中 文 信 息 学 报
第 20卷 第 4期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 20 No. 4

文章编号: 1003 - 0077(2006)04 - 0049 - 07

基于概念匹配的中文问答处理模型核心问题探讨 *

吴 晨[1,2],张 全[2]

(1. 中国科学院 研究生院,北京 100039; 2. 中国科学院 声学研究所,北京 100080)

摘要:为了解决问答处理系统中的语义模糊问题,提高问答处理的性能,研究人员尝试采用概念作为系统处理的对象,而不再是语言表层符号,然而,在引入概念进行处理的同时引来了一些新的问题,如概念的抽取、概念关联计算以及特定于问答系统的问题理解、问题求解、答案生成等问题。在概念抽取、概念关联计算方面,已有一些比较成功的算法。本文将在此基础上,针对实现这样一个问答处理系统所存在的一些未涉及的核心问题进行一个探讨,同时提出解决以上问题的方法。实验及实际应用表明基于所提出算法的概念问答系统具有较强的性能,系统总体自动处理准确率将近达到 40%。在实际应用中也表现出较高的应用价值。

关键词:计算机应用;中文信息处理;中文问答系统;语言概念空间;核心问题研究;概念匹配;算法

中图分类号: TP391 **文献标识码**: A

Research on the Key Problems of Concept-Based Chinese QA Model

WU Chen[1,2], ZHANG Quan[2]

(1. Graduate School of the Chinese Academy of Science, Beijing, 100039;
2. Institute of Acoustics, Chinese Academy of Science, Beijing 100080, China)

Abstract: Concept-Based Question Answering (QA) is a brand new research topic which takes concepts, instead of the lexical terms, as the processing object. Concepts, as a formalized meaning, can well help to resolve the word sense ambiguities. However, using concepts brings some new problems, such as the concept extracting; the semantic relativity calculation for concept as well as the QA-specialized issues such as how to comprehend the query; how to search the answers and how to generate the nature language answers. Most of them, especially the QA-specialized issues, have not been addressed. In this paper, we discuss these key issues for carrying out a concept-based QA system. Some algorithms will also be proposed in order to solve the problems. The experiments indicate that the concept-based QA system powered by the proposed algorithms performs very well. The precision of the system reaches almost 40%. The actual application also indicates these algorithms contribute a lot to a commercial concept-based QA setting.

Key words: computer application; Chinese information processing; Chinese question answering system; concept space of natural language; key problem; concept matching; algorithm

1 引言

问答系统是一种特殊的信息检索系统,在这个系统中,用户能够根据自己所提的要求获得更高效的、更人性化的检索结果。自从文本检索会议 TREC于 1999年第一次设立了 QA Track

* **收稿日期**: 2005 - 07 - 22 **定稿日期**: 2006 - 06 - 01
基金项目:国家 973项目资助 (2004CB318104);中科院声学所知识创新工程资助项目
作者简介:吴晨 (1979—),男,博士生,主要研究方向为自然语言理解、软件工程.

以来,在国际上,包括微软、卡耐基梅隆大学、宾西法尼亚大学在内的多家著名研究机构都不同程度的展示了自己的前沿研究成果。开发的相对成熟的问答系统包括:密歇根大学的 Answer-Bus(http://www.answerbus.com/about/index.shtml),麻省理工的 StartTV (http://www.ai.mit.edu/projects/infolab),Ask Jeeves公司的自然语言检索系统 (www.ask.com)等。在国内,在往届的 TREC QA Track评测中,中国科学院计算所[1,2]、复旦大学[3]也都取得了较好的成绩。中国科学院自动化所、哈尔滨工业大学、复旦大学等也在中文问答技术领域进行了非常有益的探索。

通过考察不难发现,以上涉及的问答系统都是从表层语言符号入手来实现相关的检索、匹配任务的。这样做有其内在的好处:方便、直观,并且具有很好的研究基础。但词语本身无法克服的会存在一些语义模糊,比如同义、多义模糊。抛开上下文关联性,我们无法知道“jajuar”这个词指的是一辆汽车 (英国产的“美洲虎”汽车)还是一只动物。于是,我们开始尝试从语言概念层面去解决问题,从而使得用户的提问更明确,答案更具体并易于发现。本文的主要内容将围绕实现这样一个系统遇到的主要问题展开。问题主要集中在三个方面:一是如何理解用户的提问;二是如何获取答案;三是如何进行生成。在基于概念的理解方面,HNC[11,12]已经进行了相当多的工作,这将作为我们解决其他核心问题的基础。

