# 面向中间语义表示格式的汉语口语解析方法*

解国栋　宗成庆　徐波

中国科学院自动化所 模式识别国家重点实验室　北京 100080

e-mail:{gdxie, cqzong, xubo}@nlpr.ia.ac.cn　tel:(010)82614468

**摘 要：** 口语解析在人机对话系统和口语翻译系统中的作用是十分关键的。本文提出了一种统计和规则相结合的汉语口语解析方法，解析结果是一种中间语义表示格式。该方法分为两个阶段。首先，采用统计方法，解析出输入句子的语义信息，然后，利用规则，将这些语义信息映射到中间语义表示格式。试验证明，此方法具有较强的鲁棒性，而且避免了完全用规则方法解析的一些弊端，达到较高的解析正确率。

**关键词：** 口语解析　　统计解析模型　　中间语义表示格式（IF）

## 1、引言

口语解析在人机对话系统和口语翻译系统中的作用是十分关键的。典型的口语翻译系统和人机对话系统如图1所示。语音识别模块和语言解析模块在两个系统中是可以共用的。语音识别模块识别出用户的语音输入，并将识别的结果传递给语言解析模块，语言解析模块解析出句子的语义，并将语义表示传递给两个系统相应的模块。对于对话系统来说，语义表示传递给对话管理模块，对话管理模块根据语义表示，作出响应，然后由语音合成模块生成相应的声音。而对于口语翻译系统来说，语义表示传递给语言生成模块，语言生成模块生成相应的语言，然后由语音合成模块生成相应的声音。本文所叙述是就是语言解析模块部分的工作。

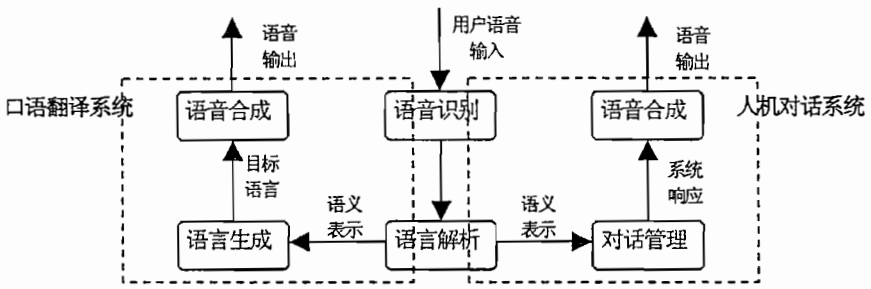


图1 口语翻译系统和人机对话系统结构框图

本文给出了一种统计和规则相结合的汉语口语解析方法，解析的结果是一种中间语义表示格式IF(Ingerchange Format)[7]。IF为C-STAR(Consortium for speech translation advanced research)[8]所采用，该组织的目标是建立一个面向旅游信息查询领域的口语多语言翻译系统，目前该组织已经将汉语，英语，日语，德语，韩语，意大利语，法语等语言包括到这个系统之中。根据需要，这些IF表达式可以转换为不

*本文受国家自然科学基金项目资助（资助号为69835003和60175012）

同的语言，从而实现多语言间的互译。这一过程如图 2 所示。

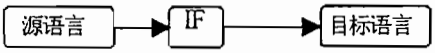


图 2 利用 IF 作为中介进行语言间的互译过程

口语解析的任务是从口语对话的句子中提取出语义表示。在口语里，句子往往不符合语法规范，句子中充满着重复、省略和颠倒等现象[6][12]，利用规则进行解析，往往需要针对这些特殊的语言现象编写大量的规则，需要花费一定的时间和成本。

近几年来，统计方法越来越表现出它在自然语言处理方面的优势。[2]和[5]利用统计的方法，进行自然语言解析，其中的语义解析器实际上是一个各态遍历的隐马尔可夫模型 HMM[1]。统计解析的特点是需要有足够的、经过标注的语料来对模型进行训练。只要有足够的时间，收集和标注出足够的语料，用这些语料对模型进行训练，便可以得到统计理解模型。如果需要移植到其他的领域，只需对新领域的语料进行标注，对统计模型进行训练，就可以得到新的领域的统计解析模型。

[2]和[5]是面向口语对话系统的，采用了格框架[3]作为最后的解析结果，我们所需要的是面向口语翻译的中间语义表示格式 IF，它有着不同于格框架的一些特点。为此，我们提出了统计和规则相结合的方法，来进行汉语口语到 IF 的解析。

本文的安排的是这样的：第 2 部分介绍中间语义表示格式；第 3 部分介绍解析方法，其中包括预处理，标注符号的定义，统计解析模型等；第 4 部分介绍实验结果；第 5 部分为结语。

