THECMT系统中规则的处理机制

苏玉宏　刘月荣

（清华大学自动化系／外语系　100084）

摘要： 本文阐述了清华英汉机器翻译系统（THECMT）中规则的处理机制，找到了比较有效的规则描述语言，提出了"模糊匹配"的想法，并对规则的优先级及规则解释器的设计作了细致的分析和深入的探讨．

关键词： 规则处理机制，规则解释器，模糊匹配，规则优先级

The Processing Mechanism of Rules in the THECMT System

Su Yuhong　　Liu Yuerong

(The Department of Automation/Foreign Languages, Tsinghua University)

Abstract: This thesis provides a systematic explanation of the rule-processing mechanism in the THECMT system. An efficient "language of rule description" has been established. As an important feature of our system, the idea of "approximate matching" has been elaborately illustrated. The thesis also includes the designing of the rule interpreter and the priority of rules.

Key words: rule-processing mechanism, rule interpreter, approximate matching, priority of rules

一　引言

规则处理是机器翻译中的核心部分，目前，无论是基于规则、基于例子、及基于智能的翻译系统，实质都是用规则来处理机译问题的，基于规则的机译系统不言而喻，对于基于例子的机译系统而言，一条条例子也就是一条条规则，基于智能的机译系统虽然利用了人工智能的一些思想和方法，但现在的人工智能，离开了规则库则寸步难行．不同的处理方法，只是出发点和侧重有所不同而已．可以毫不夸张地说，一个机译系统能否开发成功，很大程度取决于该系统的规则处理机制的成功与否．

本文着重讨论在清华英汉机器翻译系统（简称THECMT系统）中的规则处理机制．

二　规则设计标准

目前，神经元和统计学的方法用于机器翻译，取得了一些成果，例如有人利用神经网络的思想，进行英文中的句法分析，还有人利用统计方法处理歧义，这些方法都开拓了机器翻译研究的视野．但是，仅靠神经网络和统计学方法来处理机器翻译中的所有复杂问题，目前来看仍然是不现实的．至今还没有出现能完全抛开规则处理而开发成功机器翻译系统的例子．因此规则的描述及处理是机器翻译系统中必不可少的组成部分．

由于自然语言具有灵活性，复杂性及开放性等特点，机器翻译系统不可能一挥而就，为了使系统能不断完善和不断进化，规则和程序完全分离－－－这是贯穿我们研究始终的问题．要实现规则和程序的完全分离，规则的设计就显得更加重要了，如果我们的规则只能描述50% 的语言现象，即使我们系统处理完全正确，整个系统翻译的准确率也只有50%，由此可见规则设计的重要性．

由于目前计算机还不能直接处理用自然语言描述的规则，例如，在对a book消歧义时，（book有两个词性：名词和动词），倘若用自然语言来描述，规则为“冠词后面不可能跟动词原形”，根据这一规则就可以排除book的歧义，但这样的规则是计算机无法或者难以处理的．我们必须将自然语言中的规则形式化、符号化，描述成计算机容易处理的形式．

语言是有规则的，规则是可以描述和处理的，除去本征歧义，即利用语法语义分析仍然无法解决歧义，几乎所有的语言现象都可以借助于规则来进行分析和处理的，这是我们在机译研究中的一个根本的出发点．

规则设计的实质是寻找一种“规则描述语言”，自然语言中的规则都能描述成这种语言，而被计算机所处理，机器翻译中的歧义处理及语法语义分析都是通过对这种形式化、符号化的规则进行处理而实现的．

虽然，计算语言学的研究及机器翻译的实践都取得了长足的进展，目前，关于规则如何描述迄今还没有一套行之有效的规则描述语言，几乎是一个系统一个样，甚至同一个系统，不同类型的规则，其规则的描述形式都不相同，这种状况影响了机器翻译的发展．

