汉日韩-英多语机译系统的通用英语生成器

郭宏蕾　蒋建民　胡岗

IBM中国研究中心 北京 100085

E-mail {guohl@cn.ibm.com, jiangjm@cn.ibm.com, hugang@cn.ibm.com}

摘要 在汉日韩英多语机译系统的研制中，我们设计并实现了一个独立于源语言的通用英语生成器，本文主要介绍了通用英语生成器的句法依存结构描述、词汇复杂特征集及生成策略。

关键词 机器翻译 英语生成 多语翻译

Universal English Generator in Chinese-Japanese-Korean-English Multilingual MT System

Guo Honglei　Jiang Jianmin　Hu Gang

IBM China Research Laboratory, Beijing, 100085

E-mail {guohl@cn.ibm.com, jiangjm@cn.ibm.com, hugang@cn.ibm.com}

ABSTRACT In Chinese-Japanese-Korean-English multilingual machine translation system, we present a universal English generator which is independent of the source language. This paper describes the syntactic dependency structure, lexical complex feature set and the English generation strategy in the universal English generator.

Keywords machine translation, English generation, multilingual translation

1.引言

机器翻译主要包括源语言的分析及目标语言的生成两个过程，从机译系统开发者的角度看，与系统设计相关的关键问题是：1）系统的语言学基础；2）执行翻译的计算机程序的内部操作，即用于完成源语言分析及目标语言生成的内部计算；3）各种知识源的编码和使用。

在现有的机译系统中，大多数目标语言生成器都设计成为某种特定源语言服务的生成器，许多生成策略与源语言极为相关。这样的目标语言生成器对源语言具有一定的依赖性，通用性不强，n种语言间的互译至少要建立n(n-1)个语言处理部件[1]，系统的开发任

务十分繁重。出于开发周期和成本的考虑，在汉英、日英、韩英多语机译系统的研制中，我们设计并实现了一个独立于源语言的英语生成器，使之成为适用于汉英、日英、韩英多语机器翻译的通用英语生成器。本文主要介绍了通用英语生成器的句法依存结构描述及基于词项序位计算的生成策略。

2.句法依存结构的表示及词汇复杂特征集

我们的汉英机译系统采用了分析-转换-生成处理模式，主要由汉语分析器、汉语-英语转换器、英语生成器组成（参见图1）。

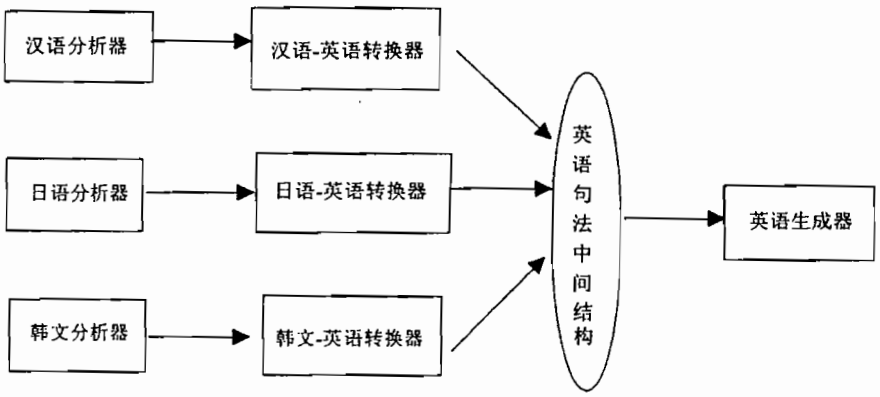
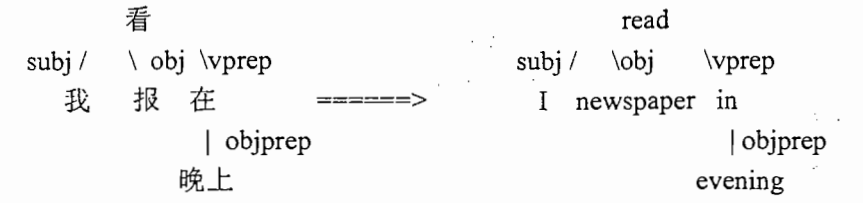


