汉英机译系统英文生成中特殊动词短语的处理研究*

洪青阳，李堂秋，杨晓峰
(厦门大学 计算机科学系，福建 厦门 361005)

摘　要：主要讨论汉英两种语言特殊动词短语表示方式的不同，包括动词短语的拆分、变形和辅助成分的确定。在分析模块给出的中间语言信息不够详尽的情况下，需要生成模块针对性地作进一步处理，才能得到规范准确的译文。生成模块中一个详尽的词典非常重要，提出的生成策略就是在一个完备的词典的基础上通过规则来实现的。

关键词：动词短语；中间语言；生成模块；词典；生成规则；拆分
中图分类号：TP391.2　　**文献标识码**：A　　**文章编号**：1001-3695(2001)03-0027-04

Research on the Particular VP for the English Generation of Chinese-English Translation System

HONG Qing-yang,　LI Tang-qiu,　YANG Xiao-feng
(Dept. of Computer Science, Xiamen University, Xiamen Fujian 361005, China)

Abstract: This article mostly discusses the difference of expression of special VP between Chinese and English, including the split, distortion and assistant ingredient's confirming of VP. When the informations of interlingua given by the analysis model is not elaborate, we have to deal with the generation model centralizly and ulteriorly to get the normative and correct translation. A elaborate lexicon is very important in the generation model. The generation tactic given by this article is realized by rules on the foundation of a elaborate lexicon.

Key words: VP; Interlingua; Generation model; Lexicon; Generation rule; Split

1　前言

不管是汉语还是英语中，动词短语(VP)都是很重要的成分，但两种语言对一些特殊动词短语的表示形式有一定的差别，特别是对“动词+宾语”结构的表示。如汉语中“关掉它”的中间语言如下：

```
(*A-TURN-OFF
   (TENSE PRESENT) (MOOD DEC)
   (THEME (*R-IT (CAT PRON) (SUBCAT PERSONAL_PRON)
                 (AGREE SG)(PERSON THIRD)))
   (CAT V) (SUBCAT WEIBIN_VERB))
```

我们可以发现，动词“关掉”在中间语言中被表示为“*A-TURN-OFF”，如果简单地按照中间语言生成英文，就会出现不合英文语法规范的结果“turn off it”，而正确结果应为“turn it off”。

从上面举的例子我们可以看出，生成模块不能机械地搭积木似地把中间语言转化为目标语言，有必要借助常识、上下文和英文语法对中间语言进行二次处理，对某些特殊成分做特殊的转化。本文讨论的主题是特殊动词短语的拆分、变形和补助成分的添加。

收稿日期：2000-08-14
基金项目：国家“863”资助项目(863-306-ZT03-06-1)

2　动词短语(VP)的类型

有一类动词短语可看成是由动词(V)与一个辅助成分构成的。辅助成分可以是介词短语(PP)，介词如up，in等；也可以是附加词，我们称这个附加词为PARTICLE，一般可看成对应汉语中的副词，如out，over和on等。当然这个辅助成分有时很难区别是介词还是副词，如下面的例子：

例1，look *over* the paper

这个短语有两种理解。如果把over看成是一个PARTICLE，该短语的意思是“看报纸”；而如果把over看成是一个介词，则意思是“看在报纸上的某样东西”。两种不同的理解可用如图1语法树表示。

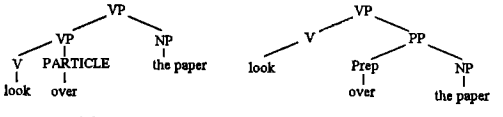

图1　“look over the paper”的两种语法树

当然，我们没必要在这里详细探讨辅助成分的鉴别，我们关心的是如何找到规律，寻找解决的办法。

动词后接介词短语和接PARTICLE有一个很明显的区别。当宾语是一个代词(PRON)的时候，“V+PARTICLE”的情况代词要放在PARTICLE前面，结构为“V+PRON+PARTICLE”，如“I found it out”。而“V+Prep”的情况代词要放在介词后面，结构为“V+Prep+PRON”，如“I looked on it”。

当后面接的宾语是NP的时候，NP可放在PARTICLE前面，也可放在后面。即有两种可选结构：“V+NP+PARTICLE”，或“V+PARTICLE+NP”。

例2，I gave *up* the game. I gave the game *up*.

而对于V+Prep的结构，NP只能放在介词后面。结构：“V+Prep+NP”。

例3，I climbed *up* the ladder. *I climbed the ladder *up*.

