基于领域语义信息的百科问答系统*

韩先培[1] 齐振宇[1] 田野[2] 王渝丽[2] 赵军[1]

1. 中国科学院自动化所模式识别国家重点实验室 100190, 2.中国大百科全书出版社 100037

E-mail: xphan@nlpr.ia.ac.cn, jzhao@nlpr.ia.ac.cn

摘 要: 本文构建了一个基于领域语义信息的百科问答系统，描述了如何在问答系统的语料预处理、问句处理和答案抽取模块中引入领域语义信息来提升问答系统的性能。实验结果表明，相比于检索系统和未加入语义信息的问答系统，基于领域语义信息的问答系统在 MRR(平均排序倒数) 性能上分别提升了 34%和 20%。

关键词: 问答，问答系统，语义元数据，语义标注

Enhancing Encyclopedia Question Answering by Leveraging Domain Specific Semantic

HAN Xianpei[1], QI Zhenyu[1], Tian Ye[2], Wang Yuli[2], ZHAO Jun[1]

1. National Laboratory of Pattern Recognition Institute of Automation, Beijing 100190

2. Encyclopedia of China Publishing House, Beijing 100037

E-mail: xphan@nlpr.ia.ac.cn, jzhao@nlpr.ia.ac.cn

Abstract: This paper constructs an encyclopedia question answering system based on domain specific semantic. We describe how to leverage the domain specific semantic in the data preprocessing module, question processing module and answer extraction module to enhance the performance of question answering system. The experiment results show that, compared to the information retrieval system and the question answering system not adding semantic information, our system achieved 34% and 20% MRR performance improvement respectively.

Keywords: question answering, question answering system, semantic metadata, semantic annotation.

1 引言

互联网的迅猛发展导致了网络信息的爆炸性增长，将大量有价值、高质量的信息淹没在信息海洋里。这大大增加了快速、准确地获得有价值信息的难度。因此，有必要开发出精确、智能的信息服务系统来满足用户的需求，将高质量的信息呈现给用户。目前的主流信息服务系统是基于信息检索技术的搜索引擎。搜索引擎在商业领域取得了空前的成功，涌现了大量商业公司，如 Google，Yahoo！等等。现有的搜索引擎在基于关键词的信息查询上面取得了极大的成功，但是不能处理更精确的信息服务需求：1. 仅仅使用关键词无法完全表达用户所有的需求；2. 关键词匹配的算法没有涉及语义；3. 搜索引擎返回的结果是文档和网页，其粒度通常过大，且结果中包含大量的重复信息，冗余度太高。

对更精确的信息服务的需求促进了问答系统的发展。相比于搜索引擎技术，问答系统在以下两个方面进行了改进：一是查询方式是自然语言描述的问句，用户能够更好的描述其信息需求；二是返回的答案能直接回答问题，用户不需要通过进一步的阅读来寻找答案。

尽管问答系统能更好的回答信息服务的需求，但是真实世界中的用户提问通常非常复杂，回答这些问题需要大量世界知识语义以及强大的自然语言推理技术。由于这些方面的限制，目前的问答系统仍然局限于处理有限的简单事实性的问题，如 Factoid 问题，List 问题等等。随着互联网的进一步的发展，互联网上涌现了许多包含大规模语义信息的系统，

*本文受国家自然科学基金项目（60673042, 60875041）、国家 863 计划项目（2006AA01Z144）和中国出版集团科技项目资助。

如 Wikipedia，Delicious 等等。大量的语义信息以链接文本、语义标签、Encyclopedia 的形式存在。以这些大规模的语义信息为基础，进一步促进和提升现有的信息服务变得可能。

为了探索如何使用语义信息来提升问答系统的性能，本文实现了一个基于领域语义信息的问答系统。通过使用领域特定的知识描述体系，我们在问答系统的三个方面引入了语义信息：1. 在预处理语料时，对文本片段的语义信息进行了分析，并在索引中加入了这部分语义信息；2. 在处理用户问句时，通过分析问句的语义关键词来更好的理解用户的信息需求；3. 在答案抽取时，利用问句和候选答案之间的语义相似度来提升正确答案的排序。实验结果表明，语义信息能有效的提升问答系统的性能。

