ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.5,No.4,February 2009, pp.833-834,885

E-mail: xsjl@cccc.net.cn
http://www.dnzs.net.cn
Tel:+86-551-5690963 5690964

基于文法规则匹配的自然语言处理系统研究与实现

丁 杰

(北京工业大学 计算机学院,北京 100124)

摘要:对对话管理系统中的自然语言理解技术进行了研究,提出了基于文法规则匹配的自然语言处理方法,给出了采用该方法实现的自然语言处理系统的结构模型。对自然语言信息通过文法规则自动机解析为参数信息的过程做了介绍,并给出了规则应用举例。

关键词:自然语言处理;分词处理;参数标注;文法规则匹配;参数提取

中图分类号:TP391 **文献标识码**:A **文章编号**:1009-3044(2009)04-0833-02

Research and Implementation of Natural Language Processing System Based on Grammar Rule Matching

DING Jie

(College of Computer Science and Technology, Beijing University of Technology, Beijing 100124, China)

Abstract: The essay research on the natural language understanding technolgoy in dialogue management system and propose a natural language processing method based on grammar rule matching. This method provide a structural model for the natural language processing system. Also, the essay introduces the process where natural language information is interpreted as parameter information through grammar rule automation while some practices are presented on the rule.

Key words: Natural Language Processing; Segmentation; Parameter Tagging; Grammar Rule Matching; Parameter Extracting

1 引言

随着社会信息化程度的不断提高,人与计算机系统之间的交互也变得越来越频繁。在不断地交互过程中,人们通常希望将自然语言作为人与计算机的主要沟通方式,这就使基于自然语言信息查询的对话系统成为了当前对话管理系统的研究热点。自然语言处理方法是对话管理系统的一个重要组成部分,从计算机的信息处理过程上看,其主要内容是建立一种计算模型,使计算机可以从自然语言信息中提取出决定机器理解的关键要素[1]。计算机通过识别这些关键信息,采取一定的策略控制,就可以引导人机交互的顺利进行。

基于文法规则匹配的自然语言处理方法,通过文法产生式将大量适用于相同人机交互过程的句子集抽象成规则,并在其中加入对关键信息的标注,使系统可以直接通过规则集生成的有穷状态自动机将语言信息转化为参数序列,引导至对应的信息处理方法中,从而提高了自然语言信息的识别效率,也体现了规则集的易扩展性。

2 系统模型的建立

基于文法规则匹配的自然语言处理系统的主要任务是将自然语言信息解析为机器可以理解的参数信息,其功能主要靠分词处理、参数标注和文法规则匹配三个模块来实现,系统结构模型如图 1 所示。

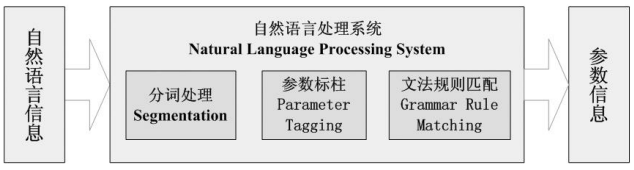

图 1 系统结构模型

3 系统实现

3.1 分词处理(Segmentation)

分词处理是通过分词算法将句子划分为词序列的过程。在英文文本中,空格是单词之间的自然分界符,无需对句子的词边界进行确认。而中文在句子构成上没有一种明显的词边界符,所以对于中文来讲,确定词的划分是理解自然语言的第一步。

3.1.1 预处理

预处理的主要任务是对源文本进行标记与拆分,以提高分词速度和准确率。

1) 预分词:在源文本中,经常会出现一些不易被分词算法正确切分的混合信息,比如浮点数、IP 地址、电子邮件地址、时间和日期等。这些信息可能是影响计算机理解的重要参数,应提前进行处理,以防被分词算法错误切分。本系统通过使用正则表达式对源文本进行匹配,将符合条件的词或子句标记为参数,分词算法不用对已标记的文本进行切分。

2) 分句:在中文文本中,汉语词是不包含符号的。将具有断句功能的标点符号作为分句依据,对源文本进行句子拆分,可以减少每次分词处理的信息量,提高分词速度。考虑到一些特殊参数(如 IP 地址)包含着影响分句的标点符号,应该将预分词处理放在分句处理之前,并将标记了参数的词或子句也作为句子拆分的依据。

