CEMT－Ⅲ 汉英机器翻译系统的设计与实现

张 民　　李 生　　赵铁军

（哈尔滨工业大学计算机系，哈尔滨，150006）

摘要：本文论述了CEMT－Ⅲ型汉英机译系统的理论设计和实现情况，主要阐述了系统的语言模型、总体结构和设计思想，并对系统的翻译处理机制和实现策略进行了详细的描述，同时，还讨论了系统在开发实践过程中所遇到的语言学工程和计算机科学等方面的实际问题。本系统已于1993·5在京通过了技术鉴定。

关键词：机器翻译，汉语，英语，自然语言处理。

一、 概述

本文论述了CEMT－Ⅲ汉英机器翻译系统的理论设计和实现情况，包括其语言学模型、总体结构、设计思想、翻译处理机制和实现策略等，并讨论了系统在开发实践过程中所遇到的实际问题，如对机译的最大难点—歧义问题的处理。

系统根据汉语的特点，采用了具有复杂特征集的多标记、多叉树结构（MMT）作为其语言模型。总体结构的设计运用知识工程和软件工程的方法，采用面向对象和数据封装等技术，做到规则和程序相互独立，形成模块化、开放式结构。系统总的设计思想是多种知识并用和面向实用化，对各种有用知识兼收并蓄，尤其是语言学和计算机科学多年的研究成果。系统在设计实现上，做到独立分析、独立生成。

从系统实现的角度看，系统主要由自动分词、查词典、兼类处理、源文分析、译文转换和译文形态生成等六个部分组成。

本系统已于1993·5在京通过技术鉴定，目前系统已拥有词汇4万余条，各类规则3600余条，鉴定测试表明，对于封闭语料，译文准确率78%，对于开放语料，译文准确率67%，翻译速度为3500字/小时（IBM PC386/33）。系统已处于实用化前期，具有很好的应用前景。

二、 汉语的特点和系统的语言学模型

机器翻译是一个由源语言向目标语言映射的过程，整个翻译的过程是在一套适合计算机的语言学规则指导下进行的。这些规则的集中反映就是语言学模型，语言模型是对自然语言规律性的形式化描述。

目前，计算语言学界的语言模型种类很多，有基于语法（或转换）的；有基于语义的；有基于知识的。那么，采用什么样的语言学模型较好呢？本系统从汉语的特点出发，采用了基于语法和语义等多种知识并用的语言学模型，具体地说就是基于复杂特征集的多标记、多叉树结构（MMT结构）和"合一"运算的语言模型。

MMT模型（Multi branch Multi labeled Tree）最早是由我国学者冯志伟提出来的，并曾有效地应用于法汉、德汉机译实验系统中。

MMT的定义如下：

1. MMT 有且只有一个根结点R；
2. 如果R有儿子，则每个儿子都是一棵MMT；
3. MMT的任一结点都有0～N个儿子；
4. MMT的每一个结点都具有复杂的特征集；

"合一"是对MMT进行运算的方法。系统之所以采用这种模型，这是由汉语自身的特点决定的。

汉语的特点决定了MMT语言模型的结构和特点。现代汉语具有如下特点：

1. 形态不发达

形态是词的句法功能的一种形式标志，对汉语来说，几乎没有形态变化，因而单独从字面上看不出该词的句法功能。

2. 语序是最重要的组合手段

语序是指词语在组合中的排列顺序。在英语中，词语的用法功能是由形态表示出来，但在汉语中，语序是区别结构和语义的重要手段。例如：“三天下一场雨”和“一场雨下三天”，前者是VO结构，后者是SV结构。

3. 语义上相联系的单位通常是相邻的

对于形态变化丰富的语言，从形态上可以看出词和词之间的一定组合关系，从而可获得某些语法意义。但对汉语来讲，语义搭配是语言组合的实质，汉语相邻词之间通常存在语义搭配。

4. 汉语句子的构成原则和词组的构造原则基本上是一致的

这是汉语独有的特点，体现出汉语结构的一种可递归性。汉语从词到短语是一种组合关系，而词组到句子却是一种实现关系。例如，句子可由NP＋VP实现，但NP＋VP不一定组成句子。如：“他打人的事实”中的“他打人”是由NP＋VP构成的主谓词组（SV），而非句子。

5. 汉语的词类或词组类型与其所承担的句法功能之间不存在一一对映关系

这说明汉语中存在着大量的歧义现象，如：词类的兼类，句法的结构性歧义等。例如：“VP＋NP”，既可表示动宾结构（如：这家公司可以“出租汽车”。），又可表示一个NP短语（如：我坐“出租汽车”回家。）

