心 理 学 报
第3期 ACTA PSYCHOLOGICA SINICA 1985年

机器理解古汉语的初步探索[1)]

李家治 陈永明
中国科学院心理研究所

摘 要

本文介绍一个理解古汉语的计算机系统——SUACH。实验材料是两个古汉语的小故事。该系统对课文进行句法和语义分析，并把它译成现代汉语。系统也能回答某些词在不同上下文中的句法功能和意义。

汉语与西方语言不同。前者没有词形变化。人们不能根据词的形态来判定一个名词的格或一个动词的时态。汉语语法也不能很好地解释这一点。文中提出和讨论了一些计算机理解古汉语的临时规则。

一 问题的提出

八十年代初，我们曾采用语义网络模型，以人机对话的形式，建立了一个机器理解汉语的程序。[(1)]当然，机器"理解自然语言"的研究工作，常常是用不同的方法，从不同的角度来进行的，而且都取得了一定的成果[(2)]。本工作则是对机器理解自然语言的另一个侧面——理解古汉语，作一初步的探索。

在我国几千年的文化发展历程中，古汉语为我们记录了丰富的历史资料。这些资料是反映我国古代文明的宝库。理解古汉语是文化事业的需要，对我国、对世界都有重要的意义。我们希望使计算机也能对此有所贡献。

我们从初中语文课本中选了两篇短文："郑人买履"和"刻舟求剑"，作为试验材料，在PDP11/23机上用LISP语言建立了一个理解古汉语的程序。

二 程序设计的依据和程序的结构

(一)**依据**：词序是我们设计程序时的第一个依据。词序是语言表达上的一种重要形式。它说明词与词的组合关系以及词在句子中的地位。词序不仅传送给我们语法信息，而且也带有一定的语义信息。不同民族的语言，往往有不同的词序。古汉语与现代汉语的词序虽有不同，但基本的词序是一致的，即主词在前，动词居中，宾词在后：

〈主语〉 〈谓语〉 〈宾语〉

限定词在前，被限定的词在后：

〈限定词〉 〈名词〉

1) 本文于1984年4月10日收到。

例如,"先自度其足……。"一句,它的基本词序是很规则的,即:自(主语)——度(谓语)——其足(宾语);其(限定词)——足(被限定的名词)。因此,词序在古汉语的理解中仍可起重要的作用。

程序设计的另一个依据,是乔姆斯基 (Chomsky)的句法分析理论[3]。对古汉语语句的分析,我们基本上采用的是乔姆斯基的生成语法。程序对语句的加工过程,就是从语句的表层结构开始,对语句的各个元素,在不同的层次上按照短语结构语法的规则加以分组。最后,生成语句的深层结构。

依据词序和句法分析理论,在分析语句时,有时会碰到一定的困难。同一个词在不同的语境条件下,往往在句中起着不同的功用,具有不同的意义。这时,就需要参考语义和上下文才能加以正确的分析。

(二)结构:程序主要由四个部分组成。它们是:句法分析程序,翻译程序,回答问题的程序和一部小小的词典。它的结构见图1。

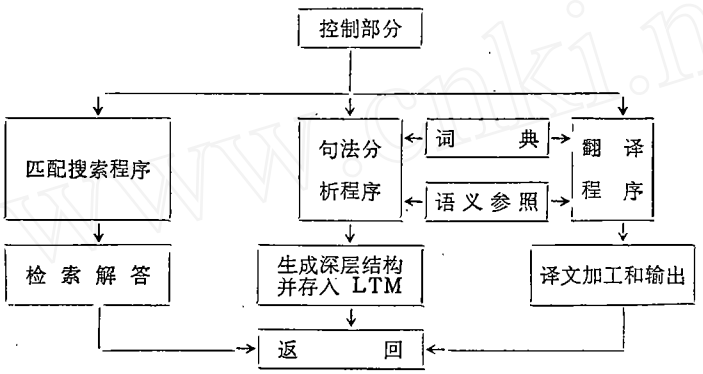

图1 程序的结构

系统在起动以后,先从课文中取出一个语句,根据所采用的理论模型,从左到右地对语句进行一系列加工:查阅词典、进行句法分析,以及可能需要的语义识别等,直至生成该语句的深层结构为止。在分析语句的不同阶段上所获得的结果,都临时存入工作记忆中。在对一个语句的分析加工完成以后,就把所得的结果转存入长时记忆中,以备课文翻译时用。然后,再从课文中取下一个语句,并作同样的加工处理。在全部课文的句法加工完成以后,可以利用词典中具有的词组成新的语句,再让机器分析;也可以让机器作短文的翻译;也可以提一些问题,让机器回答。这些是由控制部分来执行的。

