机器翻译中异化现象的分析及映射运算

郭宏蕾[1,2] 姚天顺[1]

[1](东北大学计算机系 沈阳 110006)

[2](北京航空航天大学计算机系 北京 100083)

摘要 本文从词汇-语义角度分析了机器翻译中与汉英语言相关的一些异化现象,探讨了语义-句法映射运算集的有效性和合理性.同时,本文讨论了通过变异映射集、变异类型指示器和参数合一机制解决机器翻译中异化现象的方法.这些方法提高了译准率,使生成的句子既能正确表达原中间语言的语义,又符合目标语言的表述习惯.

关键词 机器翻译,语言异化,映射运算,自然语言处理.

中图法分类号 TP391.2

在机器翻译中,有时出现译文生硬、呆板的现象,其主要原因是译文表达形式与目标语言的惯用表述风格不一致.为了使句子自然、贴切,我们有必要对具体语言的特殊表示模式进行分类研究,抽其共性,建立特殊的映照机制.

在深层概念语义到表层语言表达式的转化中,常遵循一些常规映射法则,但由于词汇、语义搭配的影响,有时需要用一些特殊规则来修正常规映射生成的表达,以符合语言表述习惯.我们称这些遵循特定映射的语言现象为翻译的异化现象.[1]异化现象是语言固有的,与采用的翻译方法(如转换法、基于中间语言法等)无关.本文以863高科技课题"汉英双向翻译系统CETRAN"为背景,从词汇-语义角度分析了语义-句法映射中存在的一些异化现象,对语义-句法映射运算集的有效性、合理性进行了分析,并通过变异映射集、变异类型指示器和参数合一机制解决机器翻译中的异化现象,以提高系统的译准率,增强语义-句法映射处理的开放性、通用性.

1 汉英语言中语义-句法异化映射

CETRAN是基于词汇语义驱动原理[2]的汉英双向翻译系统.本节我们先给出几个基本结构定义,然后讨论一些语义-句法异化现象.

1.1 几个基本定义

定义1. 语义结构(SemS).设 t 为变元,$t\in\{P,X,Y_i,Z_j\}$,$\varphi(t)$是 t 的语义类型,且 $\varphi(t)\in${事件(Event),状态(State),属性(Property),实体(Object),...};$r(t)$是 t 的语义角色.则

$$SemS::=\langle[Pred:(P,\varphi(P))],[Major:(X,\varphi(X),r(X))],[Minor:(Y_1,\varphi(Y_1),r(Y_1)),\ldots,(Y_n,\varphi(Y_n),r(Y_n))],$$
$$[Lmod:(Z_1,\varphi(Z_1),r(Z_1)),\ldots,(Z_m,\varphi(Z_m),r(Z_m))]\rangle$$

其中,(1) Pred 标识 P 为逻辑谓语.(2) Major 标识 X 为逻辑主论元,即行为、动作、状态、性质的主体;$r(X)\in${施事(AGT),当事(EXP),领事(POS),...}.(3) Minor 标识 $Y_1,\ldots,Y_n$ 为逻辑次论元序列,即行为、动作的客体——必选参与者.$r(Y_i)\in${受事(OBJ),客事(RLT),...},$i=1,\ldots,n$.(4) Lmod 标识 $Z_1,\ldots,Z_m$ 为逻辑修饰元序列,即行为、动作的情境约束——可选参与者.$r(Z_j)\in${时间(TIM),地点(LOC),工具(INS),...},$j=1,\ldots,m$.

语义结构既用于描述字典中词汇的语义配价结构,又用于表示句子的中间语义结构.前者称为词汇的语义结构(记为LSemS);后者称为合成语义结构(记为CSemS).LSemS为一个非示例SemS(即含有变元的SemS).

定义2. 深层句法结构(DSyntS).设 t 为变元,$t\in\{P',X',Y'_i,Z'_j\}$,$s(t)$是 t 的句法成分.则

$$DSyntS=\langle[Core:P'],[Strong:(X',s(X'))],[Weak:(Y'_1,s(Y'_1)),\ldots,(Y'_n,s(Y'_n))],$$
$$[Smod:(Z'_1,s(Z'_1)),\ldots,(Z'_m,s(Z'_m))]\rangle$$

---

* 本文研究得到国家自然科学基金、国家863高科技项目基金和国家教委博士点基金资助.作者**郭宏蕾**,女,1970年生,博士,主要研究领域为计算语言学,机器翻译.**姚天顺**,1934年生,教授,博士导师,主要研究领域为计算语言学,机器翻译.

