DOI: 10.16511/j.cnki.qhdxxb.2001.01.018
ISSN 1000-0054 清华大学学报 (自然科学版) 2001年 第 41卷 第 1期 18/32
CN 11-2223/N J Tsinghua Univ (Sci & Tech), 2001, Vol.41, No.1 69-72

口语对话系统中的词类概率模型和知识表示

燕鹏举, 郑 方

(清华大学 计算机科学与技术系, 北京 100084)

摘 要: 语言分析和知识库管理是口语理解与对话系统的两个重要组成部分,作者在这两方面提出了一些新的方法。一是提出并实现了词类概率模型,它具有较高的性能和较低的时间复杂度,是基于句法规则的语义分析和语言理解的基础。此外还提出了与数据无关的多叉树层次结构模型的知识表示方法,它具有很强的表达能力并易于扩展。在此基础上,实现了一个用以提供清华大学地理、办公、商业及其它一些相关信息检索、基于文本的口语对话系统 EasyNav。实验表明,上述模型和方法具有很好的性能。

关键词: 口语对话系统; 自然语言理解; 词类概率模型; 知识表示

中图分类号: TP 391.42 文献标识码: A

文章编号: 1000-0054(2001)01-0069-04

Word-class stochastic model and knowledge representation in a spoken language dialogue system

YAN Pengju, ZHENG Fang

(Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract Language analysis (parsing) and knowledge library management are the most significant parts of a spoken language understanding and dialogue system. Some new approaches to language analysis and knowledge library management are presented. The proposed Word-Class Stochastic Model (WCSM) is the basis for syntax rule based semantic parsing and spoken language understanding which has better performance. The data-independent multi-branch tree hierarchical structure is proposed for the knowledge representation, because it has strong expressing ability and can be easily expanded. EasyNav, a text-based dialogue system based on these concepts, is designed and implemented to provide users with Tsinghua University campus navigation information. Experiments on EasyNav give satisfactory performances with these two models.

Key words spoken language dialogue system; natural language understanding; word-class stochastic model; knowledge representation

自然口语对话系统是计算语音学及计算语言学研究领域内一个新兴的研究课题,它将语音识别、口语语言理解、知识表示、对话管理以及语音合成等各项技术集成起来,使各项技术整合互动,达成人机通过自然对话进行信息交流的目的。

目前已有几个效果不错的口语对话系统。LOADSTAR[1]是一个旅游信息检索系统,它采用了基于词和词类的混合语言模型,利用语义要素检出和概念组装模型来获取语义; ARISE[2]是一个欧洲的自动铁路信息系统研究计划, Els den Os等人对几个不同语种、不同应用背景的系统原型做了比较,着重论述了不同对话策略的特点和对系统的影响; 其它还有用于自动电话接线的 PADIS[3]系统以及用于航班信息服务的 ATIS[4]系统。这些系统在语料库、声学模型、语言模型、自然语言理解和对话管理等各方面均有不同的侧重和设计思路。这里介绍的 EasyNav系统,着重于将一般句法知识与领域相关的关键词知识结合起来的语言理解方法,以及结构化的知识表示方法,两者在基于文本的工作模式下表现了良好的性能。

1 校园导航系统

"校园导航系统 EasyNav"是语音实验室为研究自然口语对话系统而开发的一个研究平台,该平台以基于文本的清华大学校园信息查询为基本功能,提供语言理解和知识表示与信息查询的研究环境。系统由分词、词类标注、句法分析、句型匹配、查询表单填充、数据库查询和对话生成等 7个关键部件所组成,其运行所依赖的模型和数据库包括统计语言模型、词类概率模型、关键词库和导航知识库。

收稿日期: 1999-12-23
作者简介: 燕鹏举(1974-),男(汉),江苏,博士研究生

2 词类概率模型

词类模型是句法分析的前端处理,其性能的好坏直接影响到语言分析的性能。

在连续语音识别中,通常使用基于词 m-Gram 的统计语言模型。使用此模型,N 个词 $w_1, w_2, \cdots, w_N$ 组成的长度为 N 的词串(即句子)$W_1^N = (w_1, w_2, \cdots, w_N)$ 的出现概率用

$$P(W_1^N) = P(w_1)\prod_{n=2}^{N} P(w_n | W_1^{n-1}) = P(w_1)\prod_{n=2}^{N} P(w_n | W_{n-m+1}^{n-1}) \quad (1)$$

来估计,当 $m = 1, 2, 3$ 时这样的模型和方法分别称为 Uni-Gram, Bi-Gram 和 Tri-Gram。经验表明,Bi-Gram 或 Tri-Gram 的依赖度足以处理统计语言模型中的问题。

