基于外部语义知识补全的自然语言查询

冯 雪

(北京信息科技大学计算机学院 北京 100192)

摘 要 语义网是依托互联网技术而产生的一类非常重要的资源。目前,语义网中的用户查询仅支持形式化的查询方式,因此需要严格地遵循某种特定的语法规范,从而导致只有熟悉语义网系统和形式语言的专业人士才能正确进行查询操作。为了弥补这一缺陷,提出了一个无指导的自然语言查询系统,它能自动地将自然语言的句子转换成语义网查询支持的形式语言语句,从而方便非专业用户(即普通用户)使用。该系统首先根据语义网自动抽取给定句子中的所有实体和属性,然后将这些实体和属性关联起来形成一个语义关联图,最后通过启发式的方式从图中搜索出一条最优路径,并将这条路径转换成SPARQL语句。该系统最关键的部分在于语义网中的实体和属性覆盖度,它能直接决定语义关联图的好坏,从而影响系统的最终性能。为了提升系统的实用性,进一步利用外部语义网的知识来补全和丰富自然语言句子中所蕴含的信息,优化中间生成的语义关联度,得到更准确的SPARQL语句。最后采用美国地理问题集进行实验以验证该系统以及提出的改进方法,该数据集共包含了880个问句的人工SPARQL语句,是自然语言查询相关工作中一个被广泛认可的数据集。最终实验结果表明:提出的基准系统能够正确回答77.6%的问题,显著优于当前最好的无指导系统;当采用外部语义知识补全后,回答正确率达到78.5%。

关键词 自然语言查询,语义网,无指导学习,形式语言,SPARQL

中图法分类号 TP391 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.08.045

Natural Language Querying with External Semantic Enrichment

FENG Xue

(Computer School, Beijing Information Science and Technology University, Beijing 100192, China)

Abstract Semantic Web is one kind of extremely important resources based on Internet technique. Querying on a semantic Web only supports formal languages, which need manipulator to strictly observe certain syntax constraints, and thus only experts that are familiar with semantic Web system and formal language are capable of querying. To overcome this problem, this paper presented an unsupervised natural language querying system, which can convert natural languages into formal languages automatically, thus making common users query on a semantic web using natural languages conveniently. The system first extracts all entities and attributes in a sentence based on a specific semantic Web, then connects them to form a semantic relationship graph, and finally exploits a heuristic strategy to search for an optimum path which is used to produce the output SPARQL expression. The key of the system is the coverage of the entities and attributes from the semantic Web, which directly decides the quality of the inter-mediate semantic relationship graph, and influences the final performance of system. In order to achieve a practical system, this paper enriched a human-annotated semantic Web for a specific domain through using external semantic knowledge, so that the natural language formed languages can contain more information. By this method, better semantic relationship graphs can be obtained and more accurate SPARQL expressions for sentences are achieved. Finally, this paper used the dataset based on American geography for experimental evaluation to verify this system. The dataset is widely acceptable for related research work of natural language querying, which includes manually-annotated SPARQL expressions with 880 questions. The experimental results show that this system can correctly answer 77.6% of the natural queries, outperforming the best unsupervised system in the literature significantly. After knowledge enriching by the external semantic Web, the system reaches 78.5% in term of the correctly-answering accuracy.

Keywords Natural language querying, Semantic Web, Unsupervised learning, Formal language, SPARQL

到稿日期:2018-09-26 返修日期:2018-11-22 本文受国家重点研发计划项目(2018YFB1004100)资助。

冯 雪(1984—),女,博士,讲师,主要研究方向为数字版权保护技术、自然语言处理,E-mail:fengxue@bistu.edu.cn。

1　引言

近几年,随着大规模结构化和半结构化知识(如百度百科、维基百科等)的大量涌现,以及传统的人工构建本体(如 WordNet、HowNet、同义词林、Yago[1] 和 Freebase[2] 等)的逐步完善,语义网[3]的相关理论和技术已经引起了众多研究者的兴趣。随着这些知识库的日趋壮大,如何在语义网中进行快速查询已经成为了一个非常重要的研究问题。

目前,语义网的数据一般采用 RDF 或者 OWL 这两种形式语言进行描述。然而,查询 RDF 或者 OWL 数据所对应的形式化查询语言主要包括 SPARQL 和 RDQL 两种,这两种语言只有专业的程序员才能正确编写,因此只有熟悉这些语言的专业人士才能进行查询操作,普通用户很难使用语义网中的数据。因此,帮助普通人员方便地查询语义网中的相关信息变得非常重要。