我们认为，作为规则描述语言，应该有如下特点：

(1) 描述功能强，这是规则设计中最重要的问题，如果规则描述语言只能描述部分语言规则，而大量的语言现象不能用这种语言来描述，这样的机器翻译系统不可能有很高的准确率．一句话，描述功能弱的规则描述语言是没有价值的．

(2) 每条规则的内涵丰富，规则有一定的灵活性，应尽可能地描述语言现象，只有这样，才能大大减少规则的数量，不至于使规则数量发生“爆炸”，如果说一种规则描述语言虽然描述功能强，但很死板，数量庞大得难以接受，则这样的规则描述语言给规则的搜集、匹配、维护等都带来很多问题．下面我们举一个简单的例子：

[例1]　a man (一个人)
　　an old man (一个老人)
　　a very very old man (一个非常老的人)

如果对上面三个名词词组进行抱团处理，如果过于死板，需要三条规则：

Det + n. → NP　　(1)

Det + a. + n. → NP　　(2)

Det + ad. + ad. + a. + n. → NP (3)

但事实上用一条规则就可以描述上面三个名词词组的抱团.

Det + { ad. + {a. + n. → NP (4)

其中 Det冠词 n.名词 ad.副词 a.形容词 NP名词词组，{表示可重复，且可有可无，上面仅一条规则就可描述大量的这种名词词组了.

(3) 简明且容易理解，规则描述应尽量与自然语言的描述方法接近，如果规则描述语法的符号体系及方法过于复杂，象“天书”一样难于理解，则规则维护及扩充就非常困难，规则描述语法最好很容易被一般用户所理解，这样可以给用户更大的自由度，使整个系统更合乎实际需要.

(4) 既能描述语法，又能描述语义，只有这样，才能保证规则有较高的分辩率.由于语义处理尚处于探索阶段，因此，在规则中也应该以语法为主，语义为辅.在计算语言学的文献中，涉及到的规则几乎只描述了语义，如何将语法语义结义起来，是我们面临的一大难题.

三 语用环境的规则描述

自然语言可以近似地看作一个规则系统，大多数的问题都能通过规则处理来解决，准确地描述语用环境，是很有必要的.我们必须承认，规则的适用范围是不相同的，我们不但要处理一般的语法现象，还要处理大量的惯用法，这就要求我们对不同语用环境的规则描述有一个清醒的认识.

按规则的作用范围来看，规则可以分成四个层次

(1) 最窄描述，这种规则不含有可变部分，适应面最窄,主要针对语言中不能分析或难于分析的部分，采用直接转换的方法，这种规则与成语及搭配没有多大区别，格言、谚语及属于这种类型，请看如下例子:

[例2] Seeing is believing. （百闻不如一见）

内容 <see(j, <is, <believe(j (5)

生成 百闻不如一见 (6)

以上这种情况，当然可以放入字典中，从某种意义上看，字典也是一种规则库.

(2) 规则中有部分可变，这主要是对于特殊句型及惯用法而言的.规则中有些部分是不可改变的，这就将该规则的构形勾画出来了.这种规则的覆盖范围要比第一类规则大得多.下面的规则

内容: Nh^/Q^^,<see/hear,N^^/Q^^,V^0,(N^^/Q^^ (7)

生成: A,B,C,D,(E (8)

以上的规则可以用于解决以下两个句子

[例3] He saw a saw saw a saw.

I heard him sing.

以上的规则实际上是 see和hear的特殊用法.特殊词规则及惯用法句型规则都属于这种情况.

(3) 规则中所有的单元都是可变的.这种规则的适用面更宽，也更加抽象.

