基于 Web 的中文开放式问题回答系统*)

林旭东　彭　宏　郑启伦　陈绍坚

（华南理工大学计算机科学与工程学院　广州 510640）

摘　要　互联网正逐渐成为重要的信息资源，然而大多数搜索引擎不能处理自然语言提出的问题。基于互联网的中文问题回答系统由问题处理、信息检索、答案抽取和答案判断组成，利用命名实体识别、语义依存关系和案例规则模板实现答案抽取。实验表明：命名实体识别、语义依存关系和案例规则模板能有效地实现答案抽取，获得较高正确率。

关键词　问题回答，语义依存关系，命名实体识别，信息抽取

A Web-based Chinese Open-domain Question Answering System

LIN Xu-Dong　PENG Hong　ZHENG Qi-Lun　CHEN Shao-Jian

（College of Computer Science & Engineering, South China University of Technology, Guangzhou 510640）

Abstract　The Internet is increasingly being used as a source of reference information. Most popular search engines, however, are not designed for answering natural language questions. A Web-based Chinese question answering system is made of four parts: question processing, information retrieval, answer extraction and answer justification. It utilizes named entity recognition, semantic dependency relations and case-based rule to realize answer extraction. Experiments show that named entity recognition, semantic dependency relations and case-based rule perform very well in answer extraction.

Keywords　Question answering, Semantic dependency relations, Named entity recognition, Information extraction

1　引言

问题回答（简称 QA）系统为用户提供智能的人机界面，允许以自然语言提出问题，系统自动给出简短的正确答案。问题回答是机器学习、信息检索、信息抽取和自然语言处理等领域的研究热点[1]。虽然问题回答提出二十多年了，大规模的开放式问题回答在最近才成为一个重要的研究热点。

基于网络的开放式问题回答是一个具有挑战性的研究领域，需要对用户提出的问题进行适当的理解和表示，并从大量非结构化的文本信息中抽取答案相关的准确信息。问题回答系统克服传统搜索引擎的局限性：(1)问题回答允许用户以自然语言形式而不是关键词或其组合提出请求。(2)问题回答的输出更适合现代信息检索的需要，它返回给用户准确的答案，而传统的信息检索系统或搜索引擎返回相关文档或联接。问题回答将信息检索技术从关键词层次提高到句子的语义高度，更能满足用户的信息检索需求。

英文问题回答起步较早，MUC (Message Understanding Conferences)是早期的国际信息理解会议，设立了信息抽取专题，该会议最后一届是 1998 年的 MUC-7[2]。TREC (Text REtrieval Conference)是国际文本检索会议，1999 年 TREC-8 第一次设立问题回答专题，至 2005 年 TREC-14 连续七届，它的目标是基于大规模文本库建立能够回答实际问题的自动问题回答系统[3~8]。每年吸引了众多科研院所和公司参加，包括微软、IBM、卡耐基梅隆大学等。相对成熟的问题回答系统有密歇根大学的 AnswerBus、麻省理工的 Start、南加州大学的 Webclopedia、微软公司的 AskMSR、华盛顿大学的 MULDER 等。