本文第二节介绍了一些相关的研究工作;第三节概要介绍了整个系统,着重介绍系统中需要解决的核心问题;第四节给出了一个实验结果;第五节是结论。

2 相关工作

本文所涉及的中文问答处理系统是基于 HNC(概念层次网络)[4,5]自然语言理解框架 (由中科院声学所黄曾阳先生提出)的。HNC制定了一系列的符号体系 (语言概念空间语义表示式)来形式化的表示词语的语义及句子的语义结构。同时提出并实现了一套计算机确定词语语义、句子语义结构的软件系统,即句类分析系统[4]。通过系统处理,我们可以得到表示句子及词语语义的概念符号,从而消解句子及词语的语义模糊。HNC用句类表达式来描述句子的语义结构,共定义有 57组基本句类表达式和57 * 56组混合句类表达式。用概念来表述词语所描述对象的内涵。图 1给出的为句类分析系统处理的一个例子。通过处理,我们获得了例句的句类表达式及对象概念,例句由两个是否判断句 (基本句类)组合而成,图中 +%表示后一个句子共享前一个句子的第一个广义对象语义块 (主语)的对象部分。

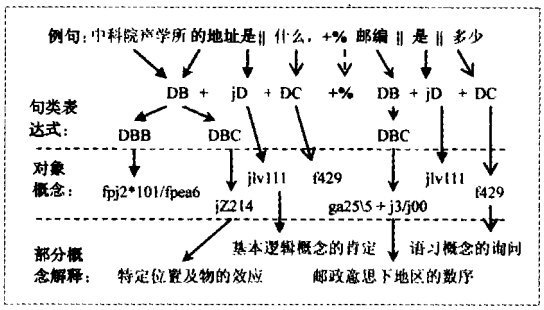

图 1 句类分析功能示意图

3 基于概念的中文问答系统核心问题

基于概念的中文问答系统 (以下称 HNCQA)存在一些特定的问答系统,在 HNC自然语言理解体系中并未系统涉及到的内容,如问题理解、问题求解、答案生成等问题,这是实现整个系统的重要环节所在。本节中将首先介绍基于概念的问答系统的总体结构,然后着重讨论实现这一系统要解决的主要问题,同时给出解决问题的方案及相关算法。

3.1 问答系统概述

HNCQA从整体上来看包含四个功能模块、一个全局知识库、一个常识知识库及其基础支撑平台,HNCQA总体结构框图如图 2所示。四个功能模块分别为:用户界面模块、问题理解模块、问题求解模块、答案生成模块。用户界面模块负责系统与用户之间的信息交互;问题理解模块负责明确用户的意图;问题求解模块则根据明确了的形式化的用户意图从常识库中提取知识;答案生成则根据获取的知识生成符合人类交流习惯的描述语句。全局知识库,即 HNC概念知识库,提供概念层面的语句处理底层的规则支持。常识知识库基础支撑平台由 HNC交互平台提供,它负责提供系统处理所必需的结构化的世界知识。

整个问答系统的设计主要基于"概念映射"思想:获取问句的语义信息,并将其映射到概念空间,通过概念空间中概念的相互关系对问题进行求解,并最终生成语言文字符号表述的问题答案返回给用户。

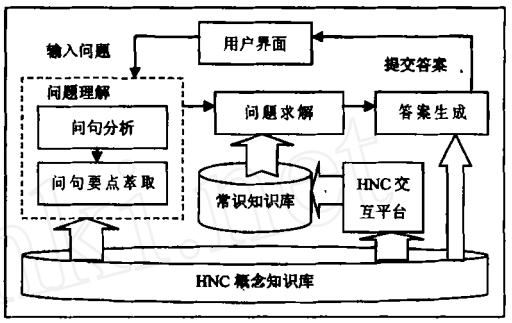

图 2 HNCQA结构框图

在 HNCQA中,常识知识库中的知识根据预先的设定划分成了与 C相映射的多个维度,每个维度对应一个语义分量,这些语义分量是我们通过对大量用户提问类型进行研究后得来的,具有一般意义,而信息的萃取则是在 HNC交互平台下,通过手工干预校正完成的,准确率达到 99%以上。

3.2 系统核心算法探讨

本文所要讨论的关键问题主要集中在图 1所示的"问句要点萃取"、"问题求解"、"答案生成"三个环节上。这些问题要解决的内容主要包括:获取了句子的句类表达式之后如何在概念层面上抽取用户提问涉及的要点;这些要点被萃取了以后如何认定哪些要点是和用户想了解的内容密切关联的;认定完用户的需求后如何从结构化数据库中获取;获取的数据如何反映射成自然语言的句子。针对这些问题,我们将在下文中进行展开讨论,并仍旧通过例句来阐述。