(4) 管理规则的规则，也就是元规则．从表面上看，这一类型的规则与语用环境的描述关系不大，但这种规则提供了规则的复合及生成新规则的能力，例如凡是有情态动词的句型，只要在该情态动词后加“not”，就可以变成否定句，这个规则就是元规则，可以作用于已存在的规则，导出新规则．元规则的存在大大提高了规则的描述功能．对于复杂的真实文本的处理，要大量研究句子的复合机制，否则，在处理大量真实文本时，陷入规则的汪洋大海之中，规则的覆盖面总是满足不了实际的需要．

本节只对语用环境的规则描述作了一些简单的讨论，我们认为，对自然语言中大量复杂句型的翻译，不深入研究出系统的元规则表示形式及处理方法，就很难有大的突破．

四 规则的优先级

规则的优先级决定了规则匹配的先后次序，前面已经谈到，规则的优先级可以先根据经验给予初值，再根据适用频率作必要调整．

在确定规则优先级时，有以下原则，越是具体的规则，其优先级越高，相反，越是抽象的规则，其优先级越低；越长的规则，其优先级越高，越短的规则，其优先级越低，即长度优先，或最大匹配．

当前语用环境可以激活多条规则时，按优先级从大到小依次匹配．

当两条规则有相同部分时，例如一条是另一条的特例，规则的优先级才有意义，没有相同部分的规则，如果属于同一类型，如特殊词规则，其处理的先后次序对结果没有影响．

下面举两个例子来说明规则中的优先级考虑

[例4] Suddenly he laughed.

该句可以写出如下的句型规则：

D^^, Q^^/N^^, Vi^ (9)

[例5] He laughed.

可写出下面的规则

Q^^/N^^, Vi^ (10)

由于描述[例4]和[例5]的规则有共同部分，但[例4]中的句型规则长度要长一些，根据“长度优先”的原则，[例4]中的规则比[例5]中的规则优先级高．

下面再看一个惯用法优先的例子

动词hear可用于如下一条规则描述的句型：

Q^^/Nh^, <hear, Q^^/Nh^, Vi (11)

我们还知道hear是一个及物动词，及物动词可用于如下句型

Q^^/N^^, Vt^, Q^^/N^^ (12)

前面的句型的优先级要高一些，因为它是针对hear的惯用法而来的，后面的句型规则的优先级要低一些，它对一切及物动词都适用．

“长度优先”及“惯用法优先”是处理规则优先级时两条最基本的规则．

五 规则的匹配模式

规则匹配有两种模式：精确匹配和模糊匹配．

所谓精确匹配就是规则中如何一个单元都必须与当前语中环境相匹配，其中任何一个单元发生差错都认为规则匹配是失败的．

例如：在处理 He saw a saw saw a saw. 这个句子时，规则内容是：

Nh^/Q^^, <see/hear, N^^/Q^^, V^0, (N^^/Q^^ (7)

假如有一个句子是

[例6] The dog saw a saw saw a saw .

上面的规则就不能处理这个句子，原因是 The dog虽然是名词词组，但属于动物的范畴，不属于人的范畴，其语法语义描述为 Nx^，与当前语义环境不符合．

在机器翻译中，我们当然希望不同的语言环境，都能与系统中规则精确匹配，这样译文的准确率也比较高，但是，事情往往不尽人意，例如句中可能有系统不认识的词汇，或者词典中有关词的释义不够完备，甚至有错误，或者在前一层次的规则处理中发生了错误，以上种种情况，都可能导致精确匹配的失败．在这些情况下，规则的模糊匹配作为一种行之有效的补充手段就显得很有必要了．

与规则的精确匹配不同，规则的模糊匹配并不要求规则中每一个单元都与当前语用环境相匹配，它只要求“大部分”规则单元与当前语用环境相匹配，这里的“大部分”就是模糊匹配时的“阈值”，只要能够匹配上单元所占的比例大于阈值，则认为规则匹配在模糊匹配的意义上是成功的．

有人又将规则的模糊匹配称为“不完全知识推理机制”，我们知道，在上面谈到的几种情况中，由于所需要的知识不完全，规则的精确匹配无法实现，模糊匹配可以用来补充精确匹配的不足．

