第4卷第6期
1993年12月

软件学报
JOURNAL OF SOFTWARE

Vol. 4, No. 6
Dec. 1993

基于自然语言计算模型的汉语理解系统

周经野*

(湘潭大学计算机科学系，湘潭 411105)

A CHINESE LANGUAGE UNDERSTANDING SYSTEM BASED ON THE COMPUTATIONAL MODEL OF NATURAL LANGUAGE

Zhou Jingye

(*Department of Computer Science, Xiangtan University, Xiangtan* 411105)

Abstract We put forward a computational model of natural language in which the process of communicating in natural language is divided into three levels: linguistic form, surface semantics and deep semantics. Based on this model we design a Chinese language understanding system which could be easily transported from one domain to another. A new grammar, Chinese Semantic Construction Grammar, is introduced, which could describes not only syntactic form but also semantic form of a sentence and performs syntacric analysis and semantic analysis simultaneouslyduring parsing. The deep semantics of a word or a phrase and the basic operations of deep semantics are formally defined. The algorithms of the parser, the understander and the generator of the system are also presented.

摘要 本文首先给出了一种自然语言计算模型，该模型把自然语言交流过程划分为三个层次：语言形式，表层语义和深层语义，从而将自然语言理解抽象为一个复合函数 UP(s, k). 依据这个模型，我们设计了一个汉语理解系统. 这个系统具有良好的扩展性和可移植性. 该系统采用汉语语义结构文法来分析汉语句子，把语法分析和语义分析有机地结合在一起. 文中形式定义了词语的深层语义以及深层语义的基本运算，给出了分析器、理解器以及生成器的算法.

§0. 引 言

自然语言理解是新一代计算机系统研制中的关键问题，并有重大的实用价值. 到目前为止，自然语言理解系统都是在给定的领域上定制的. 这种自然语言理解系统的能力都受限于

* 本文1990年4月23日收到，1991年6月9日定稿. 作者周经野，副教授，主要研究领域为计算机软件，自然语言处理.

其狭窄的领域和受限的上下文.同时,即使是在一个很小的领域上开发该系统也需要相当的工作量,而所开发的系统难于维护更难于搬迁到其他不同的应用领域.这种状况严重地阻碍了自然语言理解系统的研制和推广.我们认为自然语言理解的研制不应基于具体的领域,而应基于自然语言的计算模型.

本文中我们建立了一种自然语言的计算模型,依据这个模型设计了一种汉语理解系统,该系统可以比较容易地进行应用领域的转换.第一节中给出了模型的定义和系统的构造.在第二节中给出了一种新的描写汉语的文法,语义结构文法.它的规则式同时具有语法描写和语义描写,把句子直接转换为它的表层语义.深层语义依赖于领域,它是句子在具体语言环境中的意义.理解器将句子的表层语义转换为深层语义.第三节中我们定义了词语的深层语义,定义了语义网络的基本运算,并给出了理解器的工作过程.第四节中介绍了系统中的生成器,它生成汉语句子来回答用户;此外还介绍了系统的服务器,它为用户提供维护词典、数据库和知识库的便利.

§1. 自然语言的计算模型和汉语理解系统的构造

自然语言是由语言单位表达的.这里语言单位包括词、短语、句子、课文等.这些语言单位都是用线性符号串表示的.这种线性符号串是它们的语言形式.语言单位都有其内部结构,即其内部各成份间的关系,它不是线性的,而是较复杂的网络结构.但是这种内部结构还不是语言单位的意义.语言单位的意义是依赖语言环境的.这是因为符号要在具体的语言环境中被赋予意义以及理解者要用自己的知识对语言单位进行处理.自然语言理解的过程是一个包括知识处理在内的复杂过程.

因此自然语言交流的过程可分为三个层次:语言形式,表层语义和深层语义.它们的形式定义如下.

定义 1.1:令 W 为所有词的集合,N 为所有语言单位内部结构的集合,M 为基本意义单位集合.分析函数 P 是从 W^* 到 N 的映射,理解函数 U 是从 N 到 M 的幂集的映射;构思函数 O 是从 M 的幂集到 N 的映射,生成函数 G 是 N 到 W^* 的映射.