本文的安排如下，第二小节综述问答系统的相关工作；第三小节详细描述了基于领域语义信息的问答系统；第四小节提供了实验的结果和分析；最后以总结本文的工作结尾。

2 相关工作

目前，问答系统技术主要分为基于检索的问答技术、基于模式匹配的问答技术和基于浅层自然语言处理技术的问答技术。

基于检索的问答通过检索来抽取答案的候选，然后通过排序算法来对候选答案进行排序。目前排序算法主要依据提问处理模块生成的查询关键词。依据关键词对排序贡献的不同，算法通常把查询关键词分为几类：1）普通关键词；2）扩展关键词：对普通关键词进行扩展，通常是从 WordNet 或者 Web 中扩展同义词；3）基本名词短语；4)其他关键词。在实际排序中，使用不同的加权来指示不同关键词的重要程度。系统[H. Yang, et al. 2002]使用答案关键词和提问关键词的覆盖度来对答案候选进行打分；系统[A. Ittycheriah, et al. 2002]使用 Inverse Sentence Frequency 来进行打分。基于检索的问答技术代表系统参见新加坡国立大学的[H. Yang, et al. 2002]系统。

基于模式匹配的问答技术通过自动获取某些类型提问(如某人的出生日期)的尽可能多的答案表述模式来设计问答系统。基于模式匹配的方法先离线地获得各类提问答案的模式[D. K. Lin, et al 2001; D. Ravichandran, et al. 2002; E. Brill, et al. 2001]，在运行阶段，系统首先判断当前提问属于哪一类，然后使用这类提问的所有模式来对抽取的候选答案进行验证。

以上两种方法简单、有效。但是要更大程度地提高问答系统的性能，必须引入自然语言处理的技术[E. Nyberg, et al. 2003]。由于目前自然语言处理的技术还不够成熟，因此大多数系统都是作为对前两种方法的补充和改进。这方面代表性工作有[D. Moldovan, et al. 2001; D. K. Lin, et al 2001; E. Hovy, et al. 2001; M. Pasca. 2001]。

3 基于领域语义信息的百科问答系统

在本小节中，我们首先对领域语义信息进行描述，接着介绍了整个问答系统的框架，然后依次分析了如何在问答系统中的三个模块中引入领域语义信息。

3.1 领域语义信息

目前本文系统使用的领域语义信息包括两部分：第一是领域内实体的分类体系；第二是描述每一个类别实体的一组语义元信息。

领域的实体分类体系提供领域内实体类别的语义信息。如将中国历史里面的人物分为帝王、皇后、将军、诗人等等，将历史事件分为革命事件、起义、罢工、运动、战争等等。目前，我们的分类体系将中国历史的实体分成了 69 个类别。

同时，领域语义信息还包括描述每一个类别实体的文本的语义元信息。举例说来，为了描述一个人物，通常需要描述他的各方面的语义信息， 如*又名、字、号，国别，出生年月日，籍贯或出生地，民族，*等等。在本系统中，针对不同类型实体，我们总共定义了 600

多个的语义元信息。

3.2 问答系统整体框架

在本小节中，我们首先介绍基于领域语义的问答系统的整体框架，并对整个过程中的语义分析过程进行说明。系统的整体框架图见图 1.

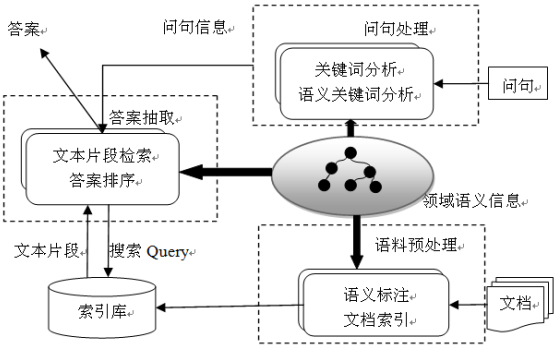

图 1. 问答系统框架图

基于检索的问答技术，我们的问答系统主要包括如下几个功能模块：1）**语料预处理模块**：对语料进行切分、语义标注并构建索引；2）**问句处理模块**：对用户的提问进行处理，生成查询关键词和语义关键词，并确定提问答案类型(PER, LOC, ORG, ...); 3）**答案抽取模块**：根据提问处理模块生成的信息构建查询 query，并从索引中检索出相关的文本片段(篇章、段落或句子)，最后对返回的文本片段进行打分并进行排序。

