第 21 卷　第 2 期　　华 侨 大 学 学 报 (自 然 科 学 版)　　Vol. 21　No. 2
2000 年 4 月　　Journal of Huaqiao University (Natural Science)　　Apr. 2000

文章编号　1000-5013(2000) 02-0205-06

基于伪自然语言理解的知识获取系统

陈永鸿①　　陈 一秀②

(① 华侨大学计算机科学系, ② 华侨大学机电工程系, 泉州 362011)

摘要　基于伪自然语言理解, 提出并实现一种高效率的知识获取方法. 首先, 知识工程师利用在结构上类似自然语言的非常简单的 BL 语言, 改写书本非常复杂的自然描述. 然后, 利用高效和成熟的编译系统处理 BL 程序, 以实现书本知识获取. 最后, 领域专家在书本知识库的基本语义框架引导下, 利用知识求精系统对书本知识库加以少许求精.

关键词　书本知识描述语言 BL, 章, 段, 知识编译, 语义图, 知识求精

中图分类号　TP 182　　文献标识码　A

知识获取是建造专家系统的瓶颈问题. 这个瓶颈的卡口处主要有: (1) 每开发一个新的专家系统, 知识工程师需要学习一门新的课程; (2) 许多专家不善于或不愿意把自己的知识用整理好的方式贡献出来; (3) 知识工程师过于紧密地依靠领域专家, 如在时间或信息交流上不能很好地配合, 即难以完成任务[1]. 一般专家系统中的大部分知识是该领域的公共知识, 可在专业书本上找到, 属于专家个人经验的比例很小. 从开发专家系统的实际情况看, 一开始为了取得共同语言, 领域专家也往往要指定几本专业书让知识工程师先作初步学习. 为此[2,3], 我们采用了一种新的高效率的专家系统开发方法.

1　总体设计

如图 1 所示, 这种方法的要点可概述为如下 5 点. (1) 设计并实现一种书本知识描述语言 BL, 使其十分接近书本自然语言. 但它又是一种计算机语言, 可以被无二义性地编译. (2) 把专家系统开发分为两个阶段, 即知识工程师 (相对) 独立工作阶段和领域专家 (相对) 独立工作阶段. (3) 在第一阶段, 由领域专家指定专业书籍, 知识工程师把专业书籍自然描述的知识, 用扫描仪高效自动地输入计算机中. 然后, 把书本语言改为书本知识描述语言. 无需理

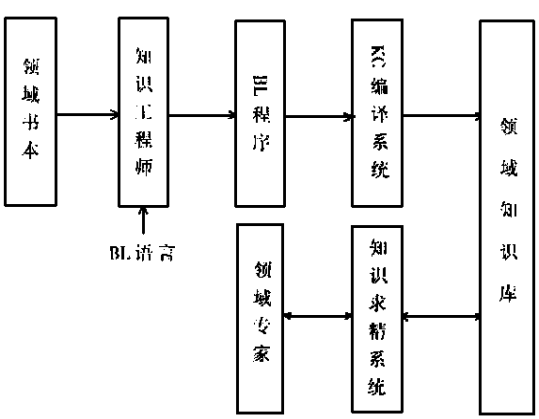

图 1　知识获取系统示意图

收稿日期　1999-09-09　　作者简介　陈永鸿(1966–), 男, 讲师

基金项目　福建省自然科学基金资助项目

解或读过这本书,只要象把英文翻译成中文那样,逐句翻译过来就可以了.由于两种语言十分接近,这种翻译比从英文翻译到中文还容易.(4) 经过翻译的书本知识已经是一个程序,可以送交计算机编译和处理.知识编译器(KC)将自动地从中抽取出领域知识来,并把它们按专家系统要求重新整理后组织成知识库.(5) 在第二阶段,领域专家在书本知识库的基本语义框架的引导下,利用知识求精系统对书本知识库加以少许求精.在此基础上建造专家系统,以符合专家个人经验.

2 书本知识获取

2.1 BL语言

系统的关键是设计书本知识描述语言BL,尤其是把书本自然描述的结构所蕴涵的基本语义体现在BL中.为此,我们首先要阅读大量书籍,总结出书本知识组织和陈述上的特点.然后,用成熟的Chomsky文法描述BL的语法.最后,在文法基础上,找出书本自然描述的结构中所蕴涵的语义描述.这些基本语义是著书者和读者之间不言而喻的、存在于书本自然描述的篇章结构中的、公理似的、非常简单的一些约定俗成.

