面向对象的专家MT系统

黄建烁
(华南理工大学)

AN OBJECT—ORIENTED EXPERT MT SYSTEM

HUANG JIANSHUO
(South China University of Technology)

Abstract This paper presents an outline of the Object-oriented Expert MT System. Some theoretical foundations of the system, its configuration features and working principles are discussed. Finally, examples are given to account for the basic principles and methods as to how an object model is established, inference carried out, and knowledge acquisition achieved.

Key words object-oriented, expert system, knowledge acquisition

摘要 本文描述一个面向对象的专家MT系统的基本模型。文章对系统的理论基础、结构特点和工作原理进行了讨论。最后通过例子对建立对象模型、实现推理和知识获取的基本原理和方法作了说明。

关键词 面向对象，专家系统，知识获取。

一、引 言

近年来，我国的机器翻译已逐渐走出实验室，开始进入实用化、商品化阶段。这是一个很大的飞跃。

然而，还不能认为机器翻译的问题已解决了。由于自然语言的复杂性及其与知识的密切关系，机器翻译至今仍是一个非常困难的课题。问题的核心是知识的表现与获取。在这一点上，现有的机译系统基本上是用产生式规则来实现的。它存在一些明显的弱点：规则间的相互关系不明显、知识的整体形象难以把握、处理效率低、推理缺乏灵活性等等。

因此，人们试图应用人工智能技术、专家知识来开发新一代的机译系统。前者着重在知识表现的基础上进行种种推理、利用语义网络、框架理论、概念依存理论等构成知识表示系统，可称之为人工智能翻译系统。后者摸拟人的翻译过程，通过例句提供的样板，提高译文的质量，可称之为面向专家的翻译系统。例如长尾真教授提出的“基于样板的方法”和“基于记忆的翻译”[1、2]，还有贝尔实验室 William A. Gale 等人使用双语语料库提供的语言样板数据进行翻译[3]，都是这方面的例子。

本文根据机器翻译的发展趋势，结合我们在机器翻译研究的经验，提出一个面向对象的专家机译系统的模型同同行探讨。不当之处，请大家指正。

二、系统的特点和理论基础

2.1 面向对象的方法的定义及其特点

面向对象的方法是80年代出现的一种程序设计方法。面向对象概念中的对象可以为客观任何客体。在系统中，它的定义为具有一定信息属性并同外界有固定接口的客体。例如，机译系统处理的就有词、短语等客体。客体的信息属性就是它们的词法、句法、语义等的表现形式和内容。它们的固定接口就是这些内容所规定的记录方法以及这些对象之间的联系。

对象实质上是将数据及相应的操作封装在一起的单元，外界只能通过向对象发送消息访问或修改其数据。在系统的构成上，类构成一个具有特定功能的模块和一种代码共享的手段，是系统的组成单元。它支持知识的层次机制。其中，继承又起重要的作用。面向对象的方法以其信息隐藏、数据抽象、类的继承等特点能较好地实现模块化程序设计为基础的软件构件重用技术，提高了软件的可维护性[4]；而且，对象间的通讯、复用等机制可为系统的知识处理提供灵活的手段。

2.2 面向对象的专家 MT 系统的知识表现

本系统的知识表现采用框架理论，具有下列特点：

1. 具有结构和功能的知识表现形式，能够恰当地表现对象的内容；

2. 不限于使用谓词逻辑之类的平面表现方式，它还可以使用某种立体数据，并使得计算机能够理解，容易管理；

3. 这种知识表现能够满足模块化、一致化以及简明性等要求。

框架理论的上述特点，不但便于表达领域概念、领域对象、对象之间的关系以及对象行为特征等信息；而且通过框架系统的分类/分解树还体现知识的层次特点、知识的继承和默认性质。

基于框架理论的对象模型是一种记录对象及其属性的数据结构。其基本形式是结构名—对象—属性—值。对象表示知识的类型，属性相当于一个槽，值就是槽值。它们组成对象的属性—值偶对。属性可以是一个对象，其值可以是一个原子符号，也可以是另一个对象。所以，知识模型的描述是递归定义的。

为了表现超知识，对象的属性、值也可以是一个默认值或附加过程。可见，框架理论提供的知识表现能力及其标准化、层次化的数据结构使它可以成为开发专家系统的有力工具。

三、系统的基本结构及工作原理

3.1 系统的基本结构

本系统由下列几个部分组成：知识库（KB）、推理机构（IE）、知识获取模块（KA）、数据库（DB）、推理过程跟踪模块（IT）、人机接口（UI）。其基本结构如下图所示：

