SinoTrans 的英文自动后编辑

童爱华
（中国计算机软件与技术服务总公司语言工程部）

Automatic Post—Editing of English (APEE) in SinoTrans

Tong Aihua
CS&S

[内容摘要] 英文自动后编辑是指对原始译文进行自动修改。本文讨论了 SinoTrans 汉英机器翻译系统中 APEE 自动编辑英译文的可行性。APEE 借助 VDL 语言定义了英语自然语言现象。APEE 基于英文词典及 APEE 定义的词和结构序列，按照推理规则，运用英文词属性序列(特征集的一种描述方式)的合一操作，对英译文进行了增、删、改及重组，提高了英译文的整体可读性。

【关键词】 机器翻译 自动后编辑 合一操作

[ABSTRACT] Automatic Post—Editing of English is a program which revises the original translation automatically. This article discusses the probability of Automatic Post—Editing of English (APEE) in SinoTrans MT system. APEE defines the natural language phenomenon of English with VDL. APEE improves the integrated readability of English translation by adding, deleting, changing and reconstructing the English translation, which uses the unification on the English words attribution sequence (a kind of description of attribution set), according to reasoning rules and based on English dictionsry and the words and structure sequence defined by APEE.

[KEYWORDS] MT, automatic post—edit, unification.

严复先生曾言："译事之难：信达雅。求其信已大难矣，顾信矣不达，虽译犹不译也，则达尚焉。"在汉英机器翻译中，由于汉语分析的困难及英语形态的多变，"求其信已大难矣"。SinoTrans 汉英机器翻译系统的目标是立足"信"力争"达"。

汉语形态标志很少，词序灵活自由。人们是靠意会、语义关系及逻辑关系理解汉语。英语则形态丰富，如：人称代词有格的区分，动词有时态语态的变化，名词有单复数形式，主谓有性数格一致的要求等等。这样，即使汉语的深层结构分析正确，英文转换生成仍会存在许多大难题。

如何从汉语分析中找到指导英译文形态变化的依据是非常困难的。现以动词为例。动词的曲折变化信息主要靠汉语虚词提供。如，"着"可以说明其修饰的动词是一种进行状态，"了"可以说明其修饰的动词是一种完成状态，"曾"和"过"可以说明其修饰的动词是一种过去状态。但是，最终应采用何种状态要视具体的英文词而定。

此外，还有英语的习惯表达方式问题。要将汉语翻译成以英语为母语的人读着顺口的英语，不仅涉及文法、选词、曲折变化，还存在词类换译、词义引申、文化差异等问题。例如：

汉语：SinoTrans 给我的印象很深。

英语：SinoTrans impressed me deeply.

名词“印象”译为动词“impress”

汉语：最后发出的命令

英语：the latest instructon issued(此句中文也可为：发出的最后命令)

副词“最后”译为形容词“latest”

SinoTrans 汉英机器翻译系统采用的基本策略是转换法[1]。英译文是按照词驱动的转换规则和转换模式即映射规则生成的。转换分别在三个平面上进行：句子平面、短语平面和短语内平面。句子平面是指主、谓、宾、定、状、补等的序列以及主动态和被动态的转换。短语平面是指作定语或状语用的短语和介词短语的嵌套结构的序列的转换。短语内平面是指短语或介词短语内的词序转换。因此转换成的英译文比较粗糙，需要通过英文自动后编辑模块进行修正，提高译文可读性。

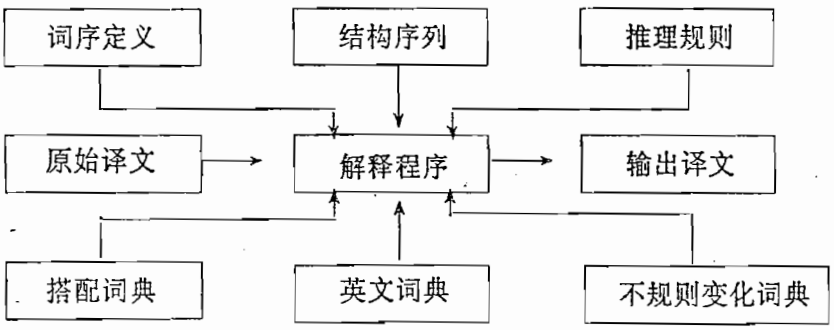
本文的目的是研究英文自动后编辑(Automatic Post—Editing of English，简称 APEE)的可行性。

