基于约束的自然语言问题到 OWL 的语义映射方法研究

高明霞, 刘椿年
(北京工业大学计算机学院, 多媒体与智能软件北京市重点实验室, 北京 100022)

摘 要: OWL 已经成为重要的知识表示和交换方式. OWL 驱动的问答系统是比搜索引擎更具挑战性的研究. 本文致力于解决为自然语言问题获取正确的 OWL 解释. 考虑到问题理解的复杂性, 影响因素的多样性, 本文提出一种基于约束的语义映射方法. 该方法分解问题成为变量, 索引 OWL 知识成为变量域, 抽象约束形成函数. 并建立了目标函数完成问题变量在候选 OWL 知识库中获取合理解释的过程. 该方法基于网络智能研究院知识库和问题集进行了初步评估. 实验结果表明: 和传统基于关键字匹配的算法比较, 本文提出的方法使可解释问题比例增加了 18% 左右.

关键词: 问答系统; 约束; 网络本体语言; 问题分解
中图分类号: TP312 文献标识码: A 文章编号: 0372-2112 (2007) 08-1598-05

A Constraints-based Semantic Mapping Method from Natural Language Questions to OWL

GAO Ming-xia, LIU Chun-nian
(*Beijing Municipal Key Laboratory of Multimedia and Intelligent Software Technology, Beijing University of Technology, Beijing 100022, China*)

Abstract: The goal of an on-line ontology-based question answering system is to automatically derive answers from ontology knowledge bases without demanding additional information or intervention from users. This paper focuses on solving mapping task from question variable to OWL elements, which belongs to component of question understanding in question answering system. The paper presents a principled approach, which builds on tokens of question and syntax or semantic relations from NLP into set of variables and functions of constrain, and a sound optimization based assigning mechanism to find sound substitute in OWL knowledge for question variables. A preliminary evaluation on the International WIC Institute knowledge and questions is presented.

Key words: question answering system(QA); constraint; Web Ontology Language(OWL); question understanding

1 引言

自然语言问答系统是比信息提取更具有挑战性的研究, 它的目标是不需要用户学习额外的知识表示形式, 就能获得满意的答案. 随着网络智能技术[1]的发展, 各种本体知识将成为知识表示和共享的基础. 因此本体驱动的问答系统研究已经成为网络智能的关键问题之一. 而分解自然语言问题为其获取合理的本体知识解释就成为此类研究的首要任务. 本文集中讨论自然语言问题到 OWL 知识的语义映射方法. OWL[2]是目前用于表示和交换本体知识的基本标准, 借鉴了语言发展过程中的很多思想, 具有代表性.

影响问题语义理解的因素很多、形式各异, 本文基于约束满足的处理模式, 提出一种问题语义映射方法. 该方法分解问题成为变量, 抽象约束成为限定函数; 并将问题变量到 OWL 知识库中元素的合理解释归结为约束满足过程, 定义了目标函数用于具体优化过程. 为评估该方法, 本文以网络智能研究院 Portal 为知识建立和问题收集的平台, 进行了初步实验.

2 相关工作

问答系统研究有很长时间, 主要分为数据库问答系统[3~7], 网络问答系统[8~11], 和本体知识问答系统[12~14]. 根据候选数据形式, 问题的分解方法和结果存在很大差异. 基于数据库的问题理解目标是产生格式良好的 SQL 语句. 基于网络的问题分解目标是产生满足搜索引擎的查询词组. 和其他数据比较, 本体知识最大的不同是可以推理, 所以本体问答系统问题分解的目标是产生用于推理的本体查询语言.

基于数据库的问题分解有两种方法: 早期系统[3]手工建立领域概念图映射问题中出现的概念和关系, 并为此采用基于逻辑的中间语言. 随着自然语言处理和机器学习技术的发展, 基于数据库的问题分解开创了新的模式并获得一定的成果[4~7]. PRECISE[4,5]利用语言解析插件, 数据索引字典和最大流算法为问题词和数据库元素建立联系, 并最终组合成 SQL 查询语句. 文献[6, 7]利用不同的学习算法生成一个谓词字典并借助这个字典和归纳学习器为问题获取对应的逻辑形式.