## 2、中间语义表示格式（IF）

中间语义表示格式是一种基于中间语义关系的人造体系[7]。它能够表达旅游信息咨询这一领域内的各种对话的意思，但不包含特定语言的语法特征[4]。

IF 由四部分构成。

(1)说话人标志。用来表示谁在说话。有两种，分别是'c' 代表顾客（client）和'a' 代表代理（agent）。

(2)语句意图(Speech Act)。表示句子的类型。是询问信息或者是回答问题等等。如“give-information”表示提供了某种信息；“Pardon”表示请求重复刚才所说的内容。

(3)概念(Concept)。表示句子的主题。各个概念之间按照一定的规则可以进一步组合成更加广泛的主题。概念之间用'+'连接。如 reservation 表示预订，room 表示房间，而 reservation+room 则表示预订房间。

(4)具体信息(Arguments)。表示句子的具体内容。比如要预订的房间个数、房间标准等。具体信息由 Argument 和对应的 Value 构成。Argument 和 Value 中间用等号连接[4]。比如参数“room-spec”表示房间种类，它对应的值可以是：single 表示单间，double 表示双人间等等。比如：room-spec=single 表示原来句子中所描述的房间种类为单间。

IF 通过 IF 表达式具体代表每一句话。每一个 IF 表达式都由一个说话人标志和至少一个的语句意图以及数目可选的概念和参数组成。语句意图，概念和参数之间按照 IF 的规则可以进行组合。其组合方式如下所示：speaker: speech act+concept*(argument*)，其中'*'表示可以重复出现。

## 3、统计和规则相结合的汉语口语解析

本方法的具体思路是：首先收集大量的口语对话语料。然后对这些语料进行标注，并且用这些经过标注的语料对 HMM 的参数进行训练，从而得到统计解析模型。对于需要解析的句子，经过 HMM 的解

析出标注符号序列。最后，利用规则的方法，将标注符号序列映射成 IF 表达式。

我们的解析系统是针对旅馆预订领域的，到目前我们共收集到了该领域的语料大约 3500 句，我们从中选取了 1500 句作为训练语料，进行标注和训练模型。图 3 表示的是本系统训练、解析的过程，其中目标语言生成部分不属于本文讨论的范围。

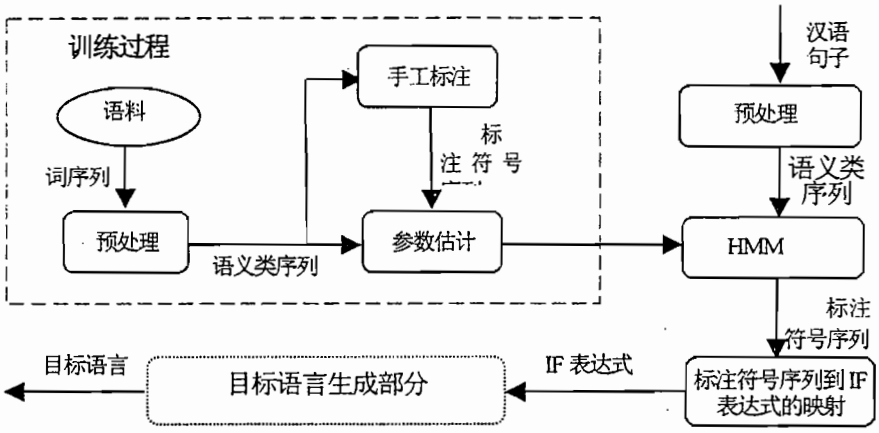


图 3　系统框架结构

## 3.1 预处理

预处理的主要功能是对词进行语义归类和语义合并，句子经过预处理以后，所有的词汇都归属到相应的语义类，预处理的输出就是语义类序列。

对于自然语言理解来说，我们所需要的最终解析结果是句子的语义表示，句子的语义表示可以是格框架[3]、语义框架[10]、语义图[11]或者是中间语义表示格式 IF。句子和语义表示之间的关系不是一一对应的。但是，类似的句子结构，往往有着类似的语义表示。例如下面两个句子，就有着相同的中间语义表示格式（句子后面的部分是它所对应的中间语义表示格式）：

我想预定一个单人间——give-information+reservation+room(room-spec(single,quanlity=1))

我想预定一个双人间——give-information+reservation+room(room-spec(double,quanlity=1))

由于我们是通过统计模型来进行解析的，而统计模型是通过训练得到的。假如在训练的过程中，“单人间”这个词汇得到了训练，而“双人间”这个词汇没有得到训练，那么，统计模型就不能够正确解析出第二个句子。类似这样的例子还有很多。我们的语料总是有限的，不可能把所有能够遇到的词汇都训练一遍，解决这个问题的方法是对词汇进行语义归类。词汇的语义归类没有一个共同的标准，总的原则是按照词在句子中的语义功能进行划分。在本系统中，我们将所有的词汇共划分为 230 个语义类。