程序中备有一本小小的词典,以供处理语句时查阅用。词典的编排参考了桑克(Schank)的词类划分的思想[4],并采用框架结构的方式。词典中放置了词的以下几个语法特征及译文词义:

1. 词类:包括名词、动词、代词、副词、助词、介词和连接词等七类。这里没有形容词,因为故事中没有真正的形容词。"郑人"的"郑"在词典里是划入名词一类的。那为什么"郑人"能译为"郑国的一个人"呢?这是因为我们有一个约定,即在名字或代词后面如果还跟着一个名词的话,这个名词或代词就作为限定词。这是在句法分析时解决的。

2. 数:即单数和复数。

3. 生物和非生物。

4. 词义。

词典的格式是：

((〈槽 1〉(〈平面 1〉 〈值〉)(〈平面 2〉 〈值〉)……))
(〈槽 2〉((〈平面 1〉 〈值〉)(〈平面 2〉 〈值〉)……))
⋮
(〈槽n〉((〈平面 1〉 〈值〉)(〈平面 2〉 〈值〉)……))

这里，在每一个槽中，放一个词；每一个平面中放该词的一个特征项，然后是该特征的值。查阅词典时，先是纵向搜索，找到该词的位置，然后在该词的位置上进行横向搜索，找到该词的特征项，再取出该特征项的值。例如，查阅“履”的词类，首先找到放置“履”的那个槽，然后横向搜索，找到放置“词类”的那个平面，再取出它的值，即“名词”。

查阅某个词的词类的LISP函数，可定义如下：

```
(DEFINE ((GETT (LAMBDA(WORD)
    (CADR (SASSOC TYPE
        (CADR (SASSOC WORD
            (GET "DIC" WORDS) NIL))
                NIL)) ))))
```

词典和程序的其他部分是相对独立的，词典中的词及其特征项，可以根据需要而加以增删，而对程序的其他部分不会产生任何影响。

三　程序的功能

(一)参照乔姆斯基的生成语法分析语句，构成语句的深层结构。下面举几个具体的例子。

句(A)：“其剑自舟中坠于水。”

计算机经过分析加工后，输出一张句法结构表：

```
(S(NP(DET其)(N剑))(PP(PRE自)(NP
    (DET舟)(N中)))(VP(V坠))(PP(PRE
        于)(NP(N水))))
```

句(B)：“人曰：何不试之以足。”

计算机对句(B)的输出是：

```
(S0(NP(N人))(VP(V曰))(S(AP(ADV何)
    (ADV不))(VP(V试)(NP(PRO之)))(PP
    (PRE以)(NP(N足)))))
```

句(B)的句法结构表可以转换成如下的树形图。(图 2)

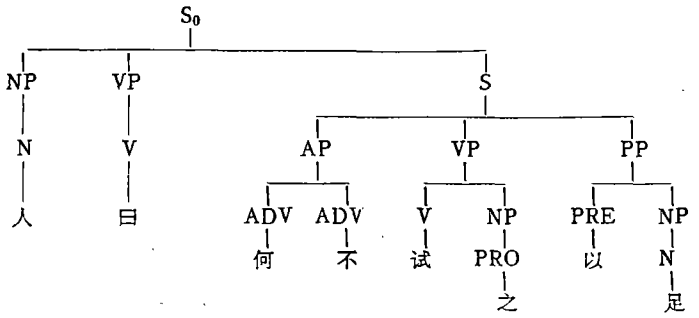

图 2 树形图

该句的语法可以表述如下：

起始符号：S_0

非终极符号：{S_0, NP, VP, S, AP, PP, ADV, V, PRO, N, PRE}

终极符号：{人, 足, 曰, 试, 之, 何, 不, 以}

重写规则：

1. $S_0 \rightarrow NP+VP+S$
2. $NP \rightarrow N$
3. $VP \rightarrow V$
4. $S \rightarrow AP+VP+PP$
5. $AP \rightarrow ADV$
6. $VP \rightarrow V+NP$
7. $NP \rightarrow PRO$
8. $PP \rightarrow PRE+NP$
9. $NP \rightarrow N$
10. N→人, 足
11. V→曰, 试
12. ADV→何, 不
13. PRO→之
14. PRE→以