基于网络的问题分解主要集中在问题到查询词组的重写[8,9]和期望答案类型分类[10,11]. 期望答案类型分类根据问题本身预测可能的答案类型; 问题到查询词组的重写方式, 从手工启发式规则[8]到行为序列的学

收稿日期: 2006-09-25; 修回日期: 2007-01-10
基金项目: 国家自然科学基金重大项目(No. 60496322, 60496327); 多媒体与智能软件北京市重点实验室开放基金

习[9]出现了很多研究.

和本文最相关的是文献[12~14]. 严格的说, 文献[13]不算一个本体问答系统, 它处理的问题是一个可控的英语子集, 有严格的语法. 可借助自身的解析器转换问题为一阶谓词的子集 DRS (Discourse Representation Structure). MOSES 主要处理丹麦语和意大利语. 它的问题分解方式涉及自然语言处理, 并需要领域本体参与. Aqualog 处理英语问题, 采用自定义三元组式的中间表示, 并对问题进行手工分类, 但是在多候选概念和关系识别过程中需要用户参与, 属于半自动过程.

和上边的研究比较, 本文的问题分解过程有以下的特点:

(1) 问题变量的分解遵循了语法、语义规则, 并应用了实体识别, 语义推理等多种语言处理技术.

(2) 将语言处理、机器学习和用户行为学习获取的辅助知识抽象为问题理解过程的定量函数, 用于约束问题变量匹配过程.

(3) 基于上述约束函数组合目标函数, 将获取问题变量解释的过程抽象成了最大值计算问题.

3　语义映射方法

问题到 OWL 知识库的解释, 本质上是组成问题的独立成分在特定语境下明确语义的过程. 为获取合理解释, 必须解决两个问题. 一是问题分解粒度, 二是特定语境. 同一个问题分解粒度不同, 变量的选择不同, 其在 OWL 知识库中的解释也不同. 从 OWL 知识库角度看, 元素是表示知识的最小粒度, 所以本文选择的问题分解粒度是能独立匹配 OWL 元素的问题成分. 特定语境是一个综合概念, 包括很多因素如: 问题领域, 领域术语, 用户习惯, 语言习惯等. 这些因素获取方式不同, 表示形式多样, 只有将其抽象为统一的约束条件, 才能限定问题变量在 OWL 知识库中的合理解释. 在此采用约束满足框架解决问题.

约束满足是人工智能中的基本问题, 它在模式识别、规划、时序推理等众多领域获得广泛的应用. 考虑到问题分解的复杂性, 限制条件的多样化, 本文将自然语言问题到 OWL 的解释归结为约束满足过程. 涉及的步骤为: 分解问题成为变量, 索引 OWL 成为变量域, 抽象多种影响因素形成约束函数.

3.1　基于约束满足的问题表示

定义 **1**　问题基础(*QuesBase*)是问题在特定语境下的形式化表示, $QuesBase:=\langle QV, QC\rangle$. 其中: $QV:=\{qv_i\}_{i=1}^{n}$是问题变量集, 它的元素是问题变量 $qv_i:=\langle ID, Term, Attribute\rangle$其中 *ID* 是标识; $Term:=\{Token_j\}_{j=1}^{r}$是该问题变量包含的经过预处理(如取词干)的 *Token* 集; *Attribute*: = {*Lemma*, *Syn*, *Pos*, *Phtype*, *Netype*} 是该变量的属性集, 用于决定涉及的约束, 每个属性的实际取值由问题变量决定. $QC:=\{qc_k\}_{k=1}^{m}$是问题约束集, 它的元素是问题约束 $qc_k:=\langle S, R\rangle$其中 $S\subseteq QV$ 是该约束涉及的变量集称为约束范围; $R: S\rightarrow D$ 是从 *S* 到变量值域 *D* 的函数集称为约束关系.