本文通讯联系人:郭宏蕾,北京 100083,北京航空航天大学计算机系

本文1996-09-17收到原稿,1997-02-03收到修改稿

其中,(1) Core 标识 P' 为句法核心,如句子的谓语;(2) Strong 标识 X' 为强句法参数,如主语;(3) Weak 标识 $Y'_1,\ldots,Y'_n$ 为弱句法参数序列,即句法核心的必选参数,如宾语、表语、补语等;(4) Smod 标识 $Z'_1,\ldots,Z'_m$ 为句法修饰序列,即句法核心的可选参数,如地点、时间、方式等修饰.

例如,在"我买了一本书"中,设 U,W 为变元,大写字母串带"*"表示概念.于是,

$LSemS$[买]=⟨[$Pred$:(BUY^*,$event$)],[$Major$:(U,$object$,AGT)],[$Minor$:(W,$object$,OBJ)]⟩

$CSemS$=⟨[$Pred$:(BUY^*,$event$)],[$Major$:(I^*,$object$,AGT)],[$Minor$:($BOOK^*$,$object$,OBJ)]⟩

$DSyntS$=⟨[$Core$:买],[$Strong$:(我,主语)],[$Weak$:(书,宾语)]⟩

若令 $Y\in\{Y_1,\ldots,Y_n\},Z\in\{Z_1,\ldots,Z_m\},Y'\in\{Y'_1,\ldots,Y'_n\},Z'\in\{Z'_1,\ldots,Z'_m\}$,则 SemS 和 DSyntS 可简化为

$SemS::=$⟨[$Pred$:(P,$\varphi(P)$)],[$Major$:(X,$\varphi(X)$,$r(X)$)],[$Minor$:(Y,$\varphi(Y)$,$r(Y)$)],[$Lmod$:(Z,$\varphi(Z)$,$r(Z)$)]⟩

$DSyntS$=⟨[$Core$:P'],[$Strong$:(X',$s(X')$)],[$Weak$:(Y',$s(Y')$)],[$Smod$:(Z',$s(Z')$)]⟩

于是,在语义-句法映射中,SemS 和 DSyntS 之间的常规位置映射定义如下:

定义 3. 常规位置映射(记为 ζ)

$\zeta=\{$(1) $Pred:P\Rightarrow Core:P'$; (2) $Major:X\Rightarrow Strong:X'$; (3) $Minor:Y\Rightarrow Weak:Y'$; (4) $Lmod:Z\Rightarrow Smod:Z'$;$\}$

即(1) 逻辑谓语 P 映射为句法核心 P';(2) 逻辑主论元 X 映射为强句法参数 X';(3) 逻辑次论元 Y 映射为弱句法参数 Y';(4) 逻辑修饰元 Z 映射为句法修饰 Z'.

1.2 题元变异映射

由于词汇、语义个性约束的影响,在许多语义-句法结构转换中,并不完全遵循语义-句法常规映射,而是通过特殊映射,生成符合具体语言表述习惯的句法结构,如在下面两个例句中,"晒"具有同样的语义.

(1)我晒鱼.　　(2)老头晒太阳.

在语义-句法常规位置映射 ζ 作用下,"晒"的逻辑主论元映射到强句法参数位置,逻辑次论元映射到弱句法参数位置.显然,例(1)遵循映射"$CSemS\xrightarrow{\zeta}DSyntS$".但在例(2)中,"晒"的逻辑次论元"老头"却映射到强句法参数位置(即作主语),逻辑主论元"太阳"则映射到弱句法参数位置(即作宾语).可见,"老头晒太阳"是一种异化作用的结果.这种逻辑主论元、逻辑次论元的句法位置互易的现象,称为题元变异映射.

定义 4. 题元变异映射(记为 α).令常规位置映射

$$\zeta=\{Pred:P\Rightarrow Core:P';\ Major:X\Rightarrow Strong:X';\ Minor:Y\Rightarrow Weak:Y'\}.$$

则

$$\alpha::=\{Pred:P\Rightarrow Core:P';\ Major:X\Rightarrow Weak:X';\ Minor:Y\Rightarrow Strong:Y'\}$$

即在 α 映射中,逻辑主论元 X 映射到弱句法参数位置,逻辑次论元 Y 映射到强句法参数位置.

