文章编号:1007-5321(2006)增-0075-04

领域语义语法的统计生成

刘建毅[1,2], 王菁华[1], 王 枞[1]

(1. 北京邮电大学 信息工程学院, 北京 100876; 2. 北京师范大学 中文信息处理研究所, 北京 100875)

摘要: 提出了一个基于统计的从未标注语料库中半自动获取语义语法算法. 该算法对特定领域的语料库进行反复的时间聚类和空间聚类,通过时间聚类发现语言片段的语法结构;通过空间聚类发现语言片段的语义类别;循环迭代,可以生成一个粗糙的文法. 最后,将这些抽取出来的粗糙文法经过人工校对,得到新领域的语义语法. 实验结果表明了该算法是有效和切实可行的.

关 键 词: 对话系统; 语义语法; K-L 距离; 互信息

中图分类号: TP929.53 **文献标识码**: A

Statistical Acquisition of Domain-Specific Semantic Grammar

LIU Jian-yi[1,2], WANG Jing-hua[1], WANG Cong[1]

(1. School of Information Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China;
2. Graduate School of Chinese Information Processing, Beijing Normal University, Beijing 100875, China)

Abstract: An approach for semiautomatic grammar acquisition from un-annotated corpus about a specific domain is presented. Its grammar is produced by an iterative procedure, it spatially and temporally clusters the words from a domain-specific corpus. Temporal clustering can discover the fragment's syntactic structure. Spatial clustering can discover the fragment's semantic category. Finally, the resultant grammar is post-processed by hand-editing. The preliminary experimental result shows that the method is effective and practical.

Key words: dialog system; semantic grammar; kullback-leibler divergence; mutual information

随着信息时代的来临,人们越来越希望能采用更加自然的方式进行信息获取、信息显示和信息交流, 使设备具有更加人性化的界面. 一种有效的方法就是让人们采用自然语言和机器交流,即人机口语对话系统(spoken dialogue systems). 人机口语对话系统的主要目的是在用户和计算机之间提供一种交互界面,使用户以语音这种非常自然的方式与计算机进行交流, 进而完成用户的任务. 为了实现此目的,口语对话系统必须要理解用户输入的语音. 口语理解的任务就是分析语音识别的输出,并从中抽取语义表示,可以说口语理解性能的好坏对口语对话系统有关键性的影响.

一般而言,语言理解包含2个过程,一是句法分析,分析语音识别输出串的成分结构,即这些词是如何结合在一起的;二是语义分析,即判断句子各个成分的意思,并根据成分之间的关系推断出整个句子的意思. 但是对口语来说, 口语的特性使得口语理解与一般的语言理解有所不同. 其中一个主

收稿日期: 2006-09-12

基金项目: 国家自然科学基金项目(60575034)

作者简介: 刘建毅(1980—), 男, 讲师, E-mail: liujy@nlu.caai.cn.

要原因是,口语和标准语言学所定义的结构合法的句子相去甚远. 一方面, 口语本身包含许多即兴口语现象,如错误修正、重复,以及不合乎语法的成分的组合;另一方面,口语对话系统中的语音识别模块很难生成合乎语法的句子. 所以在语法分析的基础上进行语义分析是非常不现实的. 因此, 目前的口语理解算法只涉及到一点甚至根本不涉及到语法分析,更多的是直接从识别串中抽取语义表示.

在这些算法中,应用最为广泛的是基于语义语法(semantic grammar)的口语理解. 基于语义语法的口语理解只考虑句子中有意义的部分,而忽略其中的随机口语现象,非常适合以语义分析为目的的口语分析. 但语义语法都是面向特定领域、特定任务的,其编写需要由了解领域特点、具备语言知识的领域专家完成,特别是其构造的语义语法基本上不能推广到其他领域,新领域需要构造新的语义语法,这就成了制约对话系统发展的主要瓶颈. 此外,领域专家编写的语义语法在某种程度上,也很难保证很好地覆盖现实中的自然语言现象. 因此, 针对特定领域自动或半自动抽取语义语法非常重要,本文采用一种基于统计的语义语法半自动生成算法.