定义 **2**　本体元素(*OntoElement*)是组成该本体的类(*class*), 个体(*individual*), 属性(*DatatypeProperty*, *ObjectProperty*)和值(*DatatypeProperty*: *rang*), $OntoElement:=\langle Type, Name, Relation\rangle$. 其中: *Type* 是元素类别, 包括 ⟨*class*, *individual*, *DatatypeProperty*, *ObjectProperty*, *rang*⟩; $Name:=\{Token_i\}_{i=1}^{r}$是元素的具体记号, *Token* 是 OWL 知识库中组成元素的词; $Relation:=\{\{property\ \ subject\ \ object>j\}_{j=1}^{r}\mid Token_i\subseteq(property\cup subject\cup object)\}$指该元素和本体中其他元素之间的关系集合, 用 RDF 三元组形式表示.

定义 **3**　变量值域 $D:=\{\{D_i\}_{i=1}^{n}\mid D_i=\{OntoElement\}_{j=1}^{s}\}$是问题变量集对应的值域, 通过 *Token* 匹配等方法获取的候选 OWL 知识库中的 *OntoElement* 集.

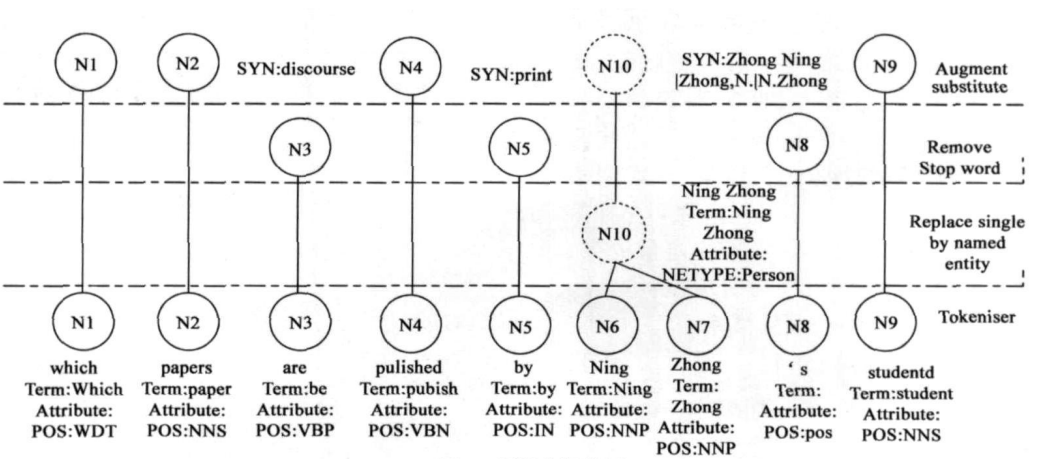

图 1　问题分解实例

3.2 形式化问题

形式化问题成为问题变量需要多种技术, 比如: 分词、词干提取、实体识别等. 基本原则是: 分解过程遵循语法语义规范, 语言常识和领域知识. 具体算法可以分层处理如图1实例. 首先利用分词和语法分析技术获取候选变量集及对应的属性集. 其次用实体识别技术获取一些特定的人名、地名、时间等作为实体变量替换候选变量集中相应的单变量. 第三步去掉 stop 词, 此处的 stop 词包括语法记号, 辅助性动词等. 从 OWL 知识库中获取变量的值域主要基于Token 匹配技术, 这种做法忽略了自然语言中存在的大量异形同义情况. 因此第四步用同义词、缩写、习惯用法等语义分析扩充变量属性, 并最终用于扩充变量值域.

获取候选值域要重新索引 OWL 库, 对元素进行标准化分割. 本文使用现存的 OWL 解析器 Jena 获取元素和 RDF 元组, 并参考自然语言的分词技术分割元素建立字典. 算法相对简单不再赘述.

3.3 形式化约束

影响问题语义理解的因素很多, 通常能够捕捉的有两类: 一类是问题自身约束; 一类是领域和用户施加的约束. 本文主要关注自身约束. 这类约束通过语法分析, 语义分析, 实体识别及机器学习等技术获取, 常见的有变量词性标注, 问题变量间的依赖关系, 期望答案类型等. 本节的形式化约束完成自身约束到定量函数的转换.