6. 依存语法研究表明，对于自然语言有四条公理（Robinson, J. J, 1970）：

a、一个句子中只有一个要素（中心词）是独立的.

b、其它要素直接依存于某一要素.

c、任何一个要素不能依存于两个或两个以上的要素.

d、如果A要素直接依存于B要素，而C要素在句子中位于A、B之间，则C或直接依存于A，或直接依存于B，或直接依存于A、B之间的其它要素。

另外，清华大学黄昌宁先生提出依存语法的第五条公理：

e、中心词左、右两边的词（中心词除外）相互不发生依存关系。

由特点6可以看出汉语句法结构构成过程的先后性和依存性，我国著名语言学家吕淑湘先生也曾在其《汉语语法分析问题》一书中指出，“任何一个语言片断都是由若干短语组成的，这些短语的合成具有先后性，是一层一层组织起来的。”。这表明汉语具有较好的层次立体结构，树形图是表达汉语这种多层立体结构最自然的方式，MMT正是一种树形图，这也就是MMT的层次性结构。特点4表明汉语结构的可递归性，层次结构是表达这种递归结构最恰当的方式。另外，层次结构还有助于分化特点5所说明的汉语的句法歧义结构。

汉语从词到短语结构到句子的构成方式都比较复杂，每次可能有多个低层短语结点一次合成一个高层短语结点，这就决定了MMT结构是多叉的而不是二叉的，即多分叉原则。多分叉可以合理地解释汉语的语言现象，把句子的格局清楚地表示出来。例：汉语的兼语式，“／让①／你②／出去③”，这个兼语结构生成高层结点时是一次完成的，①、②、③结点的提升不具有先后性。另外，当引入后文介绍的词驱动规则进行分析转换时，由于固定搭配结构不定长，而向高层的生长是一次性的，就必须采用这种多分叉原则。

自然语言是一个无限集，而且是一个模糊集，特点1表明汉语几乎没有形态变化，因此，对MMT每个结点特征的描述必然是一个“多值标记函数”，利用汉语中丰富的语法、语义特征，这也就是MMT的多标记原则。MMT正是采用了这个原则才克服了短语结构文法难以处理自然语言的歧义和生成能力过强的弱点。特点1、2、3表明，对于形态不发达的汉语，不可能从字面看出其句法功能，相邻语序的语言要素之间的语义搭配是语言组合的实质，而要表达这种语义搭配就必须利用结点的复杂特征通过上下文相关处理来解决。例如，汉语中词类和句法功能相同的结构，但他们内部结构关系却可能不同，如，“NP＋VP”，既可表示主动关系（如：“／小王／工作”），又可表示被动关系（如：“／书／买了”），还可表示工具（如：“／左手／拿笔”）。对于汉语这种复杂而又常见的“同构”现象，只有用复杂特征才能加以区别。

“合一”（unification）是对MMT复杂特征运算的方法，他判断MMT短语结点复杂特征的相容性，从而生成更高层的短语结点，并进行信息的提取、生成和回传，最终生成一个MMT的树形图结构。