中文问题回答起步相对较晚，没有统一的标准。本文提出一个基于互联网的中文开放式问题回答系统，对系统结构和主要功能模块进行讨论和分析，最后给出实验结论。

2　系统结构

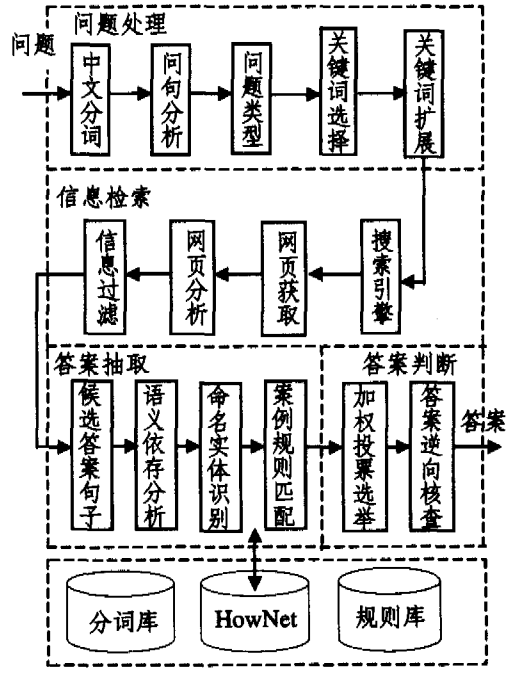

图 1　问题自动回答系统结构图

*)基金项目：广东省科技攻关项目(A10202001)，广州市科技攻关项目(2004Z2-D0091)。林旭东　博士研究生，主要研究方向为自然语言处理、Web 文本挖掘、信息检索；彭　宏　教授，博士生导师，主要研究方向为数据挖掘、智能计算、生物信息处理；郑启伦　教授，博士生导师，主要研究方向为数据挖掘、智能计算、生物信息处理、神经网络；陈绍坚　工程师，主要研究方向为自然语言处理、数据挖掘。

图1描述了问题自动回答系统的功能结构。与一般的检索系统不同，用户以自然语言提出问题，系统自动进行问题处理、信息检索、答案抽取和答案判断，最后直接返回正确答案。其中问题处理的中文分词模块使用中国科学院计算技术研究所研制的中文分词处理软件工具包2.0版及其分词库；关键词扩展基于《知网》(HowNet)的概念相关度和相似度计算；规则库包含命名实体识别规则、语义依存关系规则、案例规则等，主要用于答案抽取和答案判断。

3 问句处理

问题回答系统允许用户以自然语言形式的问句提问，系统需要理解问句的语义。问句处理需要完成中文分词、问句分析、问题类型、关键词选择和扩展。

在问句处理前，需要建立问题库。中文问题回答没有一个公认的中文问题库标准。我们使用哈工大信息检索研究室提供的问题库，包含6个大类，65个子类，共4394个问句。对其研究评估后，提取出有确定答案的2110个问句作为实验的问题库。

3.1 中文分词

中文句子是以连续的字符串的形式出现，词与词之间没有间隔，这就需要中文文本的分词处理。目前有许多中文文本自动分词方法和软件产品，我们使用中科院的中文分词处理软件工具包2.0版。它不但可以完成中文分词功能，还可以给出词性标注，分词正确率高达97.58%(973专家组评测)，未登录词识别查全率均高于90%。

3.2 问句分析

基于中文分词结果进行问句分析，结合《知网》进行问句的信息成分抽取和语义依存关系分析。首先确定问句谓词，它是句子各成分的分界点；其次提取疑问词；最后确定问句中心和约束条件。

例如：分析问句“谁是美国总统夫人?”

谓词：是；问句中心：夫人

疑问词：谁；约束条件：美国 总统

3.3 问题类型

问题类型是问句的焦点，基于问句分析的结果，主要根据疑问词，并参考问句中心和谓词确定人名、地点、数字、时间、简称等问题类型。问题类型与答案类型一致，对答案抽取规则具有重要指示作用。以下是疑问词与问题类型对应表。

表1 问题类型

类型	疑问词	问句中心/谓词
人名	谁，什么人，哪个人等	人
地点	哪里，何处，什么地方等	位于
数字	多少，多大，多高，多长	面积，长度等
时间	哪年，何时，什么时间等	发生
简称	什么	简称，缩写

3.4 关键词选择和扩展

我们需要在问句中提取出对信息检索系统有用的关键词。关键词主要由名词、动词、形容词、限定性副词等实义词组成。结合《知网》计算概念相似度，进行关键词的概念扩展。关键词扩展能提高信息检索的查全率，但扩展不当会极大地降低信息检索的正确率。因此我们对每个关键词 k，最多扩展一个与其概念相似度最高的词 ek，并且要求关键词 k 与扩展关键词 ek 的概念相似度 $Sim(k,ek)\geqslant 0.95$。

4 信息检索

信息检索的任务是根据关键词在语料库中查找相关的文档。基于Web的问题回答系统以互联网的海量信息为语料库，可以使用已有的搜索引擎完成关键词信息检索功能。信息检索功能主要包括搜索引擎、网页获取、网页分析和信息过滤。

