中 文 信 息 学 报
第8卷 第3期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol.8 No.3

一个实验性的汉语篇章理解系统

崔 耀 陈永明

(中国科学院心理研究所)

【摘要】本系统从世界现象的组成和人类的记忆结构特点出发，结合汉语的具体情况，从意义分析的角度将汉语的词汇分为描述性的词、过程性的词、辅助性的词三类。这三类词分别描述了世界现象中的事实、事件以及语言本身所具有的特性。在此基础上形成了汉语的篇章理解所依赖的知识表示和知识组织形式，即：以事实一事件网络为基本结构的记忆模型。通过这个模型建立了汉语篇章理解系统的知识库，以及与之相应的加工和管理机制。系统对汉语篇章的分析是以词为引导进行的。汉语的词直接对应于事实-事件网络中的节点和辅助词表中的词项。这些节点和词项综合了语法的、语义的、语用的知识，并且能根据处理的需要及时地为分析过程提供预期。本系统通过阅读，对自己的知识库进行动态的自我管理。在阅读了有关七种鸟类的汉语故事之后，系统能够学习到有关鸟类的一些新概念，并能回答相应的问题。

汉语篇章理解需要依赖各种知识。这些知识来自语法、语义和语用三个方面[1]。为了使计算机能够更好地处理汉语的篇章，必须对各方面的知识进行合理的组织和管理。由于语言是人们用来描述世界现象，传递信息的工具，对于自然语言理解的研究工作有必要从意义分析的角度进行。意义分析就是找出语言是如何对世界现象进行模拟，进而发现特定的言语活动所描述的有关世界现象特征及其相互关系的过程。本文从汉语的词与世界现象的对应关系出发，就汉语理解系统的建造进行了初步尝试。

一、以词为引导的记忆结构

汉语中很多词与世界现象存在着直接的对应关系，同时，世界现象又存在着静态和动态的两个方面。从这种对应关系出发，汉语的词大体上可以分为三大类：第一类是描述性的词，这类词与世界现象的静态方面相对应。它们可以描述具体的实物，如：动物、植物、太阳等等；同时也可以描述状态，如：时间、空间、生命等等。第二类是过程性的词，与世界现象的动态方面相对应。它们可以描述实物的活动，如：繁殖、飞行、生长、转动等等；同时也可以描述状态的变化，如：长、大、转化等等。另一方面，汉语中也有

①本文1993年12月27日收到

不直接代表世界现象的词，如：的、地、在、也等等。这些词为第三类词，称为辅助性的词。其功能在于标示出不同的语言单位间的相互关系及其在言语过程中构成连贯的、合法的语言单位时所起的作用。

根据上述分析，本系统将词典和存储世界知识的知识库结合在一起，形成了以词为引导的知识结构。因此，本系统的词典结构和知识库的结构是相同的。这个知识结构由节点及节点间的联系组成。根据汉语词汇的上述分类，这些节点可以分成五种：(1) f 型节点：用来表示实物。(2) sf 型节点：用来表示状态。(3) e 型节点：用来表示实物的活动。(4) se 型节点：用来表示状态的变化。(5) a 型词项：用来表示汉语中辅助性的词。

前四种节点的基本结构为：

节点名：X-{ 世界现象的名称}　其中 X={f sf e se}

成　员：{直接的下属节点的名称}

判　断：{用于组词的产生式规则}

属　于：{直接的上属节点的名称}

其中，f 型节点和 sf 型节点还具有性质槽，该槽中填充了实物或状态的特性。e 型节点和 se 型节点则具有形式槽和 e⁻判断槽。形式槽记载语义框架（语义框架的表示方法中使用了 R.C.Schank 提出的元语的基本意义[2,5]）；e⁻判断槽存储了一段用于处理过程的产生式规则。f 型节点和 sf 型节点通过所属关系形成的网络称为事实网络；e⁻型节点和 se⁻型节点通过所属关系形成的网络称为事件网络。事实网络与事件网络间通过主体、对象、行为、地点、工具等关系联系在一起，形成事实-事件网络。a⁻型词项构成一个辅助性的词表，其结构为：