1 语义语法

与一般的上下文无关语法不同,语义语法是一种以特定领域内语义范畴为结构单元的语法,它把特定领域内的语义属性作为语法的非终结符,而词作为语法的终结符[1-2]. 在为特定领域构造自然语言应用系统时,通常可以利用一些很强的约束技术提高句法、语义分析的性能. 例如,虽然一般的自然语言句子结构需要非常复杂的语法系统才能得到比较完备的覆盖,但是在特定应用中,人们可能只会用到自然语言中很小的一部分结构,而且句子结构中的每个成分可能都有十分明确的语义约束. 这样在构造语法时,语法中非终结符和产生式规则的选择就取决于特定领域的语义和文法功能. 如一般上下文中的非终结符 N(noun),在表示航班查询的语义语法中,就可能成了表示出发地、到达地、时间等语义的非终结符 DEPART_LOC、ARRIVE_LOC、TIME等. 在编写好语义语法后,便可以利用语义语法对句子进行语义分析,其分析与一般的句法分析算法相同,如自顶向下、自底向上等句法分析算法,CMU Communicator 的 Phoenix Parser 就采用了自顶向下的递归转移网络的图句法分析算法. 基于语义语法的语义分析只考虑句子中有意义的部分,忽略了其中的随机口语现象,在分析充满噪音现象的口语句子时取得了良好的效果. 但是该方法也存在2个问题,首先,构建一个完备的文法比较困难,需要由领域专家编写,有时甚至需要收集大量的实际语料,通过分析语料完成文法,以使文法尽量覆盖现实的语言现象;其次,该方法的灵活性不够,性能的好坏完全取决于语法的覆盖度,只要句子中有一点与所有的文法都不匹配,就无法得到这个句子的语义信息.

可见,为特定领域编写语义语法已经成为开发对话系统的主要障碍. 文献[3]开发了一个快速开发混合对话系统的辅助工具,但未能解决文法生成的问题. 文献[4]开发了一个修改语法的辅助工具,允许用户修改对话系统的现有语法,但仍然要求系统有初始语法为基础,而且用户也应该具备很好的语言学知识. 近年来已有不少研究者开始研究从语料库中自动或半自动获取语义语法. 文献[5]综合利用了多种先验信息从标注语义的语料库中半自动地抽取语义语法,实验表明抽取的语法比人工编写的语法获得了更好的效果. 采用表示时间、数字、地点等与领域无关的概念的语义语法库,缩短了语法生成的时间;利用特定领域内语言的语法限制关系, 减小了语法生成的搜索空间. 多种数据驱动方法也被用来从语料库中自动获取语义类别和语法结构. 文献[6]采用模拟退火算法自动获取词的语义类,文献[7]采用互信息选取语法结构. 文献[8-9]针对 How may I help you? (HMIHY)对话系统,提出了自动获取关键语法片段的框架,该框架利用 K-L 距离从语料库中自动获取关键语法片段.

2 语义语法的统计生成

2.1 算法描述

本文利用统计方法试图从一个未标注语料库中半自动获取语义语法, 其算法流程如图 1 所示. 算法的输入是关于某一特定应用领域 (如公交、天气、餐饮等)的语料库. 首先对语料进行分词、实体标注、断句等预处理,然后对语料进行反复的时间聚类和空间聚类. 在时间聚类中, 通过计算语料中2个连续词序列的互信息值, 将互信息值最高的词序列聚为一类, 这些词序列往往是常用的短语,将这些类称为语法结构类,记为 PCi,然后用 PCi 替换语料中的这些词. 在空间聚类中, 利用 K-L 距离计

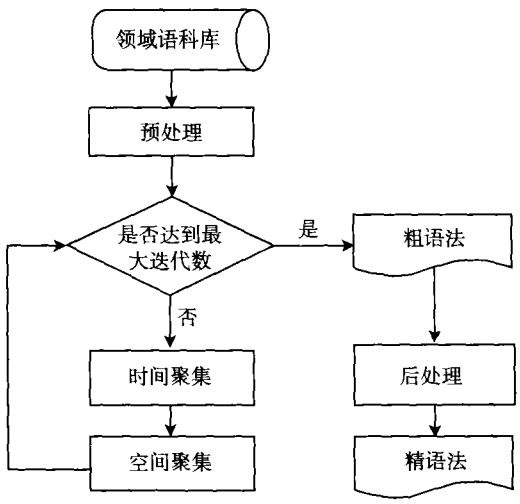

图 1　语法生成算法

算语料中任意 2 个词或短语分布相似程度，将相似的单元聚为一类，这些单元往往具有相似的语义，把这些类称作语义类，将其标记为 SCi，然后用 SCi 替换语料中的这些词. 当到达预设的最大迭代数时，就生成了粗糙文法. 最后，将粗糙文法经过人工后处理，得到该领域的精语法.