3.2.1 问句要点萃取

问句要点萃取的主要任务是如何在已获取的问句的句类表达式的基础上,抽取用户提问涉及的要点,它是概念层面上的对整个问句具有提示性的关键信息。这些信息将最终服务于问题求解模块。这些要点将以向量的形式来表示,每一个问句都将对应一个或者多个向量,于是最终的结果可能为一个向量数组。假设这个向量数组用 $Q = (Q_1, Q_2, \cdots, Q_i, \cdots, Q_m)^T$ 表示,其中 $Q_i = (q_{i1}, q_{i2}, \cdots, q_{ik})$, Q_i 表示对第 i个对象所萃取的要点。获取向量组 Q 的算法描述如 Algorithm 1。

Algorithm 1 (问句要点萃取算法)

Input: a set of sentence category expressions (SCE: 问句的句类表达式)

Output: Q

Procedure:

1. Initiate a new array Q; set integer i = 0;
2. For each SCE do
3. Obtain the B chunks and its related key contents which have inquiring characteristics according to SCE;
4. For each B do

5. Store the conceptual form of the B-related contents to Q_i;
6. End for each;
7. i ++;
8. End for each.

如例句中:根据句子的语义块可以提取出两项关键要素,依据是两个是否判断句的 DBC(内容)团块,得到其对应语义符号 jz214和 ga25\5 + j3/j00,并生成由一个向量组成的向量数组。$Q = (Q_1)^T, Q_1 = (jz214, ga25\5 + j3/j00)$。

3.2.2 问题求解

问题求解主要解决两个问题:一是认定哪些要点是和用户想了解的内容密切关联的;二是在认定用户需求后如何从结构化的数据库中获取所要的答案。

首先需要说明的是,在 HNCQA中预设了一个特征向量,该向量用来描述各种可能的用户提问类型,向量的每一个分量为解释这些类型的特征概念。系统将语义分量划分成了 15类,包括日期、时间、地点以及与应用密切结合的电话、E-mail、金额等等,并以 $Z = (z_1, z_2, \cdots, z_{15})$ 表示,z_i 为第 i 个语义分量的 HNC符号,如地点对应 j214。为了记录求解的结果,我们引入了一个新的向量 $V = (v_1, v_2, \cdots, v_{15})$,$v_i$ 与 z_i 一一对应,v_i 记录用户提问在 z_i 上的概念相关度。于是第一个问题的求解就可以归结到求极大似然概念相似度的问题:$v_j = \arg\max_k R(z_j, q_{ik})$,其中 $R(x, y)$ 表示计算概念 x 和 y 之间的概念相似度,概念相似度计算算法[6](以下称 AlgZ);$q_{ik} \in Q_i$,或者说计算以 Algorithm 1中生成 Q 的每一个行向量 Q_i 为单位进行。求得 V 以后,再将 v_j 根据预先设定的概念相似度权值 λ 进行比较,并将 v_j 投影到用户问询内容认定向量 $X_i = (x_1, x_2, \cdots, x_{15}) \in X$ 上,这样就完成了对于 Q_i 所描述问句的用户需求认定任务。整个要点生成步骤参见 Algorithm 2中所述。

Algorithm 2 (问询内容获取算法)

Input: $Q = \{Q_1, \ldots, Q_n\}$ derived from Algorithm 1; $Z = (z_1, z_2, \ldots, z_{15})$ defined in advance

Output: X

Procedure:

1. For (i = 1; i < = n; i ++) do
2. For (j = 1; j < = 15; j ++) do
3. MaxRelationScoring = 0;
4. For (k = 1; k < = kMAX; k ++) do
5. If (RelationScoring (z_j, q_{ik}) > MaxRelationScoring) then
6. MaxRelationScoring = RelationScoring (z_j, q_{ik});
7. End if;
8. If (MaxRelationScoring > λ) then
9. x_j = 1;
10. Else
11. x_j = 0;
12. End if;
13. End for;
14. End for;
15. End for;

notes: RelationScoring(x, y) is the relativity calculation algorithm which is defined in AlgZ

从 Algorithm 2的计算中,我们可以看到,向量 $V=(v_1, v_2, \ldots, v_{15})$在算法中并没有出现,实质上它是一个过渡向量,在我们的工作参数中已经进行了记录,为了算法表示的简洁,将它省去了。通过算法计算,我们可以得到结构化的用户所询问的不同对象的不同内容,这些信息记录在了 Q, Z,以及刚刚获取的 X三个向量所联合表示的数据结构中。Q记录对象, Z和 X记录针对 Q中对象想要获取的内容。