如果我们利用词典中存储的词自造一个句子，如“吾以吾足试履”，让机器分析，所得的结果是下面一张句法结构表：

(S(NP(N吾))(PP(PRE以)(NP(DET吾)
(N足)))(VP(V试)(NP(N履))))

这说明，程序对这类新组成的句子的分析也是成功的。

一个语句的句法结构表或它的树形图，可以给我们提供关于语句的以下四个方面的信息：(1)句子中各个词的顺序。如果把起始符号看作树根，而把终极符号看作树叶，那么，树形图底部的这些树叶，表示了句子中各个词的前后关系；(2)句子结构的层次。S分成NP和VP，这是一层，而VP又分成V和NP，这又是一层。这种层次排列形象地表示出句子中各成分在空间上的层次关系；(3)词组信息和词类信息。例如，“履”属于名词词组，而其本身又是一个名词等等；(4)提供了隐含的词义信息。中文和外文有点不同，有些词虽然在词形上完全一样，但由于它在句子中的位置不同，所属的词类不同，却有不同的词义。例如，课文里“……度其足”和“……忘持度”这两句中的“度”字，就有这种情况。它们的树形图分别是(图3)：

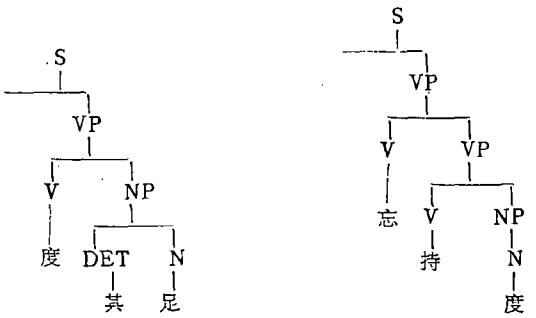

图 3　两个句子的树形图

图 3 清楚地表明，“度其足”中的度是个动词，而同一个词在“忘持度”中却是一个名词。“度”作为动词是“测量”的意思，而作为名词则是“尺寸”的意思。这就隐含着它们的词义彼此是不同的。又如，在“至之市”和“忘操之”这两个分句中的“之”，也有同样的情况。“之”在这两个地方属于不同的词类，一是动词，一是代词，因而隐含着不同的词义。

对语句的这种句法结构的分析，是理解语句的基础。

(二)问答。对机器提问，让机器回答某个词的词类和它在句中的意义。例如：

问：“吾忘持度”中的度是什么意思？

答：名词，是尺寸的意思。

问：“先自度其足”中的度当什么讲？

答：动词，是量尺寸的意思。

(三)转译成现代汉语。

“郑人买履”这个小故事的原文是：

郑人有欲买履者，先自度其足而置之其坐。至之市，而忘操之。已得履，乃曰：“吾忘持度。”反归取之。及反，市罢，遂不得履。人曰：“何不试之以足？”曰：“宁信度，无自信也。”

机器对小故事加工以后所得的译文是：

“郑国的一个人想要买鞋。他先量了自己脚的尺寸，却把尺寸放在他的座位上。他到赶集时，却忘记了带尺寸。他已拿到了鞋，才说：我忘记了拿尺寸。他返回来取尺寸。到回去时，集市散了，于是没有买到鞋。有人说：为什么不用脚试鞋呢？他却说：宁肯相信尺寸，不要相信自己。”

机器对这段小故事的译文，基本上是通顺的。

在机器对原文进行翻译和对译文进行加工时，为了使译文通顺，还考虑了以下几条：

（1） 删除：古汉语中有些词并没有实在的意义，往往只是表示语气停顿，如“也”字。对于这类词，机器在翻译时就把它删掉。

（2） 添加：这主要是指给语句添加一个主语。在古汉语中，往往在前面一句中出现过主词后，后面连续几句都没有主语。但语句所表达的行为的主体，还是前面语句中的主词所表示的主体。在这种情况下，为了更符合现代汉语的习惯，就必需给没有主语的句子添加一个主语。此外，在某些情况下，也需要适当地添加个别的词，以便语句更顺口一

