# 应用编译原理实现基于文本编码通信协议消息的解析

杭州东方通信股份有限公司网络通信研究所(310053) 刘 伟

TN91 B

**摘 要**：在介绍编译器实现原理和自动生成工具的基础上，提出设计文本编码通信协议消息解析程序的一种方法。

**关键词**：编译器 文本编码通信协议消息 词法分析 语法分析

随着技术进步，传统的通信网不断发展，逐渐与互联网融合，最终将形成新的下一代网络(NGN)。大量新的通信协议必然会伴随着新的网络结构而产生。这些新的协议不同于传统通信协议的二进制消息编码格式。新协议大都采用基于文本消息的编码方式，如 SIP、MGCP 和 H.248/MEGACO 协议等。基于文本的消息易于调试和理解，不需要专门的解码器，有较好的扩展性。传送文本消息虽然比传送二进制消息的开销要大，但对于现在的网络带宽，这种额外的开销基本上可以忽略不计。若协议本身不复杂，对文本消息的解析程序一般都很容易实现。但是，如果协议本身比较复杂，则手工编写解析程序会耗费大量的时间和人力。从本质上讲，解析文本消息其实质就是分析字符串的含义。它类似计算机高级语言编译器中词法分析和语法分析的功能。设计文本消息解析程序，应该能够借鉴编译器的实现方法。

## 1 原理简介和设想

### 1.1 编译器工作过程

编译器的工作过程一般分为以下几个阶段。

(1)扫描阶段(词法分析)：编译器阅读源程序(通常以字符流的形式表示)，将字符序列收集到称作记号(Token)的有意义单元中。记号由一个或多个字符组成，如关键字、标识符、数字常量、引用字符串及其他有意义的符号。记号是语法分析输入的最小单位。扫描阶段还可完成与识别记号一起执行的其他操作，如将标识符输入到符号表中、将文字(数字常量、引用字符串)输入到文字表中等。

(2)语法分析阶段：编译器从扫描程序中获取记号形式的源代码，分析程序的结构元素及其关系。通常将语法分析的结果表示为分析树或语法树。

(3)语义分析阶段：大多数程序设计语言都具有确定但不易由语法表示和由分析程序分析的特征。这些特征被称作静态语义。典型的静态语义包括声明和类型检查。语义分析的任务就是分析这样的语义。

(4)源代码优化阶段：该阶段对源代码进行改进和优化，输出中间代码。

(5)代码生成阶段：根据中间代码生成目标机器的代码。

(6)目标代码优化阶段：在该阶段编译器尝试改进已生成的目标代码。

在设计基于文本编码通信协议消息的解析程序中，可以借鉴编译器扫描阶段和语法分析阶段的原理和实现方法。

### 1.2 词法分析和语法分析的程序实现

在程序设计语言中，一般包括以下 4 类记号：(1)保留字，是具有特殊含义的固定字符串。(2)特殊的符号，包括算术运算符、比较运算符和赋值。(3)标识符，一般是以字母开头，字母和数字的混合串。(4)文字和常量。构造编译器词法分析程序的过程如图 1 所示。


图 1 构造编译器词法分析程序的过程

正则表达式表示了字符串的格式。根据正则表达式可以构造出非确定性自动机(NFA)。一个确定的字符串在 NFA 中存在不止一个状态序列。通过子集构造(Subset Construction)算法，可以由 NFA 生成确定性自动机(DFA)。再通过状态数最小化将生成的 DFA 化简为最小状态 DFA。根据这个最小状态 DFA 来实现词法分析程序。

编译器语法分析程序的任务是确定程序的结构。程序设计语言的语法通常是按照上下文无关方法(Context-free Grammar)规则，其方式与使用正则表达式表示记号词法规则相类似。二者的主要区别在于上下文无关文法的规则是递归的。因此由上下文无关文法识别的结构类比由正则表达式识别的结构类多，且识别这些结构的算法也与扫描算法(如 DFA 或 NFA)差别很大。此外，上下文无关文法中表示语言语义的数据结构也是递归的，经常