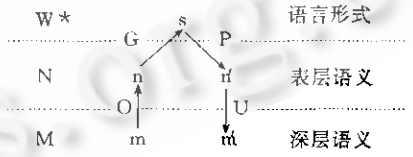

图1

定义 1.2:令 $s \in W^*$ 为一语言单位,函数值 P(s)称为 s 的表层语义,而复合函数值 UP(s)称为 s 的深层语义.

于是自然语言交流的全过程可以抽象为图 1 所示的计算模型.

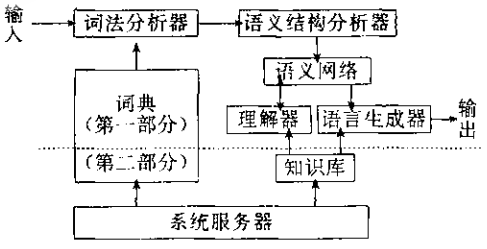

图2

图 2 中给出了根据这个计算模型设计的汉语理解系统.其各部分的功能简述如下:

词法分析器通过词典来识别词(切词)并取得词所附带的信息.词典被划分为两部分.第一部分包含词的如下信息:词形、切词信息、语法范畴和表层语义信息.这些信息基本上不依赖于领域.第二部分包含词的深层语义,它是依赖于领域的.

语义结构分析器的功能是将语言单位的语言形式转换为它的表层语义,即实现分析函数 P. 我们用语义网络表示语言单位的表层语义. 因此语义结构分析器输出的是语言单位所对应的语义网络.

理解器的任务是获取语言单位的深层语义,即实现理解函数 U. 知识同样可以用语义网络来表达,因此它是通过语义网络的运算来实现知识处理.

专门领域的知识存放在词典的第二部分,即词的深层语义和知识库中. 它们大都要根据用户来定义. 词典和知识库都必须能够扩展或迁移. 系统服务器就承担维护词典和知识库的任务,为用户提供应有的便利.

生成器的任务是生成汉语句子来回答用户的问题. 它将表示答案的语义网络进行一番组织,选定词汇,确定语序,生成各种短语并最后输出汉语句子,实现构思函数 O 和生成函数 G 的功能.

§2. 表层语义和语义结构文法

语言单位是概念的语言形式,表达了某些概念及其间的逻辑关系和语用关系. 这就是语言单位的表层语义,表示为语义网络.

定义 2.1:语义网络是有限的有向图,其中称端结点为概念结点,内部结点为概念关系,有向弧为结构关系,无前躯的结点为首结点.

将概念进行分类. 令 C 为所有概念类的集合.

定义 2.2:一个语义函数 f 是从 W 到 C 的幂集上的映射.

语义函数建立起词汇和概念类之间的联系. 文法中定义了一组语义函数,例如,CLAS 函数给出一个词所隶属的概念类的集合,EXPAGT,EXPOBJ,…等语义函数描写了动词的期望语义等. 语义函数的定义域也不难扩展到 W^* 上. 一个词汇的相应的语义函数值作为该词的表层语义信息存放在词典中.

定义 2.3:一条语义规则式是由以下四个部分构成的:

1. 模式 X,是语法范畴的符号串;

2. 表层语义条件 B,是由语义函数构成的逻辑表达式;

3. 表层语义构造 N,是一个语义网络的构造;

4. 重写式 Y,也是语法范畴的符号串,如果 Y 中有归约产生的新范畴,则还要给出新范畴的表层语义信息的计算.

具有同一模式 X 的语义规则式称为一组语义规则式,记为$\{X;B_i\rightarrow N_i,Y_i\}$.

定义 2.4:语义结构文法 G 是一个六元组:$G=(W,V,C,N,F,P)$,其中 W 是词的集合,V 是语义网络的集合,C 是概念类的集合,N 是语义网络的集合,F 是语义函数的集合,P 是语义规则的集合.

该文法的识别装置,语义结构分析器的构造如图 3 所示.

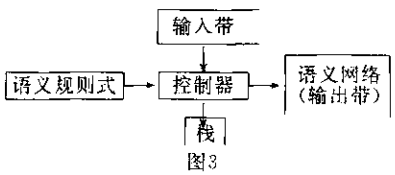

图3

输入带上是经过词法分析器处理后的语言单位. 每个单元含有语法范畴,表层语义信息和所分配的结点. 控制器是一个识别模式

的有限自动机(Q,T,q_0,F).每个终止状态与一组语义规则式$\{X:B_i\rightarrow N_i,Y_i\}$相关联.