为了描述变异映射信息,我们在字典中引入变异类型指示器来标注 LSemS.若 W 为 LSemS 中的论元,δ 为 LSemS 所包含的变异映射(即非常规映射),则 $W@\delta$ 标识论元 W 涉及 δ 变异映射,并用 $scope(\delta)$ 注明该变异类型指示器的有效作用区.例如,若 LSemS 包含题元变异映射 α,则带变异类型指示器的 LSemS 为:

$$LSemS(\alpha)::=\langle[Pred:(P,\varphi(P))],[Major:(X@\alpha,\varphi(X),r(X))],[Minor:(Y@\alpha,\varphi(Y),r(Y))],[Lmod:(Z,\varphi(Z),r(Z))]\rangle$$

$scope(\alpha)=\{C_1,C_2,\ldots,C_n\}$,$C_i$ 为诱发 $LSemS(\alpha)$ 中题元映射的词汇、语义相关条件(如逻辑主论元、逻辑次论元的语义属性等).

在语义-句法映射中,如果 LSemS 含有变异类型指示器 δ,且在有效作用区 $scope(\delta)$ 内,则激活 δ 变异映射;否则,遵循语义-句法常规映射.

1.3 升位变异映射

在语义-句法映射中,有时出现逻辑语义成分定位变异现象.例如,汉语允许形容词直接充当谓语[3],如句子(1)"他高"所示.其 CSemS 为:

$$CSemS=\langle[Pred:(BE^*,state)],[Major:(HE^*,object,EXP)],[Minor:(TALL^*,property,RLT)]\rangle$$

依据语义-句法常规位置映射 ζ,该 CSemS 应遵循映射 $\zeta=\{Minor:TALL^*\Rightarrow Strong:$高$;Pred:BE^*\Rightarrow Core:$是$\}$,即由映射"$CSemS\xrightarrow{\zeta}DSyntS$"生成句子"他是高".但汉语并不这样说,而是习惯表述为例(1)那种形式,即执行特殊映射"$CSemS\xrightarrow{\xi}DSyntS$",$\xi'=\{Minor:TALL^*\Rightarrow Core:$高$;Pred:BE^*\Rightarrow$NULL$\}$,NULL 表示为空.同样地,英语中也存在逻辑语义成分定位变异现象[4],例如,(2) *He happened to meet John.* 其 $CSemS=\langle[Pred:(MEET^*,event)],[Major:(HE^*,object,AGT)],[Minor:(JOHN^*,object,OBJ)],[Lmod:(HAPPEN1^*,manner,MANNER)]\rangle$.依据常规位置映射 ζ,逻辑修饰元 $HAPPEN1^*$(语义为"恰巧"应映射到句法修饰位置,逻辑谓语 $MEET^*$ 应映射到句法核

心位置.但在英语中,该CSemS的 $HAPPEN1^*$ 映射到句法核心位置,$MEET^*$ 映射到弱句法参数位置,即形成例句(2).通常认为,在语义结构SemS中,逻辑谓语的地位高于其它论元,在句法结构中,句法核心的地位高于其它参数[5],这类逻辑论元、逻辑谓语的句法易位现象称为升位变异映射.升位变异映射又分为0型和1型升位变异映射.

定义 5. 0型升位变异映射(记为 λ_0).令常规位置映射

$$\zeta=\{Pred:P\Rightarrow Core:P';\ Major:X\Rightarrow Strong:X';\ Minor:Y\Rightarrow Weak:Y'\},\text{则}$$

$$\lambda_0::=\{Pred:P\Rightarrow \mathrm{NULL};\ Major:X\Rightarrow Strong:X';\ Minor:Y\Rightarrow Core:Y'\}$$

即在 λ_0 映射中,逻辑次论元 Y 升级映射到句法核心位置,逻辑谓语 P 被省略.

定义 6. 1型升位变异映射(记为 λ_1).令常规位置映射

$$\zeta=\{Pred:P\Rightarrow Core:P';\ Major:X\Rightarrow Strong:X';\ Minor:Y\Rightarrow Weak:Y';\ Lmod:Z\Rightarrow Smod:Z'\},$$

则

$$\lambda_1::=\{Pred:P\Rightarrow Weak:P';\ Major:X\Rightarrow Strong:X';\ Minor:Y\Rightarrow Weak:Y',\ Lmod:Z\Rightarrow Core:Z'\},$$

即在 λ_1 映射中,逻辑修饰元 Z 升级映射到句法核心位置,逻辑谓语 P 映射到弱句法参数位置.

