中文问答系统中机构名的处理

韦向峰[1],张 全[1],吴 晨[1,2],袁 毅[1]
WEI Xiang-feng[1],ZHANG Quan[1],WU Chen[1,2],YUAN Yi[1]

1.中国科学院 声学研究所,北京 100080
2.中国科学院 研究生院,北京 100049
1.Institute of Acoustics,Chinese Academy of Sciences,Beijing 100080,China
2.Graduate University of Chinese Academy of Sciences,Beijing 100049,China
E-mail:wxf@mail.ioa.ac.cn

WEI Xiang-feng,ZHANG Quan,WU Chen,et al.Processing organization name in Chinese question answering system. Computer Engineering and Applications,2008,44(7):196-198.

Abstract: To discuss the processing organization name and improving the performance of Chinese question answering system,an approach of semantic conceptual analysis based on hierarchical networks of concepts theory is adopted to remove the words which are irrelative with organization name and to get the remain as organization name candidates.After sorting the database of the full names of organizations,the matched full names was selected out from the database and sorted according to the sameness as the candidates.The experimental result shows that the accurate rate is up to 90.6%.The approach can process the abbreviation of organization name.

Key words: question answering system;organization name;hierarchical network of concepts;index by Chinese character

摘 要:探讨问句中机构名的处理,并服务于中文问答系统。采用概念层次网络理论的语义概念分析方法分析问句,去掉与机构名无关的概念词语,得到候选机构名。对机构名全称库按字索引,在库中搜索出与候选机构名匹配的机构名全称并按拟合权值排序。实验结果表明该方法识别机构名的正确率达到90.6%,支持对机构名简称的处理。

关键词:问答系统;机构名;概念层次网络;按字索引
文章编号:1002-8331(2008)07-0196-03 **文献标识码**:A **中图分类号**:TP391.1

1 引言

经过近十年的发展,问答系统已经成为自然语言处理领域和信息检索领域的一个重要分支和新兴的研究热点。问答系统一般由三个部分组成:问题分析、问题求解和答案生成。问题分析部分需要对问句进行分词、词义、句法、类型、特征和焦点等分析,得到问句的匹配框架信息。问题求解部分根据问句的匹配框架信息在系统的知识数据库中进行搜索和匹配,在匹配标准的约束下得到符合标准的数据库内容。答案生成部分在匹配到的数据库内容的基础上根据答案类型、问题焦点和句法语义知识对内容进行组合转换,从而得到问句的答案。

本文面对的问答系统基于一个受限的某城市已注册机构名称数据库,包含2 600多个机构名称,每个机构名称又含有自己的电话、地址、联系人、交通方式等等,总共有108 735条记录。用户以问句的形式向系统询问与机构相关的信息,系统经过分析处理后返回答案。问句分析的方法一般有基于问句实例[1]、基于句法分析[2],本文在问句分析部分采用基于概念层次网络[3]的句类分析技术,通过句类分析可以得到问句的语义结构,在此基础上识别机构名。在问题求解部分,本文采用了基于汉字索引的搜索算法,可以处理机构名缩写和缺少机构名中某些汉字的情况。对于答案生成部分,将在其它文章中讨论。

2 问句分类与概念分析

问句是用于提出某种问题,期望得到回答或正确答案的句子。本文主要依托HNC(概念层次网络)理论[3],在词语概念的基础上对问句进行分析,并利用问句的分析结果找出问句中的机构名。对问句的概念分析分为以下5个步骤:第一,对语句进行分词处理,得到字词等基本构成单位;第二,依据字词的概念符号知识库把字词映射为概念基元符号体系表示的概念符号。字词库包含了41 049条词语的义项和2 333条汉字的义项,记录字词的概念类别、概念符号、使用频度、句类代码等;第三,从

基金项目:国家重点基础研究发展规划(973)(the National Grand Fundamental Research 973 Program of China under Grant No.2004CB318104);中国科学院声学研究所"所长择优基金"(No.GS13SJJ04)。

作者简介:韦向峰(1976-),男,博士,助理研究员,主要研究领域为自然语言理解、概念层次网络理论及技术;张全(1968-),男,博士,研究员,博士生导师,主要研究领域为HNC自然语言理解处理技术、计算语言学;吴晨(1979-),男,博士研究生,主要研究领域为自然语言理解处理技术、软件工程;袁毅(1967-),男,高级工程师,主要研究领域为计算机网络技术。

