北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2014.023

面向知识库的中文自然语言问句的语义理解

许坤　冯岩松†　赵东岩　陈立伟　邹磊

北京大学计算机科学技术研究所，北京 100871; †通信作者, E-mail: fengyansong@pku.edu.cn

摘要　设计从自然语言问句到结构化查询的转换框架，该方法从自然语言问句的句法结构入手，提出一套启发式识别实体与关系的方法，并利用语料库建立从实体到知识库的映射，对谓词进行消歧，进而转化为计算机可理解的结构化查询语言。从百度知道抽取人物、地点、组织 3 类一共 42 个问题作为标准测试集。实验结果表明，所提出的框架能够有效地将中文自然语言问句转换为结构化查询，为下一代智能问答系统打下良好的基础。

关键词　自然语言问句；知识库；查询语义图

中图分类号　TP391

AutomaticUnderstandingofNaturalLanguageQuestionsfor Querying Chinese Knowledge Bases

XU Kun, FENG Yansong†, ZHAO Dongyan, CHEN Liwei, ZOU Lei

Institute of Computer Science and Technology, Peking University, Beijing 100871;
† Corresponding author, E-mail: fengyansong@pku.edu.cn

Abstract　A framework to transform natural language questions into computer-understoodable structured queries is presented. The authors propose to use query semantic graph to represent the semantics in Chinese questions, and adopt predicate and entity disambiguation to match the query graph to the schema of a knowledge base. The authors collect a benchmark of 42 frequently-asked questions randomly sampled from 3 categories of Baidu Knows, including person, location and organization. Experiment results show that proposed framework can effectively convert natural language questions into SPARQL queries, and lay a good foundation for the next generation of intelligent question answering systems.

Key words　natural language question; knowledge base; query semantic graph

随着万维网的不断发展，出现了越来越多的大规模知识库，比如 Yago[1]，yago2[2]，DBpedia[3]和 FreeBase[4]等。这些知识库覆盖面广、数据量大，例如 DBpedia 有 340 万个实体，300 亿条三元组。面对如此庞大的结构化数据，让用户可以方便地使用这些知识库查询也变得越来越重要。这些知识库数据量大，但是彼此的架构并不相同，所以普通用户很难在这些知识库上轻松地搜索相关信息。

知识库通常都以 RDF①的格式存储数据，而 SPARQL[5]是目前已知的比较高效的查询 RDF 数据的语言，但是通常只有专业的程序员才能熟练掌握 SPARQL 语法。而对于其他普通用户，他们更愿意选择使用关键词或者自然语言问句来查找需要的信息。其中，基于关键词的查询应用最广泛，但是有一些语义复杂的问题不适合用关键词查询，比如用户想知道“刘德华的妻子毕业于哪里”，但是当他们在百度上输入“刘德华妻子毕业”这样的关键词，返回的结果基本都是在介绍刘德华。事实上语义复杂

国家自然科学基金(61272344, 61202233, 61370055)资助
收稿日期: 2013–06–17; 修回日期: 2013–09–30; 网络出版时间: 2013-11-07 11:30

① http://www.w3.org/RDF/

的问题更适合使用 SPARQL 等结构化查询语言在知识库上查询。所以理想的解决方法是用户使用自然语言描述问题，而系统利用 SPARQL 语言在知识库上查询。

在将自然语言问句转换为 SPARQL 的过程中，难点在于如何让计算机理解用户的查询语义。为此，本文提出查询语义图的概念，并且提出利用句法分析树构造用户的查询语义图。在本文提出的查询语义图中，顶点代表命名实体或者名词性变量，边代表顶点之间的语义关系。生成初步的查询语义图后，利用实体消歧和谓词消歧将查询语义图中的顶点与边映射到知识库中的实体与关系。在谓词消歧时，首先为知识库的每个关系收集相关度比较高的动词词组和名词词组。例如，对于知识库中的关系“毕业院校”，我们从语料库中收集到“毕业”、“考入”、“就读”、“报考”等与“毕业院校”相关度比较高的关键词。接着定义一种计算词语相关度的算法，利用收集到的词将图中的谓词映射到语义上最相关的关系。完成实体消歧和谓词消歧后，用户的查询语义图中的实体和关系被链接到知识库。最后将查询语义图转换成 SPARQL 语句并利用基于 RDF 的搜索引擎 g-store[6]进行查询。

