日汉对话翻译系统中的汉语生成

王锡江　王启祥　陈家骏　雍殿书

（南京大学计算机科学与技术系）

摘　要

本文介绍日汉对话翻译系统中的汉语生成的设计思想、系统实现中的若干技术要素以及构筑人机对话界面的环境、工具和需考虑的一些因素。

一. 前言

机器翻译的研究已经历了第一代基于词典直接转换法、第二代基于句法、第三代基于句法和语义以及第四代基于人工智能技术的一个发展过程。从这个发展过程看，为了追求机器翻译的全自动和高质量，系统需要加入更多的语义信息，语用以及语境信息；对于篇章翻译，还需要相应的上下文信息。但由于目前语言学理论的缺限和人工智能技术的限制等，使得构筑能够实用的基于知识的机译系统非常困难。

这样，迫使实用机译系统的研究朝下列方向发展：

一方面，在系统的实现目标上，开始限定其翻译领域，构筑限定翻译领域的专业机译词典，以面向要求不高的大量专业技术文献。在系统的实现技术上，加强对源语言的预处理和目标语言的后处理研究，开发适应于机译系统的编辑器。将源语言的某些语言上的难点，使用前编辑的方法将之分化，降低系统的复杂度；对于目标语言，同样也使用后编辑的方法，以提高译文的质量。在这样的系统中，都相应增加了人对系统的干予，降低系统全自动的程度，提高系统的质量。

另一方面，机译系统开始向高性能的小型机、大型机以及有着良好人机界面和系统性能的工作站上发展，求取计算机系统中的更多空间和资源，提高机译系统的效率。

在这样的一个背景下，我们在SUN 工作站上实现了一个基于对话方式的日汉机译系统，使用 X 窗口系统构筑人机界面，使系统真正朝实用化发展。

但本文仅介绍该系统中汉语生成的设计思想、系统实现中的若干技术要素，以及构筑人机界面接口的环境、工具和需考虑的一些重要因素。

二. 系统的设计原则和目标

本系统的设计目标是力求系统的简单和实用，因此在整个系统中，增加相应的对话模块，以解决语言中某些歧义现象和目前系统不能自动处理的问题，本系统主要依靠分析结果的句法信息和部分表层信息，仅添加少量的语义信息。

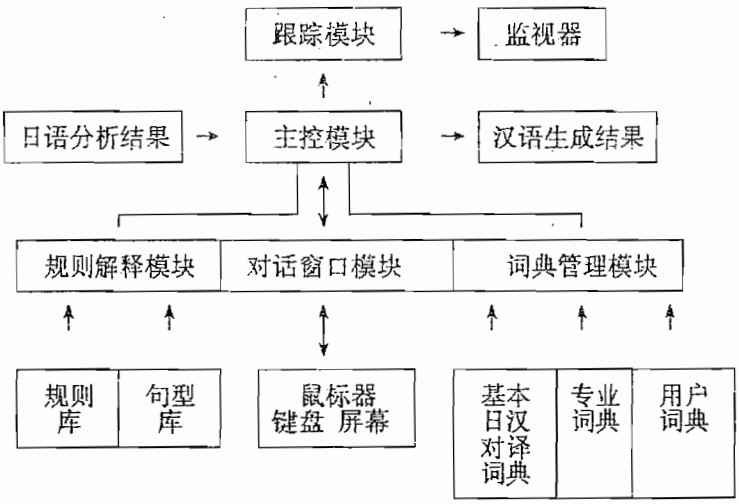
本系统的设计原则如下：

(1) 尽量使用规则和相应的语言知识解决生成要素；
(2) 不能解决的语言歧义现象采用对话方法解决；
(3) 系统以生成规则驱动为主，而人机对话为辅；
(4) 对于对话翻译方式，重视人机界面的易使用性；
(5) 系统为开放式系统，具有重组功能和可扩充性。

在这样的一个系统中，要研究系统功能的分担，弄清楚哪些语言现象由系统自动处理，哪些语言现象由对话解决。原则上，系统中不能自动处理的语言现象，均由对话处理。但在具体的实现中，对话模块主要解决具有歧义的语言现象，例如：多义译词的选择和生成，多义格关系的选择和生成以及多义态的处理等。