图1　汉日韩-英多语机译系统的体系结构

在机器翻译中，英语生成器主要对源语言-目标语言转换器输出的与原文等价的句法结构进行处理，最终形成译文。显然，为了使英语生成器独立于源语言，具有通用性，在源语言和目标语言之间必须建立能描述句子内各词项间依存关系及复杂特征的中间表达结构。

我们以原有的英语槽文法（English Slot Grammar）[2]描述为基础，给出了通用英语句法表达式的规范描述体系，共定义了近70种词项之间的句法依存关系，用于构造句法中间表达式，即ESG-Tree结构(English Slot Grammar Tree)。ESG-Tree结构是依据源语言的句法结构及词项特征，通过目标语言词汇选择以及句法结构映射转换而来的一棵英语句法依存树。

例如，给定汉语句子“我在晚上看报”，经过英语对译词选择、句法结构映射，其汉语句法结构可转换为等价的英语句法结构ESG-tree，如图2所示，其中，subj（i.e. 主语槽)、obj(i.e. 宾语槽）、vprep（i.e. 动词的介词修饰槽）、objprep（i.e.介词宾语槽）为词项之间的句法依存关系，subj表示“I”是动词“read”的主语，obj表示“newspaper”是动词“read”的宾语，objprep表示“evening”是“in”的介词宾语，vprep表示“in”是动词“read”的介词修饰。



(1) 汉语句法结构　　(2)英语句法中间结构ESG-Tree

图2　英语句法中间结构示例

ESG-Tree给出了构成目标句子的英语实词和部分虚词节点以及节点之间的依存关系，ESG-Tree中的英语虚词节点具有句法成分引导作用，多为介词、连词等，一般都是由源语言中的虚词转换而来，例如，在上例中，英语介词“in”是由汉语词项“在”转换而来的，在句法结构中引导了一个介词短语，修饰动词“read”。在英语生成阶段，除了对ESG-Tree中各词项进行线性排序和形态生成外，还将根据英语词汇间的句法依存关系，进一步添加相关虚词（如冠词、介词、关系代词、助词等），以构成符合语法的英文语句。

在英语生成中，我们利用复杂特征集描述句法树ESG-Tree中词节点的词法、句法、语义属性。词节点的复杂特征集主要包括如下几个层次：

词法层：记录词节点的词法特征，主要为：

headPOS	中心词的词性
phrasePOS	短语的词性
word	词形
MorphFeature	词法属性（如时态、语态、虚词搭配、级、数等）

句法层：描述词节点间的句法依存关系、句法特征、句式特征，主要为：

slot	词项在句子中的句法功能
optionSlot	句法依存特征
parent	句法支配者
son	句法依存者
sentFeature	句式 (如疑问/否定/祈使/感叹/陈述）

语义层：semFeature　　语义分类码

3. 英语生成策略

人们在表述思想时，总是先确定表述内容，再将其按一定的语言法则组织成句法框架，加上必要的形态变化，形成一个符合语法规范的句子。这种表述过程具有明显的层次性。与人类思维中潜在的层次性相对应，机器翻译也时常将分析、生成划分为多个环节。源语言分析是从源句子开始，通过词法分析、句法分析和语义分析几个环节，不断抽取、累加词法、句法、语义信息，找出句子的形式表示与其真实含义之间的映射关系，建立相应的句法、语义结构，从而完成对源语言句子的理解。目标语生成则是利用

分析转换过程中捕获的大量信息，通过词汇选择、句法构造、词项调序、词法生成、曲折性变换几个环节的处理，最终生成相应的目标语句子[3]。

通常认为目标语生成是源语言分析的逆过程，但生成也有其独特的处理策略。对于任何一种语言，要表达正确的语义，生成符合语法的句子，需要确定两部分内容：1）组成该句子的各词项之间的顺序，即在句中的位置；2）确定每个词项应具有的形态。因此，通用英语生成器的主要任务是依据输入的英语句法结构，计算每个词项x在句中的位置（简称序位，记为order(x)）和词项应有的形态。