采用树结构，称作分析树或语法树。

按照构造分析树或语法树的方式，语法分析算法大致可分为2类：自顶向下分析(Top-down Parsing)和由底向上分析(Bottom-up Parsing)。自顶向下分析算法构造分析树的顺序是由根到叶，而由底向上分析算法则是由树叶开始，直到构造出完整的分析树。

### 1.3 文本编码格式消息的特点

文本编码格式消息与用程序设计语言书写的源程序代码有相同之处，例如，有类似于程序设计语言中的各种记号，表现为参数名、参数值、各种含义的记号(=、>、<)等。记号之间有长度不确定的分隔符，如空格、换行、注释等。此外，基于文本编码格式的通信协议消息也有不同于程序源代码的特点：(1)文本编码格式消息结构一般是非递归的，参数不会以自身做为子参数。(2)文本编码格式消息结构一般是固定的。对于某一条具体类型的消息，消息的数据结构是明确、可知的。但是程序源代码的结构是变化的，生成的语法分析树大小也是非固定的。(3)协议消息的类型和参数的种类较多，结构的层次大小与具体的协议有关，而程序中逻辑结构种类一般比较有限。

### 1.4 编译器构造工具

语法分析的2类算法各自包含多种实际的算法，用手写程序来实现这些算法一般比较困难，通常采用自动生成工具来生成这些程序。编译构造工具中较有代表性的是PCCTS和YACC，分别用于生成自顶向下分析和由底向上分析的程序。自动生成工具根据语法描述文件，生成高级语言如C、C++或Java的源程序。根据前面的分析，解析文本编码格式消息时，采用自顶向下分析算法比较合适，所以在此介绍PCCTS工具。

PCCTS提供了一套编译器构造工具集合。PCCTS工具包括词法分析生成器DLG和语法分析生成器ANTLR。DLG根据词法描述文件生成词法分析程序。该程序可以从输入的字符流中识别出词法描述文件中定义的记号。ANTLR根据语法描述文件定义的语法规则生成编译器的分析程序。在语法分析程序执行的过程中，会调用DLG生成的词法分析程序，来识别具体的记号。

PCCTS描述文件一般包括2个部分：(1)词法描述部分。它包括记号描述列表。每个表项包含一个记号的匹配模式，用正则表达式描述匹配该记号所必须符合的规则。此外，也允许嵌入一些程序代码，用于在匹配到该记号时，完成某些操作。(2)语法规则描述部分。它是一个语法规则说明的集合。每个规则说明一个要生成的编译器对应于高级程序设计语言的语法结构。这个规则集合必须是完备的，能够描述高级程序设计语言源程序代码中出现的各种逻辑结构。在语法规则描述中同样允许嵌入用于完成各种操作的程序代码。

PCCTS描述文件中的语法规则描述部分符合EBNF规范。一个规则允许包含子规则。ANTLR将每个规则生成一个函数，在函数中调用子规则生成的子函数。由于PCCTS采用自顶向下分析方法，允许规则接收事先声明的变量，返回规则执行的结果。因此，生成的函数可以有参数和返回值，可以很方便地将事先定义的协议消息解析数据结构(解包缓冲区)传递给解析函数，并根据函数返回值来判断解析是否成功。而对于自下向上的分析程序，由于是在执行的过程中才动态地构造数据结构，因此应用在文本编码消息解析中过于复杂。

## 2 用PCCTS工具实现解析程序

H.248/MEGACO协议是ITU-T和IETF合作提出的通信标准。该标准定义的功能使得媒体网关(MG)在软交换设备或媒体网关控制器(MGC)的控制下，将传统电话业务和各种多媒体通信业务接入核心网，使语音、传真和多媒体信号在公共电话网和新兴IP网络之间进行交换成为可能。H.248/MEGACO协议中提供了2种协议消息的编码格式，一种是ASN.1描述的二进制格式，另一种是ABNF描述的文本编码格式。协议规定媒体网关至少应该支持文本编码格式，而软交换设备应该对2种格式都支持。用PCCTS工具实现媒体网关控制协议H.248/MEGACO消息的解析程序，需要以下几个步骤。

