

# 自然语言宏观规划器知识库的拟人化实现

戴炯 王纤

上海交通大学计算机工程系

**摘要:**本文以人的语言生成为参照,引出对宏观规划知识库的基本设计构造。接着描述了知识库的实现及其中对人思想的模拟。再从实践出发,在理论上归纳了动态知识库的设计要点,提出了更为实际,更为符合人习惯的方案。

**关键字:**自然语言 宏观规划 Schema 修饰谓词 知识库

## The Implementation of the Knowledge-base of the Macro-planer of the Nature Language Generation System

Dai Jiong Wang Qian

Department of Computer Science&Application, Shanghai JiaoTong University

**Abstract:** In this paper, we present a method of design for the macro-planner of the text generation system used in the nature language generation process as our model. Then we explain the implementation of this knowledge-base in detail. From this practicing, we conclude the key of design and implementation of a dynamic knowledge-base, offered a more realistic model.

**Keywords:** Nature Language Macro-planner, Schema, Predicate, knowledge-base

### 一、引言

自然语言是人类千百年劳动、发展、进化的结果,人类智慧的结晶:她是如此的复杂、奇异又是如此的精妙与美丽。

自然语言处理(NLP)就是利用电子计算机为工具对人类特有的书面形式和口头形式的自然语言的信息进行各种类型处理和加工的技术。它可分成自然语言理解(NLU)与自然语言生成(NLG)。自然语言生成是用自然语言来构造篇章(短语,句子,段落)的过程,其目标在于把一个个句子“意义”的内部表示转换成它们的表层结构,即适当的词串。

下面就一些本文用到的技术与理论作一个简介。

修辞谓词(Rhetorical Predicates)是一种讲话者描述信息的方法,它们抽象地表示各种讲话者使用的语言动作,同时还能描述文本中各命题之间的结构关系。例如:类推(analogy)、

成员或成份描述 (constituency)、属性 (attributive)。Williams (1893), Shipherd (1926), Grimes (1975)等语言学者对此进行了研究。Grimes 提出修辞谓词是可以递归处理的。

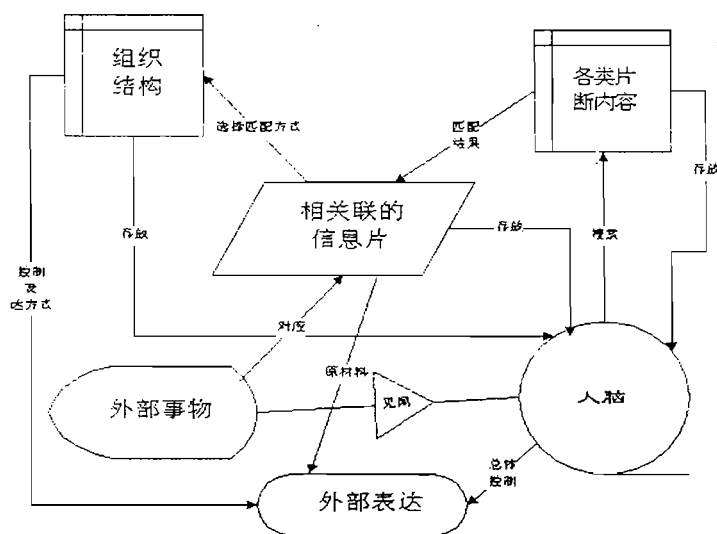
SCHEMA 就是用修辞谓词以及一些操作符来描述文本结构上的规律。Mckeown 在 1985 年提出了 SCHEMA 方法, 较为详细地讨论使用修辞谓词进行篇章描述的方法。SCHEMA 可以被递归的描述, 在一个 SCHEMA 中还可以嵌入其他的 SCHEMA 来形成对段的描述。

本文利用了修饰谓词的有限性与递归性, 及 Mckeown 的 SCHEMA 方法。在该方法上进行了改进与总结, 引进了更多的符号与属性使之更符合应用。根据语言的生成规律, 总结出一套有很强适应性的宏观规划器数据库的通用生成方案。

通常在实践中自然语言生成可分为三阶段, 本文主要讨论的第一阶段宏观规划 (Macro-planning)。主要目的是根据生成文本目标选择恰当的文本结构将选出的内容以连贯的方式组织起来。它还可后续阶段提供词法和句法选择方面的参考信息。