下面我们来看，如何用模糊匹配的思路处理[例 6]中的句子．在［例 6 ］中，规则共有五个单元，其中有四个单元能够与当前语用环境相匹配，这部分占 80%，如果阈值设定为 75%，则在模糊匹配的意义上，匹配是成功的．[例 6]这样的句子就可以得到比较好的处理． 必须指出的是，对规则首先进行精确匹配，只有当精确匹配失败时，才去考虑模糊匹配．规则匹配以精确匹配为主，模糊匹配为辅．另外，在规则匹配中，并不是每一个单元都能进行模糊匹配，例如，当规则中有特殊词时，特殊词单元不能进行模糊匹配，因为惯用法规则有些是针对某个词的惯用法而言的，不具有一般性，也就是说规则中的变元可以进行模糊匹配，规则中的不变部分不能进行这种匹配．

六 规则解释器

我们认为，规则与程序的完全分离，是机器翻译研究的大势所趋，能否实现二者的分离，除了设计出一种有效的规则描述语言外，关键在于能否实现一个功能强的规则解释器．

规则解释器是系统的核心，它就象计算机中的 CPU，而一条条不同的规则就好象不同的计算机指令序列．CPU 的好坏决定着计算机系统的优劣．任何机译系统，究其本质，都是通过对语言中的规则及用法加以处理、实现转换生成功能的，这一切都离不开规则的解释处理．如果程序与规则耦合在一起，其规则解释处理，因规则不同而异，这种规则解释，

实现要容易一些，但在规则维护及系统完善方面存在重大缺陷，规则的解释不具有一般性，不能称其为规则解释器。

规则解释器能够处理任何“合法”的规则，这里的“合法”指的是该规则的描述完全符合系统的规则格式，被系统所认可。规则解释器具有通用性的特点，因而也比较抽象，系统的语法语义处理都是通过规则解释器来实现的，规则解释器的设计与实现，非常困难，是机器翻译中最难解决的难题之一。

在设计和实现规则解释器的过程中，我们主要考虑了如下一些原则：

(1) 解释功能强，凡是系统合法的规则，都能够解释处理，这是对规则解释器最基本的要求。

(2) 语法为主，语义为辅。二者是密不可分的。

(3) 采用了Marcus确定性算法的思想，并做了一些修正，使系统具有“双向看，缓决定”的能力，当前活动窗口，可以包含两个词，也可以扩展到整个句子，一切根据语用环境及规则要求而定。激活窗口的大小不再受限制。避免了使用不必要的递归。

(4) 采用了“边分析边生成”的策略，这样就避免了并存复杂的树形结构，使生成得到了简化。

(5) 进一步可以利用“不完备知识推理”机制，采用模糊匹配的思想，以弥补精确匹配的不足，在系统开发比较完善的情况下，可以引入这一机制。

七 结束语

THECMT 系统实现了规则与程序的分离，建立了良好的规则处理机制，找到了一种有效的“规则描述语言”，该方法不仅用于本系统中，而且还用于汉英机器翻译实践中。规则的匹配以“精确匹配”为主，“模糊匹配”为辅。目前，我们的系统有特殊词规则近四百条，句型规则三百多条，时态语态及其它规则近四百余条，总共只有一千多条规则，虽然规则的数目不多，但是在规则的作用范围内，译文流畅自然，可读性好。在下一阶段的研究中，一个非常重要的任务就是利用真实的文本大量测试系统，不断添加规则，扩大知识库，不断完善系统，逐步实现实用化，商品化。

参考文献

[1] 苏玉宏 “英汉机译中规则与实例的结合与系统实现”，清华大学硕士论文，1995.6

[2] 刘月荣“利用规则消除歧义提高机器翻译的正确率”，清华大学硕士论文， 1995.3

[3] 吴蔚天，罗建林，《汉语计算语言学》，电子工业出版社，1994年7月 第一版

[4] 黄河燕，陈爱萍，“我国机器翻译技术的产品现状”，《计算机世界》月刊，1994年第2期

[5] Harold L.Somers，“Current Research in Machine Translation”，*Machine Translation*, 1993, 7

[6] Michael R.Brent，“From Grammar to Lexicon Unsupervised Learning of Lexical Syntax”，*Computational Linguistics*，Vol.19，No.2