收稿日期:2007-05-23 **修回日期**:2007-11-15

概念符号感知语义块。语义块的感知分为两个部分,一是特征语义块,相当于谓语部分;二是广义对象语义块,相当于主语、宾语部分;第四,根据特征语义块和广义对象语义块假设并检验语句的句类代码,检验通过进入第五步,否则回第三步;第五,分析语义块的内部组成成分以及成分之间的关系。

在第一步的分词处理中,依据词库的词表切词会出现当前字既与前一字成词,又与后一字成词的情况。这种切分歧义可通过一定的规则和问句分析的第四步消除大部分歧义,确定该字在句子中的组词情况:(1)与前字成词;(2)与后字成词;(3)与前后字成词;(4)与前后字都不成词,为单字。在问句分析的第三步,特征语义块的感知可从具有概念类别符号"v"的词语入手。这些词语大多数是动词,有可能构成语句的核心,从而得到特征语义块的候选集。在问句分析的第四步,对特征语义块候选集合中的词语,按优先顺序逐个得到在字词库中的句类代码,从而得到句类代码候选集。从句类代码候选集中取出一个句类代码,假设它是正确的,根据句类代码的相关知识,对各广义对象语义块及其之间的概念符号进行检验计算,若符合则通过检验,将假设的句类代码作为语句的句类代码;若不符合则不能通过检验,须回到第三步再选取下一个特征语义块候选集中的词语做假设检验。

在语言学中,问句分为是非问句和非是非问句,非是非问句又分为:特指问句、选择问句和正反问句[4]。以下是这些问句类型的一些特点:

(1)是非问句要求对问句给出肯定或否定的回答,非是非问句则不要求;

(2)特指问句中用疑问代词替代疑问焦点;

(3)选择问句并列多个项,让答者选择;

(4)正反问句并列肯定形式和否定形式,让答者选择;

(5)特指问句常用的疑问代词有:什么、为什么、哪些、哪里、谁、怎么、几、多少;

(6)选择问句中的代表词语是:"是…还是…",其中的"是"可以省略;

(7)正反问句的典型结构为"A 不 A",A 一般是动词性概念,如"喜欢不喜欢";

(8)是非问句的特点是在句末加语气助词如"吗"、"吧"等;

(9)问句末尾一般用问号,可以带疑问语气词如"吗"、"呢"、"吧"、"啊"等。

利用这些分类标准和问句类型特点,可从语言形式上对问句进行初步的分类和处理,并对问句的答案有一定的预期和形式约束。本文的主要目的是从问句中寻找并得到机构名,可以预见询问机构名的问句主要集中在特指问句中,如"什么地方"、"哪里"等,这些在概念层次网络理论中属于对基本本体概念中空间概念的询问。

3　问句中机构名的获取

机构名一般是指企事业单位、机关、团体、协会等组织机构,某个学校、研究所、公司、医院、银行、饭店、国家机关等组织的名称都属于机构名。机构名数量庞大且随时间推移某些机构名会发生变更,一个机构名可以有一个或多个简称。机构名的这些特点给计算机自动识别带来了困难:首先,不可能构造一部包括所有机构名的词典,多数机构名会成为未登录词,给问句分析中的分词处理增加了难度;其次,尽管机构名在末尾一般有特征词,但机构名左边界的确定却无规律可寻;第三,机构名可以带有下属机构形成复合机构名称,例如"中国银行股份有限公司泸州分行西路分理处"。

对机构名的自动识别有基于统计模型、基于规则、统计与规则相结合的方法。基于统计模型的方法一般先在标注了机构名的训练语料中对统计模型进行参数训练,根据训练得到的参数模型对测试语料中的机构名进行识别,通常使用的统计模型有:HMM(隐马尔科夫)[5]、最大熵[6]、SVM(支持向量机)[7]和条件随机场[8];基于规则的方法从机构名的语法、语义特性去研究和总结机构名的构成规则[9],然后在计算机中编程实现这些规则,再对文本中的机构名进行识别;统计与规则相结合的方法一般以统计模型为主先识别出机构名,然后使用规则对识别出的机构名进行筛选[10]。基于统计的方法可以方便地建模处理大规模的语料,其缺点是过分依赖训练语料库,对于训练语料库中未出现的机构名难以识别;基于规则的方法对小规模语料处理准确度较高,但规则的提炼和实现较为困难;统计和规则结合方法是先统计后规则,同样存在两者相独立时出现的问题,但是在准确率和召回率上会有所提高。

