配价制导的德汉机译系统中词素词典的微结构

许玉祥 柴佩琪

(同济大学计算机系)

摘 要:本文给出了配阶制导的德汉机器翻译系统中词素词典的一种微结构。并简要介绍了与[相关的词素词典的物理存储结构以及配价制导的德汉机器翻译方法。

Microstructures of Morphological Dictionary in Valenz-Directed Germany-Chinese Machine Translation System

Xu Yuxiang Chai Peiqi

(Department of Computer Science,Tongji University)

BSTRACT: A microstructure of morphological dictionary in Valenz-Directed German-Chinese Machine Translation ystem is given. The related physical strucrure of morphological dictionary and the method of Valenz -Directed erman-Chinese Machine Translation are briefly summerized.

1 德语配价句法概要

对德语的句法描述主要有两种体系，即成分结构语法(Constitution Grammar)和支配关系语法 Dependance Grammar)。成分结构语法常常被称为传统语法。传统语法认为德语句子由词组并列而成；而词组则由词并列而成。在对德语句子进行分析时，传统语法把句子分解为主语、谓语和状语等。动词被包括在谓语中，且分为及物动词和不及物动词两种。由于下列的两个主要缺点使得传统语法不适合于德汉机器翻译:

不能描述一个句子中的各个词之间的有主有从、有领有属的关系。

仅仅知道一个动词为及物动词或不及物动词还不足以指导对德语句子的分析，还很难写出句法分析程序。例如在句子

Ich erinnere ihn an sein Versprechen.

(我提醒他记住自己的诺言。)

中，按传统语法动词"erinnern" 是一个及物动词，人称代词"ihn"是它的宾语。介词短语"an sein Versprechen"实际上是动词"erinnern"所要求的一个成分，但按传统语法，动词只能带直接宾语或间接宾语，不能带介词短语，所以必须将介词短语"an sein Versprechen"看作为状语，这显然是不合理的。

支配关系语法则认为一个句子的各个组成部分之间有一种领属关系，就全句而言，动词是支配核心。

德语配价句法属于支配关系语法体系。法国人Lucian Tesniere 首先提出了德语的配价语法。德语配价句法不承认“主语”在句子中有特殊地位，与此相反，它认为动词是句子的核心，句子的其他成分则受动词的支配。受动词支配的成分称为补足语(Ergaenzung,Actants)。句子中除动词和补足语以外的其他成分称为说明语(Angaben),说明语表达句子中的动作发生的环境。

Engel/Schumacher在〖1〗中把补足语分为表1所示的十种。有了补足语的分类后，根据动词可以支配的补足语的种类和数量，可以进一步对德语动词进行分类。

表1 补足语的种类

代　码	名　称
0	第一格补足语
1	第四格补足语
2	第二格补足语
3	第三格补足语
4	介词补足语
5	情状补足语
6	方向补足语
7	名词补足语
8	形容词补足语
9	动词补足语

〖2〗中将动词划分为46种，并且用数码表示各类动词。例如013表示一类动词，它们支配的成分是第一格补足语、第四格补足语和第三格补足语。动词schenken是013类动词的一个实例。在句子

Er schenkt mir ein Buch.

（他送我一本书。）

中，"Er"是第一格补足语；"ein Buch"是第四格补足语；"mir"是第三格补足语。

2 配价制导的德汉机器翻译

所谓配价制导的德汉机器翻译，是指在机器翻译的各个阶段中，尽可能的利用德语的配价信息。在词法分析阶段，当识别出一个词为动词时，应找出其配价信息；在句法分析阶段，

应根据一个动词的配价进行句子的句法分析；在译文生成阶段，德语句子到中文句子的转换规则应以动词的配价为根据给出。

由于动词的配价只说明了动词与它的支配成分之间的关系，所以不能用动词的配价去制导从句的分析。从句的分析主要通过带出从句的连词的信息和德语从句的构成规则进行分析。如何在不进行句法分析的情况下进行德语从句的分析，我们将在另一篇文章中叙述。