有一种特殊的动词短语，就是不管宾语是什么成分，都得放到短语中间，如“keep...in mind”，虽然在英语中的正确表示中间应有省略号“...”，但按照中间语言的规范表示，该短语在中间语言中被表示为*A-KEEP-IN-MIND，我们在生成的时候就要进行拆分，把*A-KEEP-IN-MIND的ROOT(这里为“keep in mind”)拆分为“KEEP”和“IN MIND”两部分，而把宾语置于中间。

例4，我把它记在心里。I keep it in mind.

我把他的话记在心里。I keep his words in mind.

类似的短语还有“carry ...on the back”，“take ... into account”等。

上面考虑的都是动词后接宾语的类型，没有接宾语的动词短语也有需特殊处理的。

带系动词的特殊动词短语如“be able to”，在疑问句中作谓语成分时，要把系动词提到句首。

例5，他能做这份工作吗？*Is* he able to do this job?

在否定句中，否定词“not”要置于动词短语中系动词后。

例6，is not able to *is able to not

有些动词短语含有尚不能确定的成分，如“try one's best”，短语中的“one's”需借助上下文才能确定。

例7，我将尽力而为。I will try my best.

*I will try one's best.

这句话的主语是“I”，且没有其它可指代的对象，因此在生成中我们可断定“one's”应由“my”代替。由于人称、数和性别的不同组合，因此“one's”就可能有多种情形。

例8，她会尽力而为。

She will try her best.(第三人称、单数、女性)

小明会尽力而为。

Xiaoming will try his best.(第三人称、单数、男性)

我们会尽力而为。

We will try our best.(第一人称、复数)

上面讨论的是一些特殊动词短语。另外很多动词需要一个专门的介词短语(PP)来作补助成分。如动词“give”接的补助成分是一个NP和一个介词为“to”的PP，例“Jack gave the book to the library”。我们注意到不能用其它介词，例如下面的句子就是错误的：

例9，*Jack gave the book from the library.(其实from the library修饰the book)

但更多的动词却没有这么多的强选择性，如动词“put”就能接任何表地点的短语。

例10，Jack put the book in the box.

Jack put the book inside the box.

Jack put the book by the door.

为了表示这种情形,我们考虑用标识符来表示带特殊介词的介词短语。“give”的补助成分可表示为：NP+PP[to]。同样，动词“decide”的补助成分为：NP+PP[about]，而“blame”为：NP+PP[on]，如“Jack blamed the accident on the police”。

像“put”这样的动词，我们已经知道可接任何表地点的介词短语，地点状语其实也可以是名词短语，如home，或者副词，如back，here。在这里，我们有必要区别地点(location)短语和表动作路径(path of motion)的短语。一般来说，以to开头的介词短语表示动作路径，这种介词短语不能用于只能接地点短语的动词，如put，下面的例子是错误的：

例11，*I put the ball to the box.

表1为我们给出的一个比较完整的动词补助成分结构列表。

表1 动词补助成分的各种结构

Verb	Complement Structure	Example
laugh	Empty(intransitive)	Jack laughed.
find	NP(transitive)	Jack found a key.
give	NP+NP(bitransitive)	Jack gave Sue the library.
give	NP+PP(to)	Jack gave the book to the library.
reside	Location phrase	Jack resides in Rochester.
put	NP+Location phrase	Jack put the book inside.
speak	PP[with]+PP[about]	Jack spoke with Sue about the book.
try	VP[to]	Jack tried to apologize.
tell	NP+VP[to]	Jack told the man to go.
wish	S[to]	Jack wished for the man to go.
keep	VP[ing]	Jack keeps hoping for the best.
catch	NP+VP[ing]	Jack caught Sam looking in his desk.
watch	NP+VP[base]	Jack watched Sam eat the pizza.
regret	S[that]	Jack regretted that he'd eaten the whole thing.
tell	NP+S[that]	Jack told Sue that he was sorry.
seem	ADJP	Jack seems unhappy in his new job.
think	NP+ADJP	Jack thinks Sue happy in her job.
know	S[WH]	Jack knows where the money is.

3 特殊动词短语的生成策略

前面已比较详细地讨论了动词短语的一些特殊情形。现考虑如何在生成模块中实现。

我们研制的汉英机译系统是在Linux的环境下，用Common Lisp编写的。生成模块包括词典、生成规则和一些Lisp程序。句子的生成是以语义的概念框架表达式为基础，而不是语法结构表达式为基础。句子的语义结构直接映射到句子的线性结构。生成模块的每一条规则由一个上下文无关的短语结构描述部分和一组伪方程式组成。

词典在生成模块中的确非常重要。可以说，词典是一个不断扩充的知识库，它包括中间语言所不能给出的其它信息。一个好的词典可使生成规则写起来得心应手，既可减少工作量，又可细化某些特殊成分的生成。