最近,自然语言理解技术的发展使得普通用户的查询变得可能。如果能将普通用户输入的自然语言的句子通过特定的语义分析手段自动转换成专业的形式化查询语言,那么这一问题便能得到解决。在自然语言处理领域,已经有相当部分学者开始了这方面的研究[4-8]。主流的方法是将一些简单的固定格式的问句转换成逻辑语言,再将这些逻辑表达式进一步转换为专业的形式化查询语言。

这些主流的方法主要面临两个缺点。1)它们大部分是有指导的方法,需要创建一定规模的训练语料,在此基础上利用已有的机器翻译模型或者语义分析模型以及其他更复杂的模型来进行训练和解码。对这些训练语料进行标注是非常困难而且费时的,代价昂贵,因为标注者需要很强的专业背景知识。2)这些传统的方法都忽略了外部知识库的价值,而这些知识通常是非常有价值的,比如外部语义数据能够更全面地指出一个实体包含哪些属性,与当前句子中的实体进行结合便能对实体的属性进行扩展,同时也能在实体存在歧义时进行消岐。比如,对于"篮球明星乔丹出生在哪个城市?",可以通过职业来识别出"乔丹"是谁,从而回答这一个问题。如果没有外部语义知识库,便很难全面地知道"篮球"有哪些属性,比如它可以是职业;也很难全面知道"乔丹"具有哪些属性。

针对上面的两个问题,本文首先设计了一个无指导的自然语言查询系统,这个系统建立在一个人工设计的封闭领域的语义网数据基础之上,目前主要用来对英语问句进行分析。进一步,本文利用外部丰富的结构化知识扩展了这一语义网,使其融合了 Wikipedia,WordNet 的信息。基于该语义网,自然语言查询语句的分析结果得到了进一步的丰富和明确。

为了验证本文方法的有效性,在美国地理数据集上进行实验,该数据集中一共包含 880 个问句,它们都是针对美国地理数据进行的一些查询操作。实验结果表明,本文提出的无指导的方法在使用封闭的语义网时,能答对 77.6% 的问题;当使用了外部知识整合的语义网之后,准确率提升到了 78.5%。同时,将这一无指导的方法与其他无指导的方法进行了对比,发现本文方法显著优于已有的最好方法。

本文第 2 节介绍相关工作;第 3 节介绍无指导的分析模型;第 4 节介绍如何在语义网中融入外部知识;第 5 节介绍相关实验;最后总结全文并展望进一步的工作。

2　相关工作

语义网上的自然语言查询主要分为两种:有指导的分析方法和无指导的分析方法。与本文最相关的工作是无指导的分析方法,例如 Wang 等提出的 PANTO 算法[4,9],这类算法的一个显著优点是可以在大规模语义网上进行查询,而且不限定领域。本文方法属于这一类。

自然语言查询方面,有指导的方法已经被广泛研究[10-17],这一任务通常又被称为语义分析,即将自然语言转换成实际可以运行的语义操作。这一类方法一般能取得比较好的效果,但是它们只能局限在非常小规模的语义网上,需要标注一定规模的训练数据,标注难度和代价较大,而且领域非常受限。本文的主要目的是针对大规模的语义数据进行自然语言查询,因此并没有采用这类方法。

3　无指导的自然语言查询模型

本文提出的无指导的自然语言模型一共分为 3 步:1)实体和属性抽取,即采用模式匹配算法从句子中抽取实体以及实体的属性;2)根据所有的实体以及实体属性建立一个图,然后从图中搜索一条最优路径,并使得路径上的所有实体能够合理地串联起来;3)根据前一步的路径生成最终的形式化查询语言,本文采用的形式化查询语言为 SPARQL 语言。

3.1　实体与属性抽取

给定一个查询的自然语言句子 $w_1, w_2, \cdots, w_n$,本文采用简单的匹配算法来查询句子中的所有实体及其对应的所有可能属性。假定语义网中所包含的实体的集合为 D,在分析句子时,对于每一个位置 i,采用从左到右的顺序,遍历以该位置词开始的若干个可能实体,包括 $w_i, w_i w_{i+1}, \cdots, w_i \cdots w_{i+L-1}$($L$ 为最大实体的长度),并在 D 中进行匹配,如果匹配成功,则记录下来,并把语义网中该词的所有属性也记录下来。图 1 给出了一个具体的例子,其中给定的查询语句为"what is the population of new york city ?"。

