通用德汉机器翻译系统

谢金宝　何星　梁桂森

(上海交通大学)

摘要：　通用德汉机译系统(GCMTS)是通用的将德语自动翻译成汉语的机器翻译软件系统，本软件系统主要包含有词法分析，语法分析和译文调整生成等模块.GCMTS采用与传统的MT系统不同的有条件归约算法以及动词阶位和规则库相结合的分析算法。GCMTS将算法与规则分离。GCMTS是在XENIX 386和 ORECAL 数据库环境下开发的，容易移植和推广。

The Common German Chinese Machine Translation System

Xie Jinbao　He Xing　Liang Guaisheng

(Shanghai Jiao Tong University)

Abstract: The German Chinese Machine Translation System (GCMTS) is a common software system which translates German into Chinese. GCMTS includes lexical analysis, grammar analysis and translated text djusting modules. GCMTS adoptes conditionaly inductive methode and arithmetic of combining verb's valence with rules library which are different from trandi-tional MT system. Rules library separates from arithmetic. GCMTS is developed under XENIX 386 and CRECAL database envirement,so It is easy for immigration and wide use.

一 软件系统概述

德汉机译系统(GCMTS)是通用的将德语自动翻译成汉语的机器翻译软件系统，本软件系统主要包含有词法分析，语法分析，译文调整生成等几个功能模块和词典库，语法规则库，数据库接口等支撑模块。为了使该系统便于维护，易于修改，本系统将算法与规则分离，即算法以程序描述，而规则通过数据库及文本形式表示。本软件系统的开发及工作环境为汉化XENIX_V2. 11及关系型数据库ORACLE_V5.10,用C语言及C与ORACLE数据库管理系统接口语言PRO*.C编制而成. GCMTS系统硬件工作环境为PC386及兼容机。

GCMTS系统算法有如下特点:

(1)由于自然语言的多样性，复杂性及灵活性，企图用若干统一模式的规则去归纳其特点是不可能的，亦是徒劳的，所以本系统在考虑德语本身的特点上，提出了基于动词与规则分析相结合的语法分析算法。

(2)一般的异种语种互译系统，将翻译过程分成语法分析和译文转换二个过程，使得系统的性能大大降低，而作为机译系统，语法分析和译文生成是相辅相成的，本系统采用了语法分析与译文转换生成相结合算法，提出了条件语法分析算法。

(3)将词法分析分成二个部分，第一部分在语法分析之前进行，而将词法的深入分析穿插于语法分析之中。

本系统试运行已有一年，情况良好。

二 词典库

在机器翻译中，词典库的建立是至关重要的，它是整个系统的基础，它的建立方式，完善程度直接关系到翻译质量。词典库有常用单词2000个。GCMTS系统的词典库系统分成二级，第一级是总库，主要提供单词的词类信息，第二级是分类词典库，GCMTS系统将德语单词依翻译需要分成12类，建立了12个分类词典库。建立两级词库的目的是为了加快词典查询速度。

德语的词汇估计有五十万，传统语法按其意义和功能分为十大类：冠词、名词、形容词、数词、代词、动词、副词、介词、连词、感叹词、其中名词占总词汇量的一半以上(50-60%)，动词约占四分之一，形容词和副词占六分之一，故名词、动词、形容词在德语中称为三种主要词类，介词、连词、代词总共只有几百个。总库的功能主要是确定单词是否在词典库中，单词的类别(所属分类库名)以及单词的汉语含义个数。分类词典库依词类不同有不同的结构。下面以名词和动词为例说明分类词典库的结构。

(1)名词词典库

名词是表示生物(Mann、Frau、Hund、Baum)、事物(Haus、Tisch、Zange)和概念(Glaube、Freundschaft)的词，它们构成一个词类。名词在德语中分成三类，分别用冠词(der，das、die)表明，大多数名词在表示多于一个的生物，多于一个的事物或多次使用的概念时，可以用复数形式表明，名词在一个句子中的句法用变格形式表明，并且标志出它的句子中的功能。根据名词的特点，我们必须将其按语义分类，并且在词典中包含语义信息。量词的处理是一件很棘手问题，也无一定的规律可循。因此在词典中根据语义设置量词信息。

(2)动词词典库

动词词典库的建立必须顾及以下二点:

a) 动词是句子的核心,语法上往往有不同的要求,如价位(Valenz).从根本上说,句子中的其它成分都可以看做动词的补足语(Ergaenzungen).