(一) 英文自动后编辑(APEE)

有关后编辑的概念在机器翻译研究中早已存在。E. Reifler 于 1950 年就提出了前编辑与后编辑的概念，认为前编辑的任务是用符号标出词的语法类，后编辑的任务是选词和重排序[2]。Bar—Hill 于 1951 年提出：后编辑是不可避免的。[2]

英文自动后编辑是对英文转换模块的补充，二者共同构成了 SinoTrans 汉英机器翻译系统英文生成的完整过程。APEE 主要处理中英文差异较大的结构及一些具体的问题，生成结构与形态大致准确的英译文句子。

图 1　英文自动后编辑的框图



APEE 的框图如图 1 所示。APEE 以转换模块输出的原始英译文为输入，以数据文件记载的汉语语法、语义数据、英文树结构、英文词典信息等为依据，对原始英译文做如下的处理：

i. 修正英文词的曲折变化。根据英文语法树中英文词的形态信息及英译文分析结果，通过查找英文词典和不规则变化词典，修正英文词的形态变化。

ii. 调整英文词序。根据英文词在结构树中的位置、语法语义信息、及后编辑中定义的词序

列与结构序列推导出正确的英译文词序。

iii. 修饰英文结构，增、删、改从句引导词。

iv. 选择英文习惯表达方式。根据搭配词典选择最合适的英译文词；根据词序列和结构序列、英文习惯用法或依据推理规则增、删、改英文词。

(二) VDL 语言与 APEE 的主词典——英文词典

APEE 借助 VDL(Vienna Definition Language)形式语言定义英语自然语言现象。

1. VDL 与语言子集

VDL 是用来定义程序语言的一种程序语言，VDL 可以对语言特征进行形式化的描述，并且能够准确地描述程序语言的执行过程。[3,4]

VDL 数据分为两级：第一级为基本元素，这部分数据不可再分，但在程序执行时可通过操作改变其属性。第二级为复合元素，复合元素由基本元素通过构造算子(construction operators) 构造而成。复合元素的成分可为基本元素或复合元素。[3]

APEE 采用 VDL 的数据结构定义了词序列和结构序列对英文语法树进行重组。词序列指短语内词的序列，结构序列指句子成分的序列。APEE 为构造英文词典，还定义了语言全集和语言子集，定义如下：

1) 语言全集　　L ：所有英文词的集合

2) 语言子集

设 attribution 为属性的表达形式，w ∈L，有表达式：

$$\mathbf{ATTRIBUTION} = \{w \mid \mathbf{attribution}(w)\}$$

则称 ATTRIBUTION 为一个具有 attribution 属性的语言子集。

基于 VDL 语言的形式化描述方法，APEE 对语言子集做了如下约定：

i. 语言子集中的最小元素为英文词。

ii. 任一英文词必须隶属一个或多个语言子集。

iii. 对不同语言子集的操作可以获得新的语言子集。

现列举部分语言子集：

NOUN　表示英文名词子集。其它类似的定义有 VERB, ADV, ADJ, ARTI, NUM, PRON, PERP, CONJ

VT _ VERB表示英文及物动词子集

SUBJ _ PRONOUN = {I, YOU, HE, SHE, IT, WE, THEY}

SING _ DET = {THIS, THAT, A, AN, ANOTHER, EITHER, NEITHER, EACH, EVERY}

PLUR _ DET = {THESE, THOSE, ALL}

MODAL = {MAY, SHALL, WILL, WORLD, COULD, MIGHT, SHOULD, MUST, DARE, (<OUGHT>,<TO>), (<USED>,<TO>)}

AUXHAVE = {HAS, HAVE, HAD}

DETERMINER = SING _ DET ∪ PLUR _ DET ∪ NO _ NUM _ DET

AUX = AUXHAVE ∪ AUXBE ∪ AUXDO

(注：集合中的 WORD 表示英文词 WORD。)