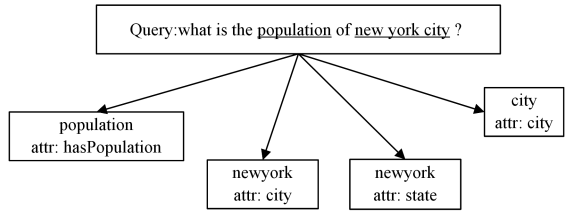

图 1　实体与属性抽取的实例

Fig. 1　Example of entity and attribute extraction

3.2　实体最优路径搜索

3.1 节的分析结果给出了实体及实体的属性,接下来便需要将这些〈实体-属性〉对串联起来,串联的结果实际上就是找出句子中所有合理的实体并对这些实体进行属性消岐。

首先对前面所产生的所有〈实体-属性〉建立一个〈实体-属性〉的关联图,这个关联图是一个有向无循环图;然后在这个图上采用柱搜索的算法搜索一条最优路径。在建图的过程中,需要遵守互斥原则:同一个实体的两个属性不能相连,同一个词所产生的不同实体也不能相连;其他的任何两个实体

都可以直接相连。对于前面的例子，通过建图，可以得到如图 2 所示的〈实体-属性〉关联图，其中〈实体-属性〉之间的路径连接参见虚线线条，start 和 end 是两个人工构造的节点，以方便后续的路径搜索。

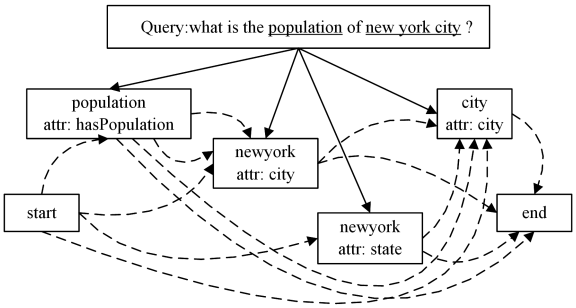

图 2 〈实体-属性〉关联图

Fig. 2 〈entity-attribute〉 relationship diagram

当〈实体-属性〉关联图建立完成之后，搜索一条从 start 节点到 end 节点的最优路径。其中每条路径的得分直接根据路径中每个节点的得分之和产生，而每个节点的得分又与该节点新产生的实体数目 N_e 和新产生的属性数目 N_a 有关，其计算分数也非常简单，如下所示：

$$score(\cdot)=\frac{N_e+1}{N_a+1}$$

直觉上，如果一条路径上的实体数目越多，且属性数目越少，则该路径越可能被选择。换句话说，如果一个新增加的实体节点能尽量不增加属性数目，则该节点更有可能被添加到最终的路径中。根据上面的路径分数计算方法可知，在图 2 所示的例子中，〈start〉→〈population，hasPopulation〉→〈newyork，city〉→〈city，city〉→〈end〉的总得分为 1+1+2=4；而〈start〉→〈population，hasPopulation〉→〈newyork，state〉→〈city，city〉→〈end〉的总得分为 1+1+1=3。图 2 最后得到的最优路径实际上为〈start〉→〈population，hasPopulation〉→〈newyork，city〉→〈city，city〉→〈end〉。

在路径搜索时，本文采用了柱搜索算法，按照从左到右的分析顺序，每次只保留固定数目的最优结果，算法 1 中给出了具体描述。虽然这一算法并不能保证找到全局最优的路径，但是其搜索效率会大大增加。本文中所采用的柱的大小为 64。

算法 1 柱搜索算法

```
输入：〈实体-属性〉关联图
输出：beam
beam←{{start 节点}}
while (beam 中存在路径不包含 end 节点) do
  newbeam←{}
  for path∈beam do
    抽取 path 中的最后一个顶点 v
    for v 从 v 到达的所有下一个顶点 p
      将 path ∪(v→p)加入到 newbeam 中
    end for
  end for
  beam←newbeam
end while
```

3.3 SPARQL 语句生成

通过前面的两步分析，最终得到了一条最优的路径，这条路径包含了查询语句中所有的实体和属性。接下来便将这些实体按照属性进行组合，生成最终的 SPARQL 语句。实际上，前面所分析的实体大致可以分为两类：一类用于生成变量，而另一类则代表了一个函数。在前面的例子中，population 的属性表明其是一个函数，而 new york 则是变量。区分这两种实体后，便可以通过一定的规则结合 SPARQL 语言的特性来生成具体的语句了。具体的规则与最优路径中的节点类别非常相关，每一个节点和其孩子节点组成一个三元组（少数节点除外），例如，对于如图 2，其最终的结果“〈start〉〈population，hasPopulation〉〈newyork，city〉〈city，city〉〈end〉”根据规则产生的 SPARQL 语句为：

```
select ?y where {
  city，new york，?x
  ?x，population，?y
}
```