b) 同一动词, 根据所带补足语的不同, 其翻译也不一样, 而且同样的名词补足语, 根据名词的语义分类不同, 译文也不同。这里的名词补足语指以名词为核心的补足语, 如第一格补足语等。

按照以上两点, 我们把动词的词条定义为如下格式:

动词词干, 动词类型, 动词译文方法, 动词价位, 是否有带介词形式

三 词法分析

词法分析是语法分析的基础。词法分析的任务是确定单词的词类以及获得有用的语法信息。单词在句子中呈现各种形态, 可以把单词看成是前缀,词干和后缀三部分组成的。前缀和后缀统称为词素, 词素是词法分析的最小单位。德语词素的形态变化是十分丰富的, 正是这些丰富的词素形态, 为语法分析提供许多有用的信息。德语单词可以分为两大类, 一类是无词形变化的, 例如介词, 连接词、副词等。另一类是有词形变化的、例如动词。名词、形容词等。通过对词素的分析可知, 对于后缀e, 可能的词类为动词、形容词、数词和名词。此时要确定单词的词类需要借助于上下文分析以及语法分析。可见词法分析虽是语法的基础, 但两者也不能绝然分开。GCMTS的词法分析获得词类等基本信息, 关于词的更多信息(格信息等)将通过语法分析获得。

在词干、词缀分析完成以后, 系统依词根将单词的所有信息从词典中取出, 并进行进一步的词法分析。

四 语法分析

语法分析为机译系统的关键之所在, 算法的好坏直接关系到系统的性能和译文的正确性。对于异种语言之间的互译问题, 不少学者对此进行了深入的研究, 提出了各自的算法, 得到了一些有益的结论。本系统结合德语的语法特点, 提出了基于动词与规则分析结合的语法分析算法及语法分析与译文生成相结合的条件归约算法。

语法分析模块由以下几个功能子模块组成:

形容词条件归约, 名词类处理, 节点调整, 时态、语态分析, 动词谓语分析和语法规则库

1 语法分析的主要算法

由于自然语言的形成与发展有其复杂的社会、文化、历史背景，造成自然语言的多样，灵活及不严格性，虽然有一定的语法规则可以遵循，但是，若企图用有限的规则来表达所有的语法现象，则是十分复杂繁锁的工作，甚至于不可能完成的。在德语的语法分析中，我们发现动词起到了决定性的作用，动词本身不仅有其含义，同时也隐含了价位信息，包含了对句型译文的许多信息。针对这种现象，我们选取了基于动词与规则分析相结合的语法分析算法，并作为DCMTS系统的核心算法，此算法的主要特点为:

先进行语法规则的归约，即依据语法规则库，完成形容词属性节点和名词属性节点的归约，在此基础上，从词典中提取句中动词的价位要求，依其对句型的要求进行匹配获取相应的译文信息进行分析，然后生成可接受的中文译文。

另外，GCMTS系统采用了有条件归约，即语法归约与译文生成相结合的分析算法。一般的机译系统，将语法分析与译文生成作为二个模块分开进行考虑，在语法分析成功的基础上，再依转换规则进行译文生成处理，这样的思想固然有其优越性，但在自然语言中，其灵活性及不严格性，使语言经常出现特殊例子，若发生这种情况，译文生成不成功，则必需回溯到语法分析中，从头开始，从而影响了系统的性能。DCMTS系统考虑上述情况，将归约与译文生成统一起来一起考虑，提出了有条件归约的语法分析算法，即在进行语法节点归约时，以译文生成为其条件，主要过程为：当出现匹配成功，即调用译文生成过程，若成功则此归约成功，否则则认为此归约失败。

2 语法分析主要模块说明

(1) 形容词条件归约

在德语中，作为形容词语法单位的类型共有二种。即：形容词+形容词和副词+形容词，其中形容词+形容词为无条件归约。当出现此类现象时，即进行合并，形成新的形容词节点，中文含义为二者合并的含义。副词+形容词为有条件归约，因为有些副词用来修饰动词或修饰整个句子，而另一些则用以修饰形容词。

(2) 名词处理

名词处理分成二个步骤来进行。

(a) 确定名词的格

由于德语中，名词的格对句子理解起着至关重要的作用，所以在对名词成份分析之前，我们必须先尽可能对格加以确认。名词的格主要由其前端的定冠词，不定冠词或形容词性代词来确定。若无法唯一确定，则我们认为此名词具有二种格。