2.1.1 *语法介绍* BL语言是非常接近自然语言的书本知识描述语言,它的语法可以被认为是汉语语法的一个子集,由程序头和程序体组成.程序头主要是程序体中的语言的特殊用法的一些说明,它包括5个方面.(1) 系统背景说明.(2) 具体说明.即系统级关键词与领域级关键词的对应关系说明.引入领域级关键词,在于扩大系统的领域适用性.(3) 同义词说明.(4) 方面说明.这是目标所带的各类属性子集的子集名的说明,如某种矿物的性质、特点、成分等.对于分类问题来说,它提供了多种角度的分类.(5) 表格说明.目的在于为目标的某方面属性子集,提供简便的描述和表示方式.程序体对应于书本自然描述的结构,由一些章组成.章由章标题和章体所组成.章标题与书本上的章节标题对应;章体对应于书本上自然描述的章节,由一些子句、段和一些更小的章所组成.

段对应于书本上的对某一方面知识的集中描述.一个段由段引导句、段体和段终结词"段止"所组成.段体由一些子句、更小的段所组成,总共有十几种段.

〈段〉= 〈对象段〉| 〈分类结构段〉| 〈原理段〉| 〈单目标段〉| 〈原因段〉| 〈可省略段〉
| 〈实例段〉| 〈方面枚举段〉| 〈方面非枚举段〉| 〈方面分类段〉
| 〈结果枚举段〉| 〈结果非枚举段〉| 〈属性分析段〉| 〈目标鉴别段〉.

我们举一个〈方面枚举段〉的例子.〈方面枚举段〉是为了某目标的方面的集中描述而引入的,如某矿物的特征、性质等的集中描述.段体中,先对属性一项项枚举,再分别说明原因.其中"后遗症"是领域级关键词,可取代关键词"方面".这可使建立在所生成的知识库之上的专家系统的界面更自然.

关于 高血压病 的 方面/后遗症状 包括:

(1) 左心功能不全:由于 心肌 日益 加重负担,造成 左心室 肥大,与 心肌 缺血,与 左心室 扩张!

(2) 冠状动脉硬化:{可通过心肌缺血而危害心肌}!

(3) 脑出血:由于 脑血管 硬化,造成 血压 过度升高,很可能 引起 脑出血!

(4) 肾功能不全:!

段止.

我们再举一个〈方面非枚举段〉的例子. 它与〈方面枚举段〉类似, 只不过属性及其原因一起描述.

关于　肺动脉瓣狭窄　的　方面/病理特征: 左心室向动脉的排血　障碍; 右心室　肥大; 若　肺动脉瓣狭窄　处于后期, 则　病理特征　左心衰竭, 引起　病理特征　右心房　淤血, 与　病理特征　大循环　淤血, 与　病理特征　右心衰竭的一系列症状和体征　出现; 段止.

每个子句中最小的语法单位是一些词, 从书本中的句子到 BL 程序中的子句是一对多的. 这对使用者非常方便. 子句包括简单句、固定句和复合句 3 类. (1) 简单句是最基本的句子. (2) 书本自然描述的句子中, 有一些结构比较固定、不可分割, 我们从中归纳出固定句. (3) 书本自然描述的句子中, 有一些是由一些描述事实的简单句子复合而成, 我们从中归纳出复合句. 复合句是通过一个大框架, 把一些简单句和固定句套起来而形成的. 固定句如各种段的不同的引导句, 简单句及复合句如上面所举的两个段的段体中描述的句子.

2.1.2　语义说明　首先我们应注意一下说明, 引导句必须被置于一条子句(关键子句)之前或之后, 或者被置于段引导句之前. 这表明所引导的关键子句或段说明一个领域知识单元. 另外, 还应注意引导句包括段引导句和说明引导句. 为了解释说明引导句、章和段如何说明书本自然描述的结构及蕴涵的语义, 我们引入如下定义.

定义 **1**　术语. 除引导句以外的子句.

定义 **2**　术语标记. 章的标题与段的引导句中的中心成分. 我们用〈α〉, 〈β〉, 〈ε〉等来表示术语标记.

定义 **3**　术语标记的作用域. 若术语标记为一个章的标题, 则其作用域为该章体; 若术语标记为一个段的引导句的中心成分, 则其作用域为该段. 对每个术语标记〈α〉, 它的作用域记为 FD〈α〉.