节点名：a⁻ {辅助词}

判　断：{用于组词的产生式规则}

e⁻判断：{用于处理的产生式规则}

事实-事件网络和辅助词表以词作为知识存储和搜索的直接线索。同时，事实-事件网络又将语义网络的思想和 R.C.Schank 提出的概念依赖理论(Concept Dependent Theory)中的语义框架表示方法[5]及动态记忆(Dynamic Memory)中的知识组织原则[4,6]结合在一起，实现了情节性知识和非情节性知识的统一组织和管理。因此，称这样的知识组织结构为以词为引导的记忆结构。

二、以词为中心的汉语篇章处理方法

本系统是在 IBM PC／AT 微机上，使用 Lisp 语言实现的。它采用了类似黑板结构的组成方法。黑板部分包括三个层次：词切分表、处理表、意义表。知识源由事实--事件网络和辅助词表构成。由于事实--事件网络内部具有较紧密的联系，并且每一个节点都表示了一个或若干个词，包含了有关的语法、语义及语用知识，因而这个知识源与典型的黑板结构知识源是不同的。网络内每个知识单元都不是相互独立的。为了能够使系统通过阅读故事学习到一些新概念，事实--事件网络不是固定不变的，可以根据动态记忆的管理方法进行重构及更新。

系统的处理和管理机制与知识的组织方式相配合。其基本策略是将汉语篇章理解作为

一个以词为中心的意义分析过程。在这个分析过程中，强调词与世界现象之间的对应关系以及词与词之间的搭配形式，并充分利用对当前处理的结果和已有的知识，对下面的处理环节做出预期。意义分析过程的结果是生成一个对整个故事所描述内容的意义结构。系统可以根据这个意义结构中提供的事实与事件之间的联系，把它整合到事实--事件网络的适当位置上去。

本系统基本结构如图 1 所示:

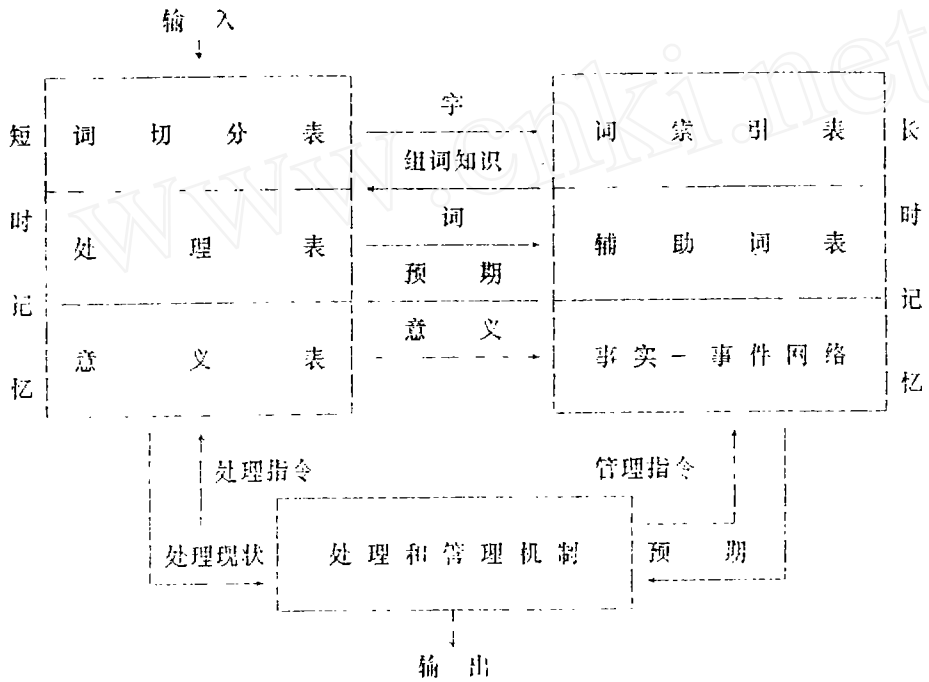

图 1　系统基本结构

词切分是第一个加工环节。为了切分的方便，在系统中设立了一个词索引表，其中记录了事实--事件网络和辅助词表中全部的词。处理过程中，汉语的文字材料从系统的输入端以自然的书写方式，一个字一个字地连续输入。每读到一个字，系统就搜索词索引表，如果找到该字，则从事实--事件网络或辅助词表中对应的节点或词项中取出判断槽的值。这个值是一段程序，用产生式规则的方式描述了以该词为中心形成双音节或多音节词的方式。例如:"群"在事实--事件网络中为一个 se 型节点，其结构如下:

```
(se-群 形式 ((pp ?pp) (state 数量( >5) 秩序( <5) ))
判断 ((read-word)
      (cond ((equal char'众)
             (setq char (read-from -string (string-append "群"char))))
            (t (setf in-stream (make-string-input-stream
                                 (string append char (read in-stream))))
               (setq char'群)))
      (when (factp (isa (first * parsing-list * )))
```

```
(setq char (read-from-string (string-append
                                (first * parsing-list * )
                                char)))
(setf (get (read-from -string
                (string-append "f-" char))
           '属于)
      'f-动物)
(setf (get (read-from-string
                (string-append "f-" char))
           '性质)
      (pp, (first * parsing-list * )) (state 数量( > 5) )))
(push (list char 'se-群) procedure)
(pop * parsing-list * )))
e-判断 (when (equal (get (read-from-string
                (string-append "f-" (first input-list)))
           '属于)
      'F-动物)
 (push (list '?pp (first input-list)) * processing -list * ))
属于 se-程度)
```

执行判断槽中的 Lisp 程序的结果是，当下一个输入为“众”时，将“群”、“众”组合在一起，成为一个双音节的词“群众”。如果下一个输入不是“众”，则将“群”放入词切分表。对于“群”的切分过程结束。

如果词索引表中没有该字，则继续读下面的一个字，并与该字组成一个双字。重复上述步骤，直到系统认识了这个组合或组合的长度超过了四个字而系统仍不认识为止。由于本系统所要阅读的故事中，词的长度最多为四个字，因此，当词切分表中形成了一个四个字的单位而词索引表中仍然不存在这样的词时，系统即确认四个中的第一个字为生字。此时，通过人的干预，系统可以直接获取有关信息，从而接纳该字。

上述词切分的主要依据是一个字在上下文中形成一个意义单元时所起的作用。。因此，词的切分过程与意义分析过程是相伴进行的。系统根据意义分析过程的进展以及系统的输入，并参照判断槽中提供的组词规则，将字组合成适当的词。因此，本系统没有提供单独意义上的词切分算法。作为一种尝试，其合理性和普遍性尚有待进一步地研究和证实。

只有词切分表中出现了一个切分好的词，系统就将其传送到处理表中，并判断该词的词性。如果它是过程性的词，系统就从该词所对应的节点中取出 e-判断槽中的值，从而对处理的下一个步骤进行预期。同时，还要将形式槽中的语义框架放入处理表中。如果这个切分好的词是描述性的词，系统就将其暂时存放在处理表中。如果这个词是辅助性的词，就可以根据其 e-判断槽中的值确定该词在形成意义结构中所应发挥的作用。例如，“在”是一个辅助性的词，其结构如下:

```
(a-在
```

```
e-判断 (let ((?time) (?modifier))
          (cond ((sonp (isa (second input-list)) 'SF-气象)
                 (push (list '?time (second input-list)) * processing-list * ))
                ((eq (string-search "e"
                                    (string (isa (second input-list))))
                     0)
                 (push (list '?time
                             (same-kind-of-word (second input-list)
                                                (nthcdr 2 input-list)
                                                'e-动物活动))
                       * processing-list * ))
                (t (modifier (second input-list)
                             (nthcdr 2 input-list)
                             'e-动物活动)
                   (push (list '?modifier ?modifier) * processing-list * )
                   (if (atom ?modifier)
                       (push (list '?time (same-kind-of-word
                                           (second input-list)
                                           (nthcdr 2 input-list)
                                           'e-动物活动))
                             * processing-list * )
                       (push (list '?time (same-kind-of-word
                                           (nth (+1 (length ?modifier))
                                                input-list)
                                           (nthcdr (+2 (length ?modifier))
                                                   input-list)
                                           'e-动物活动))
                             * processing-list * ))))))
```