MMT模型作为信息的载体，贯穿了分析、转换的全部分。“合一”是对MMT的运算，因此，系统采用的是基于MMT和“合一”运算的语言学模型。

三、 系统的总体结构及其设计思想

3.1 系统的总体结构及工作原理

系统的总体结构及工作原理如下图：

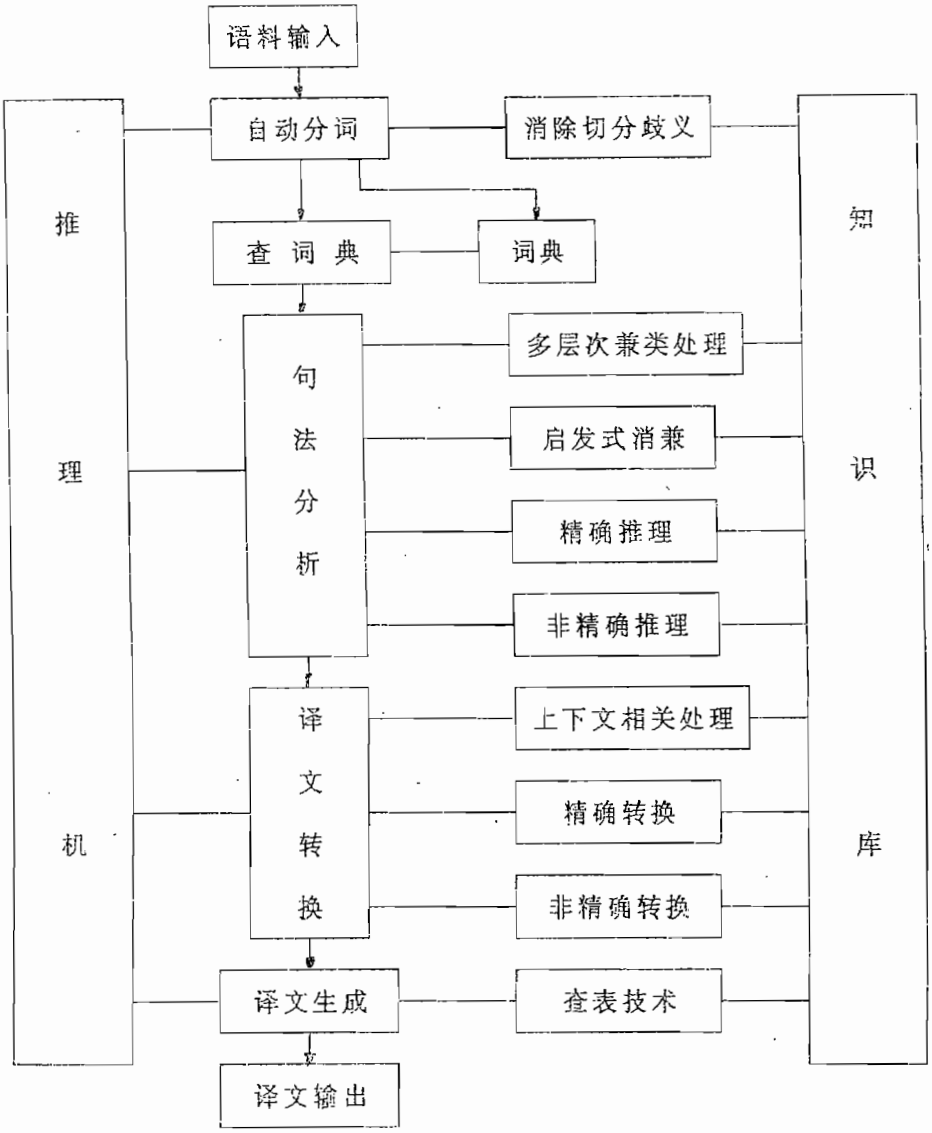

图1 系统的总体结构及原理图

系统大致分以下三个部分：

1. 系统知识部分：

由知识库和词典组成。词典是整个系统静态知识的来源，存有有关汉语词条和英语词条的各种知识，采用槽值多值化的框架结构表示；知识库存有系统各个部分

调用的规则，如消除切分歧义所用的词规则，分析、转换所用规则，译文形态生成表等，这个部分是整个系统的核心，它的能力强弱直接表现为系统的翻译处理能力。

2. 推理机：

这是一个规则解释器和控制器。它把知识库中各类规则有机地组织起来，确定在什么情况下调用哪一类规则，如在句法分析部分，它要确定高层消歧规则、精确和非精确规则何时被激活，确定调用的顺序，这是一套复杂的控制机制。

3. 设施管理程序：

由知识库管理程序、词典管理程序、语料库管理程序组成。词典采用准B+树的多级索引组织；知识库采用按中心词类型细化管理。

3·2 系统的设计思想

1. 多种知识并用

MMT是整个系统的处理平面，分析的过程就是构造一棵MMT，转换的操作也是基于MMT进行的。MMT的构造过程和结点信息的标注采用了多种知识并用原则。基于句法的系统会局限于繁琐的结构转换，基于语义或媒介语的系统既不可能，有时也不必要，况且也不易保证源文的结构特征。系统从语言的表义功能出发，对语言的语法和语义信息不加严格区别，对不同的语言现象采用不同的知识处理。纯语义描述主要体现在词汇的语义分类，对于名词分为语义大类、语义中类；对于动词体现为动词的支配者和被支配者以及动词本身的语义分类。语义、语法信息在词典中的表示方法是一样的，规则调用这些信息方法也是一样的，因此，系统体系结构的设计体现了多种知识并用的思想。

2. 面向多文种

句法分析时基本上不考虑译文的生成，二者是以MMT作为信息的载体，MMT既是分析的输出，又是生成的输入，这样充分做到独立分析、独立生成，有利于各自的模块化开发，同时，也便于开发其它汉外机译系统。