除了上述实体区分之外，查询对应还会落实到具体的查询关键疑问词上。由于本文采用的最终数据是面向美国地理的，因此本文通过观察具体数据并进行总结，专门对 5 种关键词进行了特殊处理，包括 select，where，max，min 和 count。其处理方式与前面实体的处理方式类似，这种处理方式与 SPARQL 语言的语法规范密切相关，这里不做详细的介绍。

4 外部语义网的自动构建

在第 3 节中，语义网的构建是通过人工标注完成的，而最终的保存形式使用了 RDF 语言。这种语义网很难达到比较大的规模，而且标注的领域也比较受限。本节介绍如何将外部其他形式的数据添加到一个已有的语义网中，比如外部众包形式的 Wikipedia，以及人工已经构建好的 WordNet 等。

本文自动扩展语义网方法的核心是三元组，即首先将外部结构化的数据转换成三元组，然后再将这些三元组逐个添加到已有的语义网中。这些三元组实际上就反映了两个实体之间的关系。

三元组的具体获取方法如下。

1）当外部输入单元是句子（如 Wikipedia 中的描述语句）时，首先使用 CMU 开放的 Turbo Parser[18] 和 SEMAFOR[19] 进行句法和语义分析，将其中的 Framenet-Style 形式的语义分析结果和句法分析结果相结合，通过特定的规则抽取三元组。图 3 给出了一个具体的例子，最终抽取出来的三元组为〈New York，Boundary，New Jersey〉和〈New York，Boundary，Pennsylvania〉。图 3 中的例子抽取的是两个实体之间的关系，实际上实体的属性也可以用类似的方法抽取出来，比如〈bordered，Boundary，null〉。图 4 中例子抽取得到的是〈city，Locale，null〉，也用类似的方法抽取出实体的属性。

2）对于非句子形式的外部资源，例如 WordNet、同义词典以及疑问词词典等，三元组的构造形式则更为简单，可以直接得到实体、实体属性（实体上下位关系）。

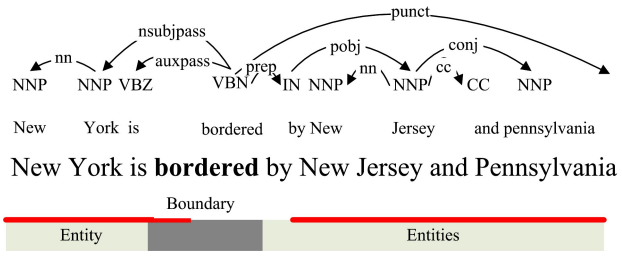

图 3　句法和语义分析结果

Fig. 3　Analysis results of syntactic and semantic

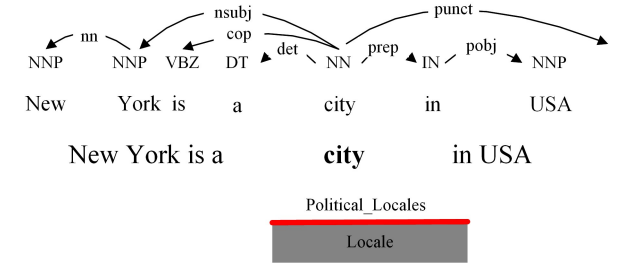

图 4　实体属性词的抽取

Fig. 4　Extraction of entity attribute word

将三元组全部被抽取出来之后进行简单的频率过滤,保留一些可信度比较高的三元组,并将这些三元组逐个加入到现有的语义网中。具体的构建方式也非常简单,即首先查找实体,如果现有语义网中没有这些实体,则将其作为一个顶点加入其中,然后再查找关系以及边,将这些顶点相连,并指定边上的关系,这些边上关系对应于一个实体和属性。

5　实验

本实验采用的语义网数据是通过对美国地理数据库进行转换而产生的。该数据库的原始数据是使用 Prolog 语言进行描述的,这里首先将 Prolog 描述的数据转换成关系数据库,然后通过开源工具将得到的关系数据库转化成 RDF 格式的本体数据。