本文对问句中机构名的识别采用的是基于规则的方法。因为在问答系统中对于问句是一句一句进行处理的,问句的长度一般都较短,问句处理的规模在一般情况下并不大。但用户对问句答案的准确度的要求却是苛刻的,如果系统给出的答案错得离谱或者用户多次在答案中找不到需要的机构名及其相关信息,那么用户忠诚度将急剧下降。而基于规则的方法具有较高的准确率,不需要事先标注好的较大规模的训练语料,正好适用于本文情况。本文对问句中机构名的识别方法如下:

(1)对问句按类型进行概念分析,获得问句的分词结果和语义结构信息,具体步骤见本文的第 1 章。问句中的词语是机构名识别的基础,因为机构名可以看成是由词语或单字构成的,在句子中充当某种语义成分。为了确定词语在句子中的地位和作用,需要对词语的概念语义进行分析,而大部分词语(单字可看成是特殊的词语)在字词库中都有自己的概念符号。根据字词的符号和知识库中句子构成的概念规则,可以确定句子属于什么语义类型的句子,各词语在句子中充当什么样的角色,从而为机构名称的识别提供合理的依据。

(2)去掉问句中与机构名无关的语义概念成分。从句法上看,句子的谓语部分不可能包含机构名,因此问句分析后得到的特征语义块(谓语部分)可以去掉;用于发问的语习类概念也可以去掉,例如"请问"、"谁知道"、"有没有"等等。从词法上看,机构名用词具有一定的规律,如不包含"的"字,一般不使用表示逻辑组合关系的"和"、"与"、"并"等,较少使用数量短语、动词,这些词语或同类概念在识别机构名时也可以去掉。

(3)去掉问句中与询问相关的词。问句的特点是带有疑问语气词或疑问代词,疑问语气词("吗"、"呢"、"啊"、"么"、"呀"等)出现在语句末尾,一般不出现在机构名中,因此可以去掉;疑问代词取代了语句中的询问对象,一般也不出现在机构名中,如"什么"、"哪里"、"哪儿"、"哪"、"怎样"、"如何"、"为什么"、"咋样"等等,也可以去掉。

(4)去掉孤立字。经过以上步骤的处理,问句中会出现孤立字(该字未被去掉、但其左右临近的字或词都已被去掉)。孤立字不能构成机构名,因此可以在问句中找出孤立字并去掉。

(5)得到机构名的候选集。问句中未被去掉的相邻的字词

就构成了机构名的一个候选,在问句中可能存在多个这样的字词组合(两个候选之间必然有被去掉的字词),多个候选构成了问句中机构名的候选集合。对集合中的候选机构名,可以根据机构名的某些特征词语如“银行”、“公司”、“店”、“厂”、“所”等进行可能性大小的排队,然后按可能性从大到小在机构名称库中进行搜索。

假设用户提交的问句是:“你知道声学所的地址和电话吗?”。该问句经分词处理和语义概念分析后的结果为:“你 知道 声学所 的 地址 和 电话 吗 ?”,其中“知道”的概念符号是“v8109”,表示动态概念,在句子充当谓语。而“你”的概念符号是“p4002”,充当主语。“的、和、吗”属于连接词和语气词,也可以去掉。此时得到问句中的机构名候选集合为“声学所 地址 电话”,由于“声学所”含有机构名特征词“所”,在机构名称库中搜索得到“声学所”的全称“中国科学院声学研究所”,搜索算法参见本文的第 4 章。使用该算法的好处是:给出简称如“声学所”也能搜索得到它的全称,此外如果在机构名识别过程中多去掉了机构名全称中的某几个字或词,依然可以在机构名全称库中找出该机构名。

4 机构名的搜索

机构名搜索算法需要解决的问题是如何根据机构名识别的处理结果(即候选机构名),从机构名全称库中快速准确地得到与候选机构名相匹配的机构名全称。机构名全称库以数据库记录的形式存储机构名全称,及其相关的电话、地址、联系人、交通方式等。由于候选机构名不一定是机构名全称,可能是简称,也可能是全称中的某些字,还可能包含了不在机构名全称中的汉字。因此,采用按字对机构名全称建立索引的方法解决这个问题。以汉字为索引可得到包含该汉字的所有机构名全称的记录号,根据记录号在数据库中很容易得到机构名全称及其相关信息。当候选机构名为机构名全称的简称时,简称的每个字都在机构名称的全称中,可完全匹配;当候选机构名为机构名全称中的某几个字时,依然可以完全匹配;当候选机构名既有机构名全称中的字、又有不在机构名全称中的字,或者根本不含机构名全称中的任何字时,则无法与某个机构名全称完全匹配。