根据第二部分的讨论，特殊动词短语包括下面六种：

I. 动介型(包括类动介型)(V+Prep)如look at
II. 动副型 (V+PARTICLE)如find out
III. 一定要拆分的短语 如keep ... in mind
IV. 短语中含特定成分的 如try one's best
V. 带系动词的短语 如be able to
VI. 对补助成分有特殊要求的动词

针对以上不同动词短语类型，采取的生成策略也有所不同。首先我们可在词典中加上分类标志，以使系统能对不同类型的动词短语进行识别，然后才能调用相应的生成规则。针对每一种动词短语，词典中有各自的特殊属性，即描述该类动词短语的独特信息。如第六种类型的动词put，其在词典中的表示如下：

```
(*A-PUT
(CAT V) (ROOT "put")
(SUBCAT NP+LOCATION)
(PRESPART-FORM "putting") (PAST-FORM "put") (PASTPART-
    FORM "put")
```

我们以(SUBCAT NP+LOCATION)来表示动词“put”必须后接一个名词短语和地点状语。当生成碰到“put”这个动词时，在词典中发现该词有这种特殊属性，就可在规则中采取相应策略。

我们接下来看一个具体的生成例子。我们举第三种类型的动词短语“keep...in mind”，这种情况比较典型。第一步是在词典中表示该VP，其相关特殊属性要详尽给出：

```
(*A-KEEP-IN-MIND
        (CAT V)
        (SUBCAT CF)
        (ROOT "keep") (EROOT "in mind")
        (PRESPART-FORM "keeping")
        (PRES3S-FORM "keeps")
        (PAST-FORM "kept")
        (PASTPART-FORM "kept") )
```

“keep...in mind”被拆成两部分：ROOT和EROOT，ROOT表示中心词，这个中心词要根据时态、人称、数等进行特殊变形，其各种变形已列在词典中。EROOT表示后一部分，是一个辅助成分，不随语境变化，因此我们在生成中保持其不变。(SUBCAT CF)表示该动词短语的ROOT在生成的时候要进行拆分。

有了词典对该动词短语的具体描述信息，我们就可写出“keep...in mind”接宾语时动词短语拆分的生成规则：

```
(<vp> --> (<v> <np> <vpend>)
        (((x0 subcat) =c cf)
        ((x0 theme) = *defined*)
        (x2 == (x0 theme))
        (x1 = x0)
        ((x3 root) = (x0 eroot))
        ((x2 obj) = +)))
```

其中((x0 cf) =c cf)表示拆分判断条件，因为动词短语“keep...in mind”的属性满足该条件，因此这条规则被激活。动词短语VP被分成两部分，前部分V(keep)和后部分VPEND(in mind)，宾语成分NP置于中间，规则中<v> <np> <vpend>正表示了生成译文中的排列顺序。

对于第四种类型，即含有未定成分的动词短语，需借助上下文来确定短语中的未定成分，其实是一个相对简单的上下文指代问题。对于“one's”的确定，也就是确定one's所指定的对象，然后根据该对象的人称、数和性来决定one's的具体形式。我们总结出几条规律，有助于不定成分的确定。

规律一：如果中间语言中有施事格AGENT，则one's倾向于指代该AGENT；

规律二：如果该动词短语出现在兼语句中作兼语的谓词，则one's应指代该兼语；

规律三：如果当前句子中找不到指代对象，则可在比较临近的上下文中找到，即就近原则。

对于第六种类型，即对补助成分有特殊要求的动词，我们引入“模板化生成”这个概念。生成模块负责将中间语言文本根据不同的概念映射到目标语言的短语或句子中，控制结构是从上到下，从左到右。映射过程分为两步：词条挑选和基于语义的句子生成。生成过程把中间语言的语义表达式作为输入，将中间语言文本逐步分解并按它们的语义功能安排顺序，一步生成译文。由于中间语言由分析模块从源语言分析得来，在目前分析技术不够透彻的情况下，或多或少会保留源语言的一些痕迹，因此得到的译文就可能不

够“纯”，带有源语言的特征，甚至会出现意思完全相反的结果。

如果采用逆向思维，直接面向目标语言做一些辅助工作，就有可能生成更加规范的译文。如对该种类型的动词，我们先分析该动词对后接补助成分的要求，在译文中就可遵照这个特殊要求，犹如有一个模板让生成模块来套用。我们要让生成规则判断中间语言中是否具备该动词所需的辅助成分，如果都具备了，就自然而然地可生成目标语言，反之，如果有些成分少了，可能是分析模块没分析全，我们要在分析模块找原因，也可能分析模块没办法给出进一步的信息。在中间语言已包括所需主要成分的情况下，我们仍需要根据“模板”添加必要的成分，以使译文更加准确完整。如动词regret后面如果接宾语从句，该宾语从句一般需带一个连接词“that”，那我们在生成的时候就要进行添加。