动词的配价信息中同样也不包含动词的时态和语态信息，在利用配价对德语句子进行句法分析前，也必须进行时态和语态的预出理。

图1给出了配价制导的德汉机器翻译系统的一般流程。

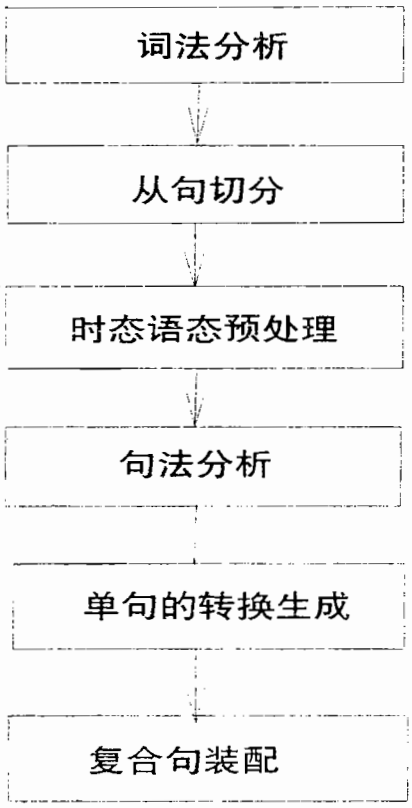

图1 配价制导的德汉机器翻译系统的一般流程

3 配价制导的德汉机器翻译系统中词素词典的微结构

从图1可以看出，配价制导的德汉机器翻译系统中，词法分析阶段和单句的转换生成阶段需要有词典的支持。考虑到实现和维护的方便性，可以为词法分析阶段建立一部词素词典；为转换生成阶段建一部转换词典。

由于翻译用的词典的规模是庞大的，它们不可能常驻内存，而只能以文件的形式驻留在磁盘上。为了支持翻译过程中的大量的实时查找，必须采用有效的直接查找技术。理想的直接访问技术是只用一次磁盘访问就可以获得词典查找的结果。对于双语词典，只用一次磁盘访问就获得词典访问的结果是容易实现的，只要在磁素词典中对于每个词存入它在双语词典中的位置即可。

德语中的一部分词在句子中要进行变位和变格。在进行变位和变格时词尾要发生变化。在词素词典中，为了节约存储空间，通常对规则变化的词只存储词干，不存变化的词尾。而对不规则变化的词，将它的每一个不规则变化的结果作为一个独立的词存入。在词素词典中查找一个词的过程为:

```
repeat
        look for word in dictionary
        if not found
                then modify the word
until word is found or
        no further modification is possible
```

要达到用一次磁盘访问就得到查找的结果比须精心地构造词素词典的索引。使用深度可变的多层次Trie索引〖3〗〖4〗〖5〗可以实现这一目标。

在深度可变的Trie索引中，索引的每个层使用的索引值是被查词的相应位置的字母。每个索引由26个记录组成，其数据结构:

```
TYPE
        letters=('a','b','c',…,'z');
        statustype=(EMPTY,EXISTS,NEXT,LEVEL);
        TRIENODE=
                record
                        status:    statustype;
                        first_block:integer;
                        last_block: integer;
                        start_byte: integer;
                        next_index: TRIE;
                end;
        index=array[letters] of TRIENODE;
        TRIE=^INDEX;
```

每一个Trie结点表示一个索引记录；域ststus表示一个字母序列（不完全的词）在词典中的状态。状态EMPTY表示该字母序列开头的词在词典中不存在；状态EXISTS表示该字母序列存在，且以这个序列开头的词可以一次读入内存；状态NEXT_LEVEL表示以这个序列开头的词很多，需要在下一层继7查找。这个序列的下一个字母的索引块由next_index指定。

下面给出在词素字典中名词、形容词、动词和其它各种词应包含的信息。

名词的信息:
词条长度
词干（即名词的不变部分）
词性
语法性（即阳性、阴性或中性）
数的限制（只取单数或只取复数）
复数后缀（构成复数时采用的后缀）
单数第二格后缀（用作第二格时采用的后缀）
在双语词典中该词的位置

形容词的信息:
词条长度
词干
词性
词干属性（原级、比较级或最高级）
变音位置（构成比较级或最高级时该位置的词要变音）
在双语词典中该词的位置

动词的信息:
词条长度
词干
词性
前缀
前缀的可分离性（即前缀是否可分）
时态
语态
完成时助动词（用haben或Sein作助动词）
配价数（可以有几种配阶形式）
配阶编码（若可以有多个配价，则给出各个配价）
在双语词典中该词的位置

其他词的信息:
词条长度
词干
词类
词性要求（只支配哪一种语法性的词）
数的要求（只支配单数或只支配复数）
在双语词典中该词的位置

4 结 论

采用配价以后，机器词典的结构可以明显地得到简化．目前我们正在实现的一个德汉机器翻译系统就是一个配价制导的系统．初步的试验的结果是令人满意的

参 考 文 献

〖1〗Ulrich Engel/Helmut Schumacher :Kleines Valenzlexikon deutscher Verben,Verlag Gunter Narr Tuebingen 1976

〖2〗Ulrich Engel: Deutsche Grammatik

〖3〗张永奎 & J James R Cowie:机器可读词典的快速查找技术，中文信息学报8:2(1994)