于是,(1)若 $LSemS$ 包含0型升位变异映射 λ_0,则带变异类型指示器的LSemS为

$$LSemS(\lambda_0)::=\langle[Pred:(P,\varphi(P))],[Major:(X,\varphi(X),r(X))],$$
$$[Minor:(Y@\lambda_0,\varphi(Y),r(Y))],[Lmod:(Z,\varphi(Z),r(Z))]\rangle$$

$scope(\lambda_0)=\{C_1,\ldots,C_n\}$. C_i 为诱发 $LSemS(\lambda_0)$ 中0型升位变异映射的词汇、语义相关条件.

(2)若LSemS包含1型升位变异映射 λ_1,则带变异类型指示器的LSemS为

$$LSemS(\lambda_1)::=[Lmod:(Z@\lambda_1,\varphi(Z),r(Z))]$$

$scope(\lambda_1)=\{C_1,\ldots,C_n\}$. C_i 为诱发 $LSemS(\lambda_1)$ 中1型升位变异映射的词汇、语义相关条件.

显然,例(1)遵循映射"$CSemS\xrightarrow{\lambda_0}$DSyntS",例(2)遵循映射"CSemS $\xrightarrow{\lambda_1}$DSyntS".

2　语义-句法映射运算集的进一步分析

在第1节中,我们从汉英语言现象中提取出与语义-句法结构定位相关的常规位置映射 ζ、题元变异映射 α、升级变异映射 λ_0 和 λ_1,它们组成语义-句法映射运算集.本节将进一步分析这些映射运算的有效性和合理性.

从语义结构和句法结构的定义出发,我们有如下的性质:

性质 1. 句法核心的唯一性:一个 $DSyntS$ 有且仅有一个句法核心.

性质 2. 管辖状态的不变性:指定行为的必选参与者(即 $SemS$ 的逻辑次论元或 $DSyntS$ 的弱句法参数)始终受选择它的对象所管辖;指定行为的可选参数(即 $SemS$ 的逻辑修饰或 $DSyntS$ 的句法修饰)不受其所修饰的词汇管辖,始终处于自由状态.

为了便于说明,我们将逻辑关系"s管辖t"记为"s $\xrightarrow{gov}t$";"s被t修饰"记为"s $\xrightarrow{mod}t$".于是,在SemS中,有两类逻辑关系:(1) $Pred\xrightarrow{gov}Minor$;(2) $Pred\xrightarrow{mod}Lmod$.在DSyntS中,有两类逻辑关系:(1) $Core\xrightarrow{gov}Weak$;(2) $Core\xrightarrow{mod}Smod$.它们具有性质3.

性质 3. SemS和DSyntS所含的逻辑关系具有反对称性,即

$$\text{if } x\xrightarrow{r}y,\ \text{then } \neg(y\xrightarrow{r}x).$$

此处,r 表示 x 与 y 之间存在的逻辑关系.

在语义-句法映射运算中,逻辑主论元及强句法参数因其强耦合性而只参与题元变异映射;逻辑次论元、逻辑修饰元、弱句法参数和句法修饰参数因其弱约束性呈现出活跃的重定位映射能力.若逐一将SemS中的逻辑谓语、逻辑次论元、逻辑修饰元随意映射到句法核心、弱句法参数和句法修饰参数3个深层句法位置之一,可形成27(即 3^3)种映射组合.由于语义-句法映射是一种受限运算,在这27种组合中只有为数不多的映射组合是有效的,即真正生成符合语言学规范的深层句法结构.因此,我们有必要讨论语义-句法映射约束,提取与句法核心、弱句法参数、句法修饰参数相关的定位映射的有效组合.若用"$Map(s,t)=1$"表示"SemS中成分 s 映射到DSyntS中的 t 位置","$Map(s,t)=0$"表示"s 未映射到 t 位置",则语义-句法映射受到如下的制约.

