实用型汉英机器翻译系统的研究与实现

孙广范

(电子部计算机与微电子发展研究中心,北京,102206)

The research and realization of a practical Chinese-English machine translation system

Sun GuangFan

(Research Center of Computer and Microelectronics Industrial Development,MEI)

我们研制的实用型汉英机器翻译系统包括汉语自动分词、句法及语义分析、转换及生成三个模块组成。这三部分具有各自的知识库，是基于规则制导的相对独立的工作模块。从另外角度看，系统也可以分成词典、程序、规则库三部分。本系统在对基本汉语语言现象进行一般处理的基础上，重点考虑了面向科技文章的翻译，同时兼顾口语翻译。我们从实际文本中选取了一百万字的典型语料并且全部进行了调试，总结出了一万条汉语分析、英语转换及生成的产生式规则，其中共性规则和个性规则各占一半。词典中有十三万词条。

词典中记录着机器翻译需要的词法、句法、语义等静态信息，是上述各模块正常工作的基础。系统采用知识库与程序分离技术，知识的表示形式采用产生式规则形式。分词部分是将中文句子切分为各个独立的词，其中涉及到切分歧义的消除问题。句法及语义分析部分对汉语句子进行分析，利用词典中的静态知识及知识库中的规则将汉语句子归并为一棵多标记的多叉句法树，其中涉及到歧义及兼类的处理。转换及生成部分以句法分析得到的句法树为基础，利用知识库中规则对其进行转换处理，将句法结构树转换成适合于英语表达习惯的句法树，然后生成英文句子。

通过对科技文章进行分析及调试，发现科技文章中的句子具有如下的特点：兼类词出现频率较高，长句子较多，括号出现较多。针对上述特点，采取了如下的处理方法：对于兼类词多的问题，通过完善及扩充分词及分析部分的规则库，改进相应的消兼机制，既能分析出正确的句子结构，又不降低系统处理的速度。如果句子中出现了括号，先将括号内的分句单独分析，再与全句一起分析。这样防止了括号内的词与括号外的词在分析时相互干扰的现象，从而保证分析出正确的句法结构。

本系统规则数量及处理能力基本覆盖汉语科技文章的各种句型及常见的表达方式，对于特殊词的特殊用法也考虑得比较全面，有五千条处理词的特殊用法的个性规则。例如，与“的”字有关的规则有时 450 多条，“是...的”结构、“所”字结构、介词与方位词搭配结构等均有相当数量的处理规则。本系统具有一套完善的规则描述语言，保证了规则与程序分离，可以方便地描述系统所需要的各种规则。

本系统的知识表示采用框架结构，易于扩充和修改，有利于研制者方便地调试系统。本系统采用基于“合一”运算的复杂特征集描述方法。

根据汉语分析需要，本系统综合运用了句法、语义信息来消除兼类和句法歧义，在句法分析深度不够时进行语义分析。