(1)定义数据结构

不管采用哪种消息编码格式，在MGC和MG中都必须将消息转换成程序可以处理的形式。这就要求根据协议消息定义消息解析缓冲区数据结构。消息解析程序将分析得到的消息参数放入消息解析缓冲区，供其他程序使用。这个数据结构类似于高级程序设计语言中源代码经过语法分析后形成的语法分析树。不同之处在于语法分析树的结构是变化的，取决于源代码的结构，而消息解析缓冲区数据结构是固定不变的，可以事先定义。

(2)编写PCCTS描述文件

H.248/MEGACO协议描述文本编码格式采用ABNF规范，与PCCTS描述文件采用的EBNF规范很类似，因此很容易写出相应的PCCTS描述文件。在书写语法规则时，将消息解析缓冲区数据结构变量地址作为参数传递到规则中去，通过嵌入代码，处理文本消息字符流，将获得的消息参数写入消息解析缓冲区相应字段，并根据处理的实际情况返回相应的值。

(3)生成分析代码

PCCTS工具支持生成C、C++、Java等高级语言的分析器源代码。以C程序为例，生成分析代码最简单的方法是执行antlr h248unpack.g，其中h248unpack.g是PCCTS的描述文件。该命令分析h248unpack.g中的正则表达式和语法规则，如果没有错误，将生成h248unpack.c和parser.dlg文件。h248unpack.c是生成的分析器源代

(下转第53页)

### 1.2 HMM 模型

HMM 采用由左到右无跳转模型，对于数字和简单的语音控制命令，模型的状态数 L 取为 6。连接词系统的 HMM 模型如图 1 所示。

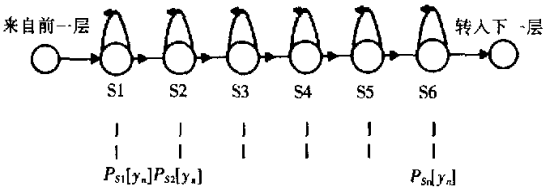


图 1 连接词系统的 HMM 模型

图中 HMM 模型系统输出矢量 $y_n$ 就是由语音特征矢量提取单元得到的 $LSP[f_1, f_2, \cdots\cdots f_{10}]$矢量。这样就为每一个数字和命令建立一套 HMM 参数 $\lambda_v=\{\lambda_{v1}, \lambda_{v2}, \cdots\cdots \lambda_{v6}\}$

### 1.3 训 练

训练使用了分段 K-平均程序 (Segmental K-means Procedure)。它有 2 个功能：(1)将成串的数字或者命令最佳分割成为孤立的词条。先把由众多说话者说出的词串存入训练词串数据库，然后使用分层构筑(Level Building) HMM 算法将所有训练的词串分割成孤立的词条。(2)将每个已分割为孤立状态的数字的训练样本汇聚起来，共有 $V$ 个词条训练集合。然后对每个词条进行 HMM 参数估计，把新得到的参数与原有的每个词条的 HMM 初始参数进行比较，如果二者差异达到某个阈值，则确认；反之，再用最终得到的 HMM 参数做为初始参数，重复上述操作。

最终的训练结果是得到了每个孤立数字或者命令词的 HMM 参数 $\lambda^v=\{A^v, B^v\}$，其中 $B^v=\{P^v_{S_1}[y_n], \cdots\cdots P^v_{S_L}[y_n]\}$。

### 1.4 识 别

由于语音识别的词串虽然不是定长的，但是可以将其分类为 4(如“call 1 1 9”)、12(如“call 010 68442345”)等不同的定长，而一条拨号指令总是由一个命令字和一串数字组成。因此可以使用分层构筑 HMM 算法。首先根据原始语音的长度确定词串大致的长度类型 K，为 K 个词构筑 K 层搜索路径。在第一层按照标准的 Viterbi 算法，搜索 $n=1\sim V$($V$ 是词汇表的大小)之间的任何一点。当第一层搜索完成时，在每个终点上可以找到一个$\delta^v_n(L_v)$的最大值及相应的词条编号 $v$，并记之为 $\delta_n[l]$和 $v_n[l]$，$n=1\sim V$。然后，以这些点为起点构筑第二个搜索层，直到全部词条搜索完成。