(1) 句法核心唯一性(即性质1)对语义-句法映射运算的约束如下:

令 $T=\{Pred,\ Minor,\ Lmod\}$,$\exists\ s\in T$,有 $Map(s,Core)=1\Leftrightarrow\{\bigvee_{t\in T-\{s\}}Map(t,Core)\}=0$　　(约束1)

即在语义-句法映射中,SemS中有且仅有一个成分可映射于DSyntS的句法核心位置.

(2) 根据管辖状态不变性(即性质2),始终受管辖的逻辑次论元不能映射到处于自由状态的句法修饰参数位置.

于是,与逻辑次论元相关的映射约束为:

$$Map(Minor,Smod)=0 \qquad \text{(约束 2)}$$

同理,与逻辑修饰元相关的映射约束为:

$$Map(Lmod,\ Weak)=0 \qquad \text{(约束 3)}$$

即逻辑修饰元不能映射到弱句法参数位置.

下面,我们集中讨论与句法核心、弱句法参数、句法修饰参数相关的定位映射的有效组合.若

$SemS=([Pred:(P,\varphi(P))],[Major:(X,\varphi(X),r(X))],[Minor:(Y,\varphi(Y),r(Y))],[Lmod:(Z,\varphi(Z),r(Z))])$,

则:(1) 当 $Map(P,Core)=1$ 时,即 P 映射于句法核心位置.

由约束 1 知,　　$Map(P,Core)=1\Leftrightarrow[Map(Y,Core)\vee Map(Z,Core)]=0$

即　　$Map(Y,Core)=0,Map(Z,Core)=0.$

由约束 2 知,　　$Map(Y,Smod)=0$

由约束 3 知,　　$Map(Z,Weak)=0$

因语义-句法映射为单值映射,故

$$Map(P,Core)=1\Rightarrow(Map(P,Weak)=0)\wedge(Map(P,Smod)=0)$$

$$(Map(Y,Core)=0)\wedge(Map(Y,Smod)=0)\Rightarrow Map(Y,Weak)=1$$

$$(Map(Z,Core)=0)\wedge(Map(Z,Weak)=0)\Rightarrow Map(Z,Smod)=1$$

从而,可推出如下有效映射组合:

$$Map(P,Core)=1;Map(Y,Weak)=1;Map(Z,Smod)=1$$

即形成基于主对角线的常规映射矩阵(如图 1(a)).

	Core	Weak	Smod
P	1	0	0
Y	0	1	0
Z	0	0	1

(a) 常规型

	Core	Weak	Smod
P	0	0	1
Y	1	0	0
Z	0	0	1

(b) Weak无关型

	Core	Weak	Smod
P	0	1	0
Y	0	1	0
Z	1	0	0

(c) Smod无关型

图1　映射矩阵

(2) 当 $Map(Y,Core)=1$ 时,即 Y 映射于句法核心位置.

由约束 1 知,　　$Map(Y,Core)=1\Leftrightarrow[Map(P,Core)\vee Map(Z,Core)]=0$

即　　$Map(P,Core)=0,Map(Z,Core)=0.$

当 $Map(Y,Core)=1$ 时,DSyntS 中有"$Y\xrightarrow{gov}Weak$"成立,SemS 中存在"$P\xrightarrow{gov}Y$"且具有反对称性(即$\neg(Y\xrightarrow{gov}P)$)(即性质 3).故在 DSyntS 中,$P$ 与 Y 之间不能存在逆向管辖关系(即 $Y\xrightarrow{gov}P$).从而,$Map(P,Weak)=0$.

由约束 3 知,$Map(Z,Weak)=0$.因语义-句法映射为单值映射,故

$$(Map(P,Core)=0)\wedge(Map(P,Weak)=0)\Rightarrow Map(P,Smod)=1$$

$$Map(Y,Core)=1\Rightarrow(Map(Y,Smod)=0)\wedge(Map(Y,Weak)=0)$$

$$(Map(Z,Core)=0)\wedge(Map(Z,Weak)=0)\Rightarrow Map(Z,Smod)=1$$

从而,可推出如下有效映射组合:

$$Map(P,Smod)=1;\ Map(Y,Core)=1;\ Map(Z,Smod)=1$$

此时,映射矩阵中 Weak 列的元素均为 0,如图 1(b)所示,简称 Weak 无关型映射矩阵.