算法 1:语义结构分析器的算法如下:

设初始时开始状态q_0已压入栈内,控制器正阅读输入带上的第一个符号,a_i为正被阅读的符号,q为位于栈顶的状态.

1. 若q in F,转第7步;
2. 取出与终止状态q相关联的一组语义规则式$\{X:B_i\rightarrow N_i,Y_i\}$;
3. 设与模式X相匹配的输入子串为z,依次用B_i来检测z表层语义条件;
4. 若B_i满足,建立语义网络N_i的一个实例,并根据Y_i来重写z;若都不满足,则出错.
5. 从栈中弹出$|X|$个状态,控制器阅读新写子串的第一个符号.
6. 若输入带上全是空白,则终止.
7. 读进符号a_i,将状态$\delta(q,a_i)$压入栈内;控制器阅读下一个符号;转第1步.

§3. 深层语义和理解器

一般来说,自然语言理解系统中总有个内部模型.令E为表示具体环境的数据库中全体实体的集合.词语的深层语义定义为下面三种类型.

定义 3.1:词语w的指称语义为$D(w)=\{e\in E:e$ 为w所指称$\}$.

定义 3.2:词语w的操作语义为$O(w)=$procedure $P(x_1,\cdots,x_n):S$;其中P是过程名,$x_1,\cdots,x_n$为其参数,S为过程体.

定义 3.3:词语w的抽象语义为$A(w)=$concept $C(a_1,\cdots,a_n):N$;其中C称为概念名,N是一个语义网络,称为概念体,$a_1,\cdots,a_n$是N中的一些概念结点,称为概念C的概念参数.

一个语言单位的深层语义是其表层语义经过处理后所得到的新的语义网络.这些处理除了调用其词语的深层语义进行的操作外还包括和知识库中的某些知识(也表示为语义网络)进行的运算.下面给出语义网络的几种基本运算.

定义 3.4:令n为语义网络中的一个结点,设其表示的命题为P.n的指称为$D(n)=\{e\in E:P(e)\}$.语义网络的指称是其首结点指称的并集.

定义 3.5:语义网络u和v的联结是合并u和v中相互等价的概念结点、概念关系后所产生的新的语义网络.

定义 3.6:语义网络u的一个限定是将u中的某个结点n用另一个结点m替换所产生的新的语义网络,其中n与m满足$D(m)\subseteq D(n)$.

定义 3.7:若语义网络v中存在一个子网络w是语义网络u的若干次限定,则称w是u在v中的投影.

定义 3.8:若语义网络u和v中有一个是另一个的若干次限定,则称u和v是匹配的.若u中存在子网络u′,v中存在子网络v′且u′和v′是匹配的,则称u和v是部分匹配的.

定义 3.9:设n是语义网络v中的一个概念结点,其对应的概念为concept $C(a_1,\cdots,a_n):u$.用概念C的概念体u替换v中的结点n所得到的语义网络称为v的一个展开.反之,若v中存在一个子网络与概念C的概念体u相匹配,则用一个新的概念n来替换这个子网络,结点n与概念C相对应.由此得到的语义网络称为v的一个卷叠.

算法 2:理解器的工作算法如下.理解器反复进行以下操作,直至不再产生新的语义网

络为止.

1. 查看是否存在可以联结的语义网络. 若有,对它们进行联结.

2. 查看具有操作语义的概念结点. 看是否可以执行. 如可,依据语义网络对其过程赋予实参,再调用该过程.

3. 就某个概念结点的指称语义进行限定.

4. 就某个概念结点的抽象语义对语义网络进行展开.

5. 查看是否存在可以卷叠的子网络. 若有,对语义网络进行卷叠.

6. 依据知识库中的规则对语义网络进行变换.

理论上,理解器的工作可以是不确定的,也可能不终止,正像人对话语也会不停地左思右想一样. 实际上,我们可以使它按某种确定的方式工作,也可以用某种强制手段,如达到某个目的或深度,使其终止.

§4. 生成器和服务器

生成器的作用是生成汉语句子来回答用户.

获取答案的工作仍然由理解器来承担. 理解器对疑问句所对应的陈述句进行运算,对于一般疑问句,判断其陈述句所表达的命题是否为真;若是选择问句,则指明哪部分为真. 当答案否定时,可附带产生相关的真命题. 对于特殊疑问句,找出与之部分匹配的语义网络;然后,对于"什么"一类问题,取其投影作为答案;对于"为什么"一类问题,取与之匹配部分的前件或后件作为答案. 最后将表示答案的语义网络提交给生成器. 生成器还要对其进行一番组织,再转化为汉语句子.

