Analysis and Improvement on Question Answering System Ramona

Xingfu YE [1], Ruiting LIAN [2], Hugo De GARIS[3]
Department of Cognitive Science, Xiamen University, Xiamen, China
E-mail: xingfuye@gmail.com;. lianlian1022@gmail.com; profhugodegaris@yahoo.com

Abstract: In general, a Question Answering system adopts a pipeline structure that contains three modules: Question Analysis, Information Retrieval and Answer Extraction. In this paper, we analyze the Question Answering system Ramona at first. Then, an improved algorithm of calculating sentence similarity has been applied to Ramona during the information retrieval stage, including the induction of keywords' synonyms, calculating word frequency and word order in the algorithm. On the other hand, we use the improved algorithm to calculate the sentence similarity between the user's question and FAQ's questions in the period of answer extraction. Then we choose the best answer from the FAQ or the best answer of Powerset's answers. At last, we select some sentences randomly to test our system, from the result can be seen that our improved system gets better results than the original one.

Keywords: Question Answering (QA); sentence similarity; Powerset; FAQ

问答系统Ramona分析和改进

叶兴甫 [1]，练睿婷 [2]，Hugo De Garis[3]
厦门大学智能科学与技术系，厦门，中国，361005
E-mail: xingfuye@gmail.com; lianlian1022@gmail.com; profhugodegaris@yahoo.com

摘　要：问答系统一般包括问题分析、信息检索和答案抽取三个主要部分。本文首先分析了问答系统Ramona，并在此基础上提出了改进的方法：在信息检索阶段运用改进的计算句子相似度算法，包括在算法中引入关键词的同义词和计算词频、词序等；在答案抽取阶段使用该算法计算出用户问题和FAQ问题的相似度，并从FAQ和Powerset的答案中选择出最佳答案。最后我们随机选择了一些句子对其进行测试，从实验结果可看出，改进后的系统比原系统获得了更好的测试性能。

关键词：自动问答；句子相似度；Powerset；常见问题集

1 引言

随着信息技术的快速发展和广泛普及，人们对于获取的信息有着更高的要求。虽然传统的搜索引擎google和百度仍发挥着重要的作用，但仍无法满足人们的某些需求，例如：在google中输入一个关键词，常常会找到成千上万个网页，用户必须逐一阅读这些网页才能找到精确的答案。这就促使了新的技术的发展——自然语言的问答系统。

问答系统是指这样一种机器系统：对于用户通过自然语言输入的问句，它能够给出简洁、准确、人性化的回答，这是传统的搜索引擎不能比拟的[1]。

现有问答系统大致可以分为：基于知识库的问答系统、问答式检索系统、基于自由文本的问答系统等[2]。这也是目前的研究状况。

基于知识库的问答系统优点在于准确回答用户提出的问题，甚至可以进行一定程度的推理计算。但此类系统受知识库的限制。

问答式检索系统的代表是Start系统[3]，与基于知识库的问答系统相似：对于知识库外的问题，Start系统只能返回网页链接给用户。

基于自由文本的问答系统被认为是最有前途的问答系统，代表着问答系统的发展方向。其特点是由于不需要建立大规模知识库，节省了大量的工作量，并且系统返还给用户的是问题的具体答案。

本文首先简要叙述问答系统的研究概况，然后介绍了Ramona问答系统并对其核心技术进行了分析，接着提出了改进的计算句子相似度的方法和答案选择机制，

基金项目：国家自然科学基金（批准号 60975084）；福建省自然科学基金（批准号：2009J01305）

978-1-935068-10-5 © 2010 SciRes.