(3) 当 $Map(Z,Core)=1$ 时,即 Z 映射于句法核心位置

由约束 1 知,　　$Map(Z,Core)=1\Leftrightarrow[Map(P,Core)\vee Map(Y,Core)]=0$

即　　$Map(P,Core)=0,Map(Y,Core)=0$

当 $Map(Z,Core)=1$ 时,DSyntS 中有"$Z\xrightarrow{mod}$Smod"成立;SemS 中存在"$P\xrightarrow{mod}Z$"且具有反对称性(即$\neg(Z\xrightarrow{mod}P)$),即性质 3).故在 DSyntS 中,$P$ 与 Z 之间不能存在逆向修饰关系(即 $Z\xrightarrow{mod}P$).从而,$Map(P,Smod)=0$.

由约束 2 知,$Map(Y,Smod)=0$.因语义-句法映射为单值映射,故

$$(Map(P,Core)=0)\wedge(Map(P,Smod)=0)\Rightarrow Map(P,Weak)=1$$

$$(Map(Y,Core)=0)\wedge(Map(Y,Smod)=0)\Rightarrow Map(Y,Weak)=1$$

$$Map(Z,Core)=1\Rightarrow(Map(Z,Smod)=0)\wedge(Map(Z,Weak)=0)$$

从而，可推出如下有效映射组合：

$$Map(P,Weak)=1;Map(Y,Weak)=1;Map(Z,Core)=1$$

此时，映射矩阵中 *Smod* 列的元素均为 0，如图 1(c)所示，简称 Smod 无关型映射矩阵.

综上所述，在各类约束下，在句法核心、弱句法参数和句法修饰定位中仅上述 3 种映射运算组合是有效的.本节推导出的这 3 种映射组合恰好与从汉英语言现象中提炼出的映射相一致，即常规型映射矩阵对应常规映射运算 ζ，Smod 无关型映射矩阵对应升级变异映射运算 λ_1，Weak 无关型映射矩阵的特例(即 P 映射为空)对应升级变异映射运算 λ_0.因此，有效运算集 $\{\zeta,\lambda_0,\lambda_1,\alpha\}$ 基本概括了语义-句法映射中定位异化处理的主流.

3　语态映射及被动变异映射

翻译的异化现象不仅存在于语义-句法映射中，也存在于其它语言学层次.我们将另文详述其它异化映射.本节仅从词汇-语义角度分析词法层的被动异化现象.

通常，一个事件既可采用主动语态陈述，又可采用被动语态陈述.例如，主动句"我打他"亦可描述为"他被我打了".前句中说话者关注的是施事者，后句中受事者被关注.可见，语态的选择与说话者对事件中各语义角色的关注程度相关.因此，在句子的中间语言表达式(即 CSemS)中，我们引入标记 Focus-role 标注说话者关注的语义角色.在句法实现时，若逻辑主论元具有标记 Focus-role，则激活主动语态映射；若逻辑次论元具有标记 Focus-role，则激活被动语态映射.从而，使生成的句子能准确反映说话者对事件中各角色的关注程度.

语态确定之后，还需进一步确定语态形式标记是否显式表达.例如，汉语表被动意义时，如果词汇语义的搭配决定句子只可能有一种理解，则可省略表示语法关系的形式标记"被"，形成无形式标记被动句，即受事主语句.当句子或结构不足以充分显示主语的受动特点，不足以表示被动动词的语义指向时，使用表被动形式的显式标记.汉语中，某些二价动词(如动态性强，有已实现/未实现范围区分的二价动词[6,7])描述省略施事者的被动句时，既可用有形式标记被动句，又可用无形式标记被动句.例如，"桥被建成了"一般习惯陈述为无形式标记被动句"桥建成了".显然，在生成省略施事者的被动句时，存在如何选择符合陈述习惯的被动语态模式的问题.我们使用偏爱度来描述动词在有形式标记被动句或无形式标记被动句中出现的频率，由语料统计获取.通过二价动词的偏爱度可以分析省略施事者的被动句习惯采用的描述形式.动词 V 对有形式标记被动句的偏爱度记为 F(被-mark)，对无形式标记被动句的偏爱度记为 F(non-被)，且 F(被-mark)+F(non-被)=1.0.当描述省略施事者的被动句时，(1)若 F(被-mark)>F(non-被)，则动词 V 偏爱使用有形式标记被动句；(2)若 F(被-mark)<F(non-被)，则动词 V 偏爱使用无形式标记被动句；(3)若 F(被-mark)=F(non-被)，则动词 V 对两类被动语态模式的偏爱程度相同，使用任一形式都可，此时，系统约定采用有形式标记被动句.根据上述统计信息，我们在汉语电子词典中对二价动词进行了此类标注.