本文针对中文自然语言查询，从问句的句法结构入手，使用查询语义图来表示用户的查询语义，并利用中文语料库进行消歧，在用户的查询意图与知识库的底层表示之间建立映射，从而获得结构化查询语句。此外，我们还收集了面向百科知识库的问题集合，作为中文问题的标准测试集，并对所提出的方法进行验证。

1 相关工作

目前，面向知识库的问题理解主要针对英文的自然语言问题。其中有一些方法利用句法分析树和知识库来启发式理解用户的查询语义，还有一些利用用户对答案的反馈来训练问答系统。Yahya et al. [10]提出一种理解自然语言问句的方法，他们首先将问句划分成短语，并将这些短语映射到知识库中的实体、类别和关系。接着利用整数线性规划实现实体消歧。而在本文中，实体消歧和谓词消歧是分开进行的，先对实体进行消歧，然后对候选谓词进行消歧。Unger et al.[11]提出一种利用模板和句法分析树来生成结构化查询的方法。本文方法不需要预先定义这样的模板。Ferrucci et al. [12]展示了一个被称作“Watson”的深度问答系统。他们将一个问句分解成一些线索和子线索，而目标就是找到这些线索的答案。Pythia et al.[13]提出一个依赖问句深度语言分析的系统，首先手动为每一个本体构造描述该本体语义的词典，然后利用该词典来处理语义上比较复杂的问题。FREyA[14]系统依赖于用户对问句的反馈来进行命名实体消歧，在该方法中，需要有监督地训练问答系统。在该方法中，用户提出一个问题后，系统返回给用户一些候选的语义，用户从中选择最合适的语义，系统则根据用户的选择进行训练。PowerAqua[15]也是一个问答系统，该系统的知识库来自于不同的资源。这样做会造成回答的答案质量比较低，数据总体的噪音比较大。本文只使用互动百科作为底层的知识库，所以答案的质量会比较好。

我们的方法与之前的工作有两点区别。首先，之前工作主要用来处理英文的自然语言问题，而缺乏对其他语言的关注，如中文。与英文不同,中文中承担语法功能的一些结构与其词性之间没有直接关联，使得很多英文中的启发式方法在中文无法直接使用。比如，在英语中，谓语结构一般都是由动词构成，而在汉语里，承担谓语功能的形式比较多，可以是形容词、动词和名词。其次，之前的工作中大多数采用很多现有的英文资源比如 WordNet[6]和 PATTY[7]，但是中文类似这样的资源少，而且覆盖面较小。这个问题也是转换中文自然语言查询遇到的一个难点。比如，在进行谓词消歧的时候，只能根据现有的语料去收集与关系相关度比较高的词汇。

此前的很多工作中都使用 QALD-1②标准测试集。对于任何一个知识库，QALD-1 都提供 50 个不同语义复杂度的训练问题及其相应的 SPARQL 查询，同时还提供 50 个类似的测试问题。训练问题和测试问题都是英文自然语言问句。而中文并没有这样的数据集。使用 QALD-1 可以更方便地比较不同的问答系统。使用该测试集的问答系统需要使用 DBpedia 或者 MusicBrainz 作为知识库。为此，我们从百度知道上抽取了人物、地点、组织三类一共 42 个询问百科知识的事实类问题，将其作为我

② http://www.sc.cit-ec.uni-bielefeld.de/qald-1

们的测试问题。

2 基本框架

给定一个用自然语言描述的问句, 首先需要对句子进行预处理, 包括分词, 命名实体识别和句法分析。得到句法分析树后, 从句法分析树中获得用户的查询语义, 并构造出查询语义图, 然后将图中的点映射到知识库中的实体, 并将图中的边映射到知识库中的关系。最后依据查询语义图生成结构化查询语句(如 SPARQL), 利用基于 RDF 的查询引擎 g-store[6]得到最终的查询结果。

3 预处理

用户提出一个用自然语言描述的问题后, 首先分词并获得问句的句法结构、POS tag 以及词之间的依赖关系。在实验中, 我们使用 ICTCLAS③来对问句进行分词, 利用斯坦福大学的工具 Standford Parser[8−9]对问句做句法分析, 得到相应的句法分析树。图 1 中包含问句“刘德华的妻子出生于哪里”的句法分析树。

3.1 查询语义图的定义

本文假设用户所问问题均基于实体的关系或者实体间的关系, 而且是事实性问题。我们认为无论一个自然语言问句多么复杂, 它都在描述实体之间的关系。比如对于问题“刘德华的妻子出生于哪里”,