"在"可能标志着地点或时间的出现，也可能引出一个伴随过程。如果"在"的后继输入中出现了表示时间、地点、气象的词，则系统执行填充或预置地点和时间的程序。如果"在"的后继输入中出现了与 e-型节点相对应的词，则要启动一个确定伴随过程的程序。在这个处理过程中出现的一切中间结果，都被顺序地放入处理表中。

如果处理表中存在含有空槽值的语义框架，系统就会在处理表和词切分表中寻找适当的槽值。每当一个语义框架填充完毕，都要将其放入意义表。而且每当系统读到各种标点符号时都要对意义表进行一次重组和精细化，删除多余的部分，合并同类。整个故事阅读完毕后，系统要对意义表再次进行重整，形成一个对故事的完整的意义结构。处理过程如图 2。

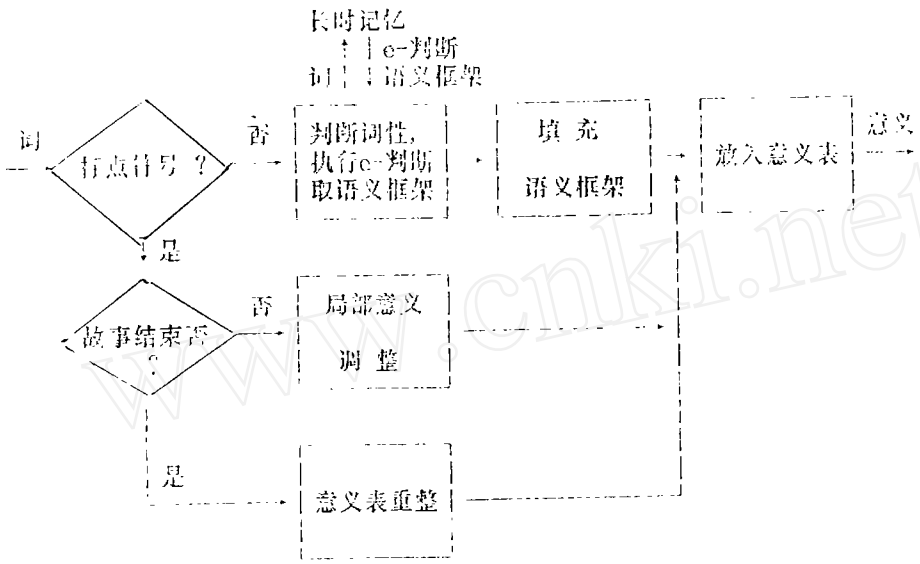

图 2　系统处理机制的工作流程

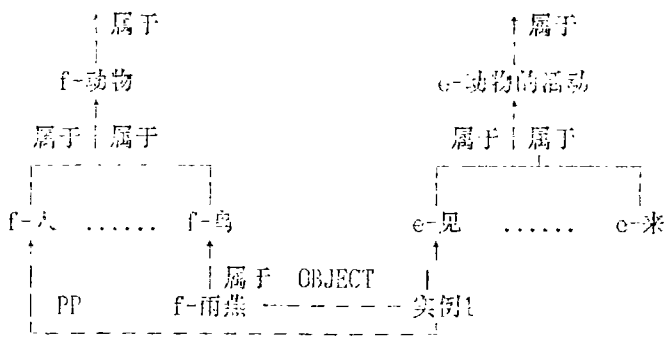

图 3　处理过程中产生的事实-事件网络片段

由于语义框架主要是由事实--事件网络中的 e-型节点和 se 型节点提供的，所以，每当一个语义框架填充完毕之后，都要相应地对提供这个语义框架的节点进行修改，在其之下建立实例节点。该实例节点的槽的设置与相应的语义框架内的槽设置情况相同。其槽值的内容，或者是处理过程中填充的内容，或者是框架中该槽原始的缺省值。当然，系统所关心的是语义框架中新加入的内容。这些内容是事实网络与事件网络相连接的过程中所需

要的联系。管理机制以这些联系为依据，确定事实网络中的那些节点应该与当前发生的事件相关。系统根据事件的实例节点内槽的名称，将事件的实例与事实网络中的相应节点连接起来。这些槽的名称所反映的就是事实与事件之间的语义关系。例如:“见”在事实-事件网络中是一个 e 型节点。它所提供的语义框架为:

((PP ?PP1) (ACT MTRANS) (OBJECT ?PP)
(FROM 眼睛 (POSS-BY ?PP1) (TO 短时记忆(POSS-BY ?PP1)))

在阅读过有关雨燕的故事后，该语义框架具有如下形式:

((PP 人们) (ACT MTRANS) (OBJECT 雨燕) (FROM 眼睛(POSS-BY 人们))
(TO 短时记忆(人们)))

图 3 是有关的网络片段:

通过建立实例节点的过程，事实与事件之间建立了暂时的联系。因此，实例节点可以组织大量的情节性的知识。这些情节性的知识是事实与事件相结合的纽带，也是通过阅读掌握新知识的途径之一。

上述实例节点的建立是一个暂时的过程。它们只是一些局部意义单元的产物。最终结构要等到故事阅读完毕之后才能确定。E.Rumelhart 提出的故事语法和 W.kintch 及 T.A. van Dijk 阐述的命题理论都强调，任何故事都有中心主题。这个主题就是故事的核心。因此，对一个故事的完整意义结构，一定是围绕一个中心意义而建立的。根据这个中心意义及其附属的其它意义结构，就可以形成一个关于这个故事的图式。如果系统阅读了若干具有相同中心意义的故事，它就能从中获得一些抽象性的特征，从而形成一个较稳定的事件脚本。这个新脚本的获得，就将在事件网络中形成一个新的 e 型节点。同时，由于 e 型节点可以通过各种语义关系与事实相联系。根据其中事实与事件之间的主体关系，有可能使系统发现若干事实同时充当某个事件的主体。这样，系统的管理机制就可根据事件中的其它一些语义关系对主体行为的影响，确定这些共同具有一种行为的事实是否可以归为一类。如果可以归为一类，便在事实网络中形成一个新的节点，从而引起事实网络的重新组合和扩充。这个新的节点即为那些可归为一类的事实节点的共同父辈节点。例如: 本系统阅读的七个故事中，有关大雁、雨燕、天鹅、杜鹃、猫头鹰这五种鸟的故事有两个副题，“越冬”和“繁殖”。这五种鸟的繁殖和越冬都借助“飞行”为运动工具。它们都是冬天南飞越冬，夏季北飞繁殖。而这两个方向、两个季节的组合所形成的自然条件都是气温适宜，食物充分，利于鸟类的繁殖。因此,“飞行”就成为这些故事所描述的事件的中心主题。围绕 e-飞行而形成的图式，就是这些故事所共有的意义结构。系统通过阅读这五个小故事，就能抽取出这些共同特征，并且在事件网络中形成一个有关“迁徙”的新脚本，也就是一个 e 型节点。同时，由于这五种鸟都具有“迁徙”这种行为特征，因而可归为一类。于是事实网络中就增加了一个新节点 f-候鸟。上述五种鸟就是它的直接下属节点。f-候鸟又是 f-鸟的直接下属节点。其它两种鸟，麻雀和鸵鸟都没有这种生活习惯和飞行特点，因而只能是 f-鸟的直接下属节点。此外，这七种鸟又各具特点，这使它们不致混同为一个节点。管理过程如图 4。