目前市场上有许多优秀的搜索引擎，其中Google是最具代表的搜索引擎，但在教育网上需要使用代理，而且速度较慢。百度(Baidu)是一个较好的中文搜索引擎，拥有最大的中文信息库，总量超过6千万网页以上。由于百度没有提供搜索的Web Service，需要根据关键词的GB2312编码构造URL，然后下载。如：世界 最高 峰，这三个关键词在百度的检索URL为：http://www.baidu.com/s?wd=%CA%C0%BD%E7+%D7%EE%B8%DF+%B7%E5。

下载检索结果后对其进行网页分析，获取相关网页的标题、摘要和URL，我们只对前100个网页进行分析处理，利用信息过滤模块将下载的网页转换成纯文本格式，形成答案候选文档集。

5 答案抽取

5.1 答案候选句子集

(1)问句与句子的关键词相似度

首先根据标点符号将答案候选文档集的每个纯文本文档切分为句子，然后将每个句子进行中文分词和词性标注，形成原始答案候选句子集。

定义1(关键词与词相似度) 任意一个词 w，关键词 k 与 w 的相似度记为：

$$KeySim(k,w)=\begin{cases}1, & \text{if } w=k\\ Sim(k,ek), & \text{if } w=ek\\ 0, & \text{others}\end{cases}$$

其中 ek 是与关键词 k 最相似的扩展关键词，关键词 k 与其扩展关键词 ek 的概念相似度 $Sim(k,ek)$ 由《知网》计算出来。

定义2(关键词与句子相似度) 任意一个句子 S，由已经切分的词串 $w_1w_2\cdots w_n$ 组成。关键词 k 与 S 的相似度记为：

$$KeySim(k,S)=\mathrm{MAX}_{j=1}^{Len(S)}(KeySim(k,w_j))$$

其中句子 S 的长度为 $Len(S)=n$。

定义3(问句与句子的关键词相似度) 假设从问句 Q 抽取出 $m=KeyCount(Q)$ 个关键词 $k_1,k_2,\cdots,k_m$。候选句子 S 与问句 Q 的关键词相似度记为：

$$KS(Q,S)=\frac{\sum_{i=1}^{m}KeySim(k_i,S)}{KeyCount(Q)}$$

显然 $0\leqslant KeySim(Q,S)\leqslant 1$，根据问句 Q 与句子 S 的关键词相似度的大小将原始答案候选句子集进行降序排列，排在前面关键词相似度最高的100个句子作为答案候选句子集。

(2)问句与句子的距离相似度

定义4(词之间的距离) 句子 S 由词串 $w_1w_2\cdots w_n$ 组成，词 w_i 的位置为 $Pos(w_i)=i$，w_i 与 w_j 之间的距离定义为：

$$Dis(w_i,w_j)=Abs(Pos(w_i)-Pos(w_j))=Abs(i-j)$$

定义5(关键词与句子的距离相似度) 关键词 k 与句子 S 的距离相似度定义为关键词 k 与 S 的问句中心关键词 $fKey$ 的距离相似度：

$$DisSim(k,S)=\mathrm{MAX}_{j=1}^{len(S)}(1-\frac{Dis(w_j,fKey)}{Len(S)})KeySim(k,w_j)$$

一般地，若问句中心关键词 $fKey$ 存在最相似扩展关键词 $efKey$，则

$$DisSim(k,S)=\mathrm{MAX}\{\mathrm{MAX}_{j=1}^{len(S)}(1-\frac{Dis(w_j,fKey)}{Len(S)})KeySim(k,w_j),\ MAX_{j=1}^{len(s)}(1-\frac{Dis(w_j,efKey)}{Len(s)})KeySim(k,w_j)*Sim(fKey,efKey)\}$$

定义 6(问句与句子的距离相似度) 问句 Q 与句子 S 的距离相似度定义：

$$DisSim(Q,S)=\frac{\sum_{i=1}^{n}DisSim(k_i,S)}{KeyCount(Q)}$$

显然 $0\leqslant DisSim(Q,S)\leqslant 1$，问句 Q 与句子 S 的距离相似度越高，说明 S 包含问句 Q 的答案的可能性越大。