定义 **4**　若对于任意〈α〉和〈β〉, 有 FD〈β〉包含于 FD〈α〉, 则〈α〉大于〈β〉, 记为〈α〉>〈β〉; 或〈β〉小于〈α〉, 记为〈β〉<〈α〉. 如果某章或段对应于〈β〉, 而另一个对应于〈α〉, 我们称后者大于前者, 或前者小于后者. 类似地, 我们可定义术语标记及章和段中的最大者和最小者.

定义 **5**　若对于任意〈α〉, 比〈α〉小的术语标记为〈β_1〉, 〈β_2〉, …, 〈β_m〉, 则 FD〈α〉去掉 FD〈β_1〉, FD〈β_2〉, …, FD〈β_m〉, 所剩区域中构成〈α〉内涵的术语称为〈α〉的固有术语, 记为 DG〈α〉; 所剩区域中构成另一些知识单元内涵的术语, 称为〈α〉的不定术语, 记为 UG〈α〉.

通过分析书本知识自然描述的特征, 从隐含的书本层次结构中, 及其上下文关系中说明关系结构上的特征出发, 赋予说明引导句、章和段在这方面的语义. (1) 对于任意 UG〈α〉, 若它的前(或后)加上说明引导子句, 则 ① 说明引导句的中心成分必须是下面两者之一: (a) UG〈α〉包含于其中的章、段的术语标记; (b) 包含该不定术语的章和段中的最小者, 对应术语标记设为〈α〉. 〈α〉的作用域且在上下文关系中, 处于该不定术语前面(或后面)的术语标记或术语对应的领域知识单元名称. ② 该不定术语所在知识单元, 说明引导句的中心成分所标明的知识单元. 对于段对应的术语标记, 若在段前加上说明引导子句, 也具有同样情形. (2) 对于某一章, 若对应术语标记为〈α〉. 含于该章中的极大的章的术语标记为〈ε_1〉, 〈ε_2〉, …, 〈ε_m〉, FD〈α〉去掉 FD〈ε_1〉, FD〈ε_2〉, …, FD〈ε_m〉. 剩下的区域中极大的段的术语, 标记为〈δ_1〉, 〈δ_2〉, …, 〈δ_n〉, 则 FD〈α〉去掉 FD〈ε_1〉, FD〈ε_2〉, …, FD〈ε_m〉, FD〈δ_1〉, FD〈δ_2〉, …, FD〈δ_n〉. 所剩区域中, 没有加上

说明引导子句的所有术语为 DG 〈α〉. 再有如 〈α1〉, 〈α2〉, …, 〈αn〉中, 对应的段前面没有加上说明引导子句的术语, 标记为 〈αi1〉, 〈αi2〉, …, 〈αit〉. 则 〈αi1〉, 〈αi2〉, …, 〈αit〉各自对应的知识单元, 都说明 〈α〉对应的知识单元. 再有 〈α〉对应的知识单元, 分别说明 〈ε1〉, 〈ε2〉, …, 〈εn〉对应的知识单元. (3) 对于某一段, 若对应术语标记为 〈α〉, 含于该段中的极大的段的术语, 标记为 〈ε1〉, 〈ε2〉, …, 〈εr〉, 则 FD 〈α〉去掉 FD 〈ε1〉, FD 〈ε2〉, …, FD 〈εr〉. 所剩区域中, 没有加上说明引导子句的所有术语为 DG 〈α〉. 再有如 〈ε1〉, 〈ε2〉, …, 〈εr〉对应的段, 前面没有加上说明引导子句的术语, 标记为 〈εj1〉, 〈εj2〉, …, 〈εjs〉. 则 〈εj1〉, 〈εj2〉, …, 〈εjs〉各自对应的知识单元, 都说明 〈α〉对应的知识单元.

2.1.3 BL 语言的特色 (1) 自然性. BL 语言结构上高度类似于书本自然描述, 只需对书本自然描述稍加修改就可以(如增删一些词). 另外, 关键词的词义也与它们在汉语中的词义基本一致. (2) 结构化与模块化. 在 BL 语言中, 通过章、段、说明引导句, 把书本自然描述的结构上的特点刻画出来. (3) 可扩充性. 在 BL 语言中可以容易地引入新的语句. 我们虽是针对分类型领域知识, 但 BL 语言对书本自然描述的结构上的特点的刻画, 并不只针对分类型领域知识. (4) 灵活性. 从书本自然描述到 BL 源程序是一对多的, 所以编辑上很灵活. (5) 容歧义. BL 源程序中一些语义相近的关键词, 在一些复合句中可混淆使用, 这方便用户使用 BL 语言. (6) 描述多级知识. BL 语言能描述表层知识、因果网等深层知识, 也能描述它们混合在一起的知识.

