汉语天气预报文本内容规划器的设计与实现*

王纤　　姚天昉
（上海交通大学计算机科学与工程系　上海 200030）

摘要：本文论述了汉语天气预报生成系统中内容规划器的设计与实现．根据天气预报文本的特殊结构，笔者采用了 SCHEMA 技术，并对 SCHEMA 的操作符进行了扩充，引入 "‖" 操作符号，使谓词的排列顺序能够根据相应的条件而改变．为了准确地表达汉语语义，笔者又对修辞谓词结构做了相应的修改，在谓词参数中加入了语义格成份．

关键词：汉语文本自动生成，SCHEMA，文本宏观规划

DESIGN AND IMPLEMENTATION OF MACROPLANNER FOR CHINESE WEATHER FORECAST AUTOMATED GENERATION SYSTEM

Wang Qian, Yao TianFang
(Department of Computer Science and Engineering of Shanghai Jiaotong University)

ABSTRACT: In this paper, we present a method of design and implementation for macroplanner of text generation used to generate Chinese weather forecast. The macroplanning approach is based on schema technology. We expand the operation of schema based on the special structure of weather forecast, introducing the "||" operation which indicates the order of predicates can be changed according to appropriate conditions. We modify the structure of rhetorical predicates and replenish semantic case as a data field of the predicate argument in order to express Chinese semantics exactly.

Key Words: Chinese text automated generation, schema, macroplanning

一、引言

天气预报是一种十分常见的书面文本，它对人们日常生活有很大的帮助．目前常见的预报文本是气象预报员根据计算机计算所得出的数据，解释成为天气预报．但是，用人工方法生成预报文本的速度很慢，不能满足人们的实时需要．一方面，随着气象和通讯技术的迅速发展，每时每刻都有大量的原始气象信息需要预报员处理．另一方面，不同的用户需要各种不同的，复杂的天气预报文本．气象员也必须尽快的生成大量的预报文本．因此，天气预报文本的生成必须采用自动化的方式进行．为了提高预报的效率，人们开始使用NLG（Natural language Generation）技术，让计算机来自动生成天气预报文本．

传统的生成理论将自然语言生成系统分成两个主要的阶段：内容决定阶段，表层形式生成阶段．随着近年来 NLG 的研究发展，为了生成高质量的文本，在内容决定阶段与表层生成阶段之间加入了一个新的阶段，称为句子规划阶段或微观规划．这阶段处理从内容到语言的映射以及语言表达的优化问题〖1〗．本文主要讨论生成的第一阶段－－内容规划阶段．在大部分应用系统中，生成系统的第一阶段都是选择和组织要被生成的信息．这一阶段可称为文本规划（TEXT PLANNING）．主要的手段是基于 SCHEMA 的规划方法〖2〗和基于修辞结构理论（RST，Rhetorical Structure Theory)的方法〖3〗．在汉语天

*本文得到国家自然科学基金（项目编号：69673008）、德国大众基金、上海市科技发展基金（项目编号：962907002）的资助．

气预报自动生成系统中，内容规划器采用的是基于 SCHEMA 的技术。同时系统根据天气预报文本和汉语的特点，对 SCHEMA 的操作符进行了一定的扩展〖4〗〖5〗。

二、文本篇章结构与 SCHEMA 技术

2.1 用 SCHEMA 来描述文本篇章结构

在设计内容规划器时，通过比较多种生成技术，结合天气预报文本的结构特点，笔者认为用SCHEMA技术是比较合适的。首先分析了气象台的多种类型的预报文本，然后又分析了新民晚报的天气预报文本。总结了这些文本结构上的特征和预报术语。目前的天气预报有几种类型，例如：天气预报的广播稿、海洋的天气预报等。这些天气报告中每一段的主题都是确定的。用词的范围也是确定的。但是有些预报文本包含了一些图表和字符串。对于相同的内容，简单与复杂的文本之间的差别是很大的。针对上述现象，决定选择日常预报文本进行分析。例如最常见的单点（如某城市）预报文本，可以看出，一般的预报文本的头部都是由--些主要的预报内容构成，如山东有大风，就要在头部出现“山东沿海海面大风警报”这样的语句。头部还应该包含发布预报的气象台名，时间等信息。从整体的结构可以看出，整篇报告是以一个标题头部开始说明该天气预报是在何时、何地、哪个气象台发布的何种类型的天气预报，接下来是根据当天的天气情况，决定应该描述什么主要的天气情况，最后是报告各地区的具体情况。在分析预报文本的过程中不但发现预报有结构的特征规律，而且还发现一些用词的规律，例如：气象台一天发布四次预报，每次对于不同的时间段的名称是不同的。在这里就不详细列出。此外，从文本生成技术来看，大部分的实际应用系统都是采用SCHEMA技术来进行宏观规划，这说明该技术是比较成熟的技术。从实现的角度看，与其它的宏观规划技术比较，SCHEMA技术比较容易设计和实现。