本系统所采用的汉语篇章处理的方法，是以词为引导的记忆组织方式以及与这种组织方式相对应的处理和管理机制。这种方法充分地利用了汉语的上述特点；同时，由于着重字词之间的搭配关系，并且将这种搭配关系以及词汇的语义功能用预期的形式表现出来，

形成一段可执行的程序，因而使得系统能够很自然地处理汉语中出现的一些多动词句、复合句和无动词句。

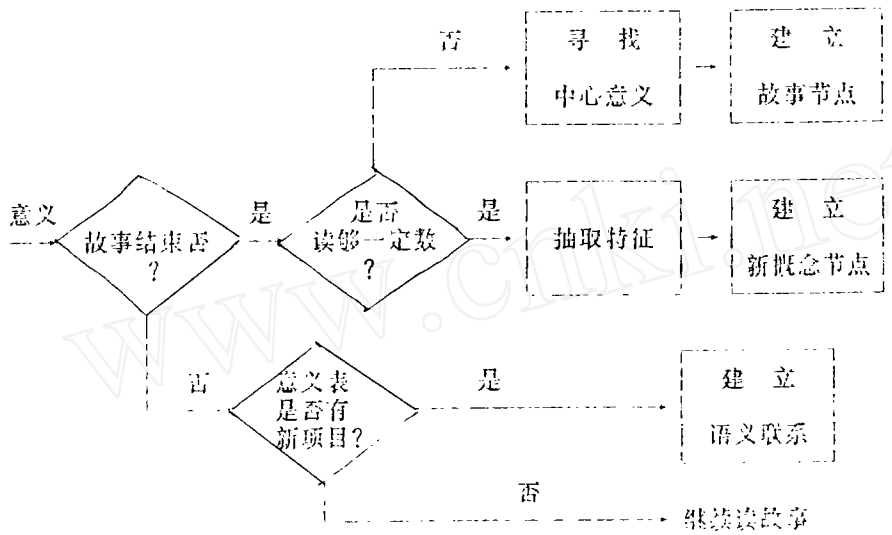

图 4　系统管理机制的工作流程

例如:“每到秋冬季节，大雁就从它们的老家西伯利亚，成群结队地来到我国南方越冬。在长途旅行中，头雁带领着雁群排成整齐的队形飞行。”

这个例子中的第一句，动词“来”和“越冬”构成连动句;“成群结队”是动词短语作状语;“西伯利亚”是“老家”的同位语;“它们”是人称代词，指代“大雁”。句中还有三个介词短语分别指出时间状语和地点状语。句子中属于描述性的词有:“秋”、“冬”、“季节”、“老家”、“西伯利亚”、“我国”、“南方”、“队”、“群”。它们都与事实-事件网络中的 sf 型节点相对应。句子中属于过程性的词有:“来”、“越冬”、“成”、“结”。其中“来”、“越冬”与事实一事件网络中的 e 型节点相对应;“成”、“结”则与 sc 型节点相对应。句子中属于辅助性的词有:“每”、“到”、“就”、“从”、“它们”、“的”、“地”。它们都与辅助性词表中的 a 型词项相对应。例子中的第二句子是兼语句。兼语结构的第一个动词是“带领”;第二个动词的位置上出现的是一个连动的结构，其中第一个动词是“排成”;第二个动词是“飞行”。例子中属于描述性的词有“途”、“头”、“雁”。其中,“头”、“雁”对应于事实-事件网络中的 f 型节点;“途”对应于 sf 型的节点。例子中属于过程性的词有:“长”、“旅行”、“带领”、“群”、“排成”、“整齐”、“队形”、“飞行”。其中“旅行”、“带领”、“飞行”对应于事实-事件网络中的 e 型节点;“长”、“群”、“排成”、“队形”、“整齐”对应于 se 型节点。例子中的“在”、“着”、“中”、“的”属于辅助性的词，对应于辅助词表中的 a 型词项。