表 1 词性和 OWL 元素类型的匹配概率

(词性标注采用的是 Penn Treebank Ⅱ Tags)

词性	OWL 元素类型				
	Class	Individual	ObjectProperty	DatatypeProperty	DatatypeProperty: rang
NNP?	0	1	0	0	1
VB?	0	0	1	1	0
CD	0	0	0	0	1
FW	0	1	0	0	1

词性约束考虑语言常识. 尽管自然语言和 OWL 是不同的知识表示方式, 但二者之间存在本质联系. 比如: 自然语言中的专有名词是指人、地方、称谓、机构、语言、国民、周日、月份、节日等的专有名称; 而这些名称通常会以 OWL 知识库中的个体和值定义. 这说明不同词性的问题变量和不同类型的OWL 元素之间存在某种联系. 为了将这种定性的语法常识抽象成定量约束, 本文定义了一个匹配代价函数如式(1) 所示, 表示某一词性变量被指派不同OWL 元素类型的概率.

$$f_{POS}: \prod_{x_i \in S_k} D_i \rightarrow [0, 1] \tag{1}$$

为简化处理, 本文只考虑三类匹配概率: 100%、0、50%, 从语义角度看, 分别代表完全匹配, 完全不匹配, 不能确定. 表 1 是部分词性和 OWL 元素类型在上述设定下的概率值.

问题在语法层的分析可以产生一些变量之间的依赖关系, 比如: 同一词组的成员, 修饰词和被修饰词, 谓词逻辑等. 这些依赖关系说明代表这些词的问题变量之间具有语义联系. 变量取值是否支持这种语义联系可以通过检查其关系集合中是否包含依赖关系中的其他变量决定. 本文使用一个二值函数表示如式(2).

$$\forall x \in S_i, f_{\{phr,\ mod,\ pre\}}(x) = \begin{cases} 1, x \in D_i \text{ and } |\exists\ x.\text{Relation} \cap \{R_{phr}, R_{mod}, R_{pre}\}| \geqslant 2 \\ 0, \quad \text{other} \end{cases} \tag{2}$$

其中 R_{phr}, R_{mod}, R_{pre} 分别指词组关系, 修饰关系, 谓词关系, 这些关系通过语言分析获取的直接表示方式是元组, 例如: 〈Gao Mingxia email〉.

期望答案类型需要大量问题实例通过机器学习技术获取, 这种约束大多用于指导答案集成, 本文关注其对问题记号的影响. 问题记号(通常指{who, which, when, what …}) 是一类特殊的变量. 他是答案在问题中的替代, 所以候选值域无法通过常规的 Token 匹配建立. 期望答案类型描述答案的具体形式, 比如“Arear-Quantity、Name-paper” 分别说明答案是“面积数, 论文名称”. 所以期望答案类型可作为对问题记号具体形式的限定. 从 OWL 知识库角度看, 问题记号的解释必然是 OWL 元素. 所以建立期望答案类型和 OWL 元素类型之间的联系用于限定问题记号, 是实现该约束的关键. 这可以看作是词性约束面向问题记号变量的扩展, 可以使用公式(1)的匹配代价函数. 表 2 是部分期望答案类型和 OWL 元素类型的匹配关系.

表 2 期望答案类型和 OWL 元素类型的匹配

(期望答案类型使用类似 webclopedia* 的标注)

问题记号	期望答案类型	OWL 元素类型				
		Class	Individual	ObjectProperty	DatatypeProperty	Rang
How big	Arear Quantity	0	0	0	0	1
Which	Name-Paper	0	1	0	0	0

上述约束抽象成定量函数后, 为问题变量集选择合适解释的过程可以看作是一个优化过程, 式(3) 是目标函数. 其中的 w_k 是约束函数的权重, 表示不同约束对变量优化过程的不同影响.

$$f(x) = \arg\max_{x \in D} \{ \sum_{k=1}^{m} w_k f_k(x) \} \tag{3}$$

4 实验

本文选择了信息检索中常用的关键字匹配方法

* http://www.isi.edu/natural-language/projects/webclopedia/

(Match) 作为实验基准, 集中验证基于约束的方法(CBM) 是否有效. 没有考虑辅助知识的影响程度, 为现存约束选择了相同权重.