(3)问句与句子的语义依存距离相似度

语义依存关系是句子中词与词之间的句法关系，这种句法关系是有方向的，通常一个词支配另一个词，这种支配与被支配的关系体现了词在句子中的语义依存关系[9]。

实验中发现，关键词之间的绝对距离不能准确地反映句子的语义依存关系。例如："王伟妻子阮国琴致信美国总统布什"，关键词"美国"和"总统"与问句中心关键词"妻子"的距离很近，但这两个关键词都不是用来限定或约束问句中心关键词"妻子"。因此有必要区分关键词与问句中心关键词存在哪种语义依存关系。

定义 7(词之间的语义依存距离) 句子 S 由词串 $w_1w_2\cdots w_n$ 组成，词 w_i 的位置为 $Pos(w_i)=i$，词 w_i 与词 w_j 之间的语义依存距离定义为：

$$Dis(w_i,w_j)=\begin{cases}Abs(i-j), & w_i\text{ 与 }w_j\text{ 存在语义依存关系}\\ Len(S), & w_i\text{ 与 }w_j\text{ 不存在语义依存关系}\end{cases}$$

将词之间的距离用语义依存距离替代后，定义 5 的关键词与句子的距离相似度变成语义相似度，定义 6 的问句与句子的距离相似度也变成语义相似度。这样可以更精确地反映问句与答案候选句子的相似度。

5.2 命名实体识别

命名实体是文本中基本的信息元素，是正确理解文本的基础[6]。狭义地讲，命名实体是指现实世界中具体的或抽象的实体，如人、组织、公司、地点等；广义地讲，命名实体还可以包含时间、数量表达式、住址、电子邮件地址、电话号码、编号、名称等。

命名实体识别就是要判断一个文本串是否代表一个命名实体，并确定它的类别。在信息抽取研究中，命名实体识别是目前最有实用价值的一项技术。命名实体识别的方法很多，有隐马可夫模型(HMM)，支持向量机(SVM)，基于规则和基于统计的方法等[3]。

5.3 案例规则模板

通过命名实体识别，根据预测的答案类型在答案候选句子中筛选出候选答案，很多情况可能筛选出多个符合条件的候选答案，需要通过案例规则作进一步判断。案例规则主要依据语义依存关系[9]，通过大量训练形成案例规则模板，主要有以下几个类型：

(1)同位语关系规则 名词词性答案类型多采用同位语关系规则，如人名、地名、缩写等。答案通常与问句中心关键词在答案候选句子中以同位语关系出现，例如：

世界最高峰 珠穆朗玛峰海拔近 9000 米。

(2)主谓关系规则 当问句的疑问词是问句的主语，且问句谓词多数情况是实义动词(在问句处理时被抽取为关键词)，可以使用主谓关系规则识别答案。在命名实体识别的候选答案基础上，选择与这样的问句谓词构成主谓关系的候选答案作为正确答案。例如：

问句：谁发明了电话？

句子：贝尔 发明电话。

(3)动宾关系规则 与主谓关系规则类似，只是问句的疑问词是问句的宾语，可以使用动宾关系规则识别答案。选择与问句谓词构成动宾关系的候选答案作为正确答案。例如：

问句：张衡发明了什么？

句子：张衡发明了地动仪。

(4)状语关系规则 通常时间、地点、方式等类型的问句多采用状语关系规则。答案通常以状语的形式出现在答案候选句子中，多为方位结构、地点结构、时间结构等。例如：

问句：秦始皇统一中国在哪年？

句子：秦始皇二十六年(前 221)，秦始皇统一中国，……。

语言千变万化，案例规则也需不断细化以适应多种变化。在实际使用中需要使用多种规则进行分析判断，例如主谓关系和动宾关系在主动句与被动句中可以相互转换；同一问句的答案也可作为多种句子成分出现。另外使用规则可能将一些正确答案也过滤掉，但筛选出来的答案正确率高。由于互联网上信息量大，冗余多，在数据密集的情况下牺牲掉覆盖率，保障了正确率的提高。

6 答案判断

通过命名实体识别和案例模板规则可以抽取候选答案句子的答案信息，从 100 个取候选答案句子中可能抽取出不同的答案，哪个答案是最正确的呢？需要算法能自动进行答案判断。