些。例如，在句首含有疑问词“为什么”时，在句末可添加一个“呢”字。

（3） 调序：古汉语的基本词序虽然和现代汉语是一致的，但是，它们总还是有差别的。如果完全依照古汉语的词序进行翻译，就会使人感到非常别扭。因此，机器在翻译过程中，还必须进行“调序”。例如，“何不试之以足”这一句，如果我们顺译为“为什么不试鞋用脚”，就感到句子不通顺。因此，必须给它调序，再在句末添加一个“呢”字。这样，计算机就输出“为什么不用脚试鞋呢？”，这就符合人们日常的语言思惟的习惯了。

四 讨 论

“机器理解自然语言”这个词的含意是多方面的，因而，机器理解自然语言的表现形式也是多样的。当然，把古汉语译成现代汉语或其他语言，或许是最能代表对古汉语的理解程度的一种形式。我们在把“郑人买履”一文译成现代汉语的过程中，在以下几个方面作了若干临时性的约定。当然，很难说这些约定是否有普遍意义。

（一）古汉语大量省略主语，译成现代汉语时，必须适当地添加主语。但是，什么是适当？有些什么规则能使机器给一个省略主语的句子适当地加上主语？我们根据具体情况作了如下一些临时规定：

（1） 在主句中如果没有主语，一般就添上一个主语。如：“已得履，乃曰：吾忘持度。”一句，在译成现代汉语时，加上一个主语“他”，则成：

他已拿到了鞋，才说：我忘记了拿尺寸。

（2） 如果主句中有主语，而分句中却没有，一般就不再加主语了。例如，“人曰：何不试之以足？”一句，译成现代汉语时，不再在“何不……”之前加什么主语了。

（3） 有时，一句话有两个并列的动词，实际上是一个联合复合句。在这种情况下，只在前面一个分句中加上主语，而在后一个分句中不再加同样的主语，以免累赘。如“至之市，而忘操之。”一句，情况就是如此。此句译成现代汉语为“他……，却忘记了带尺寸”，而不是译成“他……，他却……”。

（二）古汉语的时态问题。古汉语的时态往往以主要动词前的时态副词或含有时态意义的动词为标志的。如“欲买”是指想要买但还没有买。“已得”是指已经得到了。“至之市”是指当去赶集的时候等等。可是在许多情况下，动词并没有时态标志，但人们还是能正确地加以理解，这是中国人的语言习惯。我们针对汉语的时态表达问题，制订了一些临时性的约定。

（1） 在古汉语中，凡动词前有时态副词或含有时态意义的动词时，就按后者的含意去理解和翻译。

（2） 有些动词，如“曰”，如果前面没有表示时态的词，在翻译时就不去考虑它的时态表达问题。“曰”在任何情况下都译作“说”，而不加“了”字。如故事中的“人曰”，虽然这是过去完成的动作，但只能译作“有人说”，而不译作“有人说了”。

（3） 有的动词，如“忘”，如果它前面没有含有时态意义的其他词，在任何情况下都译作“忘了”或“忘记了”。如“吾忘持度”就译作“我忘记了拿尺寸”。

这些约定是否可行，尚待进一步探讨。

(三) 在古汉语中，应用代词的地方较多。虽然在一般的情况下，机器理解和处理代词不很困难。但是，在某些特殊的情况下，代词代表的不是离它近的前面那个名词，这时，代词的处理就会遇到困难。例如，“先自度其足而置之其坐”中的代词“之”。按照代词应用的一般规则，即代词代表的是在它前面最近的那个名词，那么，“置之其坐”中的“之”是代表它前面的“足”了。而机器会把此句译成“……把脚放在他的坐位上”。再翻译下去，就会把“至之市，而忘操之”译成“……却忘记了带脚”。这当然是笑话了。可是，机器怎么知道“之”代表的是“尺寸”而不是“脚”呢？具有尺寸含义的词，是在几句话后才出现的。一直到“……吾忘持度”时，其中的“度”才表示尺寸的意思。这种代词先于它所代表的名词之前的情况，计算机是很难正确地理解的。为了解决这一具体困难，我们作了一种临时性的安排：