## 二、宏观规划器知识库的人性化设计

设计的灵感来自对人类如何表达外部事物的研究。在此问题的众多答案中我认为语言框架模型是最为合理的一种。语言学家 Chomsky 通过分析儿童的语言能力, 得出结论是: 人的语言能力并不是后天通过总结语句实例形成的, 而是先天存在于大脑之中, 儿童是在此基础上发展起语言的。“人类实际讲的语句是有限的, 然而人却可以说出无穷的话语。就是因为人能够将这些语句进行反复不断的有条理组合。”这使我们知道让机器记忆人所会的所有词语是一件简单的事, 然而关键问题在于如何充分了解词语的意义与组织这些词语。



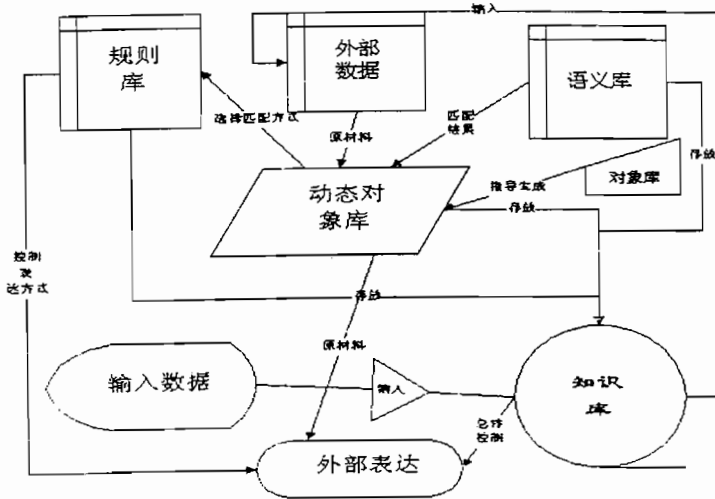
经过比较、分析、思考与归纳, 归纳出了如下的人类表达产生模型:

先是觉察到了事物的发生, 然后开始思考, 将现有的事务与记忆中的情景联系起来, 这些情景包括以前留在脑海中的片断, 所学的知识等; 而后通过先天存于脑中的经过后天不断改造的语言模式组织这些内容; 最后产生主要的篇章。其实人脑储存有大量的典型情景与片断,

当人需求时, 就从记忆中选择 (粗匹配) 一个可以称为框架的基本知识结构, 这个框架是以前记忆的一个知识空框, 而其具体内容依照新的情景内容改变。对这个框架的细节加工、修改和补充, 形成对新情景的认识在储存于人脑中。根据这个模型, 我们可以把各类事物状态、属性、发展过程和相互关系归纳成一定规律性的东西, 存于数据库中。当我们有想法需

要表达，我们不必自低向上的，从一点一点的细节作起，最后才确定全貌。而实行自顶向下的匹配，如果匹配成功，则框架中存放的便是我们所需要的表达内容，如果匹配不成功，则寻找原因，重新自数据库中取得一个更能与事物匹配的框架，或者修改刚才那个匹配的不太成功的框架，直到最后得到一个满意的表达为止。我们可以知道人在说话前，先在脑中生成一个想要表达的基本意思，称为语言的深层结构，而他说出来的是已经组织好的语句，称为表层结构。其中深层结构通过某种特殊的转化成为表层结构。（上图为人的表达生成图）

根据上述想法，在 SCHEMA 理论与修饰谓词的理论指导下，由于 SCHEMA 和修饰谓词的有限性，可表达性与低归性设计出了我们现有的系统知识库的总体结构。如下图所示：



从左图中可看出，知识库可分为三个部分：外部输入数据库，内部数据库与规则库，当中还有许多它们的连接关系。规则库中存放着各种框架，也就是大脑中的文法控制系统。外部输入数据库，可以看作人的记忆体，将所见所闻记入脑中的地方。内部数据库是人有的一些概念，片断，知识。在图中以语义库和对象库来表示。

### 三、知识库的实现

上文给出了知识库的总体设计思想，这一节主要介绍如何实现这三个部分：外部输入数据库，内部数据库与规则库。

本系统的知识库使用关系数据库的技术实现的，所有结构均以表的方式储存。因为关系数据库可表达任何复杂的数据结构又有很好的 SQL 数据查询排序功能。通用性强，扩展方便，安全性好，可以省去很多数据库管理方面的问题。