2.2 知识编译系统

文中书本知识的获取, 局限在分类型知识. 我们利用成熟编译技术, 进行书本知识获取. 书本知识自然描述的主要结构是篇章结构, 这对应于 BL 程序的说明引导句、章和段所组成的结构. 对应地, 书本知识库的基本结构, 是由概念以及概念之间的说明关系组成的一个语义图. 每个概念还有丰富的内涵, 主要由因果关系子网和一些属性所组成. 整个编译过程紧紧围绕这个结构进行. 其基本过程, 如图 2 所示.

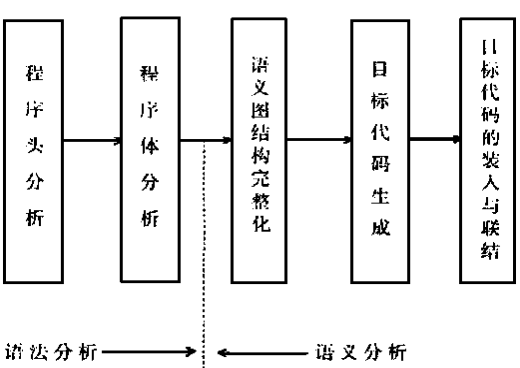

图 2 编译过程

在语法分析阶段, 首先分析程序头, 主要是分析程序头说明的语法, 以及在定义表中装入一些说明信息. 其次, 分析程序体. 分析一些章, 分析章必须被嵌套地转换成分析一些子句或段, 而分析段必须被嵌套地转换成分析一些子句. 分析一些复合句, 必须被嵌套地转换成分析一些固定句或简单句. 词法分析和语法分析交替进行, 语义图结构表、同义词表、概念的属性表、概念名表、概念的因果关系表(它和属性表一起组成每一个概念的内涵), 在同一时间必须被装入.

在语义分析阶段, 首先完整化语义图结构, 编译器检测语义图结构表. 如果语义图结构不是连通的, 知识工程师需采用交互方式把它补充完整. 然后, 生成目标代码. 此时编译器再次扫描 BL 程序, 同时不停地生成目标代码, 以及反复地调用装入和联结目标代码的模块. 如果当前的目标代码不能被联结, 编译器暂时把目标代码放入与语义图结构表对应的辅助语义图结构表. 目标代码被进一步生成后, 在辅助表中相关的目标代码才被装入和联结. 目标代码交替地被生成、装入和联结, 其过程为一旦在辅助表中, 组成一个知识单元的内涵的目标代码集被

生成, 它就被装入知识库中. 根据在 BL 程序中关键词的语义值, 这个目标代码集合与另一些目标代码集合的说明关系被建立.

3　知识求精

经过编译处理, 尤其是语义分析, 所生成的知识库及书本专家系统, 已经解决面临的大多数问题. 但是, 它仍存在是否符合专家个人经验、完备性、冗余性和一致性的问题. 因此需由领域专家在知识求精系统的帮助下, 对书本知识库进行求精. 整个求精工作(求精系统的运行) , 是在书本知识库的基本语义框架(语义图) 引导下进行. 其过程是依深度优先搜索, 由代表书名的概念(无入弧顶点) 出发, 按出弧方向遍历各孩子顶点. 同时, 对遍历到的每个新顶点, 围绕着它, 依深度优先搜索按入弧方向遍历各 "孩子 "顶点. 这两个过程是有机的统一体, 是同一个算法的两个侧面, 也求精工作的主要过程. 另外, 求精系统还配有一导航仪, 领域专家可借助此工具, 从代表书名的概念出发, 迅速搜索到需要求精的语义图的顶点. 领域专家利用此两工具, 可依次寻找到需要求精的语义图顶点, 并进行求精.

领域专家的求精工作分两个层次, 其一是对整个语义图基本框架的求精; 另一个是对语义图中顶点(概念) 的内涵进行求精. 这两个层次可同时进行, 第一层次求精工作较简单, 第二层次求精, 是在更小一级的局部子网(其中顶点主要是一些因果核和一些属性) 中进行.