获取这些知识以后,本节的问题一已经解决了,接着是问题二,实质上有一个结构化的问题表示框架,有一个受益于交互平台的结构化知识库,问题二的解决就相对简单了。在结构化的知识库中,我们已经将不同的对象按其不同的内容进行了归纳,在搜索过程中我们只需要从结构化的知识库中查找关于某个对象的某个侧面内容是否存在就可以获得最终结果。结果我们有向量 $S_i=(s_1, s_2, \cdots, s_{15}) \in S$来表示,每个 $s_j \in S_j$将记录关于语义分量 q^i_j上的一个解答,它可能是一段话。

例句通过 Algorithm 2的计算可以得到 $X=\{X_1\}$, $X_1=(1, 0, 1, 0, \cdots, 0)$,由预先定义的 Z = (j214, j1, (ga25\5 + j3/j00), j308, …),可以得知用户所提问句中涉及的问询内容,即 j214和 ga25\5 + j3/j00,最后根据问询内容所对应概念的含义,从结构化的常识知识库中获取对应答案,生成 S结构。例句最终生成的 S结构为: $S=S_1$, $S_1=\{$北京市海淀区四环西路 21号, 100080, …$\}$。

3.2.3 答案生成

答案生成的任务是将结构化的并且以概念形式表示的各种信息反映射成自然语言,这一操作主要利用 HNC句类知识以及语义块构成知识来完成。首先将答案划分成了三类,第一类是能够正确回答用户的答案;第二类是识别了用户的意图但是在知识库中不存在相应知识的情况下所要作的回答,第三类是无法识别用户意图时所要返回给用户的答案。这三类情况将根据 Algorithm 1, Algorithm 2的不同处理结果来进行分情况考虑。答案生成算法如 Algorithm 3所示。

Algorithm 3 (答案生成算法)

Input: $Q=\{Q_1, \ldots, Q_n\}$; $Z=(z_1, z_2, \ldots, z_{15})$; $X_i=(x_1, x_2, \ldots, x_{15}) \in X$; $S_i=(s_1, s_2, \ldots, s_{15}) \in S$

Output: string: Ans;

Procedure:

```
1. If Q = {Q1, ..., Qn} == nil then
2.    Ans = "Can't interpret user's intention";
3. Else
4.    For (i = 1; i < = n; i ++) do
5.       For (j = 1; j < = 15; j ++) do
6.          If (xj == nil) then
7.             Ans = "Can't interpret user's intention";
8.          Else
9.             If (sj == nil) then
10.               Ans = "The answer has not been included in the knowledge base";
11.            Else
12.               Ans = GetWord(Q1) + "'s " + GetWord(zj) + "is " + GetWord(sj);
```

```
13.              End if;
14.            End if;
15.          End for;
16.       End for;
17. End if;
notes: GetWord(x) is the function which can derive the general word from the concept x
```

从 Algorithm3中可以看出,系统统一采用 HNC句类知识中的是否判断句(pDJ)作为语句生成的基本语义结构,这一类句子常见的表现形式是传统语言学中所指的"是"字句。当然,这里面存在许多相似和不同处,不再详细论述。

根据问题求解输出的"Q-Z-X-S"数据结构,根据是否判断句生成规则,最终生成的自然语言答复语句为:中国科学院声学研究所的地址是北京市海淀区北四环西路 21号,邮编是100080。

4 算法应用与测试结果

为了验证算法的有效性,我们对基于所提出算法的 HNCQA系统进行了测试,由于目前大多数的 QA评测都是基于英文的,于是我们在不改变提问类型的情况下模仿 TREC的 QA Track(http://trec.nist.gov/data/qa/2002_qadata/main_task_QAdata/t11_500_numbered.txt)制定了一套中文的评测系统。主要用于系统的纵向性能比较。另一方面,由于 HNCQA最终将服务于特定领域的应用,所以在制定评测数据的同时我们也考虑了测试数据的实用性。同时,由于是在不改变提问类型的情况下的模仿,所以保证了测试的可行性。举例来说:

TREC的 QA Track的第 1398个 Topic: What year was Alaska purchased? 相应的中文提问改为:中国科学院成立于哪一年?