最终利用改进的方法测试Ramona系统并取得了更好的实验效果。

2 Ramona 系统概述

Ramona 是基于自由文本、以 ProgramD 为平台的问答系统，它由 Ben Goertzel 和 Murilo Saraiva de Queiroz 的 OpenCog 团队开发。其特点是以 Powerset 作为搜索引擎来查询资料，相比较传统搜索引擎更能解答用户提出的问题。简言之 Powerset 能理解用户所搜索的关键词的含义以及多个关键词之间的关联，从而找出符合人类思维的真正的相关结果。

其具体流程图如图 1 所示。

一个完整的执行过程为：当用户 User 提出一个问题，系统会将此问题交由 ProgramD 分析并将问题传输给产生答案的模块，由 Powerset 模块和 FAQ 模块分别产生候选答案，最后系统在所有候选答案中选最佳答案 BestAnswer 返回给用户。

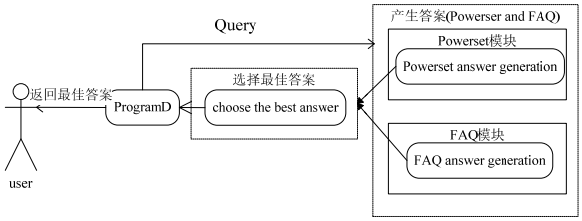

Figure 1. Ramona system flowchart
图 1. Ramona 系统流程图

接下来本文将对其核心模块进行简要介绍，并重点分析我们对其改进的模块。

2.1 Powerset 模块

Powerset 模块主要利用 Powerset 引擎获取相关的资料并选取出答案。图 2 为 Powerset 模块的数据流程图。系统的处理流程：输入问题 –> 问题分类 -> 关键词提取-> 问句扩展 ->相关文档的获取 -> 包含答案的句子排序(句子相似度计算 TFIDF 方法) –> 最终答案的获取。

Ramona的问题分析包括问题分类、关键字提取和问句扩展。其实现方法类似于文献[4]。对于用户提出的问题，Ramona将其归类。表1列出了问题类型[5]的分类。

[1] http://aitools.org/Program_D
[2] http://www.opencog.org/
[3] http://www.powerset.com/

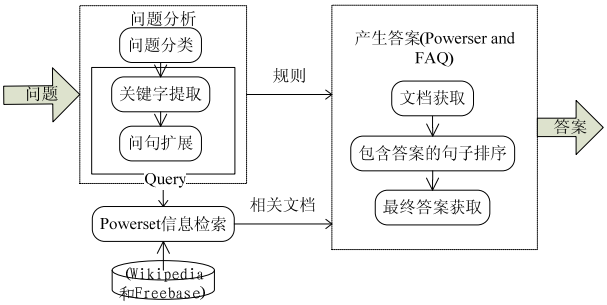

Figure 2. Powerset module flowchart
图 2. Powerset 模块流程图

Table 1. The question class definition
表 1. 问题分类定义

问题类型	模板	例子
人物	Who	Who is Bill gates?
地点	Where	Where is Xiamen University?
时间	When	When did Bill found the company?
数量	How much	How much people are t here in the world?
原因	Why	Why did Einstein go to the USA?
定义	What	What is a computer?
其他	Others	Others

2.2 FAQ 模块

FAQ 模块主要功能是把用户经常问的问题、答案保存起来。对用户提问的问题先在 FAQ 数据库中搜索是否存在相同或者相似的问题。FAQ 的存在，使系统既能提高效率又能提高答案准确性。Ramona 中问题和对应的答案的存储形式为 AIML 格式。图 3 为 FAQ 模块详细数据流程图。系统的主要处理流程：输入问题 –> 问题分类 -> 关键词提取-> 问句扩展 ->包含答案的句子排序(句子相似度计算) –> 最终答案的获取。

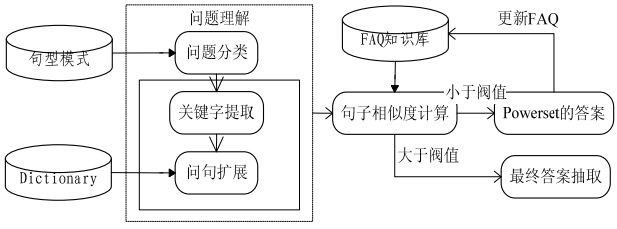

Figure 3. FAQ module flowchart
图 3. FAQ 模块流程图

3 本文对 Ramona 的改进

Ramona使用基于向量空间模型的TFIDF方法来计算答案与查询序列的相似度，文献[6]和文献[7]分别

分析了这种经典的方法并不总是工作的很好。本文使用一种新的计算句子相似度的方法并在答案选取阶段使用阀值的思想[8]，与原方法相比得到了更为满意的结果。

3.1 句子相似度计算

Powerset 搜索引擎返回的答案一般是几个连续的句子或者一篇文章，而用户通常要求的答案往往是精确的句子。系统对于返回的答案，使用最大熵模型[9]划分出每一个句子作为问题的候选答案。