4.1 实验数据

网络上存在很多 OWL 领域知识库. 这些知识库主要包括领域词汇, 缺少对应实例. 本文基于 University 领域知识库, 扩充了大量网络智能研究院** 实例知识, 形成一个相对完整的知识库 institution.owl 用于实验, 表 3 是该知识库的规模.

表 3　institution.owl

元素类型	类	数据类型属性	对象属性	个体
数目	83	20	37	90

用于实验的问题集有两个获取途径: 一是网络智能研究院收集到的实际问题; 二是通过参考 webclopedia 收集到的问题, 模拟产生和网络智能研究院相关的仿真问题. 问题集通过问题记号进行了分类如表 4. 其中答案较主观的原因问题(why) 和过程问题(how) 已经直接移去; Other 特指祈使格式问题如: "Name the person that has same advisor as Su Yila.".

表 4　问题集

	Who	Which	What	How	When	Where	Yes/no	Other	Overall
Simulative questions	4	5	9		3	4	7	3	35
Real questions	12	10	12	6	5	3	5	4	59

问题自身约束直接或者间接来源于语言处理技术, 所以语法、语义分析的正确性对算法影响很大. 为减少语言处理噪声, 本文对问题集做了手工预处理. 预处理包括: 将语法、语义分析错误的问题, 在保持等价语义的情况下, 转换成能正确解析的问题. 这项工作属于语言语义推理范畴[15].

4.2 结果分析

表 5　实际问题正确率(%)

	Who	Which	What	How many	When	Where	Yes/no	Other	Overall
Match	41.7	20.0	25.0	66.7	40.0	33.3	20.0	25.0	32.2
CBM	50.0	40.0	58.3	83.7	80.0	66.7	60.0	50.0	56.0

表 6　仿真问题正确率(%)

	Who	Which	What	When	Where	Yes/no	Other	Overall
Match	75.0	100.0	22.2	33.3	75.0	14.3	0	42.9
CBM	75.0	100.0	44.4	33.3	75.0	57.1	0	57.1

从表 5 和表 6 可见, 和 Match 相比, CBM 能正确解释的问题比率平均提高了 18% 左右, 效果很明显. 下边分析 3 个实例, 讨论一下 CBM 方法的优缺点.

例 **1:** 问题: Where does the advisor of Gao Mingxia dwell in?

CBM 解释: where/where; advisor/hasAdvisor, Gao Mingxia/GaoMingxia, dwell/dwellIn.

Match 和 CBM 方法都获得了正确解释. 但是该问题涉及的变量候选值都是一个, CBM 方法没有启用, 这种情况常见于小规模知识库. 网络智能强调的是网络, 处理的是多知识库或大知识库. 在这种情况下, 同一问题变量的候选值域增大, CBM 方法更有优越性.

例 **2:** 问题: Which students, whose advisor is Liu Jiming, are members of AskMe?

CBM 解释: which/which, student/Student, advisor/hasAdvisor, Liu Jiming/LiuJiming, member/hasProjectMember, AskMe/AskMe.

Match 解释错误, CBM 解释正确. 尽管 CBM 从变量 member 的多个候选中识别了正确解释, 却没有从 student 的多个候选中识别出正确解释, 只是按照预设程序选择了顶端匹配. 这种情况常出现在多个多候选值变量时, 原因是现有约束能力有限, 不能兼顾所有问题变量. 所以需要考虑更多约束, 以期提高解释精度.

例 **3:** 问题: Which are the publications in WICI that are related to Web Intelligence?

CBM 解释: which/which, publication/Publication, WICI/WICI, relate/null, Web Intelligence/WebIntelligence.

Match 和 CBM 方法都没有获取正确解释, 原因是 relate 变量在知识库中没有候选. 这种情况在解释失败的问题中占很大比例, 特别是仿真型问题. 尽管问题分解时考虑了相同语义不同表达的情况, 但是自然语言中某些词的涵义结合特定语境会演变成特定解释, 而 OWL 知识库不擅长处理这种模糊概念和关系. 解决这种情况需要增强预处理步骤中自然语言语义推理[15]能力, 如: 为上述问题推理出等价句"which are the publications in WICI whose area is Web Intelligence?"就能获取正确解释.