2.2　时间聚类

如果词序列在语料库中共现的次数越多，则该词序列越有可能是短语. 为此，采用目前比较常用的互信息衡量 2 个连续词或短语的共现度，互信息近似计算为

$$\mathrm{MI}(e_1,e_2)=\frac{f(e_1,e_2)}{f(e_1)+f(e_2)-f(e_1,e_2)}$$

式中，$f(e_1)$、$f(e_2)$分别是 e_1 和 e_2 在语料库中出现的频率；$f(e_1,e_2)$是序列(e_1,e_2)在语料库中出现的频率. MI 值越大，说明 e_1 和 e_2 越可能成为短语.

2.3　空间聚类

如果 2 个词或短语出现的上下文环境越相似，则这 2 个词或短语从语义的角度上越近似. 因此，采用 K-L 距离(kullback-leibler divergence)计算 2 个词或短语的上下文分布之间的距离，作为估计 2 个词或短语相似程度的量度[10-11]. 可以表示为

$$D(p_1\,\|\,p_2)=\sum_{i=1}^{V}p_1(i)\log\frac{p_1(i)}{p_2(i)}$$

式中，p_1 表示一个词或短语 e_1 的上下文分布；p_2 表示另一个词或短语 e_2 的上下文分布；V 表示所有出现在 e_1 和 e_2 上下文中的词汇集合. 因为 K-L 距离是不对称的，为了获得对称的距离量度，故将距离量度表示为

$$\mathrm{Div}(p_1,p_2)=D(p_1\,\|\,p_2)+D(p_2\,\|\,p_1)$$

因此，e_1 和 e_2 的距离可以表示为

$$\mathrm{Dist}(e_1,e_2)=\mathrm{Div}(p_1^{\mathrm{left}},p_2^{\mathrm{left}})+\mathrm{Div}(p_1^{\mathrm{right}},p_2^{\mathrm{right}})$$

式中，p_1^{left}和 p_2^{left}分别为 e_1 和 e_2 的左上下文分布；p_1^{right}和 p_2^{right}分别为 e_1 和 e_2 的右上下文分布.

e_j 的左上下文分布，可以表示为

$$p(S_i^{t-1}\,|\,e_j^{t})=\frac{C(S_i^{t-1}e_j^{t})}{C(e_j)}=\frac{C(w_1^{t-N_c}\cdots w_{N_{c-1}}^{t-2}w_{N_c}^{t-1}e_j^{t})}{C(e_j)}$$

式中，S_i 表示 e_j 的第 i 个左上下文；w_k 表示 S_i的第 k 个词；N_c 表示 S_i 中的词数；$C(\,)$表示序列在语料库中的出现次数. 因此，e_1 和 e_2 的左上下文分布的 K-L 距离为

$$D(p_1^{\mathrm{left}}\,\|\,p_2^{\mathrm{left}})=\sum_{\forall s_i\in S}p_1(S_i^{t-1}\,|\,e_1^{t})\log\frac{p_1(S_i^{t-1}\,|\,e_1^{t})}{p_1(S_i^{t-1}\,|\,e_2^{t})}$$

e_j 的右上下文分布，可以表示为

$$p(S_i^{t+1}\,|\,e_j^{t})=\frac{C(e_j^{t}S_i^{t+1})}{C(e_j)}=\frac{C(e_j^{t}w_1^{t+1}w_2^{t+2}\cdots w_{N_c}^{t+N_c})}{C(e_j)}$$

e_1 和 e_2 的右上下文分布的 K-L 距离为

$$D(e_1^{\mathrm{right}}\,\|\,e_2^{\mathrm{right}})=\sum_{\forall s_i\in S}p_1(S_i^{t+1}\,|\,e_1^{t})\log\frac{p_1(S_i^{t+1}\,|\,e_1^{t})}{p_1(S_i^{t+1}\,|\,e_2^{t})}$$

这样，如果 e_1 和 e_2 的 K-L 距离平均值越小，e_1 和 e_2 的语义相似度就越大.

3　实验和评价

实验采用的语料库是关于天气信息查询的汉语口语语料库，包含 460 个口语句子，共有汉语词汇 290 个，句子的平均长度为 5.5 个词.