2. 英文词典的结构与内容

英文词典所收录的英文词即语言全集中的元素。对 APEE 定义的每一个英语语言子集，英文词典都用一个特定的属性标志来表示，英文词典中具有相同属性标志的英文词就构成了如上所述的一个语言子集。

英文词典的主要字段有英文词字段、词性字段和属性字段。英文词字段记录英文单词的形式，词性字段记录英文词可能具有的词性，属性字段记录该词除词性以外的所有其它属性。

英文词典对每个词的属性描述都包含了词法、句法和语义三方面的信息。英文词典中记载的属性为英文词的静态属性(详见下文(三)中的第 2 段)。

例如，英文词典中"one "的信息如下：

one ：{(数词：[序数词],[单数修饰])，(代词：[单数],[第三人称])，
(形容词：[单数修饰])}

(三) APEE 的算法

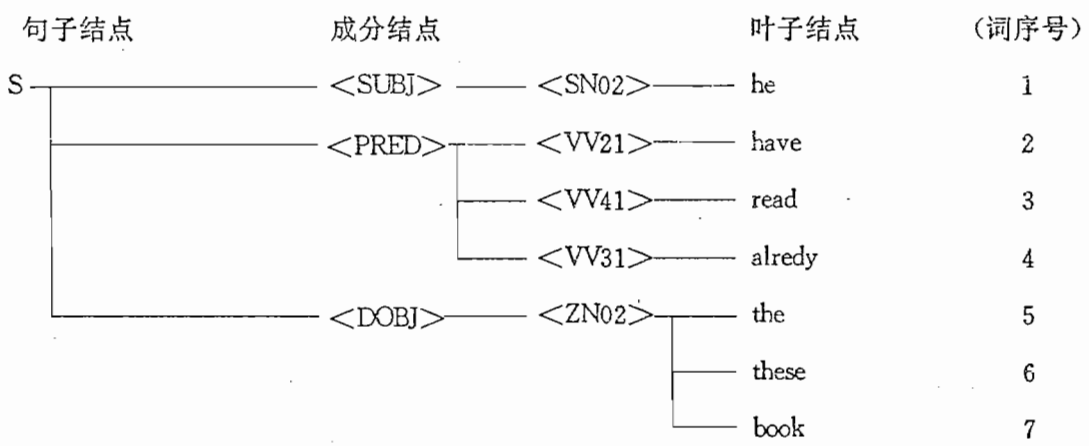
鉴于语法的复杂性和相互制约的多因素特点，自八十年代以来出现了一批采用复杂特征集和合一算法来描述语言现象的语法，诸如词汇功能语法，广义短语结构语法等。APEE 采纳了其中适宜的部分，定义了属性序列和合一操作。

APEE 处理的对象是英文语法树中的结点，因此在讨论 APEE 的算法之前需首先介绍一下 SinoTrans 汉英机器翻译系统的英文语法树。

1. 英文语法树及其结点 [5]

图 2 给出了转换模块生成的一个英译文例句的英文语法树表示。

图 2 英文语法树的表示



为了处理方便，APEE 将英文语法树的结点分为句子结点、成分结点、叶子结点三类。句子结点即句子的根；成分结点包括主部结点、谓部结点、宾部结点；叶子结点为英文语法树的终结点，即英文词。APEE 对结点的表示如下：

叶子结点表示：ni (i——结点在句中的排列序号)

成分结点表示：mi (i——语法树各成分的标志：1—主部结点 2—谓部结点 3—宾部结点...)

