TUXS 中的词法分析

(The Lexical Analysis in TUXS)

许玉祥　柴佩琪　臧德滋　赵晓东

同济大学计算机系

摘　　要

德汉机器翻译系统TUXS(Tongji University Translation System 暂定名)是由国家自然科学基金资助的项目. 系统的设计目标是实现科技德语文章到汉语的自动翻译.

就翻译方法而言, TUXS是一个配阶制导的(Valenz Directed)的系统. 所谓配阶制导是指在对源文句子的句法分析和译文生成的过程中利用动词的配阶(Valenz)信息, 用动词的配阶信息指导句法分析和译文生成的进行.

TUXS的开发和运行的硬件环境是SUN3 微机工作站. 软件环境是SUN UNIX. 整个系统用C语言实现. 可以方便地移植到其他的软硬件环境中去.

德语单词在句子中的形态变化十分丰富. 虽然增加了词法分析的复杂性, 但却提供了一些有助于句法分析的深层信息. 将这些信息分析出来有助于提高句法分析的效率. 出于以上原因,TUXS系统中确定的词法分析的目标为:

1. 识别出源文中的每个词, 并确定其词性.
2. 根据词形变化分析出有关性、数、格的信息.
3. 对可能构成固定搭配的词给出标记.
4. 对可能是从句的句子给出标记
5. 将源文中的每个词从变化形态恢复到原形.

其中任务 1 是词法分析的基本任务; 任务 2 则充分利用了德语的词形变化特点; 任务 3 和任务 4 的设计是为了提高句法分析的效率; 任务 5 可以在任务 1 的基础上方便的实现.

TUXS中用于词法分析的字典按以下原则构造:

1. 不同的词类采用不同的词典
2. 字典中只存词干, 不存词尾
3. 为每个字典建立B 树索引

对于从事德汉机器翻译工作而言，本文的工作无疑是有意义的。本文中所采用的词法分析方法对于其他的具有类似的词形变化的语言的词法分析也具有借鉴意义。词性的自动识别不仅可用于机器翻译，也可用于单词频率统计、文体分析等语言学领域，也可用于计算机辅助外语教学。