一个从中间语言生成目标语言的原理和方法*

卞世力 姚天顺 金 鸿

(东北大学计算机系,沈阳 110006)

摘要 本文介绍了我们的汉英机器翻译系统(CETRAN)中一种从中间语言生成英语的生成系统,讨论了从中间语言图到目标语言转换的基本原理.目的在于通过解决汉英之间语法和语义方面的差异,得到高质量的机译结果.文中还介绍了基于语义驱动的由中间语生成英文目标语的计算机实现算法.为了说明清楚,整个叙述都注意列举了一些实例.

关键词 机器翻译,中间语言图,词汇语义驱动,语言生成.

传统的机器翻译系统大都采用直译的方式,即由源语言直接翻译成目标语.这对于多语种机译而言,在总体结构上是繁琐的.我们的CETRAN系统采用了以中间语言为枢轴结构的分析策略,力求这种枢轴结构能够成为独立于语种的通用性语言.尽管国际上对这种语言存在争论很大,但我们仍努力于这方面的工作,为以后的多语种翻译打下基础.

我们系统的中间语言生成是由分析系完成的.它利用词汇语义驱动原理,经过了汉语源语句的分词,词法,句法和语义多层次分析,最后从语义结构中提取了一个概念结构表达式(CSE:Conceptual Structure Expression),给出了中间语言的内部表示.

生成系统的工作主要有两方面内容:一方面是从中间语言结构的枢轴语义图出发,建立译语的表层结构;另一方面是把表层结构转换成链,并根据上下文要求,生成出每个词的词形.根据这两方面的不同要求和CETRAN系统的特点,我们把生成系统分成了两大阶段,即将中间语言图拉成线性结构(称之为线性化)和形态生成.下面就以系统组成、基本原理和具体算法等三个方面介绍CETRAN的生成系统.

1 生成系统的组成

生成系统的组成,本质上仍然是层次结构的,是分析处理的逆过程,即由深层的语义描述转换成同义的目标语言表层结构.这是一个非常复杂的转换过程.我们知道,源语言与目标语的词汇,不可能是完全一一对应的,不但词与词不归一,词性也常常不归一.

例如:汉语的"看"可以有see,watch,look等多个英语词与之对应,而英语的这几个词

* 本文1992-05-22收到,1992-07-22定稿

本项目得到机电部"七五"科技三项:电C2的资助.作者卞世力,31岁,研究生,主要研究领域为机器翻译.姚天顺,58岁,教授,博士导师,主要研究领域为计算机语言学,人机系统,机器翻译.金鸿,28岁,硕士,主要研究领域为人机系统.

本文通讯联系人:卞世力,沈阳 110006,东北大学计算机系

的用法都是截然不同的.又如:汉语的"今天"的词性是时间名词,而与之对应的英语词"today"又是副词.

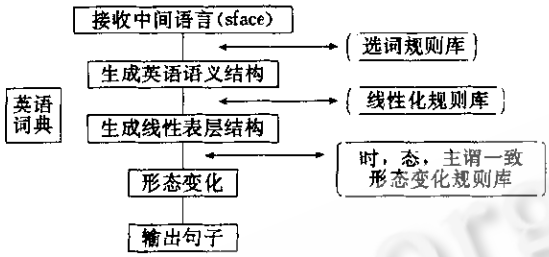


图1 CETRAN生成系统构造图

因此,我们往往不可能找到一个目标语的句子与原来源语句子完成全同义的句子,为此我们采取慎重的方法,分层次逐步地转化.主要的由3个处理模块,词典和规则库组成.其系统组成如图1所示.其中的3个处理模块分别是:

1.生成英语语义结构

通过选词规则,查词典,将中间语言的结点概念标记符转换成英语表示,寻找对应概念的英文词或词组,从而形成了英语的语义结构.

2.生成线性表层结构

由线性化规则库将中间语言图拉成线型链,同时根据各结点的语义关系添加一些虚词结点和省略词的结点,并考虑各结点线性化之后的顺序,从而形成了具有一定顺序的线形表层结构.这个过程是整个生成系统的关键所在.