三. 系统的逻辑结构

本系统主要由下列主要模块构成：主控模块，负责调度和协调整个系统的工作；规则解释模块，这实际上是一个简单的规则解释器，负责对规则和句型库进行操作，驱动相应的生成模块；对话窗口模块，这提供了一个人机接口界面，让用户可以直接干予和指导系统作相应的生成操作，使用X 窗口系统实现，可以进行多窗口的动态生成和调度；生成词典，包括基本对译双语词典，专业词典和用户词典；跟踪模块，负责监控整个系统的运行状态。系统的逻辑结构见图一。



图一. 对话翻译汉语生成的逻辑结构

四. 生成词典的结构与设计

在本系统中，生成译词的共性知识放在重写规则中，而生成译词的个性知识则放在生成词典中，包括相应的个性生成规则。这样，生成词典不仅为系统提供静态个性信息，为规则的运算提供参数，而且还能为系统提供丰富的动态知识。

将部分个性规则放到生成词典中，具有下列特点：

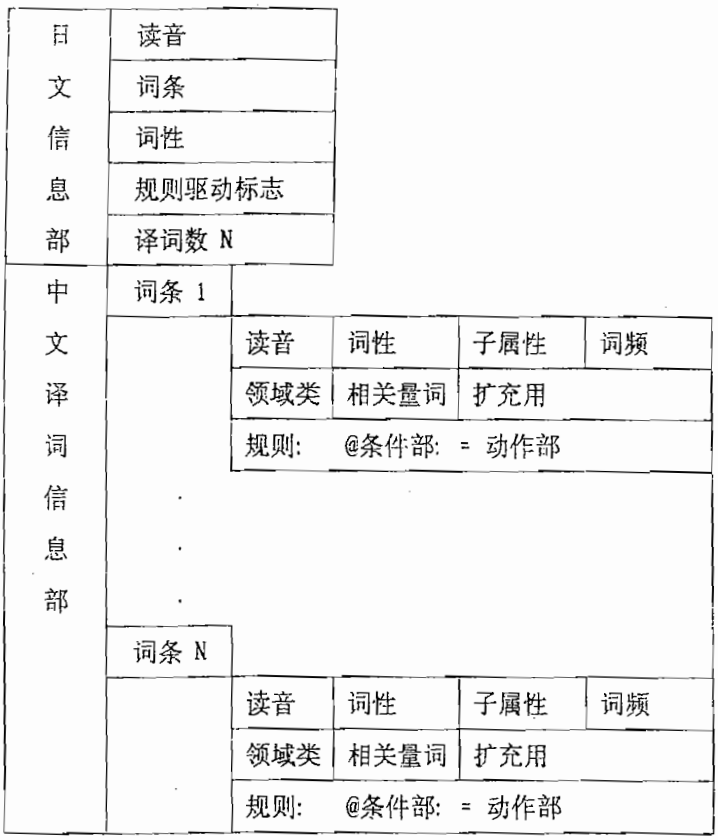
(1) 使词的具体个性静态信息与规则的结合更加灵活、准确，使每个词的信息利用更加充分；

(2) 由词制导规则，保证了规则选择的正确性，避免规则误导现象；

(3) 在规则中可以少出现甚至不出现具体参数，从而提高规则的抽象化程度；另一方面，同一规则不同的词可以选择不同的参数与之结合，从而提高了规则与参数结合的灵活性；

(4) 当前词的所有信息均可以成为元规则的隐含信息。

生成词典的逻辑结构见图二。



图二. 生成词典的逻辑结构

五. 分析与生成的界面

在本系统中，日语分析的策略是以句子中的文节为中心，利用日语格助词的表层信息，将句子分析成为一棵文节关系树。目前，定义四种集合用以表示分析结果，供生成使用:

文件信息集合A

A = {f1, s1, s2, ..., sn}

其中: A ∋ f1: 文件信息长度，且为整型

A ∋ si: 句子信息结构，且i=1, 2, ..., n

句子信息集合B

B = {sl, pn, phi, ..., phm, nl}

其中: B ∋ sl: 句子信息长度，且为整型

B ∋ pn: 句子中文节数，且为整型

B ∋ phi: 文节信息结构，且i=1, 2, ..., m

B ∋ nl: 句子信息结构终止符，且为字符型

文节信息集合C

C = {pl, wn, wdi, ..., wdm, mod, pt, ps}

其中: C ∋ pl: 文节信息长度，且为整型
C ∋ wn: 文节中单词数，且为整型
C ∋ wdi: 单词信息结构，且i=1, 2, ..., m
C ∋ mod: 文节修饰关系，且为整型
C ∋ pt: 文节类型，且为整型
C ∋ ps: 文节属性，且为整型

