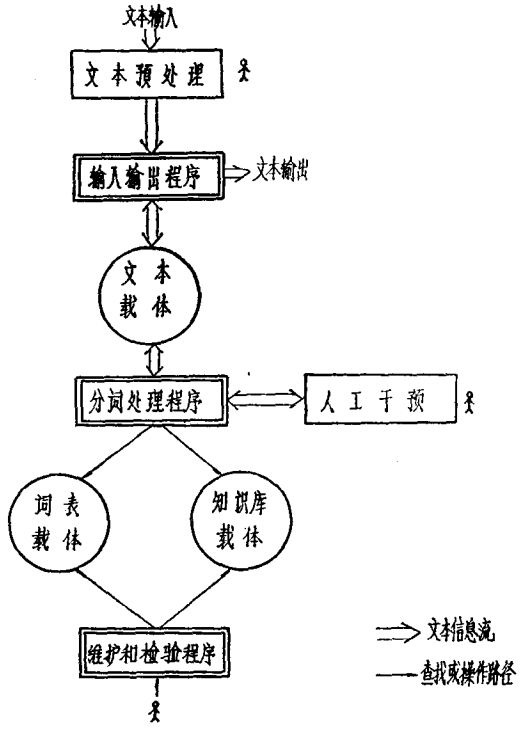

图 1　系统结构概图

特殊词的人工预标记可能有遗漏，这可在随后的人工干预中得到弥补。

3）　含有非汉字字符的词的预标记

后面要谈到，分词处理的第一步是文本切割。由于所有非汉字字符都被列为切割标志，故若不预先标记含有这种字符的词（其例如：一二·九、阿Q、维生素PP），则将造成错误的切割。

4）　文体结构的标记（如章、页）是为了跟踪和输出的方便。

三、文本的切割

对于已经存储在计算机上的文本，系统处理的第一步是把它切割成词段（纯粹由词组成的文本片断），然后针对每个词段进行分词。这里需要指出，在算法中，这两个步骤从整体上说是并行实现的。

显然，切割便是粗分。词段切割得越短，对随后的工作越有利。但是，把切割的难度限制在较低而又较为一致的水平上是合适的，复杂的工作应该集中到词段分词中。

基于上述原则，我们选择两类字符作为文本切割标志：一类是全体非汉字字符，一类是一些特殊汉字。其中，后者是具有这种性质的字：它们是某些词的词首或词尾，但不是任何词的其他部分。对于只能作词首的汉字，可从其前面切断文本，故称前位切割标志；对于只能作词尾的词，可从其后面切断文本，故称后位切割标志。但是，由于词汇的可变化性质，认为某些汉字恒具有上述特点是不合适的，于是索性放宽一些，允许这种字有少数的例外情

形。一个有例外汉字切割标志的一个或多个例外词作为一条切割知识保存在知识库中，并由该字的词表条目直接索引。每条知识都是一个逻辑蕴含式而不是等价式。

选择汉字切割标志应综合考虑下列因素：①预选汉字的使用频率；②该字作为前位（或后位）切割标志的例外词个数；③每个例外词的使用频率。一般说来，前者越大，后两者越小的汉字越宜选用。下面开列出在3775个国标一级字中初步选定的574个汉字切割标志。词的确定基本上以〔8〕为依据，选择时参考了王以德等人的工作〔4〕。

汉字切割标志：

（括号中的词是例外词）

1） 前位切割标志（共327个）

啊 阿 挨 哎 唉 艾 俺 按 肮 盎 敖 翱 懊 扒 吧 跋(序跋) 罢（也罢、作罢） 搬 颁 磅 傍 褒 鲍 悲（慈悲） 甭 迸 彼 碧（金碧辉煌） 蓖 陛 卞 扁（侧扁） 憋 濒 摈 丙 秉 玻（毛玻璃） 菠 渤 猜 踩 惭（羞惭） 沧 拆 豺 搀 阐 猖 赤（足赤） 抽 踌 稠 雌 捶 簇 崔 挫 傣 殆 逮 怠 耽 掸 掂 佃（租佃） 惦 碉 凋 掉 跌 盯 丢 恫 陡 杜 兑 吨（公吨、登记吨） 跺 讹 扼 遏 珐 诽 吩 丰 烽 孵 辐 弗 辅 釜 阜 讣 肛 搞 各 巩 蛊 剐 诡 癸 刽 憨 捍 貉 很 葫 蝴 豢 焕 幌 恍 蛔 畸 饥(充饥、朝饥) 激(感激) 讥 汲 伎 既 嘉 佳 艰 兼 缄 僵 骄 侥 狡 揭 皆 腱 粳 纠 韭 拘 狙 沮 撅 攫 抉 咖 凯 慷 磕 棵 颗 挎 夸 （浮夸） 匡 岿 傀 垃（坷垃） 喇 谰 勒(弥勒) 磊 晾 撩 撂 琳 拎 羚 另 馒 谩 枚 孟 弥 藐 哪 你 您 柠 拧 凝（冷凝） 扭(歪歪扭扭、别扭) 疟 糯 哦 沤 啪 趴 徘 磐 盼 庞（面庞、脸庞） 耪 抛 咆 呸 烹 瑟 毗 啤 疲 譬 乒 颇 （偏颇、景颇） 葡 沏 崎 祈 岂 企 迄 掐 牵 乾 谴 呛（够呛） 蔷 撬 钦 沁 琼 泅 痊 瘸 壬 妊 仍 冗 熔 揉 蠕 汝 闰 搔 啥 珊 煽 赡 稍 芍 呻 甚（过甚、太甚） 孰 曙 甩 摔 拴 谁 朔 撕 伺 怂 艘 虽 蓑 他（其他、排他）它（其它、排它） 她 檀 倘 誊 涂 拖 鸵 椭 唾 豌 完（玩儿完） 惋 我（自我、忘我、唯我、大我、小我） 牺 羡 姓（复姓、百姓） 酗 旭 绚 衙 阎 佯 耶 椰 噎 颐 矣 屹 臆 肄 赢 臃 雍 甬 永(隽永) 迂 逾 愉 驾 曰 酝 哉 再（一再） 乍 辗 崭 蟑 肇 遮 这 斟 挣 狰 帧 只（不只） 蜘 秩 皱 拽 卓 琢 苗 啄 咨 兹 滋 仔 棕 揍 最 遵 诅 柞 昨