3.形态变化

我们知道,生成同意义的目标句子是通过各词间的词序和形态表现出来的.只有当两个因素都正确时,方可把句子的意义表达出来.因此在系统的总体结构上还应该有形态变化模块.这一过程是在拉成链的基础上进行的.根据时、态、主谓一致等信息,首先确定各词的形态特征信息,然后进行形态变化,从而生成一个真正的表达句子意义的形式.

2 基本原理

我们系统的中间语言实质上是一个由语义关系图 $M=\langle V(M),R(M),\psi m\rangle$ 表示的,其中:$V(M)$是一个非空结点集合,每个结点是一个带有复杂特征集的概念;$R(M)$是边集合,每个边表示一个语义关系(共有50种关系);$\psi m(R)$是从边集合R到结点的有序偶集合上的函数.

例如:汉语句子"院里有一棵大树"经过分析系统后,生成的中间语言图,如图2.

其中:$V(M)=\{have,yard,tree,large,one,Empty,SENTENCE\}$,

$R(M)=\{SEN,EXP,LOC,QNT,NUM\}$,

$\psi m(SEN)=(SENTENCE,have)$, $\psi m(LOC)=(have,yard)$,

$\psi m(EXP)=(have,tree)$, $\psi m(EXP)=(large,tree)$,

$\psi m(QNT)=(tree,Empty)$, $\psi m(NUM)=(Empty,one)$.

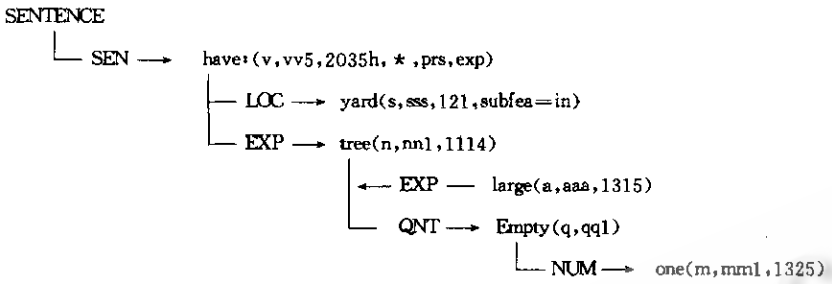


图2 "院里有一棵大树"的中间语言图(M)

说明:SENTECE 是句类标识结点.

SEN 是句类标识关系,不属于 50 种语义关系.

EMPTY 是指英语中没有对应的概念,如在英语中"棵"的形态没表现出来,即"一棵树"可转换成 a tree.

2.1 $M \overset{S-Rule}{\Rightarrow} S$ 转换

通过选词规则 S—Rule 将中间语言图(M),转换成英语语义结构图(S). 即根据每个概念结点与其关联结点的语义关系,对具有相同概念的不同词或词组的静态属性EMATCH信息进行分析匹配,选择一个恰当的词或词组,并把选中的词或词组的静态属性值从英语词中全部取出来. 图 3 是由图 2 转换过来的英语语义图. 其中概念 have 变成 there be,概念 one 变成 a.

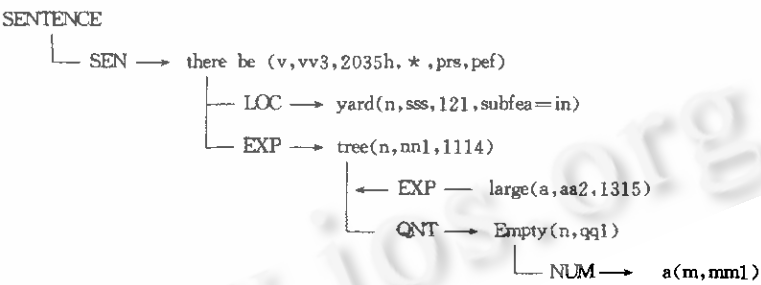


图3 英语语义结构图

从图 2 和图 3 中可以看出一些静态属性. 如 cat,subcat 已经不同了,如 yard 在英语中是个名词,而汉语的"院"又是处所词. 类似的,汉语中的时间词,方位词等,在英语中都划到了名词类中了. 所以在 M⇒S 转换中也包括了这些静态属性的转换,这样就形成了真正的英语语义结构图.