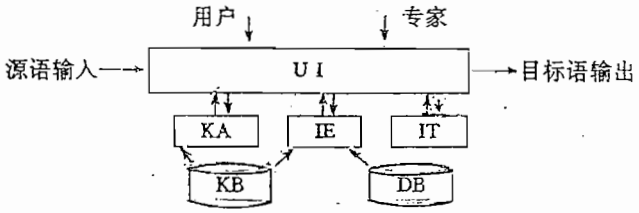

图 3.1 系统的基本结构

1. 知识库 它包含词库、规则库、对象模型、样板等有关语言事实的知识以及有关过程、判断的知识或超知识。本系统的知识库按文[5]陈述的原理建立。知识的主体是一个基于概念的抽象的层阶结构。其中，抽象度高的概念蕴涵其下层概念的属性，因此继承起着重要的作用。此外还有表现对象的分解结构。这部分用指针进行管理。对词的概念的识别，知识库按下列原则处理：

(1) 如果一个词可能对上位产生两个映射，则分别制定两个词项；

(2) 如果一个词有不同的选择限制，则有不同的义项；

(3) 如果一个词有两个不相容的同现集合，分别制定两种词义；

(4) 如果一个词有两种用法，制定两个词条。

因此，知识库可以存储有关词项的词法、词义、句法及语用等信息，这些信息或是通过本身结点获得，或是从其基类继承。

2. 推理机构　这是使用知识库的知识执行推理的控制机构，它由一个控制器和一个解释器组成。控制器负责对整个系统的元控制，解释器则负责对数据的解释和运行。在推理过程中，控制器根据加工流程控制输入数据的类型；解释器利用知识库的知识解释输入的数据，推导出结论。然后再由执行机构执行相应的操作。因此，推理的整个过程都是由知识库提供的知识控制的。

3. 知识获取模块　本模块的任务是从专家和用户那里获取知识来建立知识库。它也可以通过学习机构从输入的文本进行学习，对知识库的内容进行确认、变更或追加，从而使知识库的知识不断充实、完善和体系化。知识的自动获取是本系统设计追求的目标。它的实现有赖于知识库提供知识的完善和体系化及学习机制的建立。

4. 数据库　本系统的数据库有两个作用。一是保存作为知识的具体事实，如例子、样板等，或是保存从外部直接输入的数据及结果；二是作为知识库的管理系统。它负责对知识库中各种数据进行检索、归类、优化、合拼和更新等工作。

5. 推理过程跟踪模块　在系统的开发过程中以及系统投入使用后，开发人员或用户不单是要获得某一结论、而且还可能想知道结论的推导过程，从中检查或获取有关信息，作为修正知识、改进系统功能的依据。

6. 人机接口模块　其主要功能是使系统方便地与用户进行会话，为用户提供操作系统运行的接口。通过该模块专家可以输入有关知识；用户也可以选择适合的词义、句法形式或译语，或对译文进行编辑、打印等工作。

3. 2　系统的工作原理

系统的功能是通过各个组成部分的协调工作实现的，其中，知识库起主导作用。它不仅决定推理的方式，而且学习机制的建立、系统的性能都与它提供的知识内容有密切的关系。推理机构根据知识库的知识对输入数据进行识别、匹配，最后获得有关对象的结论。在此基础上，推理机构进入新的阶段，对新的对象进行加工，直到对系统规定的加工流程的全部输入对象加工完毕为止。输出的结果就是目标语。

在系统的工作过程中，各个阶段中对各对象加工的数据，专家或用户可以通过 IT 跟踪、检查。发现不妥当之处，就找出知识库与此直接有关的知识，对之进行改正。当知识经过定量试验证明为正确之后，就可以作为建立学习算法的依据。通过上述过程不断地对知识库中的知识进行确认、修正或补充，知识库的内容就会不断趋于完善和充实。在系统中，数据库的主要职能就是实现知识库中各种数据的有效管理。上述各组成部分协调的工作，最后可使整个系统的能力不断地提高。

四、系统的实现方法

专家系统也是一种计算机程序系统。它与一般的软件开发并没有本质的不同。但是在系统的开发过程中，需要得到专家的协助，并分阶段地实现。在这方面，还没有现成的技术。与一般的软件开发比起来，在开发方法上有“试行错误”的特点。下面，我们就对象模型的建立、推理过程和知识获取等几个主要方面，讨论系统的实现方法。

4. 1　对象模型的建立

系统实现的第一步是如何确定并建立加工对象。对于对象的数据结构，已在 2. 2 作了说明。但具体对象需要根据加工流程来确定。对机译系统来说，对象自然离不开词、词组、成语和句

子等。它们具有共性一面，也有个性一面。因此，建立对象要注意对象模型的通用性和特殊性。为此，采用多段型的知识表现是极为重要的。所谓多段是指将一个复杂的对象分解为若干部分处理。下面是以词为对象的词典结构的模型。设对象的属性有六个。