在分离出句子的关键词的基础上，对句子中关键词（关键词和其所有同义词视为等价）的个数、句子长度、关键词之间的距离的有关信息进行整合，计算出查询序列和答案之间的相似度。具体的实现方法是：

（1）关键词词形相似度

计算公式为：

$$word\_similar(S_1,S_2)=\frac{same\_word(S_1,S_2)}{word(S_1)+word(S_2)-same\_word(S_1,S_2)}$$

其中 $same\_word(S_1,S_2)$ 表示句子 S_1 和 S_2 相同关键词个数， $word(S_1)$ 表示 S_1 句子中关键词的个数，$word(S_2)$ 同理。词形相似度主要反映的是句子形态上的相似性，起关键作用的在于两个句子关键词相同的个数。

（2）关键词个数相似度

计算公式为：

$num\_similar(S_1,S_2)=1-\left|\frac{num(S_1)-num(S_2)}{num(S_1)+num(S_2)}\right|$，其中 $num(S_1)$ 表示 S_1 中(关键) 词的个数， $num(S_2)$ 同理。关键词个数相似度从一定意义上也反映了两个句子之间形态上的相似。如果两个句子中关键词数量完全相同，并不能认为这两个句子就是完全相似的。定义（2）只是一个辅助的判别条件，还要求和其他特征结合起来。

（3）关键词距离相似度

计算公式为：

$$dis\_similar(S_1,S_2)=1-\left|\frac{same\_dis(S_1)-same\_dis(S_2)}{dis(S_1)+dis(S_2)}\right|$$，其中 $same\_dis(S_i)$ 表示 S_1，S_2 中相同的关键词在 S_i 中的距离，i =1,2。若关键词重复出现多次,以产生最大距离为准。距离相似度也从一定意义上反映了两个句子之间形态上的相似，其关键在于两个句子中相同关键字之间的距离来判定句子 S_1 和 S_2 是否相似。

（4）句子相似度

在计算出两个句子 S_1 和 S_2 关键词词形相似度、关键词个数相似度和关键词距离相似度之后，我们根据上述三种特征定义句子相似度计算公式：

$sen\_similar(S_1,S_2)=k_1*m_1+k_2*m_2+k_3*m_3$，其中 m_1、m_2、m_3 分别表示：

$m_1=word\_similar(S_1,S_2)$，

$m_2=num\_similar(S_1,S_2)$，

$m_3=dis\_similar(S_1,S_2)$，

$k_1+k_2+k_3=1$ 且 $k_1>k_2>k_3$，改进的系统中我们使用 $k_1=0.5$， $k_2=0.3$， $k_3=0.2$。

利用 $sen\_similar(S_1,S_2)$ 公式计算出 Powerset 返回的答案中排名前 30 个网页所有句子的相关度并将这些句子按照相关度排序，最后取前 300 个句子作为用户问题的候选答案，以保证答案的全面性。

3.2 Powerset 模块的答案抽取

对于上一步取得的 300 个候选答案我们使用下列方法来取得 Powerset 模块的最佳答案。

(1)使用信息抽取中命名实体的抽取方法对每个候选答案进行分类。

抽取答案时，首先用IdentiFinder[10]算法对候选句子进行处理找出所有的命名实体并识别其类型。该算法基于隐马尔科夫模型，由Daniel. M. Bikel和RichardSchwariz首先提出，通过机器学习对人名、日期、时间、数字等进行正确地识别和分类。然后系统将这些命名实体中与用户查询类型相同的句子作为问题的候选答案。