本实验使用的自然语言的问题集来自于德克萨斯州立大学自然语言处理团队开发的针对美国地理数据库的问题集,它一共包含 880 个自然语言问句。这一问题集已被广泛用于语义分析、问答以及自然语言查询等相关任务中。本文进一步对这些问句进行了人工标注,使得每一个自然语言的问句都对应一个 SPARQL 查询。

在实验评价时,由于很难直接采用 SPARQL 语句来进行评价,因此使用系统生成的 SPARQL 语句和正确 SPARQL 语句在语义网数据中的结果来进行评价,如果两者一致,则认为查询结果正确,否则认为查询结果错误。最终的评价指标为准确率,即对 880 句自然语言问句进行自动分析,最终得到正确查询结果的百分比。

5.1　实验结果

本文将所提方法与目前性能最好的一种无指导的方法进行了对比,这一方法是 2007 年由 Wang 等提出的。最近也有一些其他相关工作,但是由于这些工作使用的方法都是有指导的方法,因此这里不将本文工作与这些工作进行比较。

表 1 列出了最终的实验结果。从中可以看出,本文提出的基准系统比 PANTO 高近 2 个百分点;当使用了外部语义知识之后,系统性能得到了进一步的提升。为了表明提升结果是有意义的,本文利用基于成对(pair-wise)的 t-test 方法进行了显著性检验,检验结果发现融入了外部语义知识之后,系统性能得到了显著的提升(p 值小于 1×10^{-4})。无论对测试集如何抽样,本文系统的最终性能都能大概率优于基准系统。本文提出的两种方法都没有使用任何语义信息,均利用语义网数据直接进行解析和查询。另一方面,从结果中显示的数据可以看到,当使用了规模更大的语义网数据之后,查询正确率没有下降,这表明本文方法能够应对大规模的语义网数据。

表 1　实验结果

Table 1　Experimental results

模型	问题数目	正确数目	准确率/%
PANTO	880	663	75.6
本文基准系统	880	683	77.6
本文最终系统(+外部语义知识)	880	691	78.5

5.2　错误分析

本文进一步通过实验错误分析来观察模型的优点和缺点。

首先,对比本文提出的基准模型引入外部知识和不引入外部知识所产生的差异。通过对结果的观察可以发现,引入外部知识后,对实体和属性的识别在召回率上有了较大的提升,召回率从原来的 82.7% 提升到了 89.5%,这一点与人们的认知相符,因为大规模的语义网数据接收了更多的实体和属性。另一方面,实体和属性的识别准确率从原来的 86.8% 下降至 84.4%,这也是合理的,因为大规模的语义网数据可能会把一些与本领域无关的实体也加入识别结果中。

其次,本文提出的系统很难对带有约束性句法信息的问句进行正确分析,比如问句中包含与“与”“或”“非”等相关的信息时。当然,这种问题也可以简单地通过规则的方式进行解决。事实上,通过实验也可以发现,对这些问句进行规则处理之后,系统的性能可以进一步提升至 79.2%。但是还有更多一些例子,很难一一采用规则进行处理,比如比较形式的问句——“how many states border the largeststate ?”。

最后,本文采用人工的方式对生成的 SPARQL 语句进行评价。以语义的等价为基本出发点,如果预测的 SPARQL 语句和答案在语义上是一致的,则认为分析正确,否则认为分析错误。实际上,这一评价方案比本文中所用的标准评价方式更为严格。表 2 列出了最终的分析结果。整体趋势上,这一评价方式与查询准确率趋势在表现上基本一致,在性能提升上更为显著(p 值小于 1×10^{-5})。这一方面验证了本文提出模型的优点,另一方面也表明了本文所采用的查询准确率也是一种合理的自动评价手段,因为 SPARQL 语句的人工评估所耗费的代价太昂贵。

表 2　基于 SPARQL 语句的人工评价结果

Table 2　Artificial evaluation results based on SPARQL

模型	PANTO	基准系统	最终系统
语义匹配准确率/%	60.0	65.2	70.6

结束语　针对语义网中的查询,本文尝试使用基于自然语言的问句进行查询,这样可以使得普通用户都可以方便地

对语义网数据进行查询而不再需要了解语义网查询语言的专业知识。本文提出了一种无指导的系统,它不需要任何人工标注的训练语料;同时,本文还提出了一种利用外部大规模半结构的知识库来自动扩充语义网的方案,基于这一扩充的语义网,本文方法能取得更好的性能。最终的实验结果表明,本文提出的方法是非常有效的,针对美国地理数据库,在一个人工构建的自然语言查询问句集上,能够取得 78.5%的查询准确率。下一步将集中研究如何能够更好地利用结构化的句法以及语义信息提升系统的性能。本文实验部分的结果分析中已经初步表明,如果能够很好地运用结构化的句法以及语义信息,系统的性能将会接近甚至超过 80%。

参 考 文 献

[1] FABIAN M. SUCHANE K, KASNEC G, et al. Yago: A Core of Semantic Knowledge[C]// Proceedings of WWW. New York: ACM, 2007: 697-706.