从上可以得出这样的结论：天气预报文本的结构具有一定的模式。但内部具体的细节又有一定的变化。这种文本如用模板生成，则生成的文本过于死板，而且要定义较多的模板，不利于系统的维护和扩充。用RST的方法虽然使系统有较好的灵活性，但必须花相当的精力来建立文本规划库，影响整个系统的开发速度。因此只有SCHEMA技术是最合适的。既能生成具有一定灵活性的预报文本，满足生成的需要，又能使系统比较容易建立和实现。

2.2 SCHEMA的建立

建立SCHEMA时，遵循两条准则。第一，在SCHEMA的高层，尽量使之与应用领域分开，脱离具体应用，只反映语言内容的性质，以适应对各种文本内容的描述；而在低层则比较接近具体的应用领域，使之满足应用文本生成的需求。第二，SCHEMA定义都是对用户开放的，开放的含义是：系统允许用户修改SCHEMA的结构，使它能够根据实际要求进行适当的改动。按照第一条准则，使SCHEMA具有高层的抽象性和底层的具体性。按照第二条准则使之具有较强的灵活性，生成系统能适应各种应用领域。

SchemaBegin
Name:GeneralForecast
HeadIllustration
ForecastIllustration+
TemperatureStatus
SchemaEnd

例如，一般日常预报文本，可以用SCHEMA抽象地表示为如下结构：“SchemaBegin”、“SchemaEnd”是关键字，表示一个SCHEMA的开始和结束，“Name”后面跟着是SCHEMA的名字，随后是三个SCHEMA项，它们既是谓词又是SCHEMA。“HeadIllustration”是抽象的描述气象预报头部的情况。后面的“ForecastIllustration”是用于描述具体的单

点气象情况，它后面有“+”操作符号，说明它能够重复进行1到N次的描述。最后是对实况温度描述的SCHEMA。

在设计SCHEMA时，笔者不仅让SCHEMA 具有递归性。而且对原有SCHEMA的操作符进行了改进，引入了“‖”操作符〖5〗，该符号的含义是，在一个SCHEMA中的几个有关的谓词可以根据相应的条件进行前后次序的组合排列。这一操作符对于天气预报文本生成有很大的作用，能大大增强预报文本的灵活性。例如：在日常的预报文本中，常会出现这样的情况，随着四季的不断变化，不同时间段中强调的天气情况是不同的。在夏季需要强调高温、雷雨，而在冬季需要预报低温、大雪等。

2.3 适合汉语特征的修辞谓词表达

汉语一个最突出的特点是语法缺乏形态，特别注重合意，谓词在本系统中基本上可以和句子相对应，谓词的主要作用是告诉系统该句话的语义，所以在建立谓词的时候必须考虑到汉语的特点。一般语义由谓词、项、修辞成分和连接成分四部分组成。注意这与规划器中的谓词含义不同。Fillmore 提出的格语法属于生成语义学理论，它恰恰说明了谓词与项之间的各种复杂的语义关系。在句子语义结构的描述中就是要将句子内部的各种语义关系表示出来。笔者认为，在句子语义结构的表达中应包含以下三个方面的内容：深层格关系、修辞关系和情态关系。以中文天气预报文本为分析对象，可将格分为15类，例如：施事格、受事格等〖6〗。项的修辞成分有许多种，根据所处理的内容不同，将其分为以下两类：(1)表示领属，修辞语只能是名词和代词，一般加“的”字。如：今天的天气情况。(2)表示数量，一般不能加“的”。如：最高温度3度。