3.1 词项序位计算

我们的英语句法树以层次形式描述句子的具体组织，记录各词汇项（即实词、虚词）的句法依存者、支配者以及依存关系。因此，句法层生成主要包括：（1）句法成分线性化；（2）句法成分的相关语法标记生成，即根据各句法成分的语法功能及特征，添加相关的语法标记，如关系从句中的关系代词的生成，时间、地点状语的引导介词的添加，不定式引导标记“to”的添加，并计算相应的序位；（3）句子表述结构的转换，即在疑问句、特殊疑问句、否定句、感叹句等特殊表述形式的构造中，添加相关词汇，并形成符合句式表述约定的词项序位。 贯穿句法层处理的核心任务是词项序位的计算，词项序位计算主要包括常规序位计算、序位变异计算、序位内嵌计算几类基本操作。

1）常规序位计算

依据通用的句法调序规则，对各词项间的句法依存关系及复杂特征集进行合一和扩展运算，确定词项在目标语句中的基本序位。

例如：给定句子“他喜欢书”，英语句法结构如下所示，

like

subj/ \obj

he book

在生成中，各词项的序位排列为： He like book, “like”的序位计算为

order(like)=order-Relation(he,like) ∧ order-Relation(book,like)

2）序位变异计算

在生成中，受语句的特殊句式、强调焦点的制约，某些词项的序位将发生变异，由其对应的常规序位变换到某一特定序位，这类序位计算称之为序位变异计算。

引发序位变异计算的语言现象主要有如下两类：

（1）基于语法结构的序位变异：在疑问句、感叹句、特殊疑问句等句式表述结构中，助词、疑问词等将由其所在的常规序位变换至句首。

（2）基于表述焦点的序位变异： 为了尽可能保证原文与译文在描述风格上的一致性或相似性，以源语言表述风格为基础，对一些焦点词项的序位进行调整。 通常，如果原文将时间、地点等放在句首加以强调，在译文生成时，这些被强调的成分将尽可能放在句首。 例如，在句子“*秋天，苹果熟了*”中，时间状语“ 秋天”被放在句首加以强调。在生成中，如果不考虑原文的描述风格，译文应为“*Apple has been ripe in autumn.*”，但为与原文描述风格一致，在英语生成中，我们通过序位变异计算，将时间

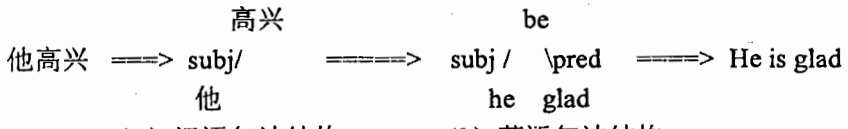
状语“*in autumn*”由句尾前移到句首，生成最终的译文“*In autumn, apple has been ripe.*”，以体现原文的风格。

序位变异计算主要涉及两个内容：1）界定将发生序位变异的子树的边界，2）计算应放置的新序位。在序位变异计算中，一般根据句法、词汇约束，对相关成分加以变异标记，根据相关的变异迁移原则，给出最终的变异序位。在变异序位计算中必须遵循一定的邻接约束原则，一般限制在同一句子中，迁移距离不宜过远，最好不要跨越子句，以免造成混乱。同时，参与变异迁移的通常不仅仅是一个词项，时常包含同该词项具有语义依存关系的子树，以保持语义群的完整性。例如，给定句子“*你喜欢这些书中的哪一本*？”，在生成特殊疑问句时，将发生序位变异的不仅仅是特殊疑问词“*which*”，“*which*”所在的英语子树“*which of these books*”都应参与序位变异计算，前移至句首，形成译文“*which of these books do you like ?*”

3）序位内嵌计算

在生成中，有时需要将某个词项或句法子树强行嵌入到一些英语习惯搭配内部，这时需要重新计算被捆绑对象在该模式中的内嵌序位。例如，在汉语句子“*他们公开了这个专利*”中，与“公开”相关的英语对译词为“*make ...public*”，这是一个分离式习惯搭配，要求将“公开”的宾语内嵌到短语内部，形成“*make this patent public*”。因此，在英语生成中，必须根据“*make ...public*”的强制捆绑信息，对被捆绑对象“*patent*”及其子树加以嵌入标记，并计算“*patent*”及其子树在“*make ...public*”中的序位，这种内嵌序位的优先级比常规序位高，这样，在表层词串生成中，可形成“*make this patent public*，而不是常规序位（即SVO）“*make public this patent*”。