实验中，把聚类的迭代次数定为 100，在每次迭代中，选取互信息值最高的那个词序列聚为一类，记为 PCi，用 PCi 替换语料中的这些词序列；将 K-L 距离值最小的那 2 个单元聚为一类，记为 SCi，用 SCi 替换语料中的这 2 个单元. 语法在迭代到 60 次时达到饱和，此时生成的语法包含 60 个短语结构类和 16 个语义类，覆盖了 143 个词汇. 生成的语法片段如下：

SC6：后天 | 大后天 | 明天 | 今天

PC0：什么　天

SC1：温度 | 气温

PC1：我 想

SC2：云 | 太阳

PC2：PC1 知道

SC5：大连 | 长春

PC5：刮 大风

PC6：PC2 SC5

SC7：热 | 冷

在生成的语法中，SC6 表示时间语义类，SC5 表示地点语义类. 一方面，这 2 个语义类与天气信息查询领域相关性不大，也可以用在其他领域内；另一方面，表达时间和地点的词数非常丰富，但是在语料库中出现的次数较少，会造成严重的数据稀疏而无法聚类，如生成语法中未覆盖的 147 个词汇中包含 102 个地名词汇. 为此，为地点语义类和时间语义类编写了初始文法作为先验信息，分别记为 CITY 和 TIME. 在预处理时，将语料库中符合初始文法的字符串用 CITY 或 TIME 替换. 生成的语法片段如下：

SC10：没有 | 有没有 | 有

PC0：什么 天

SC7：下雨 | 降温

PC1：我想

PC2：PC1 知道

PC3：CITY TIME

SC6：湿度 | 最高气温 | 温度 | 气温

PC4：是 多少

SC12：热 | 云 | 冷 | 刮 大风

PC7：SC6 PC4

语法在迭代到 43 次时达到饱和，此时生成的语法包含 43 个短语结构类和 12 个语义类. 从生成的语法片段可以看出，经过预处理后，既可以减少语法生成时的错误，也可以减少迭代次数.

对生成的语法进行手工编辑后处理，主要包括用特定领域中的意义标签替换 PC*i* 和 SC*i*，如在生成的语法片段中，用 TEMPERATURE 替换 SC6；去除一些与特定领域无关的终结符和非终结符，如 SC12 中涉及了风力和温度 2 个语义类，将其去除；合并某些 PC*i* 和 PC*i* 以使意义完整，如将 PC1 和 PC2 合并. 经过后处理的语法，就可以应用于基于语义语法的口语理解.

4 结束语

本文对自动获取语义语法进行了研究，利用统计方法从一个未标注的领域语料库中半自动获取语义语法. 该算法对特定领域的语料库进行反复的时间聚类和空间聚类，通过计算语言片段的 MI 发现片段间的语法连接程度，通过计算语言片段的 K-L 距离发现片段间的语义相似程度，进而生成一个粗糙的语义语法. 该算法是半自动的，这是因为生成的文法还不能体现领域的语义类别，要经过人工修改，从而得到该领域的语义语法. 将该算法应用在天气信息查询的汉语口语语料库中，实验结果表明，该算法是有效和切实可行的.

参考文献：

[1] Allen J. Natural language understanding[M]. 2nd ed. Redwood City: The Benjamin/Cumming Publishing Company, 1994:332-334.

[2] 王小捷，常宝宝. 自然语言处理技术基础[M]. 北京：北京邮电大学出版社，2002:78-79.

[3] Glass J, Weinstein E. SPEECHBUILDER: facilitating spoken dialogue system development[C] //Proceedings of Eurospeech. Aalborg:[s.n.], 2001: 1335-1338.

[4] Gavalda M. Growing semantics grammar [D]. Pittsburgh: Carnegie Mellon University, 2000.

[5] Wang Yeyi, Acero A. Grammar learning for spoken language understanding [C] // IEEE Workshop on Automatic Speech Recognition and Understanding. Trento:[s.n.], 2001: 292-295.

[6] Smaili K, Brun A, Zitouni I, et al. Automatic and manual clustering for large vocabulary speech recognition: a comparative study[C] // Proceedings of the 6th European Conference on Speech Communication and Technology. Budapest:[s.n.], 1999: 1795-1798.

[7] Giachin E, Baggia P, Micca G. Language models for spontaneous speech recognition: a bootstrap method for learning phrase bigrams [C] // Proceedings of 3rd International Conference on Spoken Language Processing. Yokohama:[s.n.], 1994: 843-846.

[8] Wright J, Gorin A, Riccardi G. Automatic acquisition of salient grammar fragments for call-type classification [C] //Proc of Eurospeech. Rhodes:[s.n.], 1997: 1419-1422.

[9] Arai K, Wright J, Riccardi G, et al. Grammar fragment acquisition using syntactic and semantic clustering [J]. Speech Communication, 1999, 27(1):43-62.

[10] Meng H, Siu K C. Semi-automatic acquisition of domainspecific semantic structures [J]. IEEE Transactions on Knowledge and Data Engineering, 2002, 14 (1): 172-180.

[11] Siu K C, Meng H M. Semi-automatic acquisition of domain-specific semantic structures[C]// Euro Speech '99 Proceedings. Budapes:[s.n.], 1999: 2039-2042.