汉语中，形式上没有标明句子的时态、语态、语气这些表示情态的信息，而是用对应的词来直接表示。因此可将单句看作是“情态+命题”的形式。句子语义结构的形式描述：

S(语义) :: = M(情态) + P(命题)　　M(情态) :: ={时态、语态、语气}

P(命题) :: =V+A_1+A_2+…+A_n

Ai :: =Name+DataType+Semantic+Case+Optional+Modify　　i=1,2,…,n

Name ∈ {String}，DataType ∈ {Character,Digital}，Semantic ∈ {String}，

Case ∈ {AGT,OBJ,EXT,EXP,ATT,VAL,NUM,REA,RES,DUR,TIM,DIR,LOC,SCP,OTH}

Optional ∈ {True,False}，Modify ∈ {True,False}

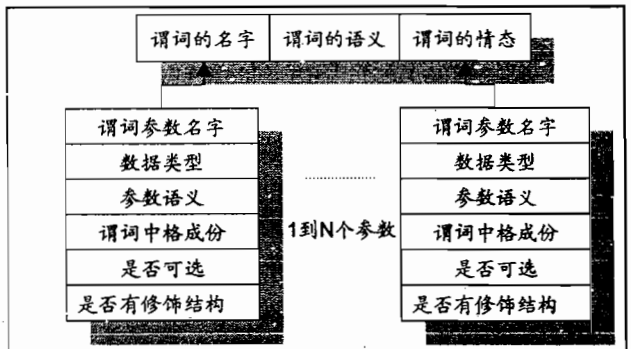

图1：谓词结构图

谓词的主要建立方法是以句子的中心词作为谓词的名字，一般情况就是以动词作为中心词进行扩展，每个谓词的参数都是句子的主要成分，例如：动作的发出者、动作的承受者、时间状语、地点状语等。每个参数包含的内容有：谓词参数的名字、参数的数据类型、参数的语义、格成分、是否可选、是否带有相应的修辞结构。谓词结构如图1所示。

三、SCHEMA 的选择和填充

3.1 SCHEMA 的选择

宏观规划的第一步就是进行SCHEMA的选择。SCHEMA是对预报文本结构的抽象，由系统人员或用户自己，根据不同的预报文本，预先写好SCHEMA，将其放在SCHEMA库中。库中包含两种信息，一种是定义好的SCHEMA，另一种是选择SCHEMA的规则。选择SCHEMA主要依靠用户的要求或根据输入数据产生的大颗粒度的信息。选择SCHEMA的方法有两种，一种是根据用户信息选择相应的SCHEMA，然后根据大颗粒度信息再次缩小选择的范围。第二种方法采用回溯递归测试方法。这种方法遵循一条基本原则：哪个SCHEMA所能覆盖的命题多，那个SCHEMA就是最优的。因为能够覆盖较多的命题，那么就是能够较清楚的讲清主题，就能生成较好的预报文本。在本系统中，同时采用这两种方法。

3.2 SCHEMA的填充

系统确定一个SCHEMA之后，它通过匹配谓词的形式进行填充。对于每一个SCHEMA的填充过程，笔者采用焦点控制技术〖2〗，来控制命题选择过程，消除歧义。在具体设计中，简化了焦点控制复杂度。如在一个谓词中，并不是所有的成分都能有资格作为焦点，例如：副词。在天气预报文本中，一般歧义出现时，主要是对象的歧义。也就是要分清信息到底属于哪个对象。因此在候选焦点堆栈中只是保留谓词的对象，而把其他的谓词参数去除。通过简化，不但降低了算法的复杂度，而且能够节约不少内存空间。

另外在填充过程中还需要考虑在同一层次上，几个匹配成功谓词的排序问题。在生成过程中会有两种排序问题，第一种是描述不同对象的语句顺序；第二种是描述相同对象的语句顺序。第一种语句的顺序问题，一般可以在定义SCHEMA的时候决定。系统以谓词的先后顺序来定序。第二种就比较困难。如果对于描述相同对象的语句采用不同的谓词进行表达，那么系统可以按第一类的处理方法。但是如果谓词是一样的，只是表达的内容有所不同，那么系统必须采用其他方法。本系统中，采用在外面排序的方法。基本的思想是，当出现相同的谓词允许出现N次，并且有M个已经匹配成功，那么系统将会按照这条谓词所对应的排序规则对匹配成功的谓词进行排序。如果该谓词没有对应的规则，那么将按照其原有的顺序排列。至于排序的规则，笔者认为大都是与应用领域有关，需要用到领域知识。

