汉语篇章生成系统中的用户模型和文本规划*

吴 华　　黄泰翼
中国科学院自动化所模式识别国家重点实验室 北京 100080

摘要：文本规划是篇章生成的一个不可或缺的组成部分，它的主要作用是确定所要生成的内容以及生成内容之间的逻辑关系，而规划的内容又受到用户模型的影响。因此，本文首先建立了用户模型，并根据用户模型采用了 Schema 方法和 Process 方法相结合的混合文本规划策略，它和句法实现系统一起用于作者设计的一个花卉知识查询实验系统，实现篇章的生成。实验结果表明：输出的篇章语法正确，文字通顺，逻辑严密，可以满足用户的要求；同时也表明，根据用户模型采用相应的文本规划策略，能够使生成的内容既满足用户需要又不累赘，大大地提高了生成内容的质量。

关键词：用户模型 文本规划 汉语生成

User Modeling and Text Planning in a Chinese Text Generation System

WU Hua　HUANG Taiyi
National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Science, Beijing 100080, China

ABSTRACT: Text planning is a necessary part if a generation system. The main task of text planner is to determine the contents to be generated and to decide the logical relations among sentences, which is affected by the user model. This paper first built a user model and then used the hybrid planning method: Schema and Process according to the user model. This text planner and the Chinese syntactic realization system were successfully used in our flower knowledge retrieval system. Experiments showed that the generated texts were grammatically correct, coherent, logical and therefore able to satisfy the needs of users, and also showed that using suitable planning methods according to the user model could make the generated contents understandable and formative to users and therefore improved the generation quality.

Keywords: user modeling, text planning, Chinese generation

1. 引言

自然语言生成是七十年代开始发展并逐渐活跃起来的领域，是自然语言处理的一个重要分枝，它研究怎样利用计算机从某种中间形式（如符号系统、数字系统）生成自然语言的过程。一个完整的自然语言生成系统通常由三部分组成：文本规划、句子规划和句法实现〖1〗。其中文本规划确定所生成的内容和结构；句子规划解决句间合并、成分省略和指代等问题；而句法实现利用语法规则和词典等知识把句子规划的结果转化为语法正确的文本。本文只涉及到文本规划，因为文本规划是确定生成内容的关键，它直接受用户模型的影响。目前文本规划的方法主要有两种：Schema 方法和修辞结构理论（Rhetorical Structure Theory, 简称 RST）〖2〗。

用户模型一直是人们关心的问题，以前的工作主要集中在推断用户目的、用户态度等方面〖3〗。从七十年代末期开始，出现了一些探讨用户模型在问答系统中的作用的论著，但主要探讨用户模型对指代、生成内容的多少的影响；1988 年 Paris 开始注意到用户对领

*本论文得到了国家 863 高技术研究项目(863-306-ZT03-02-2)和国家自然科学基金重点项目(69835030)的资助。吴华已转入微软中国研究院工作，有关文章的具体讨论请以下列方式联系：i-hwu@microsoft.com

域知识的掌握程度不但影响生成内容的多寡，而且影响生成风格，但他主要针对个人建模，没考虑群体建模〖4〗。本文不但讨论个人建模，而且讨论群体建模。

在问答生成系统中，用户对知识的了解是至关重要的。它关系到生成的内容是否能被用户理解，而又不累赘，即不生成用户已知的内容。本文通过对成人和儿童百科全书分析，发现成人百科全书的内容大多是说明性，而儿童百科全书的内容是过程性的，它对整个过程进行详尽的描述。这说明对于用户比较理解的概念和过程，无需进行过程描述。相反，就必须对概念以及各种概念之间的联系进行详尽的描述。根据这种观点，本文采用两种生成策略：对于用户了解的概念或过程采用 McKeown 的 Schema 方法，而对用户不熟悉的过程采用过程描述。在一段生成内容中，用户可能同时有了解和不了解的概念，我们就把这两种方法结合起来混合使用。本文结合花卉查询系统，对两种方法进行了详尽的描述。

2. 用户模型

一个用户模型包括四个方面的内容：用户目的、用户可能性、用户态度和用户知识〖5〗。用户目的表示用户查询系统的目的，一般可根据用户的问题推断出来；用户可能性是指用户完成某一过程的现实可能性或理解某一概念或过程的可能性；用户态度是指用户的主观意向；用户知识包括用户的领域知识以及与此相关的通用知识。并不是所有的用户模型都必须包括这四个方面的内容，用户模型的内容与具体的应用领域相关。本文认为用户的目的是了解更多的知识，不考虑用户的主观态度，并一致肯定用户完成某一动作的可能性，而用户理解可能性和用户知识密切相关。因此，这里的用户模型主要建立用户的知识模型。

