汉英机器翻译中的简单语境处理

刘海军　陈肇雄　黄河燕

中国科学院计算技术研究所机译中心　100080

A Simple Semantic Environment Processing System in Chinese-English Machine Translation

Liu Haijun　Chen Zhaoxiong　Huang Heyan

汉英机译的研究工作已有许多，发展了许多机译系统。但现有的机译系统都是集中在句子层面上进行翻译，没有利用句子存在的具体篇章的环境信息，导致许多问题如汉语句子的分词、译文时态的确定等不能处理或处理不好。本文介绍基于汉英智能机译系统构造的语境系统 IMTENV，用该系统可以处理一些句子机译系统所不能处理的问题。

篇章机译中重要的是构造一个篇章环境，句子翻译时和这个环境进行交互以去除句子分析的多义性问题。IMTENV 中，语境信息的表示采用局部信息表示和全局信息表示相结合的方式。局部信息按句子功能成分保存有篇章中每句话的详尽信息，如句子的主语、宾语、谓语、状语以及各部分中词的语法、语义、语用等信息。这部分的表示中，要考虑到真实自然语言中句子的各种情况，如并列主语、主语从句等。以主谓宾形式表示句子的原因是：便于处理翻译中遇到的问题（很多语言现象都是承前句的某功能成分而进行的）、和中心理论（Centering Theory）的要求相一致、便于从句子的分析结果树中提取等。全局信息中保存有语言段开始标志、句子的局部中心、简化的句子表示以及环境的时态等。智能汉英机译系统的句子分析结果是一棵用复杂特征集标注的分析树，句子分析完后，IMTENV 对句子分析树进行再处理提取句子的各功能部分的详尽信息，然后从局部信息中精炼出全局信息，包括确定新句子是否开始了一个新的语言段。篇章翻译开始时，IMTENV 的环境信息库是空的，随着翻译的进行，系统不断从句子分析中提取信息来充实环境信息库。

词法分析、语法分析、语义分析和译文生成等阶段都有单句分析不能处理而只有利用环境信息才能解决的问题。分析时单句不能处理的问题主要是组合歧义问题，IMTENV 中的处理分成两步：考虑到语言应用中的线性同现现象，首先判断环境中是否出现过和歧义组合之一相同的形式；第二步判断环境中是否出现过可以和歧义组合中某组合语义相容的词。有些省略问题，如果不用环境上下文就不能处理，如“我明天打”，没有作用对象就不能确定“打”的译文，利用保存的环境信息可以解决这类问题。汉英机译中，译文时态的确定是个重要但困难的问题。英语中要求前后文的时态相一致，否则译文难以理解甚至会造成曲解。根据汉语的特点，在 IMTENV 中，以语言段为单位保存有三种环境时态：过去时、现在时和将来时。完成时、进行时等时态在汉语的每句话中都有明确的标志，如“了”、“正在”等。以语言段为单位，新句子的时态是环境时态和句子时态的交。