## 2 语音识别模块的嵌入

典型的嵌入式手机系统中的软件包括支持多任务的实时操作系统 RTOS(如 RTLinux、VxWork 等)、中断处理程序、完成特定功能的各项任务(Task)以及一些公用的支持库(如标准 C 运行库)等。每个任务完成一项特定的功能，任务间通过 RTOS 提供的通信机制 (如 Message、Signal)进行通信。在嵌入式应用中，RTOS 仅提供任务调度、任务间通信、中断向量表管理和定时服务等极为有限的功能。绝大多数的系统功能是在微内核以外的任务中实现的。因此语音识别模块也被作为一个 Task 存放于软件系统中。

语音识别模块是以静态库的形式作为 RTOS 的一个 Task，事先训练好的数据存放在嵌入系统的 ROM 中。MMI 界面控制 Task 进行消息调用，并将识别结果返回 MMI，由其电话本程序找到入口，完成整个语音拨号过程。

## 3 经验值选取

(1)用 HMM 模型计算 $P_Y[a, A, B]$中 $B$ 的最大似然率的迭代算法时，$A$ 的初值选择不好会导致迭代计算收敛到非全局最优点。实验中发现，被识别的对象从 200~230 的 31 个汉语发音连续数字，状态数为 L=9，当 $A_{i,i+1}$ 的初值为 0.111=1/9，即等于状态数的倒数时，可以得到最好的结果。

(2)实验结果表明，当训练的人数从 30 增加到 100 时，识别率有明显提高。

## 4 小 结

基于连接词语音识别的研究正随着嵌入系统硬件水平的提高和市场的迫切需求而升温。本文讨论了基于 HMM 系统的语音拨号系统的实现方法，并将低速语音编码的 LSP 系数和 HMM 模型结合使用，证明了 HMM 模型的优越性。系统的实现使嵌入设备的使用变得更加简便。

### 参考文献


1 杨行峻，迟惠生.语音信号数字处理(第一版).北京：电子工业出版社，1995

2 胡广书.数字信号处理—理论、算法与实现.北京：清华大学出版社，1997

3 柴海薪.连续语音识别的研究和汉语数字连接系统的实现.成都：四川科学技术出版社，1994


(收稿日期：2002-10-28)

(上接第 38 页)

码。对 parser.dlg 还需执行以下命令：

dlg-C2-i parser.dlg scan.c

该命令生成 scan.c 和 token.h。scan.c 为词法分析程序，由分析器程序调用。token.h 中对描述文件 h248unpack.g 中说明的记号进行宏定义。

其他程序可以调用 h248unpack.c 中的程序解析文本编码消息。

### 参考文献


1 *Louden K C.*编译原理及实践.北京：机械工业出版社，2000

2 ITU-T Study Group.Gateway Control Protocol.ITU-T H.248，2000

3 Network Working Group.Augmented BNF for Syntax Specifications：ABNF.RFC2234，1997

4 Terence J P.Language Translation Using PCCTS and C++.Automata Publishing Company，1993


(收稿日期：2002-11-28)

# 应用编译原理实现基于文本编码通信协议消息的解析

作者： 刘伟
作者单位： 杭州东方通信股份有限公司网络通信研究所,310053
刊名： 微型机与应用
英文刊名： MICROCOMPUTER & ITS APPLICATIONS
年，卷(期)： 2003,22(4)
被引用次数： 1次

## 参考文献(4条)

1. Terence J P Language Translation Using PCCTS and C++ 1993

2. Network Working Group Augmented BNF for Syntax Specifications :ABNF 1997

3. ITU T Study Group Gateway Control Protoeol ITU-T H 248 2000

4. Louden K C 编译原理及实践 2000

## 引证文献(1条)

1. 李峻 无线网络前台测试中事件合成技术的研究与设计[学位论文]硕士 2006