5　结论

本文考虑了影响问题理解的众多因素, 提出一种基于约束的语义映射方法. 该方法基于约束满足框架, 一方面将问题分解形成变量, 并将 OWL 知识重新索引, 建立变量的值域字典. 另一方面将自然语言处理技术获取的语法、语义知识形式化为定量函数. 最后组合约束函数, 将获取问题变量在 OWL 中对应元素的过程转化为最大值的计算. 本文以网络智能研究院为例做了初步评估. 实验结果表明: 和传统基于关键字匹配的算法比较, 该方法使可解释问题比例大大增加.

** http://www.iwici.org/

参考文献:

[1] Zhong N, et al. Web Intelligence[M]. Heidelberg: Springer-Verlag, 2003.

[2] Deborah L M, et al. Web ontology language overview[EB/OL]. http://www.w3.org/TR/owl-features/. 2006-09-20.

[3] Androutsopoulos I, et al. Natural Language interfaces to databases—an introduction[J]. Journal of Natural Language Engineering, 1994, 1(1): 29-81.

[4] Popescu A, et al. Towards a theory of natural language interfaces to databases[A]. IUI'03[C]. New York: ACM Press, 2003. 149-157.

[5] Popescu A, et al. PRECISE on ATIS: semantic tractability and experimental results[A]. Proceedings of the Nineteenth National Conference on Artificial Intelligence[C]. San Jose: MIT Press, 2004. 1026-1027.

[6] Cynthia A, et al. Automatic construction of semantic lexicons for learning natural language interfaces[A]. Proceedings of the Sixteenth National Conference on Artificial Intelligence[C]. Orlando: AAAI Press, 1999. 487-493.

[7] Lappoon R, et al. Using multiple clause constructors in inductive logic programming for semantic parsing[A]. Proceedings of the 12th European Conference on Machine Learning[C]. Freiburg: Springer, 2001. 466-477.

[8] Brill E, et al. An analysis of the AskMSR question-answering system[A]. Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing[C]. USA: Association computational linguistics, 2002. 257-264.

[9] Dragomir R, et al. Mining the Web for answers to natural language questions[A]. Proceedings of International Conference on Information and Knowledge Management[C]. New York: Association for Computing Machinery, 2001. 143-150.

[10] Li X, Roth D. Learning question classifiers: The role of semantic information[J]. Natural Language Engineering, 2006, 12(3): 229-249.

[11] Pomerantz J. A linguistic analysis of question taxonomies[J]. Journal of the American Society for Information Science and Technology, 2005, 56(7): 715-728.

[12] Bernstein A, et al. Talking to the Semantic Web—a controlled English query interface for ontologies[J]. Ais Sigsemis Bulletin, 2005, 2(1): 42-47.

[13] Paolo A, et al. Ontology-based question answering in a federation of university sites: the MOSES case study[A]. Proceedings of the 9th International Conference on Applications of. Natural Language to Information Systems[C]. Berlin: Springer, 2004. 413-420.

[14] Vanessa L, Michele P, Enrico M. AquaLog: An Ontology-Portable question answering system for the Semantic Web[A]. ESWC 2005[C]. Heidelberg: Springer-Verlag, 2005. 546-562.

[15] Braz R, et al. An inference model for semantic entailment in natural language[A]. Machine Learning Challenges Lecture Notes in Artificial Intelligence 3944[C]. Heidelberg: Springer-Verlag, 2006. 261-286.

作者简介:

高明霞 女, 1973 年生于河北, 北京工业大学多媒体与智能软件北京市重点实验室博士研究生. 研究方向为 Ontology 管理, 知识问答系统. E-mail: gaomx@emails.bjut.edu.cn

刘椿年 男, 1944 年生于江苏, 北京工业大学教授, 多媒体与智能软件北京市重点实验室主任, 博导. 研究方向为人工智能、知识工程、数据挖掘等.