如图 1 中所示，本文系统的支撑是特定领域语义信息，它在以下三个方面起重要作用：1）在对语料进行预处理时，依靠领域语义信息来对文本片段（段落，句子）进行语义标注，由此进一步理解文本片段的语义；2）在处理用户提问时，还提取与领域语义信息相关的语义关键词，由此来理解用户提问的目标语义。3）在答案抽取模块，除了传统的打分标准之外，还使用了领域语义信息来帮助对候选答案的排序，使得候选答案在粒度、准确率和召回率上都能有一定的提升。下面，我们依次对在各个模块中集成语义信息进行描述。

3.3 语料预处理：语义标注和语义索引

为了让系统在抽取答案时，能够有效地利用语义信息，我们在语料预处理阶段使用语义标注来引入语义信息，并将把这部分信息存储在索引中。

语义标注的任务是，基于给定的领域语义信息，确定给定的文本片段描述了语义分类系统中的那些语义信息。举例来说，给定如下描述白居易的一段文本：*唐朝中叶的大诗人。字乐天,晚年自号香山居士。祖籍太原，后迁居下邽。*使用百科领域语义信息，文本包含的三个句子的语义标注如下：*唐朝中叶的大诗人。*→ 人物.定义;*字乐天,晚年自号香山居士。*→ 人物.字、号、又名; *祖籍太原，后迁居下邽。*→人物.籍贯和出生地。由此，我们能理解文本片段中所包含的语义信息。

本文采用分类的方法对两个粒度的文本片段（段落和句子）进行了语义标注，详细过程参照我们之前的相关工作[韩先培，赵军, 2009]。通过选取文本片段中的词语及实体的类别作为特征，语义标注在段落级别及句子级别文本片段的准确率可分别达到 90%及 77.9%。

语义标注通常是一个耗费大量计算资源的过程，因此需要离线进行并存储其结果。本文使用语义索引来存储语义标注的结果。语义索引的主要思想是，不按照文档来进行索引，

而是对不同粒度的文本片段单独进行索引，并在索引中加入指示其粒度和语义信息的域。一个文本最主要的信息被保存在三个域中：第一是文本粒度，主要分为三级（篇章、段落和句子）；第二是文本片段的内容；第三是文本片段的语义。

3.4 问句处理模块：语义关键词分析

用户的信息需求多种多样，仅仅使用关键词通常难于完整进行描述。考虑如下的问句：*长恨歌的作者是谁？*一个用户通常会使用*长恨歌*和*作者*作为关键词，但是，在百科的文本中，描述相关信息的句子如下：1）*白居易与陈鸿、王质夫同游仙游寺，作《长恨歌》。*2）*白居易的代表作有《新丰折臂翁》、《卖炭翁》、《秦中吟・重赋》、《琵琶行》及《长恨歌》等诗。*在相关的句子中并没有出现*作者*这个关键词，因此同时使用长恨歌和作者作为关键词将不能检索到相关的文档。但是仅仅使用*长恨歌*作为关键词又会引入大量的噪音信息。

因此，本文引入了语义关键词的概念。语义关键词指的是描述用户的语义信息需求但本身通常不会在一段文本直接出现的词语。在上面的例子中，其语义关键词就是*作者*。下面可以看到几个其它的语义关键词的例子：1）*白居易的成就*→ 成就；2）*李世民与李治的关系*→ 关系。识别问句中的语义关键词需要大量语义知识。基于已给定的领域语义信息，本文识别问句中与领域语义元信息相关的词语作为语义关键词，识别方法如下：