3. 基于完备和不完备知识相结合的推理机制

自然语言是一个无限集，也是一个模糊集，其信息量无限丰富，不可能用二值逻辑加以准确描述，因此，系统在推理策略和规则描述上采用了精确、非精确和确定、非确定推理相结合的策略，先精确，后非精确，先确定，后非确定。对于规律性很强，比较明确的语言现象，采用精确和确定推理，对于一些模糊现象采用非确定、非精确推理，对于个别现象采用词驱动规则推理。这样做使系统的鲁棒性更强，扩大了系统的覆盖范围，增强了系统的适应能力，避免了不可归约的恶性失败现象。

4. 采用知识工程的设计思想，做到规则与程序相独立

在系统的各个部分，规则和程序都完全独立。各部分规则都按照知识工程的方法加以组织和管理，这样有利于系统的不断完善。另外，词典采用开放式的槽值多值化的框架结构加以表示，这样，词典完全开放，对词典可任意增、删、改，而系统其它部分可完全不动。

5 模块化、开放式结构

系统采用软件工程的方法，运用面向对象、数据封装和继承等技术，使系统的各个模块对内封闭，对外采用参数传递，这样分工明确，有利于系统的发展。由于系统的模块化，使系统在加入一些新思想、新功能时，不会影响到其它部分，这样的开放式结构可以使系统不断吸收新的营养，做到系列化的发展。另外，机译系统的开发是一个浩大的工程，需各方面专家共同努力，要保证语言学家和计算机专家之间以及他们各自之间协调工作，就必须在系统设计上采用开放式结构，通过系统提供的一致化操作，实现整体的协调。

四、 系统的翻译处理机制

翻译处理机制是建立在语言模型基础之上的，相当于整个系统的控制流。它负责实现源文的分析，译文的转换生成，规则的收集，词典信息的表示和应用，知识库和翻译处理环境的维护等等。

从功能和实现角度看，翻译处理机制如下：

1. 语法、语义知识并用

系统始终坚持的原则就是多种知识并用和面向实用化。我们认为语法、语义界线不易确定，对二者也不必严格区别开来，只要从语言表义功能的角度出发，对不同的问题采用不同的手段加以解决，就会发现二者互相联系，互相补充，共同满足于表义功能。系统的文法规则描述了句法规律，但规则产生的结构不仅与句法相关，而且很多与语义关系相关。在本系统中，把语法、语义统称为复杂特征。

2. 复杂特征描述与“合一”运算

整个翻译的过程是以MMT为平台的。句法分析不仅要建立一棵句法树，而且要计算出每个结点的一系列的特征值，转换的过程不仅要进行结构的转换，还有特征的转换，“合一”运算是对复杂特征的操作，这些操作包括提取、生成、回传、删除等。复杂特征又进一步分为：静态特征和动态特征，全局特征和局部特征。静态特征是单词本身固有的特征，它在词典中静态表示出来。动态特征是分析过程中获得的，只有在词与词、短语结构和句子之间发生联系时，才表现出来。分析的过程是逐步检查静态特征或已生成的动态特征之间的相容性，进而生成新的动态特征，转换的过程是结构和特征的转换，这一过程体现了“合一”操作。全局特征主要用来描述谓语，因为谓语是句子的核心；同时，也反映了主、谓、宾三者之间的关系，这些特征都是动态的，主要包括描述形态和整个句法结构的特征；局部特征用来描述局部的词汇或短语结构内部的特征，如定语形容词语义大类。

3. 基于不完备知识的启发式推理机制

这一机制包括：精确与非精确推理；确定性与非确定性推理。自然语言的现象无限丰富，在基于规则的系统中，无论规则设计得多么细致合理，总有一些现象没有被考虑进去，规则和词典不可能完备且总会有错误存在，因此，为了增加系统的柔性，采用了这种基于不完备知识的启发式推理机制。其原理是，对于确定的可以很好形式化的语言现象，采用精确和确定性推理；对于模糊的现象，通过忽略严格的上下文条件，利用逐步减少约束条件的方法，尽可能利用相近的语法、语义约束，来得到译文；对于无法解决的歧义结构或多义现象，采用非确定性推理—启发式回溯机制求解。不完备知识的引入扩大了规则体系的覆盖范围，能够对未遇到的语言现象实施正确或相对正确的翻译，提高了系统的适应能力。

4. 规则的通用和个性组织机制

自然语言现象的种类很多，有些现象普遍存在，不是针对某个词，而是对一类结构都适用，而有些现象则是一种固定结构，只是针对某个词或某个搭配结构。因此，对于这两种情况，在规则设计时采用通用和个性两级组织原则。对于通用的语言现象采用通用规则加以描述，对于汉语的固定搭配结构采用词驱动的个性规则加以描述。这些个性规则的引入，由于其针对性强，很好地解决了固定结构无法准确地翻译的问题，