6.1 加权投票选举

一个简单直接的办法是进行投票选举，哪个答案出现的次数最多就是正确答案。但这等于假设每个答案具有相同可信度，实际情况可能更加复杂。因此我们采用加权投票选举，对每个答案进行评估，给出评分。我们利用问句 Q 与候选答案句子 S 的语义相似度给答案进行可信度评估。

定义 8(答案可信度) 基于问句 Q 与候选答案句子 S 的语义相似度，给出从 S 中抽取的答案 a 的可信度：

$$Score(a)=\alpha^{DisSim(Q,S)}$$

其中 α 是常数，由于 $0\leqslant DisSim(Q,S)\leqslant 1$，故 $1\leqslant Score(a)\leqslant\alpha$，实验中我们设置 $\alpha=10$。

求出单个答案的可信度后，将相同答案的可信度相加得出同一答案的总评分，获得最高评分的答案作为正确答案。

6.2 答案逆向核查

对于有确定答案的问题类型，可以使用答案逆向核查方法进行自动答案判断。即以答案为关键词进行检索，从检索出来的信息中过滤出包含答案的句子，然后利用答案抽取算法判断句子中包含的问句关键词与答案的语义相似度。用答案逆向核查方法获取的答案语义相似度作为判断答案的可信度依据。结合加权投票选举和答案逆向核查，可以有效地自动判断正确答案。

6.3 实验结果

表 2 实验结果表明我们的中文问题回答原形系统获得了较高正确率。

(下转第 226 页)

的关联表示盱眙是十三香龙虾的原产地。

(4)该认知活动中,小张是单一学习者,无需与他人交互,进而图 8 即为其对南京这次认知的最终结果。当然,如果他还有更多的认知需求,可以再次向知识网格提出请求,对南京进行更进一步的了解。

其中部分主题、关联以及合并操作的 XTM 实现如下,其它的 TMs 相关实现均与此类似:

```
〈! - 小张的主题地图中的"南京"主题 -〉
〈topic id="南京"〉
    〈instanceOf〉
        〈topicRef xlink:href="#city"/〉
    〈/instranceOf〉
    〈baseName〉
        〈baseNameString〉
        南京〈/baseNameString〉
    〈/baseName〉
    〈occurrence〉
        〈resourceSet〉
    〈/occurrence〉
〈/topic〉
〈! - 图 6 中"中山陵位于南京"的关联 -〉
〈association〉
    〈instanceOf〉
     〈topicRef xlink:href="#is located in"/〉
    〈/instanceOf〉
    〈member〉
     〈roleSpec〉
     〈topicRedf xlink:href="#some place"/〉
     〈/roleSpec〉
     〈topicRef xlink:href="#南京"/〉
    〈/member〉
    〈member〉
     〈roleSpec〉
     〈topicRedf xlink:href="#scenery"/〉
     〈/roleSpec〉
     〈topicRef xlink:href="#中山陵"/〉
    〈/member〉
〈/association〉
〈! - 将图 6 与当前持有知识即图 5 合并 -〉
〈mergeMap xlink:href="#图 6"〉
    〈resourceRef xlink:href="#图 6"/〉
〈/mergeMap〉
```

结束语 知识表示和传播对于知识网格环境下的协同认知至关重要。本文利用 TMs 技术这一知识表示的有效手段,对协同认知进行了深入的研究。基于网格的特点,提出了对 TMs 的扩展,使其不仅能提供对信息资源的访问,还能使引用某主题的实体获得该主题所拥有计算资源的使用权。并根据 TMs 的特点提出了有别于传统信息查询的新型查询机制,使得用户能够更准确地得其所需。本文针对协同认知的不同形式,定义了元学习和群学习过程,将机器学习和人脑学习结合进行,以更好地完成认知,有效地引领了认知的顺利开展。后续研究中,我们拟将对网格环境下基于 TMs 知识表示的内容查询、认知过程中的学习机制、协同方式,以及通过主体自省以提高学习能力的具体实现做进一步的研究。此外,我们会试图将 TMs 模式应用到更多的网格引用中,特别是知识网格,尽管会伴随更多的挑战。