上述生成策略是基于我们研制的汉英机译系统的，其它类型动词短语也是按照这种思路来生成的。

4 实验结果及结论

本文阐述的特殊动词短语生成策略，已应用于我们研制的汉英机译系统，取得了令人满意的生成效果，译文质量有了明显的提高。

汉英两种语言在词法、句法甚至句法与语义的关系上都存在极大的差异，动词短语的差异只是一个方面。我们期望能在生成这一块做更多的工作，面向目标语言，引入更多的知识库，弥补分析模块的不足，以翻译出更规范更准确的结果。对于上面提出的“模板”生成，我们将做更深入的研究。

参考文献：

[1] 徐广联. 短语动词[M]. 大学英语语法(讲座与测试), 上海: 华东理工大学出版社, 1998.

[2] 李堂秋, 卢伟. 一种高效的通用型自然语言语法分析器系统分析[J]. 厦门大学学报(自然科学版), 1997, 36(6).

[3] 李堂秋, 卢伟. 基于语义的中文句子的直接生成方法[J]. 厦门大学学报(自然科学版), 1998, 37(5).

[4] 周会平. 基于中间语言的汉英翻译系统ICENT的研究与实现[D]. 工学博士学位论文, 国防科技大学研究生院.

[5] 王凌飞. 汉英机译系统中上下文处理的研究[D]. 2000'计算机系硕士学位论文, 厦门大学研究生院.

作者简介：

洪青阳(1977-)，男，计算机应用硕士研究生，研究方向为人工智能、机器翻译；李堂秋，男，教授，主要研究方向为人工智能、自然语言处理、Linux应用与研究.

(上接第13页)

证明：

(1)n=1时，只有一个字串，显然(5)式成立。

(2)令n=N时，$T_N=c\times\sum_1 m_i^2 = O(n)$成立。

(3)n=N+1时，参加排序的总字符串数增加1，此时可分两种情况证明。

1)新增字串无法归入已有的k个队列中，而单独形成一个队列，则总工作量

$T_{N+1}=T_N+T_{新增}=O(n)+T_{新增}$，

根据步骤1(n=1)得，　T=O(n)。

2)新增字串被归入k个队列中的任一个，假定归入第i队列，此时队列i中汉字串的个数$m_i'=m_i+1$，总工作量

$$T_{N+1}\le c\times m_1^2+c\times m_2^2+\ldots+c\times(m_i+1)^2+\ldots+c\times m_k^2$$

$$=c\times m_1^2+c\times m_2^2+\ldots+c\times(m_i^2+2m_i+1)+\ldots+c\times m_k^2$$

$$=T_N+2c\times m_i+c$$

由于T_N为O(n)，$2c\times m_i+c$也为线性的，所以T_{N+1}的复杂度为O(n)，(5)式得证。

由于在Step16前形成的队列之间已经有序，因而在经Step17的内部排序后，Step18输出的汉字串即是全部排序结果。在最坏情况下，即参加排序的全部汉字串的首字符相同，而被分配到一个队列中，算法的复杂度与传统的排序方法相同。

6 结束语

由于汉字字符集属于大字符集，由汉字符组合而成的汉字串的数量更为庞大，因而汉字串排序算法的研究受到普遍关注。本文提出了一种汉字串的快速排序方法，它具有以下特点：

- 能对任意汉字串进行排序，算法的复杂度为O(n)。
- 传统的排序方法是以在全部数据之间相互比较为基础的，算法的复杂度最低为O(nlogn)；而本文采用将数据先分组，再在组内比较的策略，使算法的复杂度降低，达到快速排序的目的。
- 本文所提策略无需将汉字进行类似文献[2]中的变换，仅基于汉字的存储结构就可对汉字串进行分组，降低算法的复杂度。
- 本文提出了一种利用汉字串比较位的合一逻辑运算来判定两个汉字串大小的方法。

参考文献：

[1] 周培德. 算法设计与分析[M]. 北京: 机械工业出版社, 1991.

[2] 周建钦. 关于汉字的分组排序算法及其复杂性[J]. 中文信息学报, 10(3).

[3] 王晓龙, 李红斌. 计算机汉字处理实用技术[M]. 哈尔滨: 哈尔滨工业大学出版社, 1993.

作者简介：

李建华，博士研究生，主要从事人工智能、自然语言理解和文本校对等领域的研究；王晓龙，博导，香港理工大学高级研究员，主要从事人工智能、自然语言理解和中文信息处理等领域的研究.