另外，对于数字我们也做了一定的处理。在汉语中，数字的表达方式十分的灵活，数字的表达往往需要一个一个或多个词组成一个词组。例如：“一千一百二十”，这个词组由五个词组成。但是，作为一个词组，它在句子中的语义功能和“五千三百”以及“二十”在句子中所起到的语义功能是没有区别的。为此，有必要对表达数字的词汇进行语义合并。我们采用了一些简单的规则，对数字进行合并，无论是“一千一百二十”或者是“五千三百”，经过合并以后，都只是一个数字，我们用一种语义类来代表它们，在训练或者解析的时候，把它们都看作是一个语义类。

## 3.2 标注符号的定义

对语料进行标注，是建立统计模型的第一步。经过预处理后，句子由词序列变成语义类序列，对应于句子中的每一个语义类，都应当标注出它在最后的语义表示中担当的角色。我们要求的最终理解目标

是中间语义表示格式 IF，所以希望解析的结果能够很方便的映射到 IF。因此，根据 IF 的特点，我们从 IF 中抽出一些 Speech Act， Concepts 和 Arguments 作为标注的符号，另外结合汉语口语的的特点和 IF 的规则，又定义了一些符号。主要有以下三类：

(1) 代表 IF 中的 Concept 或者 Argument。如 who,room-spec 等。

(2) 代表 IF 中某种 Concept 或者 Argument 的标志。如 family-name=,room-spec=等。

(3) 代表 IF 中的多个 Concept 或者 Argument。在汉语中，某些词汇的含义需要用多个 IF 中的 Concept 或者 Argument 才能表示。比如“过来”表示“来”这个动作，同时又表示目的地是“这儿”。因此，我们定义了符号+trip:destination=here 来对这个词汇进行标注。类似的还有 price:quantity 等。

## 3.3 统计解析模型HMM

每一个句子都有他的语义，每一个语义都通过一定的句子来体现，句子和它的语义之间的关系可以用隐马尔可夫模型来体现。句子中的词汇相当于 HMM 的观察状态，而句子的语义相当于 HMM 的内部状态。HMM 的参数可以通过标注的语料进行训练。

在我们的系统中，我们采用了一个各态遍历的二阶 HMM 作为解析模型，它允许所有标注符号之间能够相互转换。预处理的输出语义类序列作为 HMM 的输入也就是观察状态，而标注符号序列作为 HMM 的输出，也就是内部状态。我们标注了 1500 句语料，涉及到的标注符号有 171 个，词的语义类有 190 个。

下面我们给出一个隐马尔可夫模型的解析的例子。

需要解析的汉语句子：“你好，你们还有没有房间？”。

预处理的结果为：“GREET P_PEOPLE advOther V_Q_AVAILABILITY N_C_ROOM”。

HMM 解析的结果：“greeting who nul +availability +room”。

## 3.4 从标注符号序列到IF的映射

HMM 的输出是标注符号序列，而我们解析的目标是 IF 表达式。IF 表达式有着严格的形式，为了将标注符号转换成 IF，我们采用了规则的方法。通过分析 IF 文档和统计解析的结果，从中总结了 60 多条规则，然后用这些规则分析标注符号序列并将他们映射到 IF 表达式，试验证明这个方法是行之有效的。下面我们给出一个 IF 表达式映射的例子。

汉语句子：“您好！您这还有没有房间？”，经过HMM 的解析之后，对应的标注符号序列为：*greeting=begin, availability=question, room.* 系统利用了三条规则，将这个语义序列映射成 IF 的表达式，这三条规则是：

*greet=?→greeting(greeting=?)* (1)

*?=question→request-information+?* (2)

*availability, room→availability+room* (3)

“？”表示该条规则对应的位置可以是任何任何词汇。当系统分析语义序列的时候，它首先会发现 *greeting=begin* 和规则(1)匹配，因此用 *greeting(greeting=begin)* 替换 *greeting=begin*，接着会发现 availability=question 和规则(2)匹配，相应的用 *request-information+availability* 代替 availability=question，最后发现 *availability,room* 和规则(3) 匹配，相应的用 availability+room 替换 *availability,room*，最后生成了 IF 表达式：

greeting(greeting=begin), request-information+availability+room。

下表给出了一些标注符号到 IF 表达式映射的例子。.

| 标注符号序列 | IF 表达式片段 |
| --- | --- |
| quantity=1, time-unit=day | duration=(quantity=1,time-unit= day) |
| who=I, disposition=desire | disposition=(desire,who=I) |
| room-spec= double, or, room-spec=single | room-spec= (operator=disjunct,[double,single]) |