在进行词类标注时,将词和词类随时间变化的过程看成是一个隐马尔可夫过程,词类视为状态,词视为输出,即假设: a) 当前词类出现的概率只跟前 $m-1$ 个时刻的词类相关,而与更以前时刻的词类无关; b) 当前词出现的概率只与当前词类相关,而与以前时刻的词类和词无关。在此假设下,词串 W_1^N 及其相应的词类串 $C_1^N = (c_1, c_2, \cdots, c_N)$ 的同现概率可以用词类 m-Gram 以及词类属概率估计如下:

$$\begin{aligned} P(W_1^N, C_1^N) &= P(w_1, c_1)\prod_{n=2}^{N} P(w_n, c_n | W_1^{n-1}, C_1^{n-1}) = \\ P(c_1)P(w_1 | c_1)\prod_{n=2}^{N} P(c_n | W_1^{n-1}, C_1^{n-1})P(w_n | W_1^{n-1}, C_1^n) &= \\ P(c_1)P(w_1 | c_1)\prod_{n=2}^{N} P(c_n | C_{n-m+1}^{n-1})P(w_n | c_n). \end{aligned} \quad (2)$$

可见在我们的词类模型中,词类 m-Gram 和词类属概率构成两类重要的模型参数。

2.1 模型训练

在缺乏充足的领域标注语料的情况下,模型的训练使用基于 50 624 词和 91 词类的通用标注语料[5]。词类概率模型的训练(估计)公式可表达为:

$$\begin{cases} C(R,S,T) = \sum_{A,B,C:\ [(A,B,C),(R,S,T)]} C(A,B,C), \\ C(R,S) = \sum_{A,B,C,T:\ [(A,B,C),(R,S,T)]} C(A,B,C), \\ C(R) = \sum_{A,B,C,S,T:\ [(A,B,C),(R,S,T)]} C(A,B,C), \\ C(A,R) = \sum_{B,C,S,T:\ [(A,B,C),(R,S,T)]} C(A,B,C). \end{cases} \quad (3)$$

$$\begin{cases} P(T|R,S) = C(R,S,T)/C(R,S), \\ P(S|R) = C(R,S)/C(R), \\ P(R) = C(R)\sum_{R_i} C(R_i), \\ P(A/R) = C(A,R)/C(R). \end{cases} \quad (4)$$

其中: $C(\cdot)$ 表示事件频度; $P(\cdot)$ 表示事件概率; A, B, C 表示词; R, S, T 表示相应的词类。当前时刻对应的"词和词类对"为 C/T,前一时刻对应 B/S,再前一时刻对应 A/R; $[(A,B,C), (R,S,T)]$ 表示由词三元组 (A,B,C) 和词类三元组 (R,S,T) 组成的同现标注。

从式(3)中可以看出:

$$\begin{cases} C(R,S) = \sum_{T} C(R,S,T), \\ C(R) = \sum_{S} C(R,S) = \sum_{S,T} C(R,S,T), \\ C(R) = \sum_{A} C(A,R). \end{cases} \quad (5)$$

因之式(4)满足归一化要求。

本模型的统计方法需要的数据量比较小,计算简单。

2.2 词类标注算法

词类标注的目的是对词串中的每个词标注上最大似然词类,选择式(6)作为词类标注准则

$$\tilde{C}_1^N = \arg\max_{c_1^N} P(C_1^N | W_1^N). \quad (6)$$

在词串已知的前提下,式(6)可以改写为

$$\begin{aligned} \tilde{C}_1^N &= \arg\max_{c_n^N} P(C_1^N | W_1^N) = \\ &\arg\max_{c_n^N} P(W_1^N, C_1^N)/P(W_1^N) = \\ &\arg\max_{c_1^N} P(W_1^N, C_1^N). \end{aligned} \quad (7)$$

延用式(2)的假设,设 $m = 2$(Bi-Gram),作符号替换 $T(c_1) = P(c_1)$, $p_a(c_n | c_{n-1}) = P(c_n | c_{n-1})$, $p_b(w_n | c_n) = P(w_n | c_n)$,上式改写为

$$\tilde{C}_1^N = \arg\max_{c_1^N} T(c_1)p_b(w_1 | c_1) \cdot \prod_{N=2}^{N} [p_a(c_n | c_{n-1})p_b(w_n | c_n)]. \quad (8)$$

词类标注可以看成是在"词串-词类串"网格中的搜索过程,搜索网格如图 1 所示,图中某个 n ($n = 1, 2, \cdots, N^{-1}, N$)对应第 n 个词 w_n,节点 $c_{n,i}$ 表示词 w_n 可能属于的第 i 个词类。使用 HMM 中 Viterbi 解码算法,递归表达式描述如下:

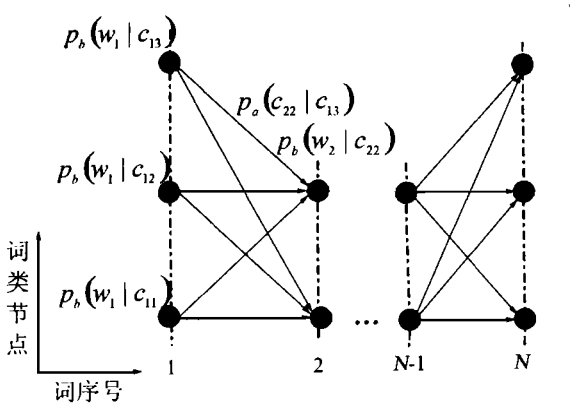

图 1 搜索网格

$$\begin{cases} s_{1i} = \pi(c_{1i})p_b(w_1|c_i), \quad i = 1,\cdots,M, \\ s_{ni} = \max\limits_{1\le j\le M_{n-1}}[s_{n-1,j}p_a(c_{ni}|c_{n-1,j})p_b(w_n|c_{ni})], \\ \qquad n = 2,\cdots,N;\ i = 1,\cdots,M_n, \\ B_{ni} = \arg\max\limits_{1\le j\le M_{n-1}}[s_{n-1,j}p_a(c_{ni}|c_{n-1,j})p_b(w_n|c_{ni})], \\ \qquad n = 2,\cdots,N;\ i = 1,\cdots,M_n, \end{cases} \tag{9}$$

其中: s_{ni}为节点 c_{ni}处的局部最大得分, B_{ni}指向使当前节点取最大得分的前一节点, M_n为词 w_n可能属于的词类的个数。用下式来回溯最佳路径:

$$\begin{cases} c_N = \arg\max\limits_i s_{Ni}, \\ c_n = B_{n+1,c_{n+1}}, \ n = N-1,\cdots,1. \end{cases} \tag{10}$$

如句子中各词所属于的可能词类数最大为 M, 从式(9)及式(10)可看出, Viterbi解码算法的时间复杂度为 $O(NM^2)$。

有些情况下,要求给出前 n 名路径,这就是 N-Best问题。Viterbi算法可以说是 N-Best搜索算法的特殊情形,因为后者在每一个节点处保留前 n 名分数和相应的 n 条前指指针, $n\geq 1$。N-Best算法的时间复杂度为 $O[n^2NM^3\cdot \mathrm{lb}(nM)]$。

在领域特定的本系统中, N 一般不大于 10, M 一般不大于 3,在保证搜索速度比较快的前提下进行了全搜索(保留所有局部分数),这同时也保证了标注的准确性。特定领域内关键词的词类有唯一标注,这使搜索速度和标注准确性得到了进一步提高。

3 知识库管理

知识表示是对话系统的重要组成部分,它是领域知识的表达手段,为对话系统提供领域数据检索支持。在 EasyNav系统中,知识库的构建基于一种与数据无关的多叉树层次结构模型。

3.1 数据结构

根据语料,可以归纳出模型框架必须处理的数据。下面以句子"周六医院有牙科门诊吗?"为例说明归纳思路。大多数查询语句的中心均围绕特定的主体 sites("医院"); site信息分为一些大类 items("牙科"), item有时含有 sub-items; 最底层为细节信息 attributes("开放时间"); attribute必须有具体内容 attribute-values("周六")。

基于以上分析,设计了如图 2所示的数据结构。该数据结构有以下几个特点:

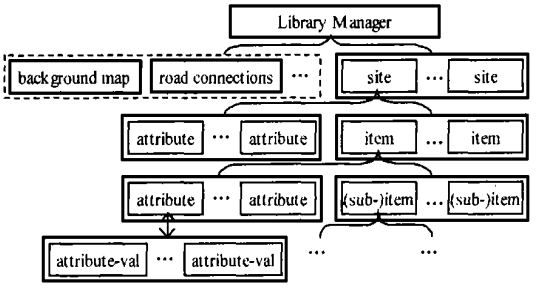

图 2 知识库的数据结构

1) 递归的多叉树结构,其广度和深度均可随数据的需要而扩展;

2) 层次结构,各层次对象有不同的表达功能,一条从根节点到叶子节点的路径(下称为贯通路径)就是一条现实信息实例。

3) 它是一种与数据无关的结构,数据的含义和相互间的关系不由数据结构确定。

本结构模型的一般性,使其具有良好的可扩展性和可移植性,随着语言理解模块性能的增强,知识库也可以方便地增强。

3.2 知识库的构建-KLib子系统

知识库由模型框架和数据含义及其相互间的关系所组成。子系统 KLib被用来对知识库的维护作独立管理。知识库的构建过程中,图 2模型框架中各层次对象的实例被赋于领域内特定含义,数据间的相互关系由附属于本位节点的属性(子)节点确定。