用户模型的另一方面是用户模型的性质，主要包括模型的特定性、动态性和时效性。特定性表示模型是针对一类人还是单独的一个人；动态性是指模型参数是否随时变更；而时效性与模型的动态性密切相关，是指模型是否随着对话的进行变更。一个静态的模型必定是长期的，而一个动态的模型也一般是短期的，随时间变更的。

本文讨论的生成方法的实验基础是花卉查询系统。用户和系统的交互是通过对话框进行的，用户可以查询十种名花的形态特征、繁殖方法以及每种花的介绍（包括其科属，类别，栽培历史以及应用等）。用户的问题以及用户模型都可以通过对话框来获得，生成系统根据用户模型搜索知识库，选择恰当的内容回答问题。用户模型的获取是在生成过程开始之前进行的，因此，我们的用户模型是静态的、长期的。如果用户想了解花的形态特征和每种花的介绍时，我们采取群体建模，通过对用户职业、受教育程度、对花的结构了解程度的询问，系统可以确定用户类型：行家还是门外汉。而当用户询问关于花卉的繁殖方法时，我们采取个人建模，因为繁殖方法多，每个人的知识相差很大。因此，我们的用户模型是一个静态的、长期的、个人与群体相结合的模型。

3. 生成策略

在用户模型中，如果用户表示了解某种物体，我们就认为用户了解其基本结构以及基本原理，如用户表明他了解插扦繁殖方法，那么我们认为他知道插扦原理、插扦类属和一般插扦过程，此时我们采用 McKeown 的 Schema 方法〖6〗；相反，采用过程描述方法(Process 方法)。

3.1 Schema 方法

McKeown 认为，句子某些文本结构是有规律可循的，这些结构就表示成 Schema，每个 Schema 由许多修饰谓词（Predicate）组成，每个谓词可能对应多个实现方法，由焦点控制算法〖7〗选择其中的一个，并控制句子之间的连贯性。Schema 是可以循环嵌套的，理论上，每个谓词都可以扩充为对应的 Schema。本文利用最多的 Schema 是 Constituency，其原始定义如图 1 所示。其中“{}”表示括号内的内容可选，“/”表示有多个选项，“+”表示前面的修饰谓词可以出现 1~n 次，“*”表示前面的谓词可选，并可出现 0~n 次。

```
Constituency
Cause-effect* / Attribute* /
{Depth-identification / Depth-attribute
  {Particular-illustration / evidence}
    {Comparison / analogy }}+
{ Amplification / Explanation / Attribute / Analogy}
```

图 1　Constituency 方法

为了使 Constituency 方法更适合花卉的描述，我们把它改写为如图 2 所示。Constituency 方法主要用来描述具有子类或需要描述其各组成部分的物体。比如描述花卉形态特征时，我们需要描述各个组成部分，而且必须对其类属进行说明，所以我们采用图 2 的方法。其中 Identification 是对所描述的物体进行定义；Attribute 是对物体的属性进行描述，Renaming 是对物体的别名等进行说明；Constituency 是就物体的组成部分或子类进一步对物体进行阐述；Cause-effect 是事物之间的因果关系；Depth-attribute 是指对物体组成部分的属性进行描述；Amplification 是对物体的属性进一步进行阐述；Analogy 是通过类比的方法阐述物体的特点，使用户更容易理解生成内容。

```
Identification*
Attribute*/ Renaming *
Constituency
Cause-effect* / Attribute* /
{Depth-identification}+
{ Amplification / Attribute / Analogy}
```

图 2　修改后的 Constituency 方法

在实际应用中，一个 Schema 有三个部分：名称、应用条件、谓词。每个谓词都带有五种信息：一是语义类型，这些语义类型与具体的知识库对应，生成系统根据这些语义类型从知识库中提取内容；二是可扩展性，表示谓词是否可以扩展成同名字的 Schema 或新的 Schema；三是优先级，确定同时有多个谓词可选时的选择顺序，数量越小，优先级越高；四是可选性；五是表示谓词在 Schema 中的顺序。其描述方法如图 3 所示。其中“[ ]”表示一个谓词所包括的内容，“+”表示一个 Schema 可以由多个不同的谓词组成。

```
SCHEMA BEGIN                    // 标志 Schema 开始
    SchemaName                  // Schema 的名称
    Condition                   // 应用条件
    [ PredicateName             // 谓词名称
      SemType                   // 语义类型
      Extension                 // 表明谓词是否能扩展
      Priority                  // 谓词优先级
      Option                    // 谓词的可选性
      Order ]+                  // 谓词在 Schema 中的位置
```