定义 7.　被动变异映射(记为 μ).令被动常规映射 $\eta=\{mark(被动态)=显式标记\}$，

则　　$$\mu::=\{mark(被动态)=隐含标记\}$$

即在 μ 映射中，被动形式标记被隐含.mark(被动态)为被动标记生成函数.

若动词的 LSemS 包含被动变异映射 μ，则带变异类型指示器的 LSemS 为：

$$LSemS(\mu)::=\langle[Pred:(P@\mu,\varphi(P))],[Major:(X,\varphi(X),r(X))],$$
$$[Minor:(Y,\varphi(Y),r(Y))],[Lmod:(Z,\varphi(Z),r(Z))]\rangle$$

$scope(\mu)=\{C_1,C_2,\ldots,C_n\}$，$C_i$ 为诱发 $LSemS(\mu)$ 中被动变异映射的词汇、语义相关条件(如受事者的生命度取值[7]等).

4　CETRAN 中语义-句法映射的实现

如第 1 节所述，在语义-句法映射中常常出现异化现象.为了生成真正符合语言表述习惯的句法结构，语义-句法映射必然涉及具体的词汇、语义、句法实现信息.异化现象遵循的特殊映射组成变异映射集.变异映射集仅与通用的语义、句法范畴相关，可采用与常规映射集相一致的建构方式，使之适用于含有相应异化现象的任何语言.这样，常规映射、变异映射的处理机制可独立于具体词汇，且变异映射集的优先级高于常规映射集；变异类型指示器则成为与词汇相关的特殊语言信息编码.因此，语义-句法映射依赖于 3 类信息：(1) 常规映射集；(2)变异映射集；(3) 变异类型指示器.CETRAN 的变异映射求解没有使用转换机制，而是借助变异类型指示器激活相关映射集.CETRAN 由语义处理、深层句法处理、表层句法处理、深层词法处理和表层词法处理五大部分组成.在句法处理中，我们增强了对特征词(如"把"、"被"等)的处理能力，利用特征标识词优先获取一些特殊句法语义信息[8]，然后进行常规的词法、句法处理，实现

表层句法结构和深层句法结构之间的相互转化.在语义处理部件中,利用词汇、语义信息激活相关的语义-句法映射集,实现中间语言表达式和深层句法结构之间的相互转化.语义-句法映射由下述操作实现:

(1) 词汇选择:将 CSemS 与 LSemS 相匹配,选出恰当词汇.在 CETRAN 的目标语词汇选择中,LSemS 与 CSemS 之间的匹配遵循完全覆盖约束,即:

LSemS B 与 CSemS C 匹配当且仅当 B 完全覆盖 C　　(约束 4)

LSemS B 完全覆盖 CSemS C 当且仅当　　(约束 5)

① C 与 B 间的兼容性检查成功(即 B 中没有不匹配 C 的部分);

② 或者 B 完全匹配 C(即 C 中没有不匹配 B 的部分);或者 B 匹配 C 中除 C'(即 C 的子部分)以外的部分,且 C' 由某一个 LSemS B' 完全覆盖;

(2) 深层句法示例:根据各种由变异类型指示器标记的变异映射和常规映射确定适宜的句法结构.一旦找出与 CSemS 相匹配的 LSemS,则利用与 LSemS 相关联的参数信息生成深层句法结构.由顶层 CSemS 结点产生一个结构,并根据 LSemS 的参数定位要求,挂接它的句法实现参数.在深层句法示例中,语义-句法映射集缺省值为各常规映射集.同时,生成器将考虑第 1 节描述的语义-句法映射变异类型指示器.例如,变异类型指示器 α、λ_0、λ_1 与依据常规定位映射 ζ 定位句法核心及其参数的操作相关联.α 修正 ζ 的弱句法参数和强句法参数定位操作.λ_0,λ_1 修正 ζ 的句法核心、弱句法参数和句法修饰的定位操作.

下面,以"*He happened to meet John*"的英语生成为例,说明基于变异类型指示器的异化求解过程(参见图 2).令 *LSemS*(w)表示词汇 w 的语义配价结构.给定输入 $CSemS=\langle[Pred:(MEET^*,event)],[Major:(HE^*,object,AGT)],[Minor:(JOHN^*,object,OBJ)],[Lmod:(HAPPEN1^*,manner,MANNER)]\rangle$.