[2] BOLLACKER K, EVANS C, PARITOSH P, et al. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge [C]// Proceedings of the SIGMOD. New York: ACM, 2008: 1247-1250.

[3] BERNERSLEE T, AHENDLER J, LASSILA O. THE SEMANTIC WEB[J]. Scientific American, 2001, 284(5): 28-37.

[4] WANG C, XIONG M, ZHOU Q, et al. PANTO: A Portable Natural Language Interface to Ontologies[C]// The Semantic Web: Research and Applications, ESWC 2007. Berlin: Springer, 2007: 473-487.

[5] TROELS A. An approach to knowledge-based query evaluation [J]. Fuzzy Sets and Systems, 2003, 140(1): 75-91.

[6] ZHANG Z R, YANG T Q. SPARQL ontology query based on natural language understanding[J]. Journal of Computer Applications, 2010, 30(12): 3397-3400. (in Chinese)
张宗仁,杨天奇. 基于自然语言理解的 SPARQL 本体查询[J]. 计算机应用, 2010, 30(12): 3397-3400.

[7] LI H, TIAN J W, WANG H H, et al. Ontology-based Natural Language Interface to Relational Databases[J]. Computer Science, 2010, 37(6): 200-205. (in Chinese)
李虎,田金文,王缓缓,等. 基于 Ontology 的数据库自然语言查询接口的研究[J]. 计算机科学, 2010, 37(6): 200-205.

[8] XU K, FENG Y S, ZHAO D Y, et al. Automatic Understanding of Natural Language Questions for Querying Chinese Knowledge Bases[J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2014, 50(1): 85-92. (in Chinese)
许坤,冯岩松,赵东岩,等. 面向知识库的中文自然语言问句的语义理解[J]. 北京大学学报(自然科学版), 2014, 50(1): 85-92.

[9] LINCKELS S, MEINEL C. Semantic Interpretation of Natural Language User Input to Improve Search in Multimedia Knowledge Base [J]. Information Technology, 2007, 49(1): 40.

[10] BERANT J, CHOU A, FROSTIG R, et al. Semantic Parsing on Freebase from Question-Answer Pairs[C]// Proceedings of the EMNLP 2013. USA: ACL, 2013: 1533-1544.

[11] LIANG P, JORDAN M I, KLEIN D. Learning dependency-based compositional semantics[J]. Computational Linguistics, 2013, 39(2): 389-446.

[12] KWIATKOWSKI T, CHOI E, ARTZI Y, et al. Scaling Semantic Parsers with On-the-fly Ontology Matching [C]// Proceedings of the EMNLP. USA: ACL, 2013: 1545-1556.

[13] WONG Y W, MOONEY R J. Learning Synchronous Grammars for Semantic Parsing with Lambda Calculus[C]// Proceedings of ACL 2007. USA: ACL, 2007: 960-967.

[14] JONATHAN H, BERANT J. Neural Semantic Parsing over Multiple Knowledge-bases[C]// Proceedings of the ACL 2017. USA: ACL, 2017: 623-628.

[15] ALON T, BERANT J. The Web as a Knowledge-Base for Answering Complex Questions[C]// Proceedings of the NAACL-HLT. USA: ACL, 2018: 641-651.

[16] SUHR A, IYER S, ARTZI Y. Learning to Map Context-Dependent Sentences to Executable Formal Queries[C]// Proceedings of NAACL-HLT. USA: ACL, 2018: 2238-2249.

[17] CHEN B, AN B, SUN L, et al. Semi-Supervised Lexicon Learning for Wide-Coverage Semantic Parsing[C]// Proceedings of the COLING 2018. USA: ACL, 2018: 892-904.

[18] MARTINS A F T, SMITH N A, XING E P, et al. Turbo parsers: Dependency parsing by approximate variational inference [C]// Proceedings of the EMNLP 2010. USA: ACL, 2010: 34-44.

[19] DAS D, CHEN D, MARTINS A F T, et al. Frame-semantic parsing[J]. Computational Linguistics, 2014, 40(1): 9-56.