如果“度”后面有名词或名词短语，“度”就作为一个动词，在词典中查找“度”作为动词时的意义，即“量……的尺寸。”紧跟在后面的代词，就是代表“尺寸”这个名词。

如果“度”前面有动词而后面又没有名词或名词短语，则“度”就是一个名词，在词典中查找其作为名词时的意义，即“尺寸”。

当然，诸如此类的临时性安排，只是为了解决上述那样的具体问题，不一定适用于其他情况。就上例而言，在一般情况下，“度”作为动词，恐怕就是“测量”的意思。

此外，古汉语中有大量虚词，如：之、其、而、以、则、者、所等等。它们有的可作代词，有的可作连词，有的可作介词，等等。虚词的应用，使古汉语的表达方式比现代汉语简练。而且，古汉语和现代汉语句子的结构，也往往因此而很不相同。例如：

“……何不试之以足”。

这句话的介词短语是在动词短语之后，而现代汉语必须把介词短语放在动词短语之前。再如：

求剑若此，(不亦惑乎？)

是吾剑之所从坠。

在现代汉语中也没有这种表达方式。有趣的是，这样的表达方式却同英语相似。上面的三句话译成英语将是：

Why don't you try on them with your feet?

Seeking the sword this way, (isn't it stupid?)

This is where my sword was fallen from.

从这几个例子来看，英语颇具有古汉语的韵味。这可能是由于英语的许多代词和关系代词、介词、连词(如Who, which, that, it, when, where, by, with)等，同古汉语的虚词的功能有许多相似之处，我们并没有从事古汉语英译的探索，可是我们感到，古汉语译成英语或许并不比译成现代汉语更困难。

以上是我们对机器理解古汉语的一些初步探索。

我们研究机器理解古汉语，也是为计算机理解现代汉语寻求适当的方法。我们要求机器把古汉语陈述的事实和概念用现代汉语表达出来，就是为了这一目的。

汉语，无论古汉语或现代汉语，与西方语言的明显不同，在于它没有词形变化，不能依靠词形来判断一个名词属于什么格，一个动词属于什么时态。汉语虽然有一个基本词序，但灵活多变，只可以说是聊胜于无。可是，我们中国人，包括儿童在内，理解汉语并未感到

多么困难。我们是依靠什么来理解汉语的呢？我们还缺乏一个适宜的汉语语法来说明这一问题。源自西方的现代汉语语法，并不能很好地揭示汉语表达的规律。因此，在我们的程序中，不得不用一些临时约定来帮助机器理解古汉语，并造出现代汉语的句子。我们希望，这种临时约定的积累、修改和补充，或许会有助于为机器理解汉语制定出一套适合于中国人的思惟习惯和语言习惯的汉语语法。

参 考 文 献

〔1〕李家治、郭荣江、陈永明，机器理解汉语——实验Ⅰ，心理学报，1，1982年。
〔2〕范继淹、徐志敏，RJD-80型汉语人机对话系统的语法分析，中国语文，3，1982年。
〔3〕Chomsky, N., Syntactic Structures. Mouton, The Hague, 1957.
〔4〕Schank, R. C., Conceptual Information Processing, 1975.

AN EXPERIMENT IN UNDERSTANDING ANCIENT CHINESE THROUGH THE COMPUTER

Li Jiazhi Chen Yongming
(*Institute of Psychology, Academia Sinica*)

Abstract

The paper describes a computer system of understanding ancient Chinese (SUACH), using two stories written in ancient Chinese as experimental material. The system can make syntactic and semantic analysis of every sentence contained in the two stories and translate them into modern Chinese. It also has the ability to answer questions about the syntactical function and meaning of certain words used in different contexts.

Unlike Western languages, the Chinese language has no declensions. For instance, one cannot determine the case of a noun or the tense of a verb according to the morphology of Chinese characters. Neither does Chinese grammar contain adequate rules governing inflections. A tentative plan for the programming of ancient Chinese is presented and discussed in this paper and offered as an aid to students of ancient Chinese.