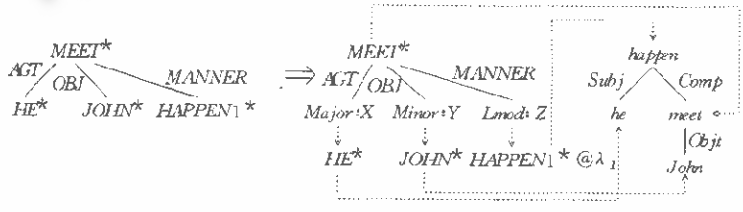

图2　1型升位变异映射处理示意图

在生成深层句法结构时,词汇-语义部件为该 CSemS 选择适宜的英语词汇.因其与 *LSemS*(*meet*)相匹配,词汇"*meet*"将成为其句法结构的组成部分,系统可根据 *LSemS*(*meet*)的参数定位要求挂接所需句法结点;同时,因逻辑修饰元 $HAPPEN1^*$ 与 *LSemS*(*happen*)相匹配,词汇"*happen*"也被选用.由于 *LSemS*(*happen*)含有 1 型升位变异类型指示器 λ_1(即 $LSemS(happen)=HAPPEN1^*@\lambda_1$),在句法实现中,系统应根据 λ_1 的定义,修正常规映射 ζ 的句法核心、弱句法参数的定位操作.因此,在该 CSemS 的深层句法示例中,必须先实现 $HAPPEN1^*$ 所对应的词汇"*happen*",将其升位映射到句法核心位置;然后,将逻辑谓语所对应的词汇"*meet*"及其相关参数"*John*"映射到弱句法参数位置;最后,生成强句法参数"*he*".从而,完成 1 型升位变异映射处理,生成相应深层句法结构,句法部件进一步处理后,输出英语句子"*He happened to meet John*".

5　结束语

本文考察了语义-句法映射中的异化现象,并分析了语义-句法映射运算集的有效性和合理性.在 CETRAN 中,引入带变异类型指示器的异化映射处理机制,实现了机器翻译异化现象的一致化处理,提高了系统的译准率,使生成的句子既能正确表达原中间语言的语义,又符合目标语言的表述习惯,增强了知识的可重用性,使 CETRAN 具有极大的灵活性和可扩展性.

参考文献

1　Cardfurd J C. The Linguistic theory of translation. Beijing: Beijing Travel and Education Press, 1991

2　Yao Tian-shun *et al*. Natural language understanding——a study of making a machine understand human languages. Beijing: Tsinghua University Press, 1995

3　Li De-jin, Cheng Meizhen. A practical Chinese grammar for foreigners. Beijing: Sinolingua, 1994

4　Ren Xue-liang. The comparision between Chinese grammar and English grammar. Beijing: Chinese Social Science Press, 1981

5 MelCuk I A. Depend ency syntax: theory and practice. New York: State University of New York Press, 1989. 103～105
6 Li Shan. Studies on bei-sentence in modern Chinese. Beijing: Beijing University Press, 1994. 3～9
7 Shen Yang, Zheng Ding-ou. Studies on valent grammar in modern Chinese. Beijing: Beijing University Press, 1995. 90～118
8 Guo Honglei, Yao Tianshun. The architecture of Chinese-English bi-direction MT system CETRAN. In: Proceedings of the ICCC'96. Singapore, 1996. 218～226

An Analysis of Divergences and Mapping Operations in Machine Translation

GUO Hong-lei[1,2]　YAO Tian-shun[1]

[1](*Department of Computer Science　Northeastern University　Shenyang*　110006)
[2](*Department of Computer Science　Beijing University of Aeronautics and Astronautics　Beijing*　100083)

Abstract　This paper observes some divergences associated with semantic-syntactic mapping across Chinese and English from lexical-semantic viewpoint. And further analysis of the validity and reasonability of mapping operations in semantic-syntactic mapping is given. Divergence mapping sets, divergence indicators and parameter uniform mechanism are provided to resolve the machine translation divergences. The application of the divergence processing mechanism with divergence indicators improves the accurateness of machine translation, and makes the target-language sentence natural and perfect.

Key words　Machine translation, linguistic divergence, mapping operation, natural language processing.
Class number　TP391.2