SCHEMA END

图 3　Schema 描述

3.2 Process 方法

对于用户不知道的概念或过程，系统采用过程描述方法。例如，在描述花卉的栽培繁殖方法时，如果用户不清楚花卉繁殖方法，那么系统就向用户阐述花卉插扦原理和插扦过程，此时采用过程描述方法(Process 方法)。其具体过程如图 4 所示。

（1） 对于每个物体或每个过程，重复步骤(2)~(4)；
（2） 根据知识库中的关系链，依据事物之间的因果关系、时间关系或空间关系选择生成内容；
（3） 如果步骤(2)中涉及新的过程，转(1)进行循环操作，否则转(4)；
（4） 重复(2)中的直到遍历完所有关系链。

图 4　过程描述方法

3.3 Schema 方法与 Process 方法的结合

在实际应用中，用户可能只了解一个物体的某一部分知识，我们对已了解的知识采用 Schema 方法，而对那些用户不了解的概念采用 Process 方法。在描述同一个物体时，同时采用两种生成策略，下面详细介绍两种方法的结合以及相互之间的转换。

在 Schema 方法中，如果我们涉及到新的部分是用户不了解的，为了使用户能清楚地理解这个概念或物体，系统必须向用户简述与此概念相关的原理、方法等，这时就必须切换到 Process 方法，现以 Constituency 为例说明，如图 5 所示。

Identification（如果所描述的概念有上位类，且用户没有关于它的知识，则转入过程描述）
Attribute
Constituency (对于物体的每一个组成部分，如果用户没有关于它的知 识，则转入过程描述)
……

图 5　Constituency 方法中的转换点

同样，在过程描述中，如果涉及到描述过程或物体的属性时，生成策略转换为 Schema 方法。其切换点如图 6 所示。

（1） 对于每个物体或每个过程，重复步骤(2)~(4)；
（2） 根据知识库中的关系链，依据事物之间的因果关系、时间关系或空间关系选择生成内容；如果生成内容中需要详尽描述物体的属性时，转入 **Schema** 方法，否则转(3)；
（3） 如果步骤(2)中涉及新的过程，转(1)进行循环操作，否则转(4)；
（4） 重复(2)中的直到遍历完所有关系链。

图 6　Process 方法中的切换点

4. 生成策略的实现

Schema 是采用扩充转移网络（Argumented Transition Network，简称为 ATN）实现的，代表 Schema 的 ATN 是一张有向图，它的结点代表起始状态、填充过程的中间状态或终点

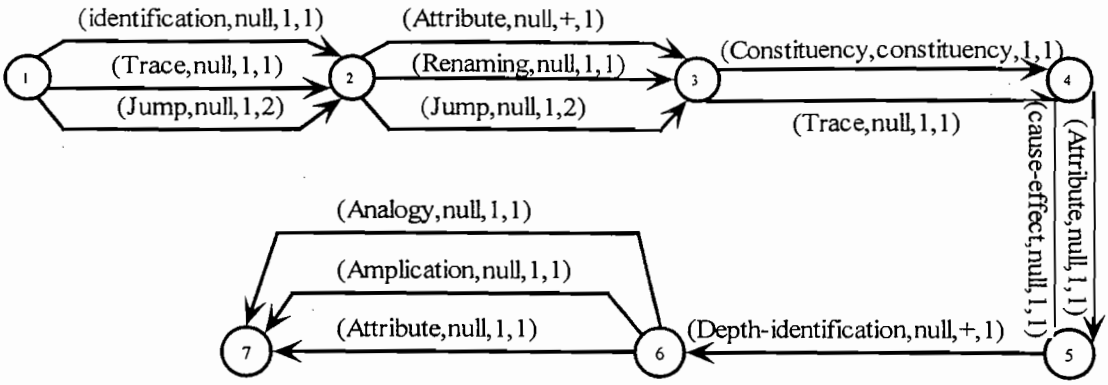

图 7　Constituency 的所对应的 ATN

状态，Schema 的递归嵌套相当于 ATN 的递归嵌套。ATN 弧共有三种：一是 Schema/predicate 弧，代表 Schema 中的一个符号；二是 Jump 弧，代表状态之间的无条件转移，即转移时无需匹配任何 Schema 或 Predicate；三是 Process 弧，表明切换至过程描述。每条弧由一个四元组表示：（弧名，扩展性，重复特性，优先数），弧名可以是以上三种的任意一种，重复特性代表此弧的重复次数，优先级是指弧上有多个 predicate 实现时，Predicate 之间的优先级，数越小，优先级越高。此表达方法于〖7〗相同，Constituency 的实现如图 7 所示。

