汉英机器翻译系统THCEMT VERSION 1.0 的设计

（摘要）

严浩，顾宇杰，陈圣信

清华大学自动化系、外语系

清华汉英机器翻译系统（THCEMT 1.0）是由清华大学自动化系和外语系共同研制开发的小型实验系统．该系统拥有一个近七万条记录的汉英电子主词典，一个计算机专业词典和三个形态变化词典；拥有准确率较高的汉语分词系统和初步的兼类词排歧、短语抱团和句型匹配三个分析层次．规则库系统采用有效的统一的规则描述语言，句法和逻辑语义信息可以分享和传递．现在已经对受限的频率最高的16种汉语句型的160个汉语句子进行了测试试验，取得了较好效果．本文介绍THCEMT Version 1.0 实验系统的系统设计．

系统的设计以将来的实用化为目标，把汉语句法结构的易实现性和系统可扩充性放在首位，进行模块化设计．要求各子系统功能完备，分析能达到一定精度．整个系统能演示翻译全部过程．系统的机制是"基于规则"为主，适当辅以语料统计方法．系统由五个子系统组成，各子系统任务明确，功能互补，它们是：（1）文本预处理子系统．包括文本分析，标点符号处理，以及自动分词；（2）词处理子系统．包括汉语词查字典和兼类词消歧；（3）分析与生成子系统．包括短语合并、句型匹配及统一生成；（4）词典及其维护子系统；（5）规则及其维护子系统．

系统词典采用 Foxbase 数据库文件格式，每条记录包括四个域：汉语原词(HZ)，控制符(X)，特征字(G)，英文译文(E)，其中'X'、'G'字段包含语法、语义信息和词处理控制信息．词典收词原则有三条：(1)相同汉语词，只能在 'X' 或 'G' 字段符号上有区别时才能分为多个记录；(2)相同汉语词，对应多条记录时，应将最常用的一项放在前面；(3)不收生僻词，不收不能独立成词的单字．

系统规则设计原则是规则与程序完全分离和规则的通用性、易维护性． THCEMT 系统中，规则采用统一格式，统一符号．规则库同样采取 Foxbase 数据库格式，每条规则有十个字段和三个描述字段．为清楚描述句子结点和结点左右的上下文环境，同时兼顾语法、语义信息，制订了一套标准符号来描述．书写规则时，这些符号可任意组合，系统对这些符号的组织是递归的．所有规则存放在同一数据库中，各类型的规则通过' TYPE' 字段区分开．同一类型规则按优先级排序，优先级确定原则：（1）越是具体，描述越是详细的规则，其优先级越高；（2）越长的规则，优先级越高；（3）特殊词优先原则．

汉语句子数据结构贯穿系统处理过程始终．本系统设计了以下的数据结构：从句子平面看是一条双链表，双链表的每个结点代表一个汉语词或词组．而这每个结点都是一条单链表，对应单词的多个词条，每个词条用复杂特征集 Phrase 来存储．

THCEMT系统选用面向对象的编程语言Borland C++实现．由于C++特有的性质，使得系统的数据封装性好，代码的可重用性强，便于系统的模块化开发和升级．