(2)原问题和分类的答案的相关度计算使用下列计算公式计算。为了防止出现一个句子异常而导致答案不准确的情况，系统整体考虑所有句子的序号位置的思想，即第 k 个句子的实体类别在类型相同的候选答案中的影响力计算公式[11]为：

$$\mathrm{Re}\,sult[k]=\sum_{i=1}^{n}\frac{1}{num[i]}$$，

其中 $num[i]$ 是所有候选答案的句子的序号。例如句子的总数为 100 个，第 k 个句子的实体在 10、20、50 个句子中出现，则 $\mathrm{Re}\,sult[k]=\frac{1}{10}+\frac{1}{20}+\frac{1}{50}=0.17$。系统最终选择 $max\{\mathrm{Re}\,sult[k]\}$ 那个句子作为 Powerset 的答案。

3.3 系统最终答案的获取

对系统的最终答案来源于两个模块，即 Powerset 和 FAQ 模块。在应用新的方法计算出 FAQ 和用户问题句子相似度的基础上，使用阀值的思想，选择最佳

978-1-935068-10-5 © 2010 SciRes.

答案。其中 s_sim 表示用户问题和 FAQ 中所有句子的句子相似度。

(1)当 Max{s_sim} ≥ 阀值(系统阀值取 0.7)时，系统把 FAQ 模块的答案作为 BestAnswer 呈现给用户。

(2)当 Max{s_sim} ＜ 阀值，系统把 Powerset 模块的答案作为 BestAnswer 呈现给用户。然后把用户的问题和对应的答案加入 FAQ 库进行数据库的更新。

使用这样的方法既保证了 FAQ 领域内的准确性，又使得当用户问题不属于 FAQ 库范畴时 Powerset 模块保证 BestAnswer 的全面性。

(3)为了使答案更加的人性化，系统有一个名字数据库。当用户的问题是关于人名时，比如 Who is Bill Gates？，系统根据人名数据库自动识别性别，然后用 He 代替 Bill Gates 使句子更加的简练。另外，当用户询问原因时系统会在答案开头随机加上 I suppose、I think that 等语句以显得系统人性化。

4 实验结果

系统比较了 Ramona 系统改进前后的性能。具体方法是我们随机在互联网上选取 100 个问题，然后在改进前和改进后的系统上分别对这些问题进行测试并由 5 个人对答案进行评价，其评价标准为：

(1)如果系统的答案是正确的答案，那么评测的结果为“Excellent”；

(2)如果系统的答案是个可以接受的答案（比如基本关键词相同或者相似等)，那么评测结果为“Good”；

(3)如果答案不正确，则为“Bad”；

(4)满意度 = (Excellent + Good) / 问题数量。

表 2 是具体的测试结果。作为示例我们选择了五个问题来反映改进前后的差异。其中 Q 代表用户问题，R1 代表原 Ramona 系统，R2 代表改进后的系统。

Q1: Do you know Bill Gates?
R1: You know that he is the head-guy at the world's leading software company, but he was the one who developed the first software program that could run BASIC programs on micro computers, That's why he called his company Microsoft.
R2: William Henry "Bill" Gates III is an American business magnate, philanthropist, author, and chairman of Microsoft, the software company he founded with Paul Allen.

Q2: Where is Xiamen University?
R1: Xiamen, also known as Amoy, is a coastal sub-provincial city in southeastern Fujian province, People's Republic of China.
R2: Xiamen University, colloquially known as Xia Da, located in Xiamen, Fujian Province, is the first university in China founded by overseas Chinese.

Q3: What is NCURSES?
R1: It's a programming library providing an API, allowing the programmer to write text user interfaces in a terminal-independent manner.
R2: NCURSES is a programming library providing an API, allowing the programmer to write text user interfaces in a terminal-independent manner.

Q4: Who is the tallest person in NBA?
R1: I remember faces better than names. You said nothing.
R2: Yao Ming is a professional basketball player who plays for the Houston Rockets of the National Basketball Association.

Q5: What is a reading machine?
R5: This was one of Raymond Kurzweil's neat inventions, it reads printed material for the blind.
R5: A reading machine is a piece of Assistive Technology that allows blind people to access printed materials.