共涉及 3 个实体,设为 E_1、E_2 和 E_3, 其中 E_1 代表刘德华, E_2 是一个类型为人的实体, E_3 是一个类型为地点的实体。它们之间的关系是实体 E_2 是实体 E_1 的妻子, 实体 E_3 是实体 E_2 的籍贯。根据关系的复杂程度将自然语言查询分为两种: 单关系查询与多关系查询,定义如下。

1)单关系查询。如果一个查询只涉及了两个实体, 即查询只涉及一种关系, 那么这种查询为单关系查询。

2)多关系查询。如果一个查询涉及的实体数目超过两个, 即查询涉及多种关系, 那么这种查询就是多关系查询。

在查询语义图中, 将查询中涉及的每个实体视作一个点, 将实体之间的关系视作边。那么对于任意一个查询, 都可以构造出一个查询语义图。如果查询是单关系查询, 查询语义图就只包含两个点, 一条边。而如果查询是多关系查询, 语义图则会包含多个点, 多条边。图 1 包含多关系查询“刘德华的妻子出生于哪里”的查询语义图。总的来说, 查询语义图是用来描述用户的查询中实体关系的一张图。

3.2 构造查询语义图

从查询语义图的定义中, 我们可以发现, 查询语义图主要描述的是问句中实体之间的关系, 而问句的句法分析树不仅能精确地描述句子成分之间的语法关系, 更可以准确地表达用户的查询语义。本文提出一个基于句法分析树的算法。

由于中文本身的复杂性, 实体与关系的表达方式会有很多, 但我们发现实体在句法分析树中一般都以名词的形式出现, 而关系则一般以名词或者动词词组的形式出现。例如, 在图 1 中, 动词性结构“出生于”描述“刘德华”和“哪里”之间的关系。名词性成分“刘德华”和“哪里”则代表实体。为了验证这个观点的正确性, 我们从百度知道上随机抽取了 20 个问题, 分析它们的句法分析树, 经过统计, 发现 90%的问句中的实体都以名词的形式出现, 80%关系都以名词或动词词组的形式出现。

首先介绍从句法分析树构造查询语义图的算法, 算法描述如下。

在构造查询语义图时, 需要分析句法分析树中的每个节点。如果节点属于名词性节点(构造查询语义图算法第 3 行), 那么在查询语义图中构造一个点代表该名词性节点。例如, 在图 1 的句法分析树中, “刘德华”、“妻子”和“哪里”3 个节点都是名词性节点, 因此查询语义图中有 3 个点分别代表这 3 个节点。句法分析树中的名词性节点可以分为两类: 命名实体和名词性变量。图 1 中, “刘德华”是一个命名实体, 而“妻子”和“哪里”是名词性变量。如果一个名词性节点有其他的名词性节点对其修饰, 就在查询图中建立一条从修饰词到被修饰词的有向边。这条边描述这两个词之间的关系, 如果被修饰的词是一个名词性变量, 则该名词描述这个关系(构造查询语义图算法第 9 行)。例如, 在图 1 中, “妻子”是一个名词性变量, 而“刘德华”修饰“妻子”, 所以这两个节点之间的边描述“妻子”这个关系。

如果句法分析树中的节点属于动词性节点(构

③ http://www.ictclas.org/

造查询语义图算法第 5 行), 确定该节点所描述的动词以及该动词的主体与客体, 在主体与客体之间建立一条有向边。例如, 在图 1 的句法分析树中, “出生于”是一个动词, 主体是“妻子”, 客体是“哪里”, 故在查询图中存在一条从“妻子”指向“哪里”的有向边。该边描述“出生于”这个关系。

4 消歧

生成初步查询语义图以后, 已经大致清楚用户的查询意图, 但是还需要将查询语义图中的点与边映射到知识库的底层表示。

这个问题从本质上是一个消歧问题, 可以形式化为: 给定一个字符串 s 和一个知识库 K, 从 K 中找到一个与 s 最相似的实体或者关系。这个问题中, 关系消歧相对于实体消歧更加复杂。因为在关系消歧时, 一个关系可以有很多种的自然语言描述方式。所以对关系消歧和实体消歧采用不同的方法。