图 2 中符号解释：

S ：句子结点，树的根

SUBJ ：主部结点	SN02 ：主部的主语部分
PRED ：谓部结点	VV21 ：谓部的助动词部分
	VV41 ：谓部的主动词部分
	VV31 ：谓部的副词部分
DOBJ ：直宾部结点	ZN02 ：直宾部的宾语部分

2. 属性序列与属性子序列

SinoTrans 用属性说明词法、语法和语义信息[5]，因此属性的集合就等同于复杂特征集。APEE 仍沿用属性描述英文语法树中的结点。在 APEE 中计有三种属性，即静态属性、结构属性和动态属性。静态属性即词典信息，为该词所固有的词法、语法和语义信息；结构属性指从英文语法树中获得的信息；动态属性为在处理过程中可被增加、修改或删除的信息。APEE 将三种属性的属性项按一定的次序排列，构成了英文语法树结点的有穷有序属性序列。下面给出属性序列的表达式。

所有属性项的集合 ：C

结点 ni 的有穷有序属性序列： A(ni) = (a1, a2, ..., aj, ...)

j：属性项在属性序列中的排列序号

aj ∈C

结点 mi 的有穷有序属性序列： B(mi) = (b1, b2, ..., bj, ...)

j ：属性项在属性序列中的排列序号

bj ∈C

根据不同的处理目的，APEE 抽取属性序列中的某些属性项构成一系列的属性子序列：

结点 ni 的属性子序列： Ap(ni) = (a1, a2, ..., ak, ...)

p ：属性子序列的标志

ak ∈A(ni)

k ：属性项在属性子序列中的排列序号

结点 mi 的属性子序列： Bp(mi) = (b1, b2, ..., bk, ...)

p ：属性子序列的标志

bk ∈B(mi)

k ：属性项在属性子序列中的排列序号

APEE 对属性子序列约定：

i. 属性子序列的属性项是有穷有序的

ii. 同一属性子序列中的属性项不允许重复

iii. 不同属性子序列中的属性项允许同名

3. 合一操作

APEE 的合一操作 U(X,Y)定义如下：

对两个有穷有序子序列 X = (x1, x2, ..., xi, ...)

Y = (y1, y2, ..., yi, ...)

进行如下操作：（设 nil 为无定义的元素）

i. 依次比较 xi, yi；

若存在相异项 xi/yi

若 (xi not= nil) && (yi not= nil)
则 结束合一操作；
ii. i 重置值 1；
iii. 依次比较 xi，yi；
若存在相异项 xi/yi
若 yi = nil
则 yi = xi
否则若 xi = nil
则 xi = yi；

4. 除合一操作外，APEE 还定义了其他一些操作。例如：

lookup _ lexicon(WORD)：查找 WORD 的词典信息，获取 WORD 的静态属性初值。
lookup _ shu(WORD)：从英文语法树中获取 WORD 的结构属性初值。
p _ agree(P) (P——短语)：匹配短语中的词序与词序列中定义的序列。
s _ agree(S) (S——句子)：匹配语法树中的结构与结构序列中定义的序列。
sort(ni)：以新获取的属性项值置换 A(ni) 中相应的属性项值。
sort(mi)：以新获取的属性项值置换 B(mi) 中相应的属性项值。
form(S)：根据属性项的值，改变英文词的形态，包括大小写、单复数、动词进行式、过去式和分词形式等；并输出修改后的英译文。

(四) APEE 的处理实例

现以图 2 所示的例句为例，说明 APEE 的处理过程。处理该例句时使用到的属性子序列的属性项如下：

A1(ni)：用于比较所属的成分结点的属性子序列
若 ni 为主部结点的中心词叶子结点
则 A1(ni) =(词性大类，小类，数，时态 1，时态 2，语态)
若 ni 为谓部结点的主动词叶子结点
则 A1(ni) =(词性大类，小类，数，时态 1，时态 2，语态)
若 ni 为谓部结点的助动词叶子结点
则 A1(ni) =(小类，数，时态 1，时态 2，语态)
若 ni 为谓部结点的副词叶子结点
则 A1(ni) =(时态 1，时态 2)

A2(ni)：用于确定可否前继冠词的属性子序列。
若 ni 为前继冠词 则 A2(ni) =(冠词保留标志) 否则 A2(ni) = (前继冠词标志)

A3(ni)：用于比较名词和动词的数的属性子序列。
A3(ni) =(数)

A4(ni)：用于比较动词数、时态和语态的属性子序列。
A4(ni) =(数，时态 1，时态 2，语态)

B1(mi)：用于主部成分与谓部成分比较的属性子序列。
B1(mi) =(成分的数，成分的时态 1，成分的时态 2，成分的语态)