本系统处理完第一句后处理表中形成如下结构:

((PP 大雁) (ACT PTRANS) (OBJECT 大雁)
(FORM 西伯利亚) (TO 我国南方) (TIME 秋冬季节)

(GOAL ((PP 大雁) (ACT PTRANS)
(OBJECT 大雁) (FROM PLACE (温度(<5) 食物数量(<5)))
(TO PLACE (温度(=5) 食物数量(>5))) (GOAL 生存)))
(INSTRUMENT ?INSTRUMENT))

这是个不完整的意义结构，其中的 INSTRUMENT 槽尚未填充。当系统处理完第二句时，处理表中形成如下结构:

((PP 头雁) (ACT DO)
(ENABLE ((PP 大雁) (STATE 数量(>5) 秩序(>8))
(ACT PTRANS) (OBJECT 大雁)
(FROM ?FORM) (TO ?TO)
(INSTRUMENT ((PP 大雁) (ACT PROPEL) (OBJECT 空气)
(INSTRUMENT ((PP 大雁) (ACT MOVE)
(OBJECT 翅膀)))))))

经过意义结构的精细化，在意义表中得到如下结构:

((PP 头雁) (ACT DO)
(ENABLE ((PP 大雁) (STATE 数量(>5) 秩序(>8)) (ACT PTRANS) (OBJECT 大雁)
(BETWEEN 西伯利亚 我国南方)
(INSTRUMENT ((PP 大雁) (ACT PROPEL) (OBJECT 空气)
(INSTRUMENT ((PP 大雁) (ACT MOVE) (OBJECT 翅膀))))))
((PP 大雁) (ACT PTRANS) (FROM 西伯利亚) (TO 我国南方)
(GOAL (ACT PTRANS) (FROM PLACE (温度(<5) 食物(<5)))
(TO PLACE (温度(=5) 食物(>5)))
(GOAL 生存))
(INSTRUMENT 飞行))

意义分析过程到此结束。

三、结束语

本系统从意义分析的角度对汉语的词汇进行了分类。在此基础上提出了汉语的篇章理解所依赖的知识表示和知识组织形式，即以词为引导的记忆结构。系统通过这个记忆结构向以词为中心的自然语言处理过程提供知识。这些知识综合了语法的、语义的、语用的信息。同时，本系统还建立了在预期驱动下的处理机制和记忆的动态管理机制。通过初步的实验证明本系统能够处理一些汉语所特有的语言现象，能够通过阅读，学习到一些有关世界实物和这些世界实物活动特点的知识。

本文中词性的分类方法所要突出的是汉语的词与世界现象之间的对应关系，并试图通过这种对应关系，将汉语篇章理解系统的词典和知识库有机地结合在一起。这样做的目的是尝试寻找一种能够适合于实现汉语篇章处理中，从词的切分、句子分析、推理直到建立

篇章意义结构整个过程的、统一的知识表示结构以及相应的处理机制。本文报告的只是一个初步的实验和基本的思路。而且，目前处理过的汉语篇章非常有限。因此，上述分类和处理方法还有待进一步地完善，并希望得到同行们的宝贵意见。

参考文献

[1] 陈永明　罗永东　现代认知心理学：人的信息加工　团结出版社,1989

[2] 黄昌宁　计算语言学　清华大学计算机系教材，1990

[3] 吕淑湘　现代汉语八百词　商务印书馆，1991

[4] Kolodner, J.L. Retrieval and organizational strategies in conceptual memory: A computer model. YALEU／DCS／TR-187, Yale University, 1980.