Process 方法也是用 ATN 实现的，在 Process 方法中，弧也有三种：一是 Link 弧，代表 Process 中的一条关系链；二是 Jump 弧，与 Schema 中的 Jump 弧定义一样；三是 Schema 弧，表示切换至 Schema 方法中。每条弧也是由一个四元组表示，如图 8 所示：

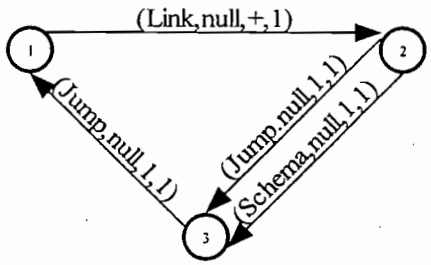

图 8　Process 方法的 ATN 实现

5. 实验系统设计

实验系统的应用领域是花卉知识查询，可查询范围为国内十大名花的形态特征、繁殖方法和简单介绍。实验系统的总体框图如图 9 所示，它由文本规划、句子规划和句法实现三部分组成〖8〗。系统首先通过对话框获得用户模型和交际目标，然后根据用户模型和交际目标选择文本规划方法：Schema 方法、Process 方法或者两者的结合，并根据文本规划的填充算法和焦点控制算法从知识库中选择相应的知识填充 ATN 中的状态，完成文本规划过程；然后进入句子规划进行句法合并或句子成分省略的操作，使规划的文本更加

简洁和通顺；句子规划的输出送入第三章建立的汉语句法实现器，最后生成汉语文本。

现举例说明整个生成过程，假设用户想了解梅花的繁殖方法，但对繁殖方法的知识很少，系统向用户详细阐述繁殖原理、过程等等。根据用户模型，在 Schema 中搜索匹配的 Schema，发现 Constituency 匹配。其整个过程如下：

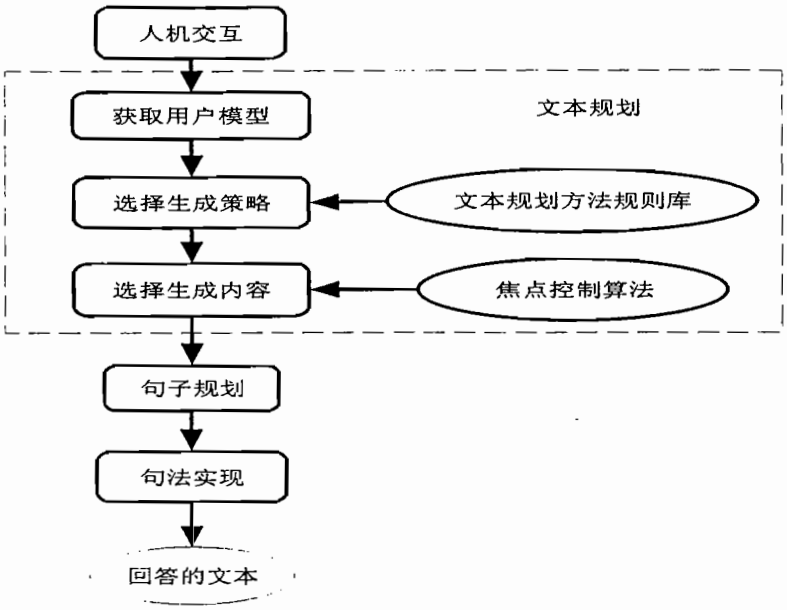

图 9 实验系统总体框图

Constituency:

Predicate: REPRODUC: 插扦&嫁接&压条&播种

*对 **Constituency** 进行扩展，分别讨论四种繁殖方法，现以插扦繁殖方法为例*

Identification:

Predicate: SUPERCLASS: 无性繁殖

Attribute:

Predicate: CALL: 插穗(used-for: 插扦的枝条）

*转入 **Process** 方法中对插扦原理进行描述*

Link 1: 有(植物的一部分,再生作用)

Link 2 : 有(支配植物再生作用的物质,两种)

Substep 1:

Cause Link: material: 创伤激素

Cause Link: 产生(原生质(mod:死细胞&伤细胞(mod: 营养器官被切伤部分)),创伤激素)

CauseLink: 促使(,细胞,分裂, result:产生(,新组织))

Substep2:

Cause Link: material: 内源激素

Cause Link: 产生(植物体内芽及叶柄,内源激素)

Cause Link: 促进(内源激素,再生力(mod:植物))

*插扦原理叙述完毕，返回 **Constituency** 方法继续对梅花插扦方法继续描述：*

Depth-Attribute

Predicate: TIME: 11 月

Depth-Attribute

Predicate: SUI: 一年生枝条(length: 10-15 厘米)
Depth-Attribute
Predicate: SOAK: 浸泡(,插条,material:吲哚丁酸,duration:5-10 秒)

知识库中有关梅花插扦繁殖的知识已遍历，梅花插扦过程结束，接着继续描述梅花的其它繁殖方法，其过程与上述方法类似。文本规划的结果再经过句子规划和语法实现，得到最终的文本〖8〗。其生成结果如下：（只包括插扦部分）

插扦、嫁接、压条和播种是梅花常用的繁殖方法。插扦繁殖属于无性繁殖，用于插扦的枝条称为插穗，其成活原理是：植物的一部分有再生作用，支配植物再生作用的物质有两种，一是创伤激素，营养器官被切伤部分的死细胞和伤细胞原生质产生创伤激素，促使细胞分裂产生新组织，二是内源激素，植物体内芽及叶柄产生内源激素，内源激素促进植物再生力。梅花插扦时间为 11 月，选 10-15 厘米的一年生枝条作插穗，用吲哚丁酸浸泡插条 5-10 秒，然后插扦。

从上面生成过程可知，Schema 方法和 Process 方法在生成过程的任意地方都可以切换，对用户了解的概念或过程，我们采用 Schema 方法，而对用户不了解的过程采用 Process 方法。两种方法的结合使生成内容更能满足用户的需要，提高了生成质量。

6. 结论

本文讨论了汉语篇章生成系统中的文本规划以及用户模型，实验结果表明根据用户模型选择相应的文本规划方法是提高生成质量的有效途。依据建立的用户模型，Schema 方法和 Process 方法可以在生成过程中的任意必要的地方切换，对用户了解的概念或过程，我们采用 Schema 方法，而对用户不了解的过程采用 Process 方法，因此，用户模型不但影响生成内容的多少，而且决定采用何种文本规划方法。如果没有建立用户模型，篇章生成系统就必须把知识库中所有的内容全部生成出来，不能做到因人而异。有了用户模型，对同一个问题的回答是多样的，而且能够满足用户的需要，做到不生成用户已知的内容，又能给用户提供有用的信息，从而提高生成内容的质量。与不采用用户模型的篇章生成系统相比，我们的系统更好地满足了用户的需要。尽管本文的实验领域是针对具体领域的，但生成策略是与领域无关的。

参考文献

[1] Cole A et al. Survey of the State of the Art in Human Language Technology. http://www.coli.uni-sb.de/~hansu, the Web Page for the Department of Computational Linguistics at the University of the Saarland, 1996

[2] Johanna D. Moore, Cecile L. Paris. Planning Text for Advisory Dialogues: Capturing Intentional and Rhetorical Information. Computational Linguistics, 1993, 19(4):651~693

[3] Sandra Carberry. Modeling the User's Plans and Goals. Computational Linguistics, 1988, 14(3):23~37

[4] Cecile L. Paris. Tailing Object Descriptions to a User's Level of Expertise. Computational Linguistics, 1988, 14(3): 64~78

[5] Robert Kass and Tim Finin. Modeling the User in Natural Language System. Computational Linguistics, 1988, 14(3): 5~22

[6] Kathleen R. McKeown. Discourse Strategies for Generating Natural-Language Text. Artificial Intelligence, 1985, 27: 1-41

[7] 姚天昉，汤学彦. 文本规划中焦点移动控制算法的研究. 软件学报，2000, 11(2):277~284

[8] 吴华. 汉语自然语言生成的理论、方法的研究及系统实现. 博士学位论文，中国科学院自动化研究所，2001.3