B2(mi)：用于比较下属中心词的属性子序列。

B2(mi) =(中心词的词类，中心词的小类，中心词的数，中心词的时态 1，中心词的时态 2，中心词的语态)

B3(mi)：用于比较下属助动词的属性子序列。

B3(mi) =(助动词的小类，助动词的数，助动词的时态 1，助动词的时态 2，助动词的语态)

B4(mi)：用于比较下属副词的属性子序列。

B4(mi) =(谓部时态 1，谓部时态 2)

(注：时态 1—— 完成时，进行时，将来时 时态 2—— 现在时，过去时)

当输入的英文词为原形时，APEE 假定成分结点的属性序列中所有属性项的初值为 nil，叶子结点的属性序列中所有需经推理才能获得的属性项的初值也为 nil 。

APEE 的合一操作主要在三个层次上进行：i. 叶子结点间的合一操作，ii. 叶子结点与成分结点间的合一操作，iii. 成分结点间的合一操作。三个层次的合一操作可以重复进行。每进行一次合一操作，都要调用 sort(ni)或 sort(mi)来修改相应的 A(ni) 或 B(mi) 的值。

下面依次列举 APEE 处理该句时的主要步骤。

1. 获取属性初值

处理对象：英文语法树的所有结点

操作内容：lookup _ lexicon(WORD)；lookup _ shu(WORD)

处理结果：该句的所有结点的属性序列具有了初值

2. 叶子结点间的合一操作

i. 处理对象：the these book 结点序列：n5 n6 n7

操作内容：

A2(n5) = (nil) ; A2(n6) = (noarti)

(A2(n5), A2(n6)) ==> A2(n5) = A2(n6) = (noarti)

A3(n6) = (plur) ; A3(n7) = (nil)

U(A3(n6), A3(n7)) ==> A3(n6) = A3(n7) = (plur)

处理结果：the 要求删除 book 赋予复数形式

ii. 处理对象：have read 结点序列：n2 n3

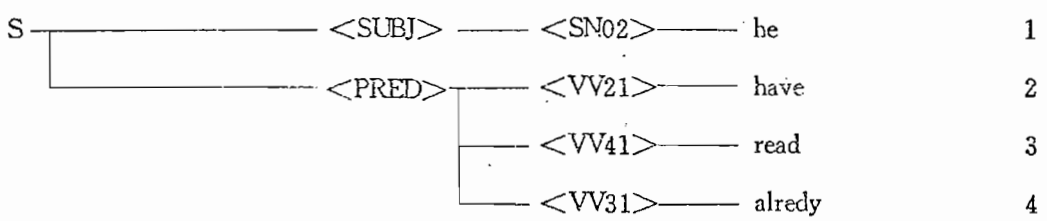
操作内容：

A4(n2) = (nil,pp,nil,nil) ; A4(n3) = (nil,nil,nil,nil)

U(A4(n2), A4(n3)) ==> A4(n2) = A4(n3) = (nil,pp,nil,nil)

处理结果：read 赋予分词形式

以下 3——5 步的处理对象：



3. 叶子结点与成分结点间的合一操作

当成分结点的叶子结点数为1时，成分结点可直接抽取叶子结点的属性值。当叶子结点数大于1时，各成分结点抽取属性值的方法不同。

操作内容：

A1(n1)=(pron,subj_pron,3ps,nil,nil,nil) ; B2(m1)=(nil,nil,nil,nil,nil,nil)

U(A1(n1), B2(m1)) ==> A1(n1) = B2(m1) = (pron,subj_pron,3ps,nil,nil,nil)

A1(n2) = (aux_have,nil,nil,nil,nil) ; B3(m2) = (nil,nil,nil,nil,nil)

U(A1(n2), B3(m2)) ==> A1(n2) = B3(m2) = (aux_have,nil,nil,nil,nil)

A1(n3) = (verb,vt_verb,nil,pp,nil,nil); B2(m2) = (nil,nil,nil,nil,nil,nil)

U(A1(n3), B2(m2)) ==> A1(n3) = B2(m2) = (verb,vt_verb,nil,pp,nil,nil)