#### 3.1 语义库的构造

首先要做的是将我们脑海中的的情景包括以前的所见所闻留在脑海中的片断，所学的知识等转储到电脑中。语义可以说是人脑中的一些概念，片断，也就是人所知道的东西，如小鸟、跑步、粉红等等。。语言可以不同，但同一事物的语义是相同的。我们一般只关心信息的真正的含义或者内容是什么，而对其表层的形式是不管的，所以要建立这个语义库。

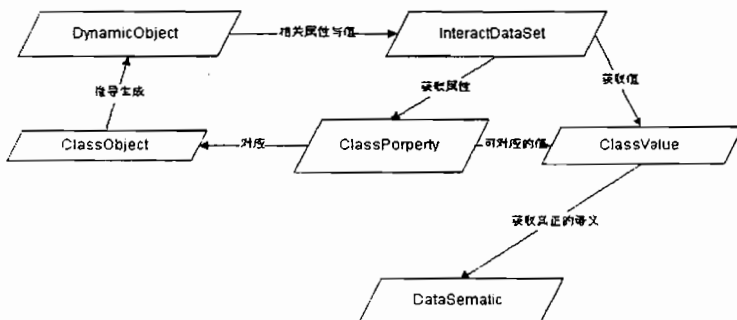
对相同的一个含义的表达的形式不同，即使在同一语种内部，它的表达的词也是可能不同的，因此我们在对语句做深层次的处理时。就像人脑中同一意思会有不同的词语，相同的词语也会有不同的意思。如果在系统内部做到，一个意思一个语义码。这样可以不管各种

自然语言在表层形式上的千差万别，它们表达的意思或内容是一样的。系统采用语义编码后，使生成系统至少有两个好处。第一个优点是解决多义词问题。系统不会把词的含义和表层形态、具体词的用法混在一起处理。这样有利于简化处理复杂性，同时也是生成各部分的任务分开。第二个好处是，用语义信息来表达谓词参数的内容时，较为容易的正确表达句法结构。例如明天有雨。该句英语的表达是“*It is going to rain tomorrow*”。不难看出，如果你直接使用表层的单词，那么中文中主要动词就是“雨”，而英文中就是“rain”。造成不一致。

有了语义库，就像一个人的大脑中有了许多概念。如他知道了什么是下雨，什么是晴天，什么是多云。

### 3. 2 内部数据库构造

我们是采用对象的方式管理数据的。因为人对事物就是采取对象的方式来认识的。对象是现实世界中的一切物理实体、抽象概念和事件。在本系统的一个对象是用四维来确定的，是时间+地点。



一个对象中可以有多多个属性，各种属性所包含的各种信息可以分成两类：字符信息和数字信息。这些属性是生成系统所处理的最小颗粒度信息，每一个对象属性都有一个明确定义的语义。对

象结构定义了一类对象的框架。每个属性都有语义，对象内的属性依靠属性语义确定它们之间的关系。它们的内容就是属性值。对象又分静态与动态，静态对象是上图中的对象库，它规定了对象的结构等，使得对象可以动态的生成。动态对象是将外部的实例在内部具体的存放形式。关系如上图。

就好比，一个人，知道了下雨，天晴。（语义库中）。1999年5月10日，他到了上海，见到天在下雨，想告诉家人。于是，他需要在脑海里记住这么一个事物（存放一个动态对象），他首先要记住这是1999/5/5/上海的天气。（在classobject中确定这是一个天气的对象）。然后他要记住两个属性：天气状况：雨；风力大小：3级。于是他将语义与现实情况联系起来，所以他要搜寻所知的概念（语义库），与对象的属性（对象库），然后找到雨和天气状况，将其关联起来存入脑海中的事物下，然后记住这个事物（内部知识库）。这个系统拥有适应性，因为是完全仿照人的认知模型，所以当语义库里有股票的概念，对象库里加入了股票的对象，然后加入相关的属性，和取值范围，便可记忆股票信息。所以很容易将本系统移植到其他用途。移植的关键就是添加相关语义，与加入相关的外部数据结构及算法。