算法 3:生成器的工作算法如下.

1. 根据用户的身份和所问的问题,确定回答方式. 回答方式有:详细回答,部分回答,间接回答,反问和婉言拒绝.

2. 根据用户、问题和答案,确定是否省略和省略哪些部分.

3. 根据回答方式和省略的情况,对表示答案的语义网络进行处理,产生答句的语义网络(素材).

4. 根据用户的问句和答句的素材,确定答句的句型,在素材上增添表示语用的标志,确定答句的基本语序.

5. 填词. 如果一个概念结点有多个词汇供选择,选定一个词填入素材中. 对于重复出现的词语,确定可否替代和如何替代.

6. 对名词短语,根据"的"字规则添加"的"字,根据定语规则排好修饰成份的顺序,构成名词短语.

7. 根据结构关系、语序和选定的动词,选择必要的介词,构成介词短语.

8. 确定状语位置,添加必要的助词和副词.

9. 根据最后确定的语序从左到右把各部分连接起来,输出句子.

服务器提供的主要服务功能有:

1. 用户可直接对词典、数据库和知识库进行管理维护.

2. 自动将用户输入的信息转换为系统内部的表达形式和组织形式.

3. 将用户在维护知识库或数据库时产生的词语信息自动加入到词典中去.

4. 支持理解器的运行,并为系统自身的维护提供良好的支撑环境.

服务器创造的良好环境是系统具有较好的扩展性和可移植性的重要保证.

§5. 总 结

最初,我们在苹果2e型微机上用BASIC语言设计了一个汉语问答系统.其分析器和理解器是基于自然语言计算模型设计的,脱离具体领域.词汇和知识都是后装入的.该系统可回答关于四则运算和算盘操作的问题,有良好的扩展性.但是该系统没有服务器,词汇和知识得由设计者直接装入.

后来我们应用这一思想在IBM/XT型微机上设计了关系数据库的可移植的汉语接口.它通过服务器与不同的关系数据库挂接,使用户能用汉语进行数据库的各种操作,并有一定的推理能力.目前我们正在完善该系统,争取做成一个实用软件.

以基于自然语言计算模型的汉语理解系统为核心,我们还设计了计算机辅助程序设计系统.该系统在理解器后面接上一个程序生成器,能理解用户对问题的汉语描述,自动生成相应的PROLOG程序.它还能接受用户用自然语言描述的概念,即具有一定的被告知的学习能力.

以上实践使我们认识到基于自然语言计算模型的汉语理解系统具有良好的扩展性和可移植性.

理论上,自然语言计算模型将自然语言理解过程抽象为复合函数UP(s,k),其中k为具体领域的知识.这样具体领域的知识被处理为复合运算中的一个变元.从而系统的移植只是变元k的一个替换.分析函数与领域的联系主要表现在词汇.我们用词汇的语义函数来建立这种联系,并通过系统服务器在背景知识更换或扩充时同步实现词典的维护和各种语言信息的获取.这样,分析函数(文法规则)和理解函数(语义网络的运算)自身是不依赖领域的.所以,在一定的意义上可以说,这种基于自然语言计算模型的汉语理解系统是一种具有通用性的汉语理解系统.

参考文献

1 Zhou Jingye, The Semantic Levels of Natural Language and Semantic Construction Grammar, 北京国际计算机和通讯会议论文集, 1986, 386—393.

2 Zhou Jingye and Chang Shikuo, A Methodology for Deterministic Chinese Parsing, Computer Processing of Chinese & Oriental Language, Vol. 2, No. 3, 1986, 139—161.

3 Bruno G. Bara and Giovanni Guida eds., Computational Models of Natural Language Processing, Elsevier Science Publishing Company, INC., 1984.

4 M. P. Marcus, A Theory of Syntactic Recognition for Natural Language, MIT Press, 1980.

5 D. G. Bobrow and A. Collins eds., Representation and Understanding: Studies in Cognitive Science, Academic Press, New York, 1975.

6 W. A. Woods, Transition Network Grammar for Natural Language Analysis, Communication of the ACM 13, 1970, 591—606.