2.2 $S \overset{L-Rule}{\Rightarrow} L$ 转换

通过线性化规则(L—Rule)将英语语义结构图转换成有序线型表层结构图:

$$L=\langle V(L),R(Ord),\psi m\rangle$$

在 L 图中,50 种语义关系已经消失,只有前后的 Ord 序关系,同时也增加了一些 S 图中所没有的虚词结点和省略结点. 因此,在将树形结构拉成链同时,还要考虑两方面问题:一方面是增加结点,另一方面是拉链后的结点顺序.

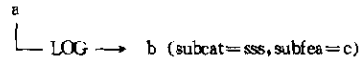
2.3 下面讨论这两方面的几种主要操作

2.3.1 增加结点操作

在增加结点操作中主要有两种情况：

a. 在关联结点的 dmch，subfea 等动态属性中提供了要增加的结点信息.

如：设有 ψs(LOC)=(a,b)

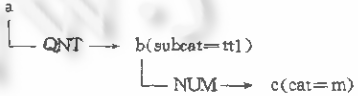


其中，subfea 为中间语言提供 b 结点的下位特征. 则转换成如图 $a \xrightarrow{Ord}\ c \xrightarrow{Ord}\ b$.

即：$\psi s(LOC)=(a,b) \overset{L-Rule}{\Rightarrow} \psi l(Ord)=(a,c), \psi l(Ord)=(c,b)$.

b. 没有提供要增加的结点的信息，只有根据相关联结点的语义关系，确定要增加的结点.

如：设有



则转换成 a→for→c→b(W_FORM=add_s).

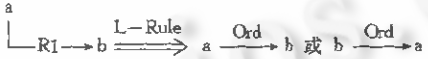
即：$\psi s(QNT)=(a,b), \psi s(NUM)=(b,c) \overset{L-Rule}{\Rightarrow} \psi l(Ord)=(a,for), \psi l(Ord)=(for,c), \psi l(Ord)=(c,b)$

其中 W_FORM=add_s 表示结点要变为复数形式.

2.3.2 建立全序关系操作

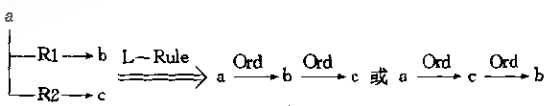
所谓建立全序关系操作，就是将语义图拉成链，使各结点顺序符合与源语句同义的目标语句子有正确的顺序. 这一操作同样有两种情况：

a. 两个结点由上下(父子)关系，变成左右(兄弟)关系，如下图.



即 $\psi s(R1)=(a,b) \overset{L-Rule}{\Rightarrow} \psi l(Ord)=(a,b)$ 或 $\psi l(Ord)=(b,a)$.

b. 在同一结点的不同子结点线性化后，在同一侧时的操作顺序. 假设在右侧，就有下图的转换.

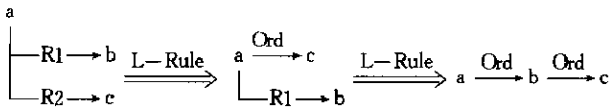


实际上这种转换是 a 种情况的复合操作，对两种不同结果的结点操作次序分别定义如下：

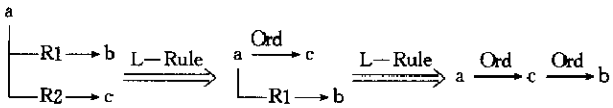
·若要转换成 $a \xrightarrow{Ord}\ b \xrightarrow{Ord}\ c$

则操作次序为：先 c 后 b(先远后近)