A1(n4) = (nil, p) ; B4(m2) = (nil, nil)

U(A1(n4), B4(m2)) ==> A1(n4) = B4(m2) = (nil, p)

4. 成分结点间的合一操作

操作内容：

B1(m1) = (3ps,nil,nil,nil) ; B1(m2) = (nil,pp,p,nil)

U(B1(m1),B1(m2)) ==> B1(m1) = B1(m2) = (3ps,pp,p,nil)

5. 叶子结点与成分结点间的合一操作

操作内容：

B3(m2) = (aux_have,3ps,nil,p,nil) ; A1(n2) = (aux_have,nil,nil,nil,nil)

U(B3(m2),A1(n2)) ==> A1(n2) = B3(m2) = (aux_have,3ps,nil,p,nil)

以上3——5步的处理结果：<SUBJ>与<PRED>中的每个结点的属性值被修改

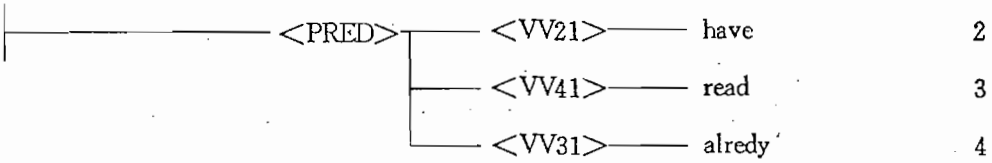
6. 结点位置的确定

处理对象：英文语法树

操作内容：p_agree(P) ; s_agree(S)

该例句匹配的结构序列：is_aux(WORD) is_2v_adv(WORD) is_verb(WORD)

处理结果：



7. 英译文输出

处理对象：英文语法树

操作内容：form(S)

处理结果(即APEE的输出译文)：He had already read these books.

上述规则中所使用的属性项说明：(参照上文(二))

subj_pron：主格代词　　3ps：第三人称单数　　vt_verb：及物动词

aux_have：助动词HAVE　pp：完成时态　　p　　：过去时态

2v_adv：有助动词时，该副词位于助动词之后，主动词之前；

无助动词时，该副词位于主动词之后。

(五) 讨　论

APEE 解决了英文转换模块中不易处理的一些问题，在原有基础上提高了英译文的整体可读性。APEE 能比较成功地处理：

1. 有明显时态标志时的动词曲折变化(时态、主谓一致和主从句一致)；
2. 谓语部分的词序(情态动词、助动词、副词和主动词的序列)；
3. 动介搭配和惯用法(如：'...进行+VERB'，'...得到+VERB'等的转换)；
4. 从句引导词(WHICH,THAT)。

由于 APEE 只能根据汉语分析的信息进行编辑，所以许多由于两种语言的习惯用法造成的差别，APEE 无能为力。例如，

1. 名词的数

   汉语名词的数常常是意会的，所以在没有标识数量的修饰词的情况下 APEE 无所适从，

2. 动词的时态

   汉语动词的时态信息常在虚词中表达或隐含在语义中。由于 APEE 不能作隐含语义的逻辑推理性编辑，所以有些情况不能处理。如"到 1997 年，香港的主权已经收回"。

3. 目前处理不了同一种汉语句型因语义不同而要求变换英语句型的问题，例如：

   汉　语：参与社会生活的要求增加了，

   英译文：The demand to participate the social life was increased.

   汉　语：参与社会生活的人增加了，

   英译文：The number of people who participate the social life was increased.

参考文献

1. 吴蔚天，"SinoTrans 汉英机器翻译系统"，《MMT'91 论文集》，1991. 8，北京.
2. W. J. Hutchins, Machine Translation, Past, Present, Future, Ellis Horwood limited, Halsted Press. 1986.
3. P. Wegner, The Vienna Definition language, Computing Surveys 4(1) (1972) 5—63.
4. Dale Johnson and Barrett R. Bryant , Formal syntax methods for natural language, Information Processing Letters 19(1984) 135—143.
5. 吴蔚天 罗建林，"关于建立汉语形式语法体系的探讨"，《机器翻译研究进展》，电子工业出版社，1992，135—144.