### 3、3 外部数据库构造

人看到事务，通过脑子分析，记入脑海。现在通过上述的知识库，计算机知道如何记忆对象了，但如何处理外部事物也是个问题。例举天气预报系统风的外部数据储存结构如下：

字段名称	数据类型	说明
RecordNo	Number	记录的唯一编号、
DirectionName	Text	风向的中文表示

WinForce	Number	风力的中文表示
----------	--------	---------

也就是，当我们看到风以后，我们在脑海里对风速，风力等数据的储存。

### 3、4 规则库的结构

这里所谓的规则库也就是人脑中的框架、规则结构。由于使用的是 SCHEMAL 技术，本系统的规则库可分为两类，一类是谓词的匹配规则，另一类似 SCHEMAL 的匹配规则。

#### ● 谓词匹配规则库

构造谓词规则库如下：

##### 1. Argument table, 存放 argument

字段名称	数据类型	说明
ID	Number	记录的唯一编号
Name	Text	Argument 的名字
Semantic_ID	Number	语意代码 Semantic 有关

##### 2. Predicate table, 存放修饰谓词的构成规则

字段名称	数据类型	说明
ID	Number	记录的唯一编号,
Name	Text	Predicate 的名字
DatSematic_ID	Number	语义代码 DataSematic 有关

谓词在本系统是由 Augument 构成的。比如某人要说：“我吃饭。”他的大脑里必须有这样的句子结构：此句子结构由三个 Augument 构成，Augument1 是一个表示施者的参数，Augument2 是一个表示动作的参数，Augument3 是一个表示受者的参数，当出现这三个参数“我”“要”“吃饭。”这个命题就可匹配成功了。

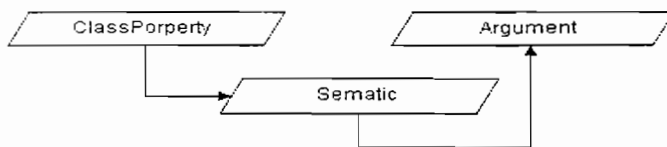
#### ● SCHEMA 匹配规则库

SCHEMA 在知识库总分两张表实现：SCHEMA 库主要用来存放 SCHEMA 自身的一些信息。而 SCHEMAITEM 库是用来描述 SCHEMA 的树形结构的，是典型的用关系型数据库来表示树的体现。

我们在建立SCHEMA规则库时，遵循了两条基本的准则。第一，在SCHEMA的高层，应尽量使之与应用领域分开，脱离具体的应用。第二，所有的SCHEMA规则库都是对用户开放的。

人的思维模式是多样性的，产生表达并不是只根据一个框架，而是一种综合分析，在框架上加有很多辅助信息。系统为了加快选择SCHEMA的速度和向后面提供有关文本布局的信息，为SCHEMA加入一些辅助信息，可以在前面列出的SCHEMA表中看到，目前包括选择条件（SelectCondition），排序条件（OrderCondition），描述风格（DescribeStyle），优先级（Priority），布局信息（LayoutInfo），焦点变量（FocusVariable）。

下图为规则库与内部数据的联系：

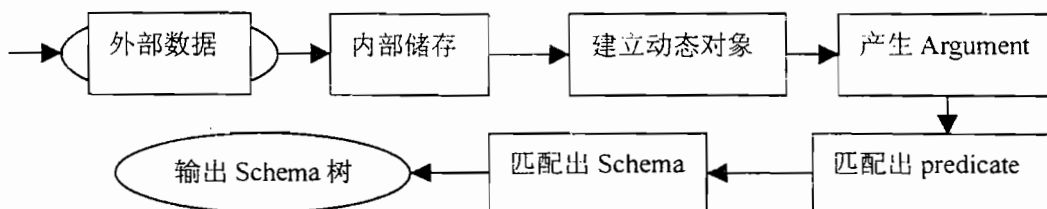


在规则库中Predicat定义了它的形成可能，SCHEMAItem则使用了树形结构，凡是匹配到树叶的都算是成功的。

有了我们现在的知识库，当一个人记忆下1999/5/5日上海的天气对象后，通过中间桥梁SEMATIC表可将所知道的对象属性转换成实例化的Argument，挂在该动态对象下。然后给这些Argument匹配出可以形成的所有句子，如：今天上海有雨。近日上海风力5级。等等。

## 四、知识库使用简介

有了上述知识库，还需要知识库控制器，来控制语言的生成，就像大脑可以控制其记忆的东西一样。每个过程都要用到相应知识库中的关系表。本系统设计了一个完整的控制器，大体过程如下：



本系统可以生成天气预报与银行的自然语言文本，现简单举例普通天气预报文本生成过程：

1、读取气象运图中的信息，放入设计好的天气预报数据库。

比如：时间:1999.5.5 地点:shanghai 风力:3.4m/s 风向:278度 ...