[5] Schank, R.C. Conceptusl information processing. North-Holland Publishing Company, 1975 Schank, R.C. Dynamic memory: A theory of reminding and learning in computers and people. Cambridge University Press, 1982

附录　故事

故事 1　大雁

每到秋冬季节，大雁就从它们的老家西伯利亚，成群结队地来到我国南方越冬。冬去春来，它们又飞回老家繁殖幼鸟。在长途旅行中，头雁带领着雁群排成整齐的队形飞行。

故事 2　雨燕

北京有一种雨燕。它是人们在夏季能够经常见到的鸟类。每当雷雨将至，雨燕就特别活跃，经常擦着地面掠过。它们是长途旅行的冠军。随着天气转凉，雨燕就离开北京，飞回温暖的南方。

故事 3　天鹅

天鹅全身长满洁白的羽毛。它们与雁生活习惯相同。越冬结束后，也要从我国南方飞回西伯利亚进行繁殖。那里的夏季日照时间长，花草茂盛，昆虫繁生，利于雁与天鹅觅食和育雏。

故事 4　杜鹃

鸟几乎都会孵蛋。杜鹃的繁殖习性却与众不同。杜鹃自己不筑巢，不育雏，而把蛋产在其它鸟的巢中。它希望这些鸟替它孵蛋，育雏。每年冬季，杜鹃都要飞到南方以这种方式繁殖幼鸟。春暖花开，小杜鹃就会离开养育它们的义亲鸟，飞向北方，开始独立生活。

故事 5　猫头鹰

猫头鹰是一种益鸟。每年冬季，猫头鹰都会飞到南方越冬。第二年春天又飞回北方繁殖。不论飞到何处，它主要的食物都是鼠类。它的繁殖率随鼠类多少而变动。鼠类多，猫头鹰产蛋也多。因此，猫头鹰飞回北方后，不是很快就筑巢育雏，而要先看看食物是否丰富。

故事 6　麻雀

麻雀是一种常见的鸟类，遍布各地，常在晒谷场，稻田等处活动．这些地方有大量谷类，是麻雀觅食的主要场所．麻雀对稻谷的破坏很大．即将成熟的稻谷常被成群结队的麻雀吞食．

故事 7　鸵鸟

鸵鸟是一种不会飞的鸟类．它们生活在沙漠和草原中．它们的脚长且粗壮，只有两个粗大向前的脚趾，脚趾底部还有厚皮．这样，走在沙漠中既不会下陷，也不会烫脚，可以方便地行走．因此鸵鸟虽不会飞，却跑得很快．

An Experimantal System For Chinese Text Understanding

Cui, Yao　　Chen, Yong-ming

Institute of Psychology, The Chinese Academy of Sciences

Abstract

Acorrding to the component parts of the world phenomena and the character of human memory structure, the reasearch examined Chinese usages, and classifies the vocabulary again. From the view of meanning analysis, the vocabulary can be divided into three groups: description, procedure and auxiliary. The description and the procedure correspond to facts and events in the world phenomena respectively. The auxiliary describes the information of function word usages in Chinese. Based on these examination and classification, the basic structure of knowledge representation and organization have been formed. The core of knowledeg base in the system is a memory model, cassed Facts-Events Network whose nodes represent facts and events separatedly. The auxiliary is organized in an auxiliary list which is another part of the knowledge base. The vocabulary is the centrel clue in processing Chinese texts and managing the knowledge base. All activities carried on by the system are based on Facts-Events Network and driven by expectations some of which are stored in the knowledge base, the other is produced during text processing. Through reading, the system can dynamicly self-manage its memory. After reading some stories describing seven birds, the system can learn some new concepts about the birds that it did not know before.

Key Word: natural language understanding, knowledge representation, memory model.