四、内容规划器的实现

4.1 气象数据的处理

气象预报处理包括三个阶段。第一阶段是通过气象的仪器设备收集气象实况数据，根据数学模型生成数字天气预报，第二阶段是根据前一阶段的数字天气预报以及观察数据生成天气描述图。最后一步是将天气描述图转变成预报文本和预报图形〖7〗。根据上海中心气象台的情况，目前第一步已经做到了计算机处理，能够根据实际的观察数据，经过计算机处理生成数字天气预报。由于种种原因，现在该预报数字一般只作为参考。而从第二步开始的预报工作都需要人工的大量干预，往往需要多个具有气象专业训练的专家，经过复杂的计算和分析才能得出较为准确的天气预报。

系统目前以日常单点预报文本作为输出文本的样本。主要有两个原因：第一，该预报文本是气象台每天都要用到的，使用频率很高。第二，该预报文本具有较好的篇章结构。容易为其建立SCHEMA。系统首先从气象台获得数字天气预报的输出。现在气象台是每天

发布四次天气预报，每次预报都是对接下来的三十六个小时内天气情况进行预报，每隔六个小时就有一次数字预报的结果。每次数字预报有四到五个地点的天气形势描述和两个气象站的温度实况报告，每个地点的天气形势描述包括五个要素：天气状况、风向、风力、最高温度、最低温度。此外，输入信息还包括预报发布的时间和气象台名。所有这些信息将作为系统的原始输入数据。

由于原始的数据都是数字信息不能作为文本内容规划的输入，系统必须把这些数字信息，利用气象的领域知识，通过计算、分析后，转换成概念数据（Conceptual Value）才能作为输入。概念数据具体表达形式如下：（数据内容，数据类型，数据语义，数据的标识符，数据的修饰对象标识符），数据的内容字段对应于原始数据的数字，数据类型对应于原始数据的类型，语义信息是对应于谓词的语义信息。数据标识符用于唯一地识别该数据或对象。数据的修饰对象标识符是表示该数据是修饰哪个对象。例如：上海市有五级风的数字信息，转变成概念数据的表达是（5，数字类型，风力，ID，上海市对象ID）。所有这些信息组成宏观规划的最小输入信息，称为概念数据元。

在生成系统中，除了进行一般性赋予语义信息的转换外，更主要的是需要进行空间与时间上的合并转换[8]。所谓空间的合并就是将天气情况相似或者接近的若干个地区结合起来，作为一个整体进行处理。而什么样的天气状况可以说是相似或者接近，需要用大量的气象知识进行判断。例如：即使在简单的海洋天气预报中，对于不同的地域一般采用不同的合并规则。在近海区域，认为天气情况相似的条件比较严格。在远海地区，条件就比较松。时间合并也是一个很重要的转换。预报的数值是连续的一个序列，需要根据一定规则，将相同的情况合并起来，例如5点钟是多云，11点钟还是多云，那么必须按照时间将他们合并，表示今天早晨到中午是多云。