(b) 名词归约处理

名词归约包含有名词词组(NP)和介词词组(PP)归约两种类型，对于名词词组归约(NP)，主要分成以下几种类型:

i). 含形容词词组修饰

由于形容词对不同语义类型的名词具有不同的汉语翻译方式，我们将此类情况独立出来，对名词的语义类型进行判别，并由此语义类型得到此处形容词的译文，若匹配失败，则返回为归约失败.

ii). 由冠词、数词、及形容词性代词修饰名词词组

此类问题较为简单，主要注意量词的翻译。

iii). 由第二格名词、介词词组修饰名词词组

对于此类情况，仅是进行匹配归约。

iv). 同位语归约

对于同位语这类语法现象，多发生于称谓名词，系统考虑在两个名词若格相同，语义类型也相同时，即为归约成功，否则归约失败。另外，在进行名词归约时，若对句子进行正向扫描，会出现岐义。如 das Maedchen im Zimmer meines Freundes.（这位在我朋友的房子里的小姐）为了预防此类错误的发生，我们采用反向扫描归约的方式，进行名词性词组归约.

3 节点调整

这一过程主要是为进行动词分析之前的预处理，将一些有明显特征的节点进行调整，以便使译文顺利生成。下面列出第一位节点为非名词词性节点处理过程。

当第一位节点为非名词词性时，一般为疑问句形式或倒装形式，我们将此类情况分成以下几类情况处理:

a) 若为一般性副词，则与句子中的第一格名词进行互换。

b) 若为疑问代词，则不进行调整。

c) 若为 wessen ，则与所修饰的名词结合，

4 时态、语态处理

德语的时态、语态一般依动词来体现，主要包括过去时处理，将来时处理，现在完成时处理，被动态处理，情态动词处理等。

5 动词分析及译文生成

正如前面所提及的，在德语中，动词对语法分析及译文生成起着决定性的作用，在动词词典库中，我们可以提取出动词对句子价位要求，依不同的句子语法信息、组成，可以给出不同的译文。如对于动词 trink. 有二种译法:

a) trink+A(00)　　即若有第IV格名词,译为“喝+A”

如　Ich　trinke　etwas　Wasser　　(我喝一些水)

b) trink+#　　即若无名词,译为“喝酒”

如　Wir　trinken　jetzt　　(我们正在喝酒)

6 语法规则库

为了易于系统维护,GCMTS系统将语法规则库与算法分离,规则库以文件形式存储。系统在进行语法规则库修改时,首先对此文本文件进行修正，然后与系统连接,再进行编译则可以进入系统内部使用。

五 译文调整和生成

由于节点调整时已经使每个节点的译文作了调整，因此对于完成了语法 分析步骤后的句子来讲,一般只包含有以下几个部分:

[NP1] + VP + [副词 | 介词词组 | 其他状语成份]

这样我们认为只有三个部分组成,即:

主语 + 谓语 + 状语

对于一般疑问句,应在句末加语气词“吗”?

而对于一般陈述句,则依据 [时间状语 + 介词词组 + 地点状语] 的顺序与谓语合并,形成“主语 + 谓语” 而进行直接归并而生成句子。

如:　Wir　lesen　heute　abend　im　Klassenzimmer　Buecher.

在这里,句子成为

Wir　lesen　Buecher　heute　abend　im　Klassenzimmer.

而译为:　　我们今天晚上在教室里看书。

六 测试结果

对100个句子进行了测试。句子类型包括陈述句，疑问句，否定句，祈使句和感叹句。动词时态包括现在时，过去时，将来时，现在完成时和过去完成时。从测试结果看，译文意思正确，语句通顺的占百分之66.7%,译文意思可以理解，需要少作修饰的占23.3%，译文意思不正确 ，或不能翻译的占10%.下面是部分翻译的句子:

Wir verreisen im Sudeamerika.

我们在南美洲旅游。

Wohin geht unsere Lehrerin immer um 5 Uhr 30?

我们的女教师大约5点30分经常走往哪儿?

Ich will heute abend die Hausaufgabe machen.

我想要今天晚上做家庭作业。

Er schickte mir eine Gitarre.

他送给过我一个吉他。

参考文献

[1] 谢金宝，梁桂森，　<<德汉机器词典设计>>，<<软件应用与开发>>，1995年第2期

[2] 谢金宝、葛显平，　语法分析器JGCYACC　<<复旦大学计算中心通讯>>，1993年第3期