例：

汉语短语　　<NP1+对+NP2+的影响>　　对应的英语结构为

<Influence of NP1 ON NP2>

其中“对”和“ON”的使用很难从语法、语义上解释，这只是“对”和“影响”的一种约定俗成，直接在表层利用词驱动规则进行分析、转换即可，根本不必上升到语法、语义，这样不但效率高而且译文更地道。

系统消除切分歧义的规则均是个性的，因为在错误切分的局部其语法、语义信息是不能搭配的，呈随机状分布。

5. 上下文相关分析和转换

语义组合是汉语的实质，要很好地解决语义问题，就必须采用上下文相关分析和转换策略。在基于产生式规则的系统中，很难做到上下文相关处理。本系统引入一个通用操作原语（系统称之为SEEK()函数），它能根据需要准确地对MMT的每个角落进行信息搜索，从而建立起上下文联系，完成上下文相关分析。系统把SEEK()函数嵌入到规则中，SEEK()函数根据丰富的入、出口参数完成搜索工作。

6. 复杂问题的分级处理

对于系统中一类复杂问题，系统采用“化整为零、分级处理”的策略。这主要体现在兼类处理和译文转换上。系统对兼类处理有了突破性进展，采用多级渐进演绎推理策略，将确定性推理和非确定启发式推理相结合，首先是低层线性消兼，然后是高层树形结构消兼，最后是启发式回溯消兼，实现了汉语词兼类的全自动消除，且避免了组合爆炸。系统对转换的处理采用了第一层直接转换、第二层通用转换、第三层上下文

相关转换等三种转换方式。系统的分级处理策略换策略从不同的角度、不同的层次保证了复杂问题处理的正确性和准确性。

7. 规则冲突的消解

规则冲突在整个系统中普遍存在，特别是转换部分要进行不可预测的树重构，冲突表现得更为突出。系统采用了很多机制来处理这个问题，如：根据规则的功能，对规则进行细分类；通过调序确定规则激活的优先级；通过规则的动作部分破坏其它规则的生存环境；根据规则之间的相关性，做到规则间的制约等等。

8. 框架的引入和槽值多值化表示

系统的主数据结构、知识库、词典均采用框架结构表示，这样从本质上确保了系统的开放性和可扩展性，非常有利于系统的调试工作。另外，框架槽值的多值化表示，有效地解决了自然语言固有的不确定性和模糊性给计算机处理带来的困难，同时，也给词典的编写提供了极大的灵活性。

五、 结束语

机器翻译，特别是汉英翻译，是一个研制周期长、涉及学科领域较广的工程性难题，系统经过多年的努力，取得了一定的进展，这在１９９３·５的鉴定会上，国内专家也给予了一定的肯定。但我们认为工作还很粗糙，还有很多工作要做：

１、机译理论的提高和系统知识处理水平的加强；

２、大规模真实语料的调试和系统语义刻化和处理的加强；

３、开发智能接口与机译的连接，组成完整的机译系统；

参考文献

[１] 赵铁军，李生，周明：一种生成复杂特征集句法树的汉语句法分析方法与系统实现，中文信息学报，１９９２，Ｖｏｌ·６，Ｎｏ·４，Ｐ１１－２３

[２] 周明：汉－英机器翻译的研究与实践，哈尔滨工业大学计算机系工学博士论文，１９９１年６月

[３] 冯志伟：机器翻译中汉语分析和生成的四个原则，全国机器翻译学术会议，北京，１９９２年１０月

[４] 陈肇雄等：智能型机器翻译研究进展，全国机器翻译学术会议，北京，１９９２年１０月

[５] 哈工大，航空航天部一院１０３所：科技文章汉英机器翻译系统鉴定材料，北京，１９９３年５月

[６] 黄昌宁，苑春法，潘诗梅：语料库、知识获取和句法分析，中文信息学报，１９９２，Ｖｏｌ·６，Ｎｏ·３，Ｐ１－６

The Design and Implementation of CEMT-Ⅲ Machine Translation System from Chinese to English

Zhang Min, Li Sheng, Zhao Tiejun

(Computer Department, Harbin Institute of Technology, Harbin 150006)

Abstract: In this paper, the author discussed the theory and implementation of the CEMT-Ⅲ Machine Translation from Chinese to English, and explained the linguistic model, the overall architecture and the ideas of design. In addition, the practical problems of the liguistic project and the computer processing in the development of the system have been demonstrated. In the end, the author described the mechanism and the tactics of the system. The system has been appraised in Beijing in May in 1993.

Key Words: MT, Chinese, English, Natural language processing