- Library Manager作为根节点,除了以 site链表表示的核心数据外,还包含 background map, road connections等部件。
- site对象分为 Building, RoadCross, Sight及 Agency等 4类,除 ID及场所名等固定属性外,不同 site对象所含数据不尽相同。
- item对象间的包含关系上可以是递归的,以便表达诸如下面的复杂信息:

"校医院(site)—牙科(item)—大夫(item)—水平(attribute)—好(attribute-value)"。

- attribute统一地标识一类数据的类型。相同

标识的数据用一个 attribute-value链来表达。

° attribute-value对象是信息的最终载体。

3.3 查询的语义表示和知识库查询

知识库查询模块利用模型框架提供的元操作接口进行组合条件查询。大多数的查询请求均围绕着知识树中的一条贯通路径而提出,也即提供贯通路径中的某些节点,要求得到其余节点的信息。

如表 1所示,查询表单被设计作为查询请求的语义表示,由上层模块根据关键词信息和句法信息装配得到。查询表单分为 where, what, which, how-to等 10种类型,分别对应"哪儿"、"什么"、"哪一个"及"怎么走"等各种不同的查询请求。其中大多数域对应贯通路径中的节点; pItemArray/uItemNum提供语义理解扩展的可能性。

表 1 查询表单结构

QueryForm中的域	注释
QueryType	疑问词,如 where
AddOnInfo	疑问词附带信息,如 nearby
ReferrenceName	参照体
FocusName	查询体
FocusCategory	查询体类别
Action	行为及动作
SubObj	行为及动作的施事/受事
AttrName	属性名
AttrVal	属性值
OriVal	方向值
MarchMode	行进方式
ItemNum	上下位项目个数
ItemArray	上下位项目链
MePos	用户所在位置

对话生成任务由句型匹配和知识库查询模块合作完成,查询的过程也是对查询表单进行逐步理解和解释的过程。

4 总 结

EasyNav系统构建达成后,我们进行了几组测试,从测试结果看,系统达到了我们的设计要求,即对于符合句型模板的查询句子,系统均能作出正确的理解; 在知识库中找到相应信息的情况下,均能给出正确的应答。下面是 2个典型的例子:

例 1 用户: 大礼堂在哪儿?

系统: 大礼堂在您的西北,自清亭的东北。

例 2 用户: 我想买皮鞋,该怎么走?

系统: 您可以去照澜院商场。先向西边走大约 536 m到二校门,再向南边走大约 129 m到邮局,就到了。

实验表明,基于句法规则的语言分析器,由于其一般性(其解析的粒度为与领域无关的词类),可以得到较为精细的分析结果(句法树),这是别的方法(比如统计方法)不易做到的。关键词知识在词类标注和几个后续部件中的应用,在本系统中表现了良好的性能: 对于符合预定义句型的句子均能作出正确(且为首选)标注,理解模块的处理也变得更简便。

与数据无关的多叉树层次结构模型,具有一般性、良好的可扩展性和良好的可移植性。一般性造成缺乏针对性,表现为缺乏对领域内数据含义和相互间关系的理解机制; 但在需求明晰化的情况下,数据理解可以纳入到系统中,作为外围模块,或成为知识库架构中的一部分。

参考文献 (References)

[1] HUANG Chao, XU Peng, ZHANG Xin, et al. LOADSTAR: A mandarin spoken dialogue system for travel information retrieval [A]. Olaszy G, Németh G, Erdohegyi. EuroSpeech 99 Proceedings Vol 3 [C]. Budapest: 1999. 1159-1162.

[2] Elsden Os, Lou Boves, Lori Lamel, et al. Overview of the ARISE project [A]. Olaszy G, Németh G, Erdohegyi. EuroSpeech 99 Proceedings Vol 4 [C]. Budapest: 1999. 1527-1530.

[3] Kellner A, Ruber B, Seide F, et al. PADIS-An automatic telephone switchboard and directory information system [J]. Speech Communication, 1997, 23: 95-111.

[4] Epstein M, Papineni K, Roukos S, et al. Statistical natural language understanding using hidden clumpings [A]. IEEE International Conference on Acoustics, Speech, and Signal Processing Vol 1 [C]. Atlanta: 1996. 176-179.

[5] ZHENG Fang, SONG Zhanjiang, XU Mingxing, et al. EasyTalk: A large-vocabulary speaker-independent Chinese dictation machine [A]. Olaszy G, Németh G, Erdohegyi. EuroSpeech 99 Proceedings Vol 2 [C]. Budapest: 1999. 819-822.