4.1 实体消歧

为了在 K 中找到与 s 语义最接近的实体, 我们需要计算 s 与 K 中实体的语义相似度。出于对百科知识库覆盖面的考虑, 我们收集互动百科中实体的中文名称及其别名、别称并构造成一个词典, 接着计算 s 与词典中每个词的编辑距离, 并将编辑距离最小的实体加入到 s 的候选集里。在实验中, 与 s 编辑距离最小的实体一般都会有多个, 我们在生成 SPARQL 语句时, 会穷举 s 的所有可能性。

构造查询语义图算法

1: 令 TreeNodes 表示句法分析树中所有节点的集合
2: for 每个节点 node ∈ TreeNodes do
3: if node 是名词性节点 then
4: 调用函数分析名词性节点(node)
5: Else If node 是动词性节点 then
6: 调用函数分析动词性节点(node)
7: End For

分析名词性节点(npNode)

1: 令 npNode 的子节点集合是 Children
2: if Children 中不包含名词性节点 then
2: 构造一个节点代表 npNode 加入 G
3: if npNode 不是一个命名实体 then
4: npNode 是一个变量
5: End If
3: else
4: for 每个节点 Child ∈ Children do
5: if *Child* 是名词性节点 then
6: 调用函数分析名词性节点(Child)
7: End If
8: End For
9: 在修饰性名词节点与被修饰名词节点之间引入边 l, l 从修饰性名词节点指向被修饰名词节点。如果被修饰名词节点不是命名实体而是一个名词性变量, 则该名词是用来描述这两个节点之间的关系
10: End Else
11: End If

分析动词性节点(vpNode)

1: 令 vpNode 的子节点集合是 Children
2: for 每个节点 Child ∈ Children do

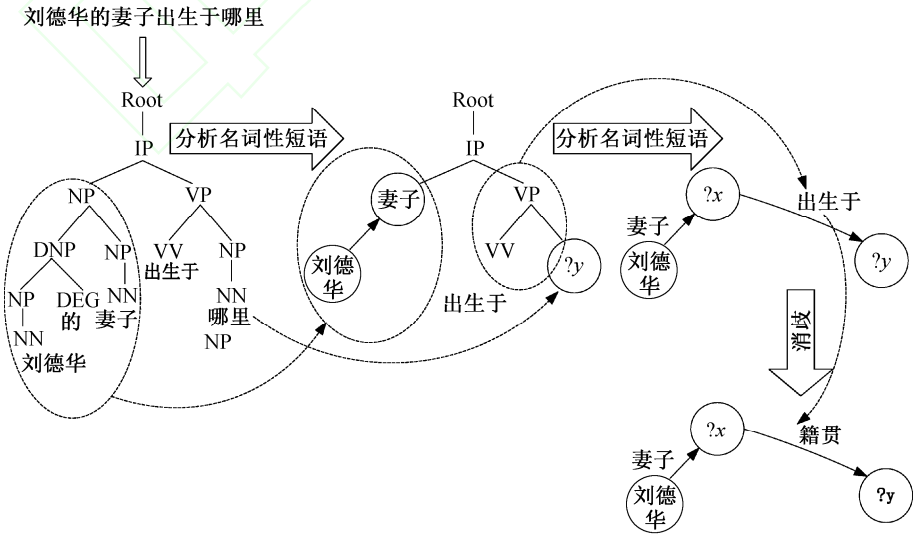

图 1 问句“刘德华的妻子出生于哪里”查询语义图的生成过程

Fig.1 Procedure of constructing the semantic graph of query “where was Liu Dehua’s wife born”

3: if Child 是名词性节点 then
4: 调用函数分析名词性节点(Child)
5: End If
6: End For
7: 在代表动词主体的名词性节点与代表客体的名词性节点之间构造一条边

4.2 谓词消歧

在我们系统中, 谓词消歧是一个比较困难的问题。主要是因为与知识库中的一种关系往往会有很多种描述方式。比如, "*Y* 的作者是 *X*"和"*Y* 是一本 *X* 的作品"这两种表达方式都表明 *X* 是 *Y* 的作者, 但是它们在字符串层面上差距很大。为了解决这个问题, 本文提出利用与关系相关度较高的关键词来进行消歧。虽然每种关系都会有很多描述方式, 但是在这些描述方式中, 有些词出现的频率会比较高。例如, 对于知识库中的关系"籍贯", "出生于"、"祖籍"、"出生"等一些关键词可能描述该关系的概率就比较高。我们认为这些出现概率比较高的关键词与关系相关程度比较高, 所以, 如果查询语义图中的边对应的词组与某个关系相关度比较高, 就认为这个词组是用来描述该关系的, 并将这条边映射成该关系。