参 考 文 献

1 Hoc J M. Towards a cognitive approach to human-machine cooperation in dynamic situations. Int J Human-Computer Studies, 2001,54:509～540

2 Garshol L M. Living with TMs and RDF. http://www.ontopia.net/topicmaps/materials/tmrdf.html

3 Hai Zhuge. A Knowledge grid model and platform for global knowledge sharing. In:Expert systems with application, Elsevier, 2002,22:313～320

4 Foster I. An anatomy of grid. Intl J Supercomputer Applications, 2001

5 Hai Zhuge. Semantics, Resource and Grid. In:Future generation computer systems, Elsevier,2004, 20:1～5

6 Hai Zhuge. A knowledge flow model for peer-to-peer team knowledge sharing and management. In:Expert systems with applications,2002,23:23～30

7 http://www.topicmap.com

8 http://www.topicmap.org/xtm/1.0

9 Sova J F. Knowledge Representation: Logical, Philosophical, and Computational Foundations. Brooks/Cole, 2000

10 Power R. TMs for context management, July, 2003

11 Pepper S. The TAO of TMs. http://www.ontopia.net/topicmaps/materials/tmrdfoildaml.html

12 Hai Zhuge. Semantic Resource Exploitation with TMs

13 Abel M H,Lenne D,Moulin C,et al. Using topic maps in an E-learning context. In:ICWE2004, LNCS3140,2004. 581～582

14 Chen Weiqin. Reuse of collaborative knowledge in discussion forums. ITS2004, LNCS3220,2004. 800～802

15 窦万春. 知识网格环境下认知协作的工作流原理、集成方法与原型系统研究. 国家自然基金资助项目 60303025 申请书

(上接第 213 页)

表 2 实验结果

类型	数量	正确数	错误数	正确率
人名	92	71	21	77.2%
地点	630	412	218	65.4%
数字	878	509	369	58.0%
时间	422	320	102	75.8%
简称	88	72	16	81.8%
总计	2110	1384	726	65.6%

结论 本文实现了一个基于互连网的中文开放式自然语言问题自动回答的原形系统,利用命名实体识别、语义依存关系和案例规则模板实现答案抽取,对有明确答案的简单问题具有较高的正确率。今后我们将不断完善问题自动回答系统,进一步深化和细化命名实体识别、语义依存关系分析和案例规则模板。同时运用范例推理,建立知识库、规则库和过程库,提高答案抽取的正确率,使系统具有在线问题回答能力。

参 考 文 献

1 Brill E, Dumais S, Banko M. An analysis of the Ask-M3R Question-answering system. In:Proc. of 2002 Conference on Empirical Methods in Natural Language Processing, 2002

2 Chinchor N, Marsh E. MUC-7 Information Extraction Task Definition (version 5.1). In:the Proceedings of MUC-7, 1998

3 Srihari R, LI Wei. Information Extraction Supported Question Answering. In:the Proceedings of TREC-8, 1999

4 Vicedo, Luis J, Ferrández, Antonio. A Semantic approach to Question Answering systems. In:the Proc. of TREC-9, 2000

5 Oh J H,Lee K-S,Chang D-S,Seo C W,Choi K-S. TREC-10 Experiments at KAIST: Batch Filtering and Question Answering. In:the Proceedings of TREC-10, 2001

6 Bellot P,Crestan E,El-Bèze M,Gillard L,de Loupy C. Coupling Named Entity Recognition, Vector-Space Model and Knowledge Bases for TREC-11 Question Answering Track. In:the Proceedings of TREC-11, 2002

7 CHANG Yi, XU Hongbo, BAI Shuo. TREC 2003 Question Answering Track at CAS-ICT. In: the Proceedings of TREC-12, 2003

8 TAN Wei, CHEN Qunxiu, MA Shaoping. THUIR at TREC 2004: QA. In:the Proceedings of TREC-13, 2004

9 LI Mingqin, LI Juanzi, WANG Zuoying, LU Dajin. A Statistical Model for Parsing Semantic Dependency Relations in a Chinese Sentence. Chinese Journal of Computers,2004,12(7)