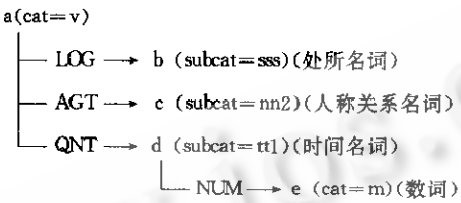
·若要转换成 $a \xrightarrow{Ord}\ c \xrightarrow{Ord}\ b$



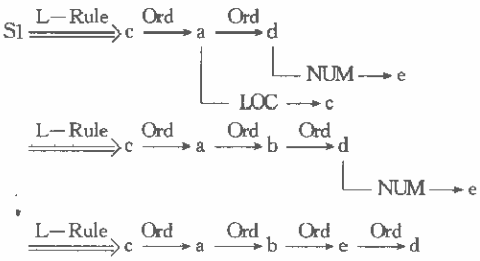
则操作次序为:先 b 后 c(先远后近)



例:设有下列中间语言图(实际上是英语语义结构图 S1).



从语义关系来看,c 结点要拉到 a 结点的左侧充当句子的 subj. b,d 结点要拉到右侧,根据英语文法要求,时间状语通常放在地点状语的右侧,所以,对 b,d 结点的处理次序为先 d 结点后 b 结点(先远后近). 即



说明:c,e 结点前可能应该增加介词结点,此处略.

从上述可知,在实现过程中,我们是以通过词汇语义驱动,中心动词为轴心,自顶向下,由远到近的原则,完成线性化操作的.

3　计算机实现算法

通过上面的阐述,我们知道,尽管生成系统主要有两个任务,即线性化和形态变化,但由于两种语言的转换是极其复杂的,完成这两个任务需要设计一套计算机可以实现的控制算法. 因此我采取了在知识规则的组织上使之分类化,和生成过程上采用层次化的方案设计计算机实现算法. 下面仍以图2为例,详细描述生成英语句子全过程.

生成算法

第1步:接收由分析系统提供的中间语言文本(sface).

这一步是从文本名为 sface 中取出,由分析系统提供的中间语言文本,其数据结构如图2所示.

第2步:将每个结点概念标记符转换成对应的词或词组.

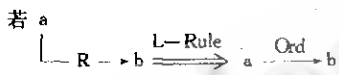
这一步的主要任务是在具有相同概念的词中,选择一个合适的词或词组. 为了完成这个

任务，在词典设计中，将一些具有相同概念的词用 EINDEX 域链接成一个环形链表，对每个词的 EMATCH 域进行匹配. 匹配成功则该词被选中，读出其全部词典的静态信息，如图3是执行了这一过程后的结果.

第3步：将树形结构图拉成线性链表结构

这一步是按着以中心动词为轴心，自项向下，从远到近的原则，逐层次地将树形结构图拉成链表，同时增加相应的虚词结点. 线性化操作，实际上是通过我们系统中的规则描述语言中的“增”，“改”操作来完成的. 在线性化的同时，确定每个结点，并由 ADD_CONST(x)函数完成句法成份域 const 的赋值.

例如：



则我们的规则是：

$$\hat{}\_\#\text{SENTENCE}.\text{R}\Rightarrow\hat{}\_\#,1$$

意思是，若当前结点 a 与儿子 b 的语义关系是 R，就将儿子结点改成结点 a 的右侧结点. 其中“\”字符为“改”操作.

因此，在完成了这一步以后，从图3转换成下列有序结点集：

SENTENCE	(,,SENTENCE)
there be	(v,vv3,2035h,*,prs,pef)
	QUANT=single,CONST=sen/vp
a	(z,zzz)
	QUANT=single,CONST=sen/subj/det
large	(a,aa2,1315)
	CONST=sen/subj/mod
tree	(n,nn1,1114)
	QUANT=single,CONST=sen/subj
in	(p.pp1)
	quant=single,CONST=sen/pp/prep
the	(z,zzz)
	QUANT=single,CONST=sen/pp/det
yard	(s,sss,121)
	QUANT=single,CONST=sen/pp

第4步：时态处理

对中心动词的 TIME 和 TENCE 测定其时态，如“there be”的 TIME =prs，TENCF =pef，ELOC=2(是指“there be”的词形变化位置是在第2个词)操作过程. 即首先以第2个词为分离点，将“there be”变为两个结点：“there”和“be”，然后，在结点“be”前加助词结点“have”，并把中心动词的 quant 复制过来，将中心动词变成过去分词信息，加到 W_FORM 上，这就得到如下动词和助动词结点.