问题分析：

（1）改进前和改进后系统对 Q1、Q3、Q5 的回答都基本回答正确，表现出了良好的性能。

（2）对于 Q2，原系统只识别出 Xiamen 为关键词，而改进后系统正确识别“Xiamen University”。

（3）对于 Q4，原系统的回答不知所云，而改进后的系统虽然没有正面回答“Yao Ming is the tallest person”，但是回答并解释了出“Yao Ming”。

Table 2. The result of the experiment
表 2. 实验测试结果

测试	测试问题数量	Excellent	Good	Bad	满意度
原系统	100	37	23	40	60%
改进后	100	48	20	32	68%

5 结束语

在实际测试中，改进后的系统无论“Excellent”数值还是用户的满意度都有所提高，比原系统表现出了更好的性能。但是系统也存在着一些不足：

（1） 命名实体识别以及分词和词性标注对问题的分类的影响。因为分词的错误导致本体库中找不到实体。

（2）句子相似度的计算。可能两个句子相似但意义却不一样。原因是虽然是同一个关键词，但是因为一词多义现象的存在，导致结果不准确。

（3）系统返回的 BestAnswer 只是句子，而句子的表达能力是有限的。当用户希望得到的答案是一个单词、段落或者一篇文章时候，系统就显得力不从心，虽然现在也有些系统能实现这一功能[12]，但并不是工

作的很好，这是本系统需要继续深入研究和改进的地方。

References (参考文献)

[1] YAO Lin, LIANG Chun-xia, ZHANG De-gan. Design and implementation of case-based reasoning human-machine conversation system[J]. Journal of Computer Applications, 2007,(03)
姚琳，梁春霞，张德干. 基于实例推理的人机对话系统的设计与实现[J]. 计算机应用, 2007,(03)

[2] Wang Shuxi. Question Answering System: Core Technology, Application[J]. Computer Engineering and Applications, 2005, (18)
王树西. 问答系统:核心技术、发展趋势[J]. 计算机工程与应用, 2005,(18)

[3] Katz B. From Sentence Processing to Information Access on the World Wide Web. In Natural Language Processing for the World WideWeb：Papers from the 1997 AAAI Spring Symposium，1997：77-94

[4] S. Harabagiu, Moldovan et al., FALCON: Boosting Knowledge for Answering Engines. TREC 2000 Proceedings, 2000.

[5] Xiaojie Yuan, Shitao Yu, Jianxing Shi, Qiushuang Chen. Question classification in question answering based on real-world web data sets [J]. Journal of Southeast University (English Edition), 2008, 24(3): 272-275.

[6] SHI Cong-Ying, XU Chao-jun, YANG Xiao-Jiang. Study of TFIDF algorithm[J]. Journal of Computer Applications, 2009,(S1)
施聪莺,徐朝军,杨晓江. TFIDF 算法研究综述[J].计算机应用, 2009,(S1)

[7] ZHANG Yufang, PENG Shiming, LV Jia. Improvement and Application of TFIDF Method Based on Text Classification[J]. Computer Engineering, 2006,(19).
张玉芳，彭时名，吕佳. 基于文本分类 TFIDF 方法的改进与应用[J]. 计算机工程, 2006,(19).

[8] ZHOU Fa-guo, YANG Bing-ru. New method for sentence similarity computing and its application in question answering system[J]. Computer Engineering and Applications, 2008,(01)
周法国， 杨炳儒. 句子相似度计算新方法及在问答系统中的应用[J]. 计算机工程与应用, 2008,(01)

[9] J. Reynar and A. Ratnaparkhi. A Maximum Entropy Approach to Identifying Sentence Boundaries, In Proceedings of the Fifth Conference on Applied Natural Language Processing, Pages 16-19, Washington D.C.1997.

[10] Bikel D，Schwartz R，Weischedel R. An Algorithm that Learns What's in a Name。 Machine Learinging-spcial Issue on NL Learning， 1999，34：1-3

[11] ZHANG Yongkui, ZHAO Zheqian, BAI Lijun, CHEN Xinqing. Internet-based Chinese Question-answering System Computer Engineering, 2003,(15)
张永奎，赵辄谦，白丽君，陈鑫卿. 基于互联网的中文问答系统[J]. 计算机工程, 2003,(15)

[12] ZHENG Shi fu, LIU Ting, QIN Bing, LI Sheng. Overview of Question-Answering[J]. Journal of Chinese Information Processing，2002，6 (16)：46–52
郑实福，秦兵，刘挺等. 中文自动问答系统综述[J].中文信息学报，2002，6 (16)：46–52