Create Lexicon　　Sname＝LX

Oname＝EN

attri1＝FM　　attri4＝AR

attri2＝ID　　attri5＝NS

attri3＝GS　　attri6＝TW

上述各个属性都是一个集合名，它们记录有关词的形态变化、成语、语法语义值、论元结构、性质结构、目标语等信息。这些值还可以进一步用其 属性的子结构加以描述。属性也可根据加工的需要进行添加、修改或删除。

4．2　知识推理

知识推理是根据框架提供的过程性知识实现的。它将推理过程转化为一棵决策树。树的结点存放对象的属性—值偶对，当某一属性确定后，其值就由解释器求出，或转入一个新的过程。上述词典信息的求解过程的决策树如下：

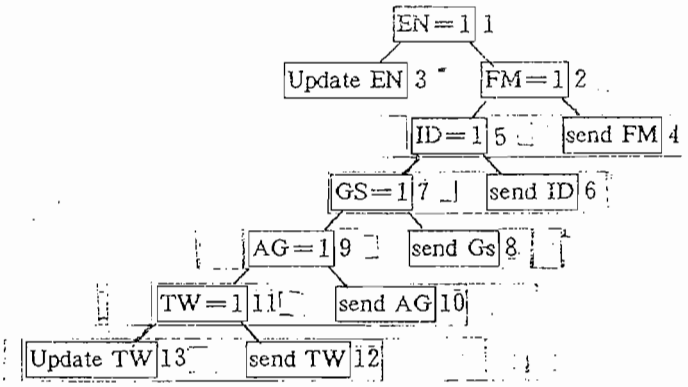

图 4．2　词典信息求解决策树

从上述决策树可以看到系统对单词的加工过程。树的结点用特征函数的值域 1/0 表示具有或不具备某一属性。决策树的子叶结点为推理的结论，而非子叶结点则反映了专家在推理过程所作的判断。由此通过寻求每一条从根结点到叶结点的路径，就可以得到专家在本加工阶段对这些问题的思维、推理过程。将这些践径按 IF—THEN 的形式转换，就得出一组产生式规则，它们构成了专家系统对这部分的加工控制机构。

经过上述决策过程获得结论之后，即可以执行相应的操作。完成一个叶结点规定的操作之后，即可沿决策树的路径返回下一个非子叶结点，继续进行余下的推理。

4．3　知识获取

知识获取是专家系统实现过程中最困难的工作。这些都与自然语言的模糊性、随机性还有语言内容及其表达方法的多样性有关。本系统可以通过知识库提供的知识建立某些学习机制，使系统具有一定的知识获取功能。下面是知识获取机构的示意图。

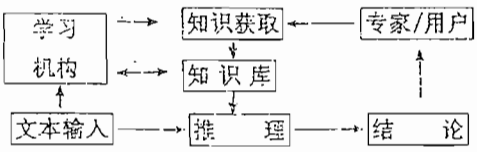

图 4．3．1　知识获取机构示意图

由上图可以看出，除了从专家那里获得知识，或由用户对加工的结论进行评估后认为需要追加有关知识外，另一个途径就是通过学习机构获取知识。学习机构含有针对不同对象建立的算法模型，它通过知识库提供的知识实现知识获取。

下面我们以求解词的搭配知识为例，说明学习机构的功能。这是根据 Velardi [6] 和 Mel'cuk [7] 等人的语义表达原理建立的一个语义偏置机制，它包括下列几个内容：

1. 一个依赖于域的概念递阶结构，这是一个由词到词义名称的多对多的映射，和一个有序的概念类别表，它和知识库递阶体系的语义类别对应；

2. 一组由域—概念关系组成的集合及句法关系与相对的概念关系之间的多对多的映射；

3. 一组用于表达概念关系的粗粒选择限制，用概念—关系—概念（CRC）三元组表示。

这一机构有两类输出：

1. 一组幼粒 CRC 三元组，它们聚集在概念或概念关系附近；

2. 一个中粒语义知识体，由 CRC 组成。

幼粒 CRC 是一些直接把概念映射入实词的三元组。例如：

[WRITE] —→ (patient) —→ [BOOK]

这种三元组都是真实的，因为它在域的子集里是可以观察到的。它也是具体某类词的搭配实例。中粒 CRC 三元组表达的概念是实词的祖先的语义值。这类 CRC 一般也是真实的。例如：

[WRITE] —→ (Patient) —→ [WORDS]

但是若果把该类某些词代入。如 WRITE newspapers。却是有问题的。粗粒 CRC 是一种分类系统中层次较高的概念关系三元组。例如：

[ACTION] —→ (benificiary) —→ [ANIMATE]