单词信息集合D

D = {wl, wd, cat, subcat, var, ext}

其中: D ∋ wl: 单词信息长度，且为整型
D ∋ wd: 单词词条，且为文字型
D ∋ cat: 单词词性，且为整型
D ∋ subcat: 单词子属性，且为整型
D ∋ var: 用言变形信息，且为整型
D ∋ ext: 扩充用，且为整型

四种集合具有下列关系:

$A \supset B \supset C \supset D$

设给定一个分析论域 P，则 $p \in P$:

由 $p \in D \Rightarrow p \in C \Rightarrow p \in B \Rightarrow p \in A$

则 $A \supset B \supset C \supset D$

再设定一个生成论域G，则分析与生成构成下列关系:

$P \xrightarrow{R} G$，即 $R \subset P \times G$

其中: R为生成映射关系子集

在本系统中，日语文节的类型暂时定义为17种类型，这17种文节的类型基本上都和相应日语格助词的表层信息相对应。从语义上讲，很多类型都具有多种深层格关系的语义属性，我们使用对话窗口，由用户参照相应的提示信息，引导用户进行唯一性的选择。

六. 对话应该处理的内容以及对话窗口的设计

原则上，系统方面不能处理的语言现象，均交给对话模块完成。但在具体的系统实现上，应尽量压缩和减少对话模块所处理的内容，以提高系统的效率。

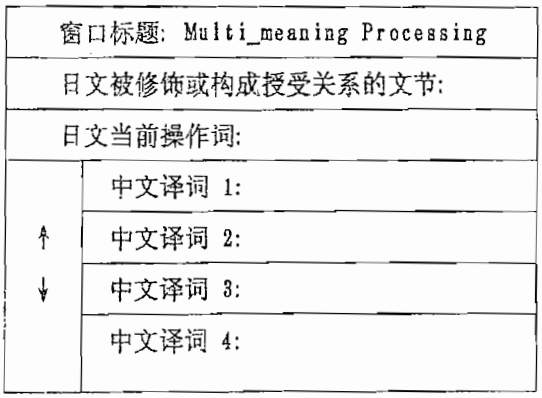
根据上述观点，本系统中由对话处理的生成要素如下:

(1) 多义译词的选择和生成;
(2) 多义格关系的选择和生成;
(3) 多义态的处理;
(4) 量词的决定;
(5) 生成句型的选择;

这里多义译词包括同音异义词，同字异义词以及同音同字的多义词。

在我们的系统中，对话窗口根据处理的具体内容而设计成各种不同风格的窗口，以方便用户使用和操作。不同的对话窗口在系统的运行过程中，可根据需要而动态生成、移动、放大和缩小以及关闭。

例如，多义译词的对话窗口格式见图三。



图三. 多义译词的对话窗口

在该窗口中，用户可使用鼠标器对照当前日文词与句中所修饰或构成授受关系的文节，进行中文译词的选择。若中文译词数大于窗口能显示的条数，则用户可使用鼠标器操作窗口中的滚动条上滚或下滚译词信息。