建立索引时, 逐个读取机构名全称库中的机构名全称,对机构名全称中的每一个汉字,按(字,记录号)的形式建立 Hash 表,如果有相同的汉字则只建一次。根据此 Hash 表,从一个汉字可以快速得到包含该汉字的机构名全称的记录号。当所有的机构名全称都建完索引后,一个汉字对应的记录号可能有多个(即多个机构名全称包含该汉字),这些记录号形成一个记录号集,可由同一个汉字索引得到。建立索引耗时较长,可以在系统运行之前进行。系统在使用索引时则是快速高效的,因为由散列的 Hash 表可从汉字直接得到记录号集。

在匹配候选机构名与机构名全称时,关键过程是寻找机构名中每一个汉字对应的所有记录号集合的交集,如果该交集不为空则找到了匹配的机构名全称,否则无法完全匹配。匹配算法描述如下:

(1)通过建立的 Hash 表,从候选机构名的汉字得到该汉字对应的记录号集;

(2)对候选机构名中所有汉字的记录号集,以记录号最少的为基准集;

(3)将某个其它记录号集与基准集比较,得到它们的交集;

(4)如果交集为空,转(7),否则转(5);

(5)如果所有的其它记录号集均比较完毕,转(7),否则转(6);

(6)以新的不为空的交集为新的基准集,取下一个记录号集作为其它记录号集,转(3);

(7)结束匹配。

经过匹配算法后,得到的交集如果不为空,那么就可以通过交集中的记录号,从机构名全称库中得到机构名全称。当交集中的记录号有多个时, 说明有多个机构名全称都包含候选机构名, 但每一个机构名全称与候选机构名的拟合程度可能不相同。

拟合程度用权值衡量,主要考虑候选机构名中首字在机构名全称中出现的位置以及字字之间的相对位置。首字在机构名全称中出现的位置越靠前,则权值越小,拟合程度越大;字与字之间的相对位置越小,则权值越小,拟合程度越大。此外,如果候选机构名中在某字右边的字在机构名全称中跑到了左边,那么权值变大,拟合程度变小;如果其它拟合程度相同,那么字多的机构名全称的权值大。权值的具体计算公式如下:

$$Weight=Origin+Offset*C1*C2+Len \quad (1)$$

在式(1)中,Origin 表示候选机构名首字在机构名全称中的位置,即在新坐标系中的“原点”。Offset 表示在机构名全称中偏离原点的距离,参数 $C1$ 用于区分两字之间距离,如果两字按原顺序相邻则 $C1$ 取 1,如果两字按原顺序但不相邻则 $C1$ 取100;参数 $C2$ 用于区分两字是否按原顺序排列,若按原顺序则 $C2$ 取 1,若按相反顺序(在右边的字跑到了左边)则 $C2$ 取 1 000。Len 用于区分机构名全称的长度,取值为机构名全称与候选机构名的长度之差。

因此, 计算出与候选机构名匹配的机构名全称的权值,可按权值从小到大进行排序。排在前面的机构名全称权值小,与候选机构名的拟合程度大,也就越有可能是用户需要的机构名。

5 实验结果分析

本文从用户提交的 4 773 句问句中选取了 302 句关于地址和机构名的问句进行分析, 并对其中的机构名进行识别处理。实验结果表明,在这 302 句中有 235 句出现了具体的机构名,有 67 句只是询问地点或某类机构。在 235 句出现了具体机构名的问句中,正确识别出来的有 213 句,机构名识别的正确率为 90.6%。

通过分析识别错误的问句, 发现产生错误的原因主要是:(1)表达时间概念的词未被去掉,例如“国家安全部今天”中的“今天”;(2)有的机构名被去掉的词过多,对后面的机构名搜索带来困难,例如“博动科技”被识别为“科技”,“毕博管理咨询”被识别为“毕博”;(3)一些地点名称被误识为机构名,例如“希格玛大厦”。

搜索匹配到的机构名全称的权值计算方法还需要进一步改进,因为当式(1)中的 Origin 较大、对权值的影响超过了 Offset 时,会出现权值大的机构名全称反而拟合程度高的情况。例如候选机构名为“国家安全部”,则“中国家电协会安全部”的权值将比“中华人民共和国国家安全部”要大,而实际应该是后者的拟合程度大,即权值要小。因此,当某两字的距离过大或者排列顺序多数不一致时应考虑加大其权值,甚至赋予该机构名全称不能完全匹配的权值。

(下转 205 页)