为了使英语生成器尽可能独立于源语言，我们将英语句法结构的规范化（缺省谓语成分、主语成分的补充等）、词汇整体属性改变（如词性的转换）、句型变换等与源语言中特殊句法现象紧密相关的部分放在转换器中处理。例如，给定句子“他高兴”.由于汉语中形容词可以做谓语，故在汉语句法结构（1）中，谓语成分“高兴”充当句法树的根。在汉语-英语转换中，在选择英语词汇的同时，对该结构进行了规范化处理，补充谓语动词“be”，形成等价的英语句法结构（2）。生成器则对规范后的英语句法结构（2）进行处理，最终形成译文“He is glad.”。（如图3所示）



（1）汉语句法结构　　(2）英语句法结构

图3　句法结构规范化示例

3.2 词汇层生成策略

作为曲折性语言，英语生成的词汇层处理主要借助组成句子的短语、词项及其线性顺序进一步给定句子的具体形式，每个词项的复杂特征集都含有该词项的基本静态属性（如词项的原形，中心词及词性，短语词性）、全部形态属性，在句中的线性位置等信息。基于语义特征的形态属性如名词的数、泛指/特指，动词的时、体、语态，形容词/副词的级等，可根据词项的语义及句法特征直接获取；基于句法条件约束的形态属性，如

名词的格，动词的人称、数等，需要依据词项间存在的支配关系、一致性句法约束，通过多次扩展、合一运算推出词项的曲折性标记。例如，在动词的表层形态特征获取中，通过对动词词项的时态、语态及句式等信息的合一运算，可获取动词词项的曲折性标记InflectedFlag。假设给定动词“*like*”,具有*vpast*(i.e. 过去时）、*vperf*(i.e. 完成式）词法特征，且句式特征为*declar*(i.e.陈述句式），则有

declar∧vpast∧vperf ==> InflectedFlag=ven(i.e. 过去分词形式）==> liked

这样，在表层词汇生成中，可根据词项的曲折性标记生成表层词形，构成表层英语词项序列。

由于英语的词序与源语言的词序并不相同，因此，在表层词汇生成中，还要依据英语文法习惯，重新定位相关标点，尽可能实现源语句的标点功能。在标点生成中，必须以语义群为单位重新计算标点的序位，才可能既保持原有的描述风格，又不破坏句子的语义整体性。例如，给定句子“秋天的晚上，....”. 时间短语“秋天的晚上”的对译短语为“*in the evening in autumn，_*”，如果仅根据词与词的对应关系将标点添加在对译词上，形成“*in the evening，_ in autumn*”，则分割了作为一个独立语义单元的“*in the evening in autumn*”的语义完整性，造成混乱。此外，在成对标点添加中，不仅要关注语义单元的完整性，还要计算标点的配对性。例如，对于短语“*_(喜欢音乐的女孩)_*”中的一对括号，在译文中的正确位置应为“*(the girl who likes music)_*”。此处，如果仅仅是在相应的对译词项上添加标点，则形成“*the girl)_ who_(likes music*”，破坏了译文的可读性。显然，标点的正确添加，不仅可增强译文的可读性，而且会对译文有一定的补充释义作用。因此，在表层词法生成中，我们根据词项间的句法依存关系和相对序位，找出标点的作用范围，计算标点的序位，并调整标点的组合，略掉冗余标点，以增强可读性。

4.小结

本文给出了一个面向汉日韩-英多语机译系统的通用英语生成器，描述了英语生成中的词项序位常规计算、序位变异计算、序位内嵌计算等几类基本操作。目前，我们的通用英语生成器已有近四百条英语生成规则，覆盖了基本的英语语法现象。该通用英语生成器已应用于汉英、日英、韩英机译系统，效果较为理想。通用英语生成器的研制，缩短了开发周期，降低了成本，是多语机译系统研制开发的一次有益的尝试。

参考文献

[1] Bonnie J. Dorr, “Interlingual Machine Translation: a Parameterized Approach”, Artificial Intelligence, Vol. 63, 1993.

[2] Michael C. McCord, “ SLF: The Slot Grammar Lexical Formalism”

[3]姚天顺等，《自然语言理解---一种让机器懂得人类语言的研究》，清华大学出版社，1995.