1） 对领域语义信息中的每一个语义元信息，人工构建一组语义关键词作为其表示。如*著作.作者*这个语义元信息的语义关键词为:作者，编著等等；

2） 对问句中每一个非命名实体名词，计算其与知识体系中每一个语义元信息的关键词的语义相似度，如果其与至少一个语义元信息关键词相似度超过一个阈值，则被认为是一个语义关键词。如在*长恨歌的作者是谁？*这个问句中，其非命名实体名词*作者*这个词由于与*著作.作者*和*人物.著作*这两个语义元的相似度超过给定的阈值，被认为是语义关键词。

由此，给定一个问句，基于通常的问句分析技术和我们提出的语义关键词分析技术，我们可以得到如下的问句信息：普通关键词，语义关键词。其中，普通关键词选取其中包含的命名实体和非语义关键词之外的名词。另外，我们还基于[Youzheng Wu，2005]的技术，分析了问句的语义分类和方法分类。

3.5 答案抽取模块：基于语义信息的答案排序

基于信息检索的问答系统利用问句的关键词来查询相关文档，然后对候选答案进行排序。其中，候选答案排序是其核心，排序的依据通常是提问处理模块生成的查询关键词。不同类别的关键词对答案排序有不同作用。在本系统中，关键词被分为如下几类：1）命名实体；2）语义关键词，通过 3.4 小节中的技术得到；3）一般名词，即除命名实体词和语义关键词之外的其他名词；4）其他词语，以上几类词之外的词语。

给定一个问句，答案抽取的第一步是构建查询的 query，并查找相关的文本片段。本文使用如下的方式来构建查询：Query = NE_1 || NE_2 || …; 即搜索至少出现一个问句中命名实体的文本片段（篇章，段落，句子）。

使用构建的 Query 搜索语料可得到 n 个文本片段 Ts = {ts_1,ts_2, …, ts_n}，其中，每一个文本片段 ts 包括如下两方面信息：1） ts.Semantic:包括文本片段的语义信息；2）ts.Content: 文本的内容，表现为一个词向量。3）我们使用如下的几个方面对每一个文本片段 ts 对问句 q 的重要性进行打分：

命名实体包含度打分 NEScore(ts, q): 当一段文本片段覆盖了越多问句中的命名实体，其可能是问句答案的可能性就越高，其计算方法为

$$NEScore(ts,q)=\|ts.NE\cap q.NE\|/\|q.NE\|。$$

内容相似度打分 ContentScore(ts, q): 计算文本片段与问句用词的一致性，其值为 ts 和 q 的词向量的 *Cosine* 相似度，其中，每一个词通过 TFIDF 计算赋予权重。

语义相似度打分 SemanticScore(ts, q)：计算问句的关键词与文本片段的语义信息的相似度，通过计算问句语义关键词与文本片段的语义信息的关键词的最大相似度得到。

目标实体包含度 TargetNEScore(ts, q)：基于[Youzheng Wu，2005]的技术，我们可以得到一个问句的目标实体类型(LOC,TIME, DATE, 等等)，如*李自成的出生日期*这个问句的目标实体为 DATE，如果文本片段中包括了 DATE 类型的实体，则其是问句的答案可能性就提高了。目标实体包含度的打分函数如下：

$$T\arg etNEScore(ts,q)=\begin{cases}1, \text{ts中包含目标实体}\\ 0, \text{不包含目标实体}\end{cases}。$$

最终，一个文本片段的打分由如下公式确定：

$$\text{Score(ts, q)}=\lambda_1\times NEScore(ts,q)+\lambda_2\times ContentScore(ts,q)$$
$$+\lambda_3\times SemanticScore(ts,q)+\lambda_4\times T\arg etNEScore(ts,q)\quad ,$$

其中，λ_1，λ_2，λ_3 和 λ_4 是实数权重，通过在一定的样本集上学习得到。最终，依照上面的方法对每一个文本片段进行打分，排序最高的前 N 个文本片段被作为正确答案。

4 实验及讨论

尽管目前有许多关于问答的评测，但是本文的问答系统主要针对百科全书领域，仍然缺少公用的评测语料。因此，我们手工按如下两个方式构建测试问句集：一是从网络中的百科问答题库（如 baidu 知道, sina 爱问知识人）中挑选问句；二是通过试用来收集问句。我们在一个 50 个问句的测试集上进行了测试。下面是一些问句的例子：

北伐战争是怎样失败的?