目前系统常识库中存在有 27887条知识,对于 35次预先设立的主题询问以及 50次随机产生的限定域下的针对地址、联系方式、人物、数量等方面信息的询问或者错误的、无结果的、过于模糊的询问输入,系统一次做出正确回答的次数达到了 32和 46次,CWS(Confidence Weighted Score) = 91.76%。单纯从绝对数据值来看,相当程度高于在 TREC会议中一般基于结构数据库问答系统 70%左右的准确率。考虑机器自动生成结构化数据库的正确率 42.32%,系统总体表现性能为 91.76% * 42.32% = 38.83%,与 TREC会议的平均水平 30%相当,或者有所胜出。由于我们的测试集并不是 TREC会议的原始英文测试集,所以与 TREC会议中基准系统之间的单纯数据比较,可比性一般。但这从直观上说明了基于本文所提算法的 HNCQA确实具有相当高的准确率。为了更客观的评价系统,我们将所实现的系统与之前完成的基于关键词的结构数据库 QA系统的性能相比,基于概念匹配的中文问答系统 CWS提高了约 20个百分点。这从纵向对比实验中对系统的性能给予支持。

基于本文所提算法的 HNCQA模型目前已经应用于一个实际的短信问答系统中。根据应用的需要要我们在原型系统的基础上,又增加了:同音字容错处理;称谓(多)识别;对象模糊查询等功能,使得 HNCQA实用性增强。在这基础上,我们也进行了限定域下随机提问的测试,结果令人满意。

5 结论

本文就"基于概念匹配的中文问答处理系统"中存在的一些关键问题进行了讨论,并且提出了解决问题的方案和关键性算法法同时参照 TREC对系统进行了评测,评测结果表明基于

本文所提算法的 QA 系统具有较高的处理准确率,可以满足实际应用的需要。同时,从另一方面也可以看到,目前的结构化数据库自动生成能力还相对薄弱,有待于在下一步工作中提高。

参 考 文 献:

[1] Hongbo Xu, Hao Zhang, Shuo Bai ICT Experiments in trer-11 QA Main Task[A]. In: the Eleventh Text REtrieval Conference (TREC l1) [C]. 2002.

[2] Yi Chang, Hongbo Xu, Shuo Bai TREC 2003 Question Answering Track at CAS-ICT[A]. In: the Twelfth Text REtrieval Conference (TREC l2) [C]. 2003.

[3] Lide Wu, Xuanjing Huang, Junyn Niu, Yingiu Xia, Zhe Feng, Yaqian Zhou. FDU at TREC2002: Filtering, Q&A. Web and Video tasks[A]. 1lth Text REtrieval Conference, Gaithersburg[C]. USA, Nov, 2002.

[4] 黄曾阳. HNC (概念层次网络)理论 [M]. 北京:清华大学出版社, 1998.

[5] 黄曾阳. 语言概念空间的基本定理和数学物理表示式 [M]. 北京:海洋出版社, 2004.

[6] 张运良,张全. 一种基于 HNC理论的语义相关度计算方法 [J]. 计算机应用研究. 2006.

[7] 郑实福,刘挺, 秦兵, 李生. 自动问答综述 [J]. 中文信息学报. 2002, 16(6): 46 - 52.

[8] 吴友政,等. 问答式检索技术及评测研究综述 [J]. 中文信息学报. 2005, 19(3): 1 - 11.

(上接第 7页)

本分词后的语料得到的,其参数空间小,数据稀疏问题不会太严重,误报率当然就很低了,而且速度也大大提高。

参 考 文 献:

[1] Chao-Huang Chang A Pilot Study on Automatic Chinese Spelling Error Correction[J]. Communication of COLIPS, 1994, 4(2): 143 - 149.

[2] 张仰森,丁冰青. 基于二元接续关系检查的字词级自动查错方法 [J]. 中文信息学报, 2001, 15(3): 36 - 43.

[3] Lei zhang, Ming zhou, Changning Huang, Haihua Pan. Automatic detecting correcting errors in Chinese text by an approximate word-matching algorithm [A]. Microsoft Research China Paper Collection [C], 2000. 9, Vol 1: 135 - 141.

[4] 罗卫华,罗振声. 中文文本自动校对技术研究 [J]. 计算机研究与发展, 2004, 41(1): 244 - 249.

[5] 骆卫华,罗振声,龚小谨. 文文本自动校对的语义级查错研究 [J]. 计算机工程与应用. 2003, 39(12): 115 - 118.

[6] 龚小谨,罗振声. 中文文本自动校对中的语法错误检查 [J]. 计算机工程与应用. 2003, 39(8): 98 - 100.

[7] Li Jianhua, Wang xiaolong Combining Trigram and Automatic Weight Distribution in Chinese Spelling Error Correction[J]. Journal of Computer science and technology. 2002, Vol 17(6): 915 - 923.

[8] 张磊,周明,黄昌宁. 中文文本自动校对 [J]. 语言文字应用, 2001. 2, (1): 19 - 25.

[9] 张仰森,曹元大. 基于统计的纠错建议给出算法及其实现 [J]. 计算机工程, 2004, 30(11): 106 - 109.