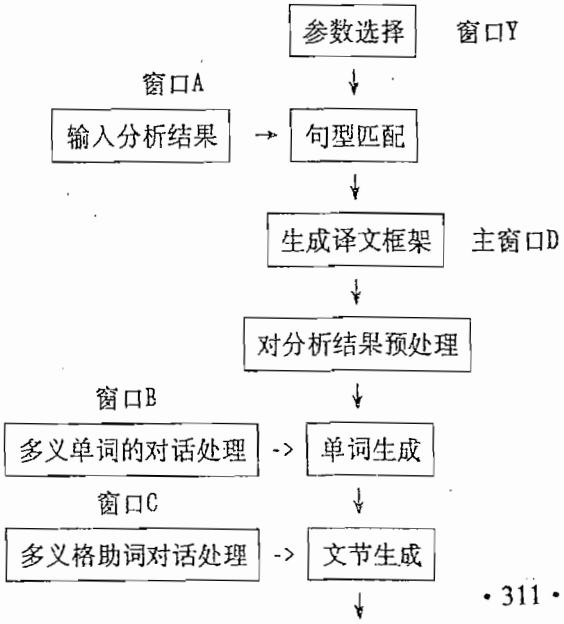
七. 系统生成的步骤和层级

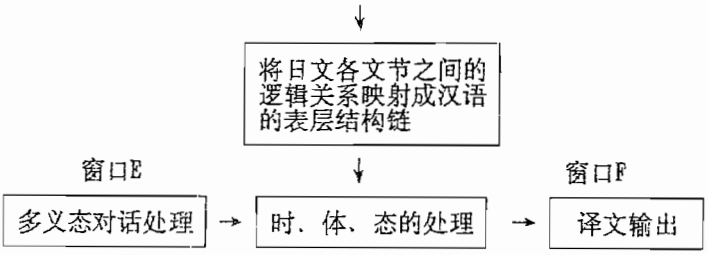
在本系统中，将分析结果作多趟扫描，作不同的生成处理，以适应不同的生成层级。

(1) 系统中的主要生成层级

第一趟扫描: 处理固定搭配，生成相应框架结构;
第二趟扫描: 对分析结果作预处理，变换分析结构;
第三趟扫描: 查对译词典，作单词生成处理;
第四趟扫描: 进行文节生成，映射到汉语的相应成分表示，并确定文节的生成位置;
第五趟扫描: 处理从句及时、体、态，完成句子表层信息的转换。

(2) 系统中的主要生成步骤





八. 生成中的几个技术要素

(1) 短语中词的定序

短语中生成词的定序主要依据修饰和被修饰关系而决定生成词在短语中的位置。对于多个定语修饰同一个词的情况下，我们给出下列的修饰顺序：领属名/代词，时间或处所词，数量词组或指示代词，主谓词组或谓宾词组或介词词组，和不带'的'的名词或形容词。

(2) 短语的定序

在我们的系统中，由用户在对话窗口中，决定多义格短语的唯一性，然而将日语的文节关系映射成汉语的主、谓、宾、补、状等成份。我们设计了一个汉语句子最大逻辑结构链，以生成汉语短语的位置。对于多个状语修饰同一个动词的情况下，我们给出下列的修饰顺序：情态状语、时间状语、地点状语、方式状语、方向状语、对象状语和描写状语。

(3) 时、体、态的生成

时、体、态的生成主要依据分析给出的附加信息和相应的表层信息。

九、结语

由于本系统是一个基于对话方式的日汉机译系统，因此在全自动翻译系统中不容易处理的语言现象，在该系统中都可以通过对话解决。为了减轻用户在对话翻译中的负担，系统还必须提供相应的支持环境或模块，帮助用户进行对话翻译的操作。例如，在多义词处理的对话操作中，系统还同时在另外一个窗口中给出分析结果的树形结构图，帮助用户理解句子中各种成份之间的关系，这相当于给用户一个句中语境的提示，使用户很容易选择正确的译词信息。

对于对话翻译系统，选择适当的构筑和运行环境非常重要，系统的运行速度和多窗口用户界面都是不可缺少条件。系统的简化和系统的效率要均衡考虑。

该系统已初具规模，还正在开发和扩充之中。

参考文献:

[1]. 李裕德，科技汉语语法，冶金工业出版社，1985

[2]. 青山升一等，对话翻译の一方式について，NLC 90-14

[3]. 王锡江等，生成自然语言的方法，《情报科学》，1988，Vol.9，No.4

[4]. 王启祥，王锡江等，从格关系生成中文文本，《日本自然语言处理》，1990，Vol.80，No.6

[5]. 王启祥，王锡江等，日汉机译系统中的汉语生成，《机器翻译研究进展》，电子工业出版社，Augest. 1992

Chinese Generation in Japanese/Chinese Dialog Translation System

Wang Xijiang, Wang Qixiang, Chen Jiajun, Yong Diansu

(Dept. of Computer Science and Technology, Univ. of Nanjing)

Abstract

The idea of Chinese generation design, some technical factors of system implementation, and the environment, tools and other factors to construct the man-machine dialog interface in our Japanese/Chinese dialog translation system are given in this paper.