2、根据语义库，内部数据结构，将天气预报库中的数据变为动态对象。

对象号:1 对象名:1999.5.5.001002001(上海的语义码) 风向:001005011(东北的语义码)...

3、通过规则库与内部数据的桥梁生成挂在动态对象下Argument。

对象号:1 有属性: Winddirect: 0010050011; Region: 001002001...

4、通过Predicate规则库匹配动态对象的Argument，获得Predicate，挂在动态对象下。

可产生修饰谓词:上海有东北风的修饰谓词方式的表达。

5、根据Schema规则库匹配Predicate，优选产生最好的Schema。

根据某几个修饰谓词可产生篇章：1999年5月5日上海气象台发布天气预报 今天天气晴.有东北风2级.的Schema方式的表达。

6、依据内部Schema的方法，输出获得的Schema。

此过程大体符合人类的思想过程。在库中还有很多的其他辅助的规则表，如：排序规则、推导规则、大颗粒度信息规则等等.用以更好的生成文本。

## 五、本系统知识库的特色

由于采取了模拟人脑的思想，本系统有五个主要的特性。

- 知识库的抽象性。所谓抽象性是指用户输入的信息是和具体领域无关的，或是关系不是很紧密。就像人脑对任何事物的记忆总是用相同的煤质与方法。系统从用户输入界面和库结构两个方面进行了较为灵活的设计。对内部数据处理，用统一的形势，达到

了抽象的目的。

- 对象的可扩展性。就像人能不断地接受新鲜事物，一个系统具有了抽象性就必须具有可扩展性。本生成系统有专门的维护对象结构库的界面。
- 对象的时序性。现有的关系模型，把四维的时态关系嵌入普通的两维关系表中。在本系统中，主要运用该特点进行时间和空间上的数据合并，以减少输入数据的冗余度。
- 知识创建的动态性。动态的建立是讲知识库的内容是随着用户输入数据自动临时生成的。我们的系统是动态的，首先系统需要用户输入大量的各种类型的数据，他们从形式到内容都不相同，其次是需要根据用户的输入信息，动态构造一个知识库。
- 多语种。本系统是多语种文本生成系统，为了更好的适应多语种的生成，在设计知识库的时候，我们把每个输入的基本信息都转换成内部的语义码，这样不管外面的语种是什么，到内部表达的样子是相同的。

## 六、测试结果与结论

采用了上述知识库后，系统很方便的从天气预报系统扩展到了银行系统。并加入了台风、雨情等特殊的天气预报，实现了智能排序与优先级算法。通过了上海市专家技术鉴定组鉴定，鉴定意见为：内容规划器采用的知识库具有较好的可扩展性和可维护性，提高了生成的有效性，并使得句子的结构组合更具有灵活性。此外，它还具有便于扩展语种和应用领域等特点。系统有很好的实用性、扩充方便灵活。基本能满足普通天气预报和银行统计文本的需要。所生成的文本基本符合语言和应用领域的表达习惯。共可产生七种天气预报文本，72种中文句型，78种英语句型，80种德语句型与两种银行文本，28种中文句型，32种英语句型，37种德语句型。产生文本全部符合要求。

最后笔者要衷心的感谢姚天昉导师的悉心指导与热情帮助，感谢张冬荣老师提出的有益建议。

注：【本系统是由上海市科学技术委员会立项、上海交通大学计算机科学与工程系承担的课题。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。】

## 参考文献

- 【1】M. W. Meteer. Expressibility and the problem of efficient text planning. St. Martin's Press, New York, 1992.
- 【2】L. Jordanskaja, M. Kim, R. Kittredge, B. Lavoie and A. Polguere. Generation of Extended Bilingual Statistical Reports. Proc. of COLING-92, AUG. 1992.
- 【3】Heidelberg. Trends in Natural Language Generation -- An Artificial Intelligence Perspective. Springer, Berlin, 1996.
- 【4】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。
- 【5】林杏光，王玲玲，孙德金等，《现代汉语动词大词典》，北京语言学院出版社，1994年11月。
- 【6】王纤。“文本自动生成器的实现技术分析”，96全国理论计算机科学学术年会，1996。
- 【7】陆汝铃。《人工智能》。科学出版社，1995。