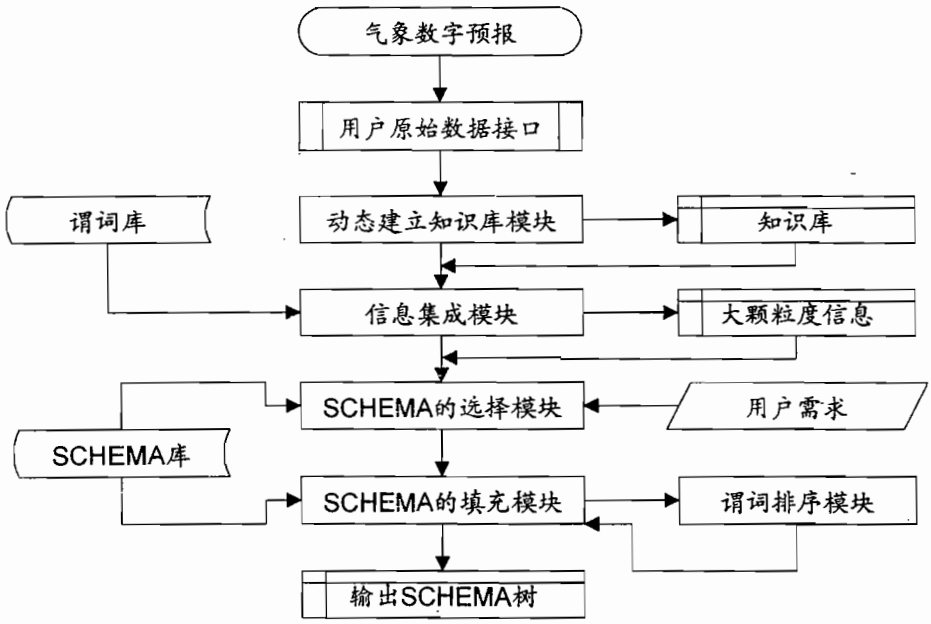

图 2： 文本内容规划器的结构组成图

4.2 内容规划器组成结构

系统内容规划器包括六个功能模块：用户原始数据接口、动态建立知识库模块、信息

集成模块、 SCHEMA 的选择模块、填充模块、谓词排序模块。内容规划器结构如图 2 所示。

下面简单地介绍规划流程，系统通过原始数据接口获得数字预报的原始数据，这部分的数据主要包括：发布预报的台名，具体时间，预报地区的名称，每个地区每隔六小时的数字预报等。接受输入后，系统就动态地建立知识库。这一部分主要包括两方面的工作：一是形成概念数据元，构造知识树，二是生成大颗粒度信息。接下来由信息集成模块，根据谓词库的修辞谓词，进行初始化谓词。然后由 SCHEMA 选择模块，根据用户输入信息和前面生成的大颗粒度信息进行选择，选出一个 SCHEMA 集合。填充模块对该 SCHEMA 集合的每个 SCHEMA 进行测试性填充。接着通过比较返回的权值，选择一个权值最大的 SCHEMA 作为最优的，最后对该 SCHEMA 进行填充。输出 SCHEMA 树。

五、测试结果与结论

宏观规划器采用两种方法进行测试。第一，设计数据随机发生器，该发生器模拟生成一组气象数据。测试的目的是观察规划器是否能够进行正常工作。结果是规划器能够进行正常工作，但是由于模拟生成的数据随机性强，且与实际的天气情况不符，使生成的文本不能反映天气预报的真实情况。第二，采用由气象台提供的 96 年 2 月到 7 月共 728 篇预报文本所涉及的气象数据进行测试，测试目的是检查规划器是否能够进行合理的规划，并且与原文进行比较。测试结果是： 90 % 生成的句子内容符合要求，其它句子内容基本符合要求，但是生成的句子与源句有差别。整个生成文本在段落和篇章一级上的顺序全部符合要求。在这七百多个文本中，基本的句型有 14 种。由于领域特点，所有类型都是陈述句。测试文本的种类是日常普通的预报文本。根据测试的情况可知，采用 SCHEMA 技术的规划器能够满足生成预报文本的一般要求，但是对于输入信息变化较大，并且生成有疑问句型结构的文本，规划效果还有待进一步的测试。本内容规划器是一个用于研究的实验系统，能够规划出正确、连贯的预报文本内容，对原有的 SCHEMA 方法进行了改进，能够让用户修改 SCHEMA ，提高了 SCHEMA 的灵活性和应用的适应性，增强了输出文本的灵活程度。

最后，笔者要衷心感谢导师姚天昉教授在研究工作中悉心的指导和热情的帮助，感谢黄小戎博士、张冬茉老师、高国栋等老师向笔者提出的有益的建议。

参考文献

〖1〗 Eduard H. Hovy. An Overview of Automated Natural Language Generation. International Symposium on Natural Language Generation and the Processing of the Chinese Language INP (C)-96, 1996. p15-33

〖2〗 Kathlean R. Mckeown. Using Discourse Strategies and Focus Constraints to Generation Natural Language. Cambridge University Press, Cambridge, U.K., 1985.

〖3〗 Johanna Doris Moore. A Reactive Approach to Explanation in Expert and Advice-Giving Systems. [Dissertation] , 1989.

〖4〗 Huang Xiaorong. The Project ACNLG. International Symposium on Natural Language Generation and the Processing of the Chinese Language INP (C)-96, 1996. p5-10

〖5〗 Xiaorong Huang, Tianfang Yao, Guodong Gao. Generating Chinese Weather Forecast with Stylistic Variations， 17th International Conference on Computer Processing of Oriental Language, Hongkong, April, 1997.

〖6〗 姚天顺等. 《自然语言理解》，清华大学出版社，广西科学技术出版社， 1995， p277-281

〖7〗 Elo Goldberg and Norbert Driedger. Using Nature Language Processing to Produce Weather Forecasts. IEEE Expert, Vol 8 April 1994. p45-53

〖8〗 E. Goldberg. FoG: Synthesizing Forecast Text Directly from Weather Maps. The Ninth IEEE Conference on Artificial Intelligence for Application, March 1-5, 1993. p1-11