李世民与李治的关系?

清末预备立宪失败的原因?

对每一个问句，我们通过人工来评估结果的准确性，并使用平均排序倒数(Mean Reciprocal Rank，简称 MRR)来衡量系统的性能，其中 N 是问句总个数：

$$MRR=\frac{\sum_{i=1}^{N}1/\text{标准答案在系统给出的排序结果中的位置}}{N}。$$

表 1. 问答系统结果

第一个正确答案的位置	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**>10**	**MRR**
检索系统	11	4	6	3	6	2	0	1	0	1	16	0.35
未加入语义信息问答系统	19	6	3	2	3	0	0	2	2	0	13	0.49
领域语义信息问答系统	**30**	**6**	1	2	2	0	0	1	1	0	7	**0.69**

为了对比系统的性能，我们构建了两个 baseline 系统。其中未加入语义信息系统是去除现有问答系统中语义信息在排序中的影响的问答系统；而检索系统则使用问句中的命名

实体和名词作为关键词检索段落和句子，使用 Lucene(http://lucene.apache.org/)自带的排序函数，对每一个问句，使用段落检索和句子检索结果中最好的结果作为最后的结果。对三个系统，我们人工评估第一个正确结果所在的位置。表 1 中是我们系统的测试结果。

从表 1 的结果中可以看出：

1. 相比于传统的基于关键词的信息检索系统，问答系统取得了更好的性能。如表 1 中所示，相比于检索系统，领域语义信息问答系统和未加入语义信息问答系统都取得了相当的性能提升，在 MRR 上分别提升了 34%和 14%。

2. 在问答系统中加入语义信息能够有效的提升问答系统的性能。相比于未加入语义信息的问答系统，添加了领域语义信息的问答系统在 MRR 上取得了 20%的性能提升。

5 结论

在本文中，我们实现了一个基于领域语义信息的问答系统。并描述了如何在问答系统的三个模块中引入领域语义信息。实验结果表明，领域语义信息能有效的提升问答系统的性能。相比于未加入语义信息的问答系统和检索系统，基于领域语义信息的问答系统在 MRR 上分别提升了 20%和 34%的性能。

目前我们的问答系统使用专家构建的领域语义信息，在下一步工作中，我们将尝试使用大规模的但高噪音的互联网上的语义信息来提升问答系统的性能。

参 考 文 献

[1] Youzheng Wu, Jun Zhao, Bo Xu. Chinese Question Classification from Approach and Semantic Views. In Proc. AIRS 2005.

[2] Hui Yang, Tat-Seng Chua. The Integration of Lexical Knowledge and External Resources for Question Answering. In Proc. of TREC11, 2002.

[3] A. Ittycheriah, S. Roukos. IBM's Statistical Question Answering System-TREC11. In the TREC11, 2002.

[4] Dekang Lin and Patrick Pantel. Discovery of inference rules for question-answering. In Natural Language Engineering, volume 7, pages 343-360, 2001.

[5] D. Ravichandran, E. Hovy. Learning Surface Text Patterns for a Question Answering System. In Proc. ACL, 2002.

[6] Eric Brill, et al. Data-Intensive Question Answering. In Proc. TREC10, 2001.

[7] D. Moldovan, V. Rus. Logic Form Transformation of WordNet and its Applicability to Question Answering. In Proc. ACL, 2001.

[8] E. Hovy, et al. The Use of External Knowledge of Factoid QA. In Proc. of the TREC10, 2001.

[9] Marius Pasca. A Relational and Logic Representation for Open-Domain Textual Question Answering. In Proc. ACL, 2001.

[10] 韩先培，赵军. 基于 Wikipedia 的语义元数据生成. 中文信息学报, Vol.23, No.2. 2009.