2） 后位切割标志（共247个）

埃 蔼 隘 坝（坝埝 坝子） 般（般配） 镑 狈 甭 蔽（蔽塞） 痹 敝（敝屣） 柄（柄子） 箔 舶（舶来品） 膊 泊（泊位） 埠（埠头） 怖 睬 踩(踩水) 舱（舱位） 蹭（蹬蹭） 碴（碴儿） 馋 攀（攀肘） 澈（澈底） 忱 骋 匙(匙子) 畴（畴昔） 蹉 础 捶 囱 瘁 粹 殆 惮 蹈(蹈海、蹈袭) 涤 甸（甸子） 睹 妒（妒忌） 盾（盾牌） 刹（刹斧石） 饵 伐 阀（阀阅、阀门） 肪 啡 吠 氛（氛围） 峰 幅 袱 弗 甫 釜 脯 傅 阜 缚 咐 溉 搞 埂（埂子） 菇 剐 貉 阂 涸 很 乎 瑚 壶 唬 猾 徊 痪 簧 凰 卉 诲 圾 箕（箕踞） 姬(姬鼠) 绩 辑（辑录） 己（己方、己任） 悸 际（际遇） 件 涧 椒（椒盐）窑（窑肥）

皆 秸 诫 烬 茎 睛 灸 咎 疚 炬 倦（倦游） 侩 眶 婪 澜 览 佬 蕾 儡 俐 哩 俩 梁 辆 晾 捺 赁 拎 龄 溜 窿 麓 戳 侣 虑 率 峦 挛 仑 纶 氓 寐 们 檬 觅 皿 馁 撵 拧 泞 你（你们） 哦 帕 琶 湃 畔 呸 坪（坪坝） 颇 粕 泣（泣诉） 讫 嵌 呛 橇 撬(撬杠) 沁 擎 榷 刃（刃具） 纫（纫佩） 蓉 蕊 涩 衫 裳 梢（梢头） 婿 蚀 驶（驾驶员） 孰 墅 帅 拴 谁（谁边、谁个） 艘 擞 他（他们、他人、他日、他乡） 它（它们） 她（她们） 蹋 汰 膛 淌 惕（惕厉、惕励） 翁 我（我们） 隙（隙地） 暇 翔（翔实） 屑 墟 讶 呀 堰 殃 奢 疡 漾(漾奶) 矣 亦 谊 诣 绎 邑 哟 泳 佑 盂（盂兰盆） 屿 域 吁 喻 誉 驭(驭手) 垣 曰 哉 蚤 燥 盏 绽 彰 蜇 辙（辙口） 帧 址 帜 炙 盅 州 洲 诌 粥 宙 拽 缀（缀文） 兹 滓 渍 揍 纂

非汉字字符和汉字切割标志的词表条目中有相应的标记，处理程序据此（有时还要涉及一次知识库查找）判断是否应在文本中的一个字符的前面或后面切断文本。

四、词段的分词

词段的分词是整个处理工作的核心。

为叙述方便，下面继续引进几个概念。

词段中的任意一串字符若是词汇中的词（不管它在文本的语义下是不是词），则称之为该词段的形式词。一个形式词若在文本的语义下是词，则称之为实际词。

在下例的词段P 中，所有的形式词下都划以底线，共八个，其中三个划以曲底线的是实际词。

P：发展 中 国家

将一个词段P按词表L分成以形式词为单位，称为按L分解P，其结果叫作 L下P 的一个形式分解。

显然，在特定的词表下，一个词段可以有零个或多个形式分解。例如对于上例的 P，若词表L中包含全部八个形式词，则P有六个形式分解；若L中不包含“发”，则P只有三个 形式分解；若“发”和“发展”都不在L中，则P无形式分解。

若一个词段的某个形式分解的所有分解单位都是实际词，则此形式分解叫作正确分解。

在一个词表下，一个词段可能有也可能没有正确分解。若有，则应是唯一的。对于上例P，下列分解是正确分解：

发展　中　国家

显而易见，对于任意文本，我们总可以通过扩充词表，使对该文本来说，它的任意词段都存在正确分解。对于这样的词表，我们说它对于这个文本是充分的。于是，我们有下述命题。

命题：若对于文本T词表L是充分的，则对于T的任意词段P，至少存在一个形式分解；当P的形式分解只有一个时，该形式分解就是正确分解。

我们对一个词段进行分词，实际上就是要求取它的正确分解。本系统采取穷尽所有形式分解而求正确分解的方法，以保证处理结果的正确性。正确分解的判定首先依据知识库中的