这种三元组只指出使用概念关系的必要条件，但还不是充分条件。因此需要制定一种算法以获取概念句法类别的搭配知识。

本系统的学习算法的基本原理是根据系统的输入与输出的域值中建立一个动态的学习模型，即在一组输入单元和一组输出单元之间设置一组可以调节的权重值。每一组输入为幼粒 CRC 单元，输出为中粒 CRC 单元；若输入为中粒单元，则输出就是粗粒单元。输入单元与输出单元都可以根据各自的语法、语义特性确定一个域值，权值可以在各域值之间作出调整，从而可以求出输入单元与输出单元之间的国值。国值表示 CRC 单元的语法，语义值范围。在某一阈值内，某一 CRC 单元被激活。工作时，一个 CRC 的特定阈值是决定性的，学习机构如下图所示：

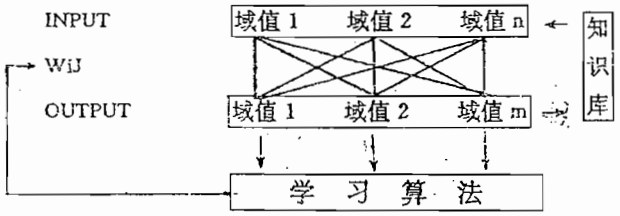

图 4. 3. 2 学习机构的原理

其中 m、n 为处理单元的所属域值，Wij 表示第 j 个域输入单元到第 i 个域输出单元的关系连接权系。Wij 的差导表明了某个关系中概念之间相互约束的强度和性质的差别。

上述加工所获得的有效数据可以形成一个权重矩阵：

$$W=\begin{bmatrix} W_{11} & W_{12} & \cdots & W_{1n} \\ W_{21} & W_{22} & \cdots & W_{2n} \\ & & \vdots & \\ W_{m1} & W_{m2} & \cdots & W_{mn} \end{bmatrix} \qquad (0<1<n)$$

其中 n 为处理单元数、通过各域之间的连接权重值的大小，可确定某一关系成立的可信度。输出函数可以用阈值函数表示：

$$f(x)=\begin{cases}1 & if\ x\geqslant\theta\\ 0 & if\ x<\theta\end{cases}$$

当然，这种方法不是用一个词项与其它词项或类别的认知内容去解释词或词组之间的结合。搭配意义在心理学上有相当于一个词必须与特定词结合使用的特征。人类可以自然地使用句法特征或其它常识来说明词之间的搭配方式。然而，这并不是概念的类别关系和词义的内在特征。词的搭配能力属于表层语义表述的推理能力。因此，用表层语义学表述词义的方法，获取词的搭配知识是一个可行的方法。

一个专家 MT 系统需要获取能解决歧义的种种知识，如等同关系、从属关系、代指关系的识别方法等等。它们都可以建立计算模型，用分析统计或人工神经网络的方法实现。当然，这些问题本身就是一个高难度的研究课题，但这是一个专家系统应该追求的目标。

五、 结 论

建造专家 MT 系统的关键是如何利用现代语言学的理论及人工智能成熟的方法来建立知识库和与之相关的推理系统和学习机构，使系统一步步地接近目标。

本系统使用面向对象的方法及框架理论建构知识库，不但有很强的知识表现能力和灵活的知识处理手段，而且较容易把握知识的全局性、完整性和一致性。对象模型统一的数据结构，以及多段的知识表现，体现了知识的通用性、可分解性等特点，还促进了知识库系统的系统化和模块化。采用多变量的解析方法对知识进行组合，可以较灵活地表现事物的个性与特点，它还便于在算法上采用统计分析方法、神经网络和人工智能分析技术。

面向对象的专家 MT 系统上述这些特点，将大大地提高它解决问题的能力，使之成为新一代的具有高度灵活性、可塑性的智能机译系统。

参 考 文 献

[1]Makoto NAGAO (1992), Some Ratonals and Methodologies for Example-based Approach. "Proceeding 1992 International Conference on Chinese Information Processing (1) 58-69

[2] Satoshi SATO and Makoto NAGAO(1990). "Toward Memory-based Translation," COLING-90 VOL. 3 247-252

[3] William A. Gale & kenneth W. Church, (1993) "A Program for Aligning Sentences in Bilingual Corpora", Computational Linguistics, 19 (1) 75-90

[4] 罗彤、钟珞 (1992)，"面向对象的专家系统设计"，〈小型微型计算机系统〉92 13 卷 8 期

[5] 黄建烁 (1992)，"一种基于框架的知识库系统"〈机器翻译研究进展〉电子工业出版社

[6] Velardi. P. M. T Pazieza & M. Fasolo (1991)"How to Encode Semantic Knowledge: A Method for Meaning Representation and Compulter Aided-Acquisition," Computational Linguistics 17(2) 153-70

[7] Mel'Cuk I (1988) "Semantic Description of Lexical units in an Explanatory Combinational Dictionary: Basic Principles and Heuristic Criteria" International Journal of Lexicograph, 1 (3), 165-188