3.1.2 中文分词

现有的分词方法可以分为基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法三类。其中,基于字符串匹配的分词方法也称为机械分词方法,是按照一定的策略将待分析的汉字串与一个"充分大的"机器词典中的词条进行匹配,若在词典中找到某个字符串,则匹配成功(识别出一个词)。按照扫描方向的不同,串匹配分词方法可以分为正向匹配和逆向匹配;按照不同长度优先匹配的情况,可以分为最大(最长)匹配和最小(最短)匹配;按照是否与词性标注过程相结合,又可以分为单纯分词方法和分词与标注相结合的一体化方法[2]。

收稿日期:2008-12-15

本系统采用双词典结构的字符串匹配分词方法，除了具有大量词条的分词词典外，还包括一个由参数表构成的参数词典作为附加词典。参数词典定义了参数表的存储格式，每一个参数表应具备参数类型、参数标记符和参数词表，如表 1 所示。

当系统加载词典时，首先会载入分词词典中的词条，随后会读取参数词典中的参数表信息，记录参数类型与参数标记符，并对每一个出现在参数词表中的词添加参数记录。通过参数词典来维护参数表，可以集中管理参数信息，便于规则集的参数扩展。

表 1　参数词典存储结构

参数词典	
参数标记符：参数类型	
1	参数词表
…	
……	
102：城 市	
1	北京
2	上海
……	
103：天 气	
1	雨
2	小雨
……	

3.1.3 后处理

后处理的主要任务是优化分词结果，提高机器的识别能力。

1) 停用词处理：停用词是指文本中出现频率很高，但实际意义又不大的词，主要指副词、虚词、语气词等。本系统使用停用词表对经分词算法切分好的词序列进行遍历，去掉其中包含的停用词。

2) 规范化处理：由于汉语语法的复杂性、词汇的广泛性和常用语的不规范性，通常会导致同一个意思的表达有多种方式，比如表示“今天”可以说：今天，今日，今儿，今儿个…这里采用同义词集生成的规范化映射表，可以将分词结果中不规范的词全部替换为标准词，便于机器识别。

3.2 参数标注(Parameter Tagging)

对于不同领域不同类别的问题，系统需要从中抽取出查询答案的关键信息，这些信息的集合就是系统预定义的参数集，比如在查询天气时候可以问“今天北京天气怎么样？”，这句话包含了两类参数，它们分别是时间参数“今天”和城市参数“北京”。由于参数对机器理解自然语言信息起到了较大的辅助作用，参数标注也就成为了分词结果进行文法规则匹配前的一个重要准备工作。

参数标注与词性标注类似，不过标注的内容不是词性，而是词所包含的参数类型。比如“天安门”在进行参数标注时，将被标记两个参数：[地点]和[景点]，分别对应“问路”和“旅游”两个领域的信息查询。一个词可能不具备任何参数类型，也可能具备多种参数类型，这是由系统的具体应用领域决定的。

3.3 文法规则匹配(Grammar Rule Matching)

在自然语言处理过程中，对语言信息的理解可以看做是有穷状态自动机的执行过程。分词处理与参数标注得到的带参词序列将作为规则匹配自动机的输入序列，序列中每一个词元的内容或参数类型则是自动机的状态转移条件，推动其执行。若分词序列输入完毕时，状态转移至终止结点，则成功识别出一条语言信息。这种方法通过文法产生式生成可以识别特定语言信息的规则，并将规则作为生成自动机状态结点和转移条件的依据，通过不断扩充的规则集来调整自动机的结构和状态，以提高对语言信息的识别效率。

3.3.1 文法设计

定义文法 G =({E, W, L},{(,), [,], {, }, <, >, $, num, word}, P , S)

其中 P ={

S→E ,

E→EE ,

E→(W) ,

W→E$E|W$E ,

E→[]|[L] ,

E→{}|{L} ,

L→num ,

E→<E> ,

E→word }

$L(G)=\{w|w \in T^*, S \overset{*}{\Rightarrow} w\}$为 G 产生的语言(language)，$\forall w \in L(G)$, w 为 G 产生的一个句子(sentence)[3]。在本系统中，将 G 产生的句子称为规则。

3.3.2 产生式解释

上述文法 G 所包含的产生式 P 可以对自然语言集进行规则抽象。对于符合文法规则的句子 $A_1,A_2,...,A_n$ ，包括以下几种格式：

1) 连接：“A_1A_2”表示两个规则 A_1 和 A_2 连接成一个新的规则 A_1A_2。

2) 选择：“$(A_1\$A_2\$...\$A_n)$”其中 $n \geq 2$,表示只需满足 $A_1,A_2,...,A_n$ 其中一条规则即可。

3) 可去除：“$<A_1>$”表示可以满足规则 A_1，也可以不满足 A_1。

4) 参数标记：“[L]”表示一个参数类型为 L 的词。

5) 任意参数标记：“[]”表示一个任意参数类型的词。

6) 组标记：“{L}”表示一组参数类型均为 L 的词。

7) 任意组标记：“{}”表示一组任意参数类型的词。

根据文法产生式的格式可以定义相应的规则表达式，比如([101][102]$[102][101])<的>天气<怎么样>。其中[101]代表日期参数，[102]代表城市参数，(E1$E2)表示在进行规则匹配时 E1 和 E2 只能选择一个来匹配，<>表示所括选内容是可以去除的。由此，前面给出的规则范例可以匹配到的句子包括：