表 1 标注符号序列和对应的 IF 表达式片段

## 4、试验结果

我们的口语解析系统是面向旅馆预订领域的。训练语料一共有 1500 句，我们用了 200 句语料对这个系统进行测试，测试结果如表 2 所示：从句子解析到标注符号，单个标注符号的正确率为 91.2%，稍微低于[2]中的方法，这是因为我们的训练语料还不够多所造成的。而从句子生成 IF 表达式的正确率为 79.2%，正确率高于方法[2]和[5]中的最终理解结果。

| 不同的方法 | 单个标注符号 | 中间语义表示格式 |
| --- | --- | --- |
| [2]中的方法 | 91.4% | 52.8% |
| [5]中的方法 | — | 72.0% |
| 本文的方法 | 91.2% | 79.2% |

表 2：几种方法在解析的不同阶段的正确率比较

从试验结果我们可以看出，本方法有如下一些优点：

(1) 增加了词汇的语义类数目。我们定义了更多的语义类，这样，更多词汇之间的语法、语义差别可以体现出来。

(2) 增加了标注符号的数目。我们根据汉语和 IF 的特点，定义了更多的标注符号，使用这些标注符号能够更详细的体现原来句子的意思。

(3) 规则的使用。从标注符号序列到 IF 表达式的映射，采用了规则的方法，这些规则保证了标注符号序列能够比较好的映射到 IF 表达式。

## 5、结语

我们利用统计和规则相结合的方法实现面向 IF 的汉语口语解析过程。在解析的不同阶段分别采用了统计的方法和规则的方法。该方法具有较高的鲁棒性。试验证明，这个方法是行之有效的。下一步，我们准备在一下几个方面进行改进：(1) 增加训练语料。(2) 采用统计与规则相结合的方法，提高 HMM 解析的正确率。

参 考 文 献

[1] 翁富良,王野翊. 计算语言学导论.中国社会科学出版社 1998

[2] W.Minker, S.Bennacef. A stochastic Case Approach for Natural Language Understanding. Proc. ICSLP ,1996

[3] B.Bruce. Case Systems for Natural Language. Artificial Intelligence. 1975. Vol.6,327-360.

[4] Lori Levin, Donna Gates. An Interlingua Based on Domain Actions for Machine Translation of

Task-Oriented Dialogues. Proc. ICSLP,1998

[5] Yunbin.Deng, Bo Xu. Chinese Spoken Language Understanding Across Domain. Proc ICSLP, Oct, 2000. 1:230-234

[6] 宗成庆,吴华等. 限定领域汉语口语对话语料分析. 全国第五届计算语言联合学术会议论文集. 1999.115~122

[7] 吴华,黄泰翼. 基于中间语义框架的系统响应生成.全国第五届计算语言联合学术会议论文集.1999.248~255

[8] Jun Park, Jae-Woo Yang, ETRI Speech Translation System, C-STAR Workshop, Schwetzingen, 1999

[9] 邓云滨. 口语对话系统领域移植—统计语言理解. 硕士研究生学位论文. 中科院 自动化所北京. 2000.07

[10] S.K. Bennacef, H.Bonnea-Maynard. A Spoken Language System for Information Retrieval. Proc ICSLP,1994.1271-1274

[11] M.Bates, R.Bobrow, R.Ingria. Advances in BBN's Spoken Language System. Proceedings of the Spoken Language Technology Workshop, Mar, 1994. 43-47.

[12] Chengqing Zong, Hua Wu. Analysis on Characteristics of Chinese Spoken Language, Proc. of 5th Natural Language Processing Pacific Rim Symposium, 1999. 358-362.

作者简介： 解国栋（1975—），男，陕西西安人，博士生，讲师，主要研究领域为口语理解；宗成庆（1963—），男，山东人，博士，副研究员，硕士生导师，主要研究领域为自然语言处理，语料库语言学与语料库建设，语音翻译技术等；徐波（1966—）男，浙江宁波人，博士，研究员，博士生导师，主要研究领域为语音识别，口语理解，知识挖掘等。

# Chinese Spoken Language Analyzing Facing the Middle Semantic Representation *

*Guodong Xie, Chengqing Zong, Bo Xu*

*National Laboratory of Pattern Recognition, Institute of Automation*
Chinese Academy of Sciences, Beijing, 100080
E-mail:{gdxie,cqzong,xubo}@nlpr.ia.ac.cn tel:(010)82614468

**Abstract:** Spoken language analyzing is a crucial part in human-machine dialog system and spoken language translation system. In this paper we present a Chinese spoken language analyzing method based on the combination of statistical and rule methods The analyzing result is a middle semantic representation. It has two stages, first, use the statistical method to analyzing the semantic information, then use the rule method to map the semantic information to the middle semantic representation. This method avoids the shortcoming of the rule and has high robustness, at the same time it achieves a lower error rate.

**Key words**: spoken language analyzing; statistical analyzing model; middle semantic representation (IF)

* Supported by the National Natural Science Foundation of China under Grant No.69835003 and 60175012