4　知识获取系统的使用

基于领域知识库, 我们构造了 CONBES 分类型专家系统外壳. 它有深层和表层两个推理机, 一起提供了 3 种推理模式. 第一种推理模式是双向的, 仅由表层推理机执行, 实现对语义图(目标网) 的搜索. 第二种推理模式仅由深层推理机执行, 实现对每个目标的因果关系子网的搜索. 第三种推理模式提供一种深表层混合推理, 实现对语义图和因果关系子网的搜索, 由深层推理机和表层推理机联合完成. 第二种和第三种推理模式对一个实际的专家系统来说是必需的, 因为它们在很多方面更加切合实际. 例如推理的聚集和切换. 在推理的每一步, 使用者能够请解释系统给出推理机活动的合理判断和解释.

我们用 KC 编译系统和 CONBES 专家系统外壳, 快速构造了两个基于书本知识的系统. 其一是消化系统疾病诊断专家系统. 它的知识库, 是从老中医吴伯平所写的一本书改写成的 BL 程序生成来的; 另一个是心血管疾病诊断专家系统. 它的知识库, 是从 《病理学》中的一章改写成的 BL 程序生成来的. 这两个系统都能自如地进行深表层混合推理, 而且由于领域级关键词成为人机界面的组成成分, 以及解释系统即时给出的推理解释. 因此, 这两个系统的运行让用户觉得诊断有说服力、切合实际、自然、友好. 我们在建造这两个系统时, 用 BL 语言把书本上的非常复杂的自然语言描述, 高效地转换成非常简单的计算机语言所写的程序. 然后, KC 编译系统利用成熟和高效的编译技术, 自动处理 BL 源程序, 在很短的时间内就建造成了两个(基于书本知识的) 专家系统.

5　结束语

利用此系统高效地开发 CAI 开发平台, 以及专家系统(分类型) 开发环境, 我们发现这种

知识获取方法有如下几方面的特色.(1) 可以减少知识工程师和领域专家合作的紧密程序.(2) 不论是知识编译,还是知识求精,都可直接获取较深层知识.(3) 可大大地缩短准备和输入领域知识源(此时还未开始知识获取)的时间和费用.(4) 不必基于背景知识.首先,知识工程师借用 BL 语言把非常复杂知识源(自然语言的描述),高效地转换成非常简单的计算机语言所写的程序.然后,知识编译系统利用成熟和高效的编译技术,自动处理 BL 程序.此时完成了绝大部分的知识获取任务,这是此知识获取方法最大的特色.最后,领域专家在书本知识库的基本语义框架(书本知识自然描述的结构及蕴涵的语义)引导下,利用知识求精系统进行少许求精工作.此一阶段也可高效完成.因此,此知识获取方法可高效地获取知识.本文仅局限于分类型知识的获取,如何获取规划型的领域知识有待进一步研究.

参 考 文 献

1 陆汝钤,庄庆雨,吴建敏.推进知识获取方法学的研究——知识工程进展[A].见:陆汝钤主编.第二届全国知识工程研讨会论文选集[C].北京:中国地质出版社,1988.66～75

2 陈永鸿,陈一秀.基于书本知识的多媒体写作系统[J].华侨大学学报(自然科学版),2000,21(1):92～95

3 陈永鸿,陈一秀.基于伪自然语言理解的 CAI 开发平台[J].计算机工程,2000,(1):63～68

A Knowledge Acquisition System Based on Pseudo-Natural Language Understanding

Chen Yonghong① Chen Yixiu②

(① Dept. of Comput. Sci., Huaqiao Univ.,

② Dept. of Electromech. Eng., Huaqiao Univ., 362011, Quanzhou)

Abstract Knowledge acquisition is the bottleneck of artificial intelligence. Based on pseudo-natural language understanding, the authors put forward and carry on an efficient method of knowledge acquisition. As the first step, the very complicated natural description in a domain book is rewritten by using the very simple BL language which is similar to natural language in structure. This has to be done by knowledge engineer. And then, the BL program is processed by using an efficient and well-considered compiling system so as to carry out the acquisition of book knowledge. And finally, the book knowledge base has to be somewhat refined by using a knowledge refining system under the guidance of basic semantic frame of book knowledge base. This has to be done by domain expert.

Keywords BL language for describing book knowledge, chapter, segment, knowledge compiling, semantic graph, knowledge refining