1) 北京今天天气怎么样

2) 明天天津的天气怎么样

3) 广州昨天的天气

4) 后天上海天气

……

通过文法规则，可以用少量的规则表达式来识别大量的句子组合。

(下转第 885 页)

2) 考虑对迭代算法的优化，确保大量主题搜索的效率。

参考文献：

[1] Page L,Brin S.The anatomy of a large-scale hypertextual Web search engine[J].Computer Networks,1998,30(1-7):107-117.

[2] Page L,Brin S,Motwani R,et al.The PageRank Citation Ranking:Bringing order to the Web [R].Technical report,Computer Science Department,Stanford University,1998.

[3] Ricardo B Y,Berthier R N.Moderninformation retrieval[M].北京:机械工业出版社,2005.

[4] Yang Y,Pederson J O.A comparative study on feature selection in text categorization [C].International Conference on Machine Learning (ICML),1997.

[5] Langville A N,Meyer C D.A survey of eigenvector methods of web information retrieval[J].The SIAM Review,2005,47(1).

张冉(1981-)，女，新疆乌鲁木齐人，讲师，硕士研究生，研究方向：网络信息检索。

(上接第 834 页)

3.3.3 状态转移

为了保证在语言信息识别成功时，句子所经历的状态转移路径只对应一条规则，由起始结点 S 出发至终止结点的自动机结构应是树状结构，其根结点为 S 结点，自动机为非确定有限状态自动机(NFA)。对于一个结点的一次词匹配，自动机会按照固定的顺序选择状态转移函数，如果一种状态转移函数的词匹配失败，将通过回溯法回到前面结点，选择新的状态转移函数继续匹配，直至匹配成功或所有结点的状态转移函数均匹配失败。

对状态转移函数的状态转移条件的选择顺序为：

1) 具有断句功能的标点符号
2) 输入词
3) 任意组标记
4) 与输入词参数类型一致的组标记
5) 任意参数标记
6) 输入词的参数类型

3.3.4 匹配成功

当自动机执行完毕时，若所在结点为终止结点，将会读取到一个规则标识。通过规则标识可以从状态转移路径经过的结点中提取出与该规则相关的参数信息，并将其保存为参数序列。至此，有穷状态自动机就完成了由分词序列到参数序列的文法规则匹配和参数提取工作。

4 规则应用举例

通过 XML 生成函数可以将定义好的规则内容、处理方法和自动机所需的参数提取信息转换为系统能够读取的 XML 格式标签。例如 “([101][102]$[102][101])<的>天气<怎么样>” 将会转换为<rule><g><o><t para = “Date”>101</t><t para = “City”>102</t></o> <o><t para = “City”>102</t><t para = “Date”>101</t></o></g><e><w>的</w></e><w>天气</w><e><w>怎么样</w></e><goto>Weather|WeathcrSearch</goto></rule>

在识别句子“北京今天天气怎么样”时，系统提取出的信息为：

1) 信息采集方法：Weather 领域的 WeathcrSearch 方法
2) 已获得参数：City|北京；Date|今天

5 结束语

本文对中文人机对话系统中的自然语言理解技术进行了研究，提出了一种基于文法规则匹配的自然语言处理方法。根据此方法实现的系统，可以通过扩充参数词典和规则集来提高对自然语言信息的理解能力，使系统有较好的领域扩展性。对于识别效果不好的语言信息只需按照其语法结构抽象出新的规则并添加到系统中，就能够改善对这类句子的识别效果，实现起来简单而有效。

参考文献：

[1] 俞士汶.关于语言信息处理技术的展望[J].计算机世界,1997(1).

[2] 湛燕,陈昊,袁方,王熙照.基于中文文本分类的分词方法研究[J].计算机工程与应用,2003,39(23).

[3] 蒋宗礼,姜守旭.形式语言与自动机理论[M].北京:清华大学出版社,2007.

[4] Allen.自然语言理解[M].北京:电子工业出版社,2005.

丁杰(1983-)，男，北京人，硕士，研究方向：计算机系统与嵌入式系统。