SENTENCE　　(,,SENTENCE)
there　　(c,cc4)
have　　(x,xx1,TIME=prs,QUANT=single)
be　　(v,vy3,W—FORM=add—ed2)

其它结点由于没有变化则省略没写.

第5步:一致性处理

一致性处理主要任务是指动词的主谓一致,和名词的数的一致性处理.如当前动词(助动词)数为QUANT=single,现在时,则该词的词尾为S操作,匹配规则为:

^(QUANT.single,TIME.prs),^(ECAT.x;ECAT.v)⇒:=(^W_FORM,'add_s')

这就确定了动词(助动词)的形态变化信息.

名词复数处理是判定其词前是否有大于1的数量词修饰,确定是否为复数变化.这个过程结束后,对每个词提供形态变化信息如下.

SENTENCE　　(,,SENTENCE)
there　　(W—FORM=(NULL))
have　　(W—FORM=add—s)
be　　(W—FORM=add—ed2)
a　　(W—FORM=(NULL))
large　　(W—FORM=(NULL))
tree　　(W—FORM=(NULL))
in　　(W—FORM=(NULL))
the　　(W—FORM=(NULL))
yard　　(W—FORM=(NULL))

第6步:对每个词进行形态变化

对每个词的W_FORM的值,先查不规则词典,再根据规则变化原则,对每个词进行词形变化,得到如下词(词组)序列:

SENTENCE there bas been a large tree in the yard

第7步:按英语文法书写规则,输出句子:

There has been a large tree in the yard.

4　结束语

由于不同的语种,其语法的风格与习俗都很不相同.生成的主要工作在于保持源语句原意的条件下,产生同义的目标语句.这是一件非常困难的任务.事实上完全同义常常是不可能的.特别是我们的系统还通过利用语义网络描述的中间语言生成目标英文句子,更是一个初步的尝试.困难很多.本文所提的思想和方法,已在几万词的词典环境下做过几百句的试验,证明是可行的.现在正在不断地扩大规则的规模,完善我们的系统.但是随着问题的不断深入,肯定还会有不少问题,请批评指正.

参考文献

1 James Allen. Natural language understand. The Benjamin/Cummings Publishing Company, Inc., 1987.

2 Simmons R, Slocum J. Generating English discourse from semantic networks. Communications of the ACM, 1972, **15(10)**.

3 姚天顺,马黎环. Generating English text from Chinese semantic representation. Proceedings of 1988 International Conference on Computer Processing of Chinese and Oriental Languages. Montreal, Canada, 1988.

4 左孝凌,刘永才等. 离散数学. 上海:上海科学技术文献出版社.

5 王宝库,张中义,姚天顺. Rule description language CTRDL in machine Translation system. Proceedings of 1991 International Conference on Computer Processing of Chinese and Oriental Languages, Taiwan, Aug. 1991.

6 殷钟崃等编. 英语语法理论及流派. 成都:四川大学出版社.

PRINCIPLE AND METHOD OF GENERATING TARGET LANGUAGE FROM INTERLINGUA

Bian Shili, Yao Tianshun and Jin Hong

(*Department of Computer Science, Northeastern University, Shenyang* 110006)

Abstract This paper describes the outline of English generation system and the basic principle of transfer processing from interlingua to target language in Chinese English Machine Translation System (CETRAN). The objective of this research is to get the high quality of translation by solving the problems of language ambiguities and difference between English and Chinese. An algorithm for generating the English target language from interlingua has been proposed and some of examples will also be reported to show the feasibility of our research.

Key words Machine translation, interlingua, lexical semantic driven, language generating.