5.2.1 收集与关系相关度高的关键词

在实验中, 我们使用最大的中文百科——互动百科收集与关系相关度比较高的词。互动百科中的一些实体页面包含一种被称作 infobox 的结构, 该结构用来描述实体的一些基本信息。图 2 就是 infobox 的一个实例。从图中可以看出, 在 infobox 中, 每一行都描述一个关系和相应的客体,而主体就是该页面所描述的实体。我们从 infobox 中抽出这些三元组, 形成底层的知识库。经统计, 互动百科共有 197 种关系。

英文名:	AndyLau
艺名:	华仔, 老大, 华哥, 华弟, 刘天王
性别:	男
血型:	AB
出生年月:	1961年9月27日
毕业院校:	可立中学、第十期无线艺员训练班
身高:	175厘米
爱好:	保龄球、羽毛球、台球、驾驶

图 1 刘德华页面的 infobox 结构
Fig.2 Infobox of wiki page about Liu Dehua

为了收集与这 197 种关系相关的关键词, 需要收集描述这些关系的句子。基于弱监督学习的假设[18],我们启发式地认为如果句子中出现符合该关系所代表关系的实体对, 则该句子有很高的概率描述了该关系。例如, 在"刘德华"页面的 infobox 中, 对于关系"毕业院校", 我们抽取到实体对 $R_{\text{毕业院校}}$<刘德华,可立中学>, 而在刘德华页面中, 我们又抽取到"刘德华在黄大仙天主教小学毕业后升读可立中学", 这句话中同时出现了"刘德华"和"可立中学"两个实体, 那么这句话有很高的概率描述了"毕业院校"这个关系。但是这个基于弱监督学习的假设并不总是正确。例如, 对于关系"籍贯", 我们抽取到"丁叮, 2005 年超级女生杭州赛区十强选手", 按照之前的假设, 这句话在描述"籍贯", 但这句话很明显与籍贯关系不大。所以这样的收集方法会引入噪声, 但百科页面主体即三元组主体, 可靠性较高, 而且当数据规模变大时, 出现频率较高的关键词应有较大概率描述该关系。

对于任意一种关系 *a*, 首先从所有 infobox 结构中包含关系 *a* 的实体页面中抽取出实体对 R_a<*s*, *o*>并放入个实体对集合 P_a, 然后收集所有出现 P_a 中实体对的句子, 并将该句子放入集合 S_a。表 1 列出了一些描述"毕业院校"的句子。

根据上面的讨论, 描述实体之间关系的词一般都是名词或者是动词。所以在收集与关系相关度较高的词时, 主要考虑名词与动词。

表 1 描述"毕业院校"语句
Table 1 Sentences which describe the relation of "Graduate From"

描述性语句	实体对
1948 年, 丁一三毕业于天津宁河中学。	<丁一三, 天津宁河中学>
丁嘉丽 1959 年 12 月生于山东, 毕业于上海戏剧学院。	<丁嘉丽, 上海戏剧学院>
十八岁的丛珊考入中央戏剧学院表演系, 学习舞台表演。	<丛珊, 中央戏剧学院表演系>
丁磊毕业于电子科技大学, 获工学学士学位。	<丁磊, 电子科技大学>

接着对描述 *a* 的句子做词性标注, 统计所有名词和动词出现的频率, 并根据出现的频率进行排序。但在实验中有些名词和动词与多个关系都有比较高的相关度。比如"出生"这个动词在关系"籍贯"、"毕业院校"和"出生年月"中出现的频率都很高, 除此之外, 常用词也很大程度地影响了最后的

排序。为了解决这个问题，我们不再简单地考虑每个词出现的频率，取而代之，考虑每个词的 tf-idf 值。把与每个关系相关的动词与名词的集合当做单个文档，并把这些文档统一起来计算每个词的 tf-idf 值，最后对与关系相关的词根据其 tf-idf 值从高到低进行排序。这样就可以得到与每个关系相关，并且按照相关程度从高到低排序的词的列表。

4.2.2 谓词映射

得到与关系相关度较高的词后，我们定义一个计算谓词与关系相关度的函数。假设待映射的谓词是 G，知识库的关系集合是 L，l 是任意一种关系，G 与 l 的词语相关度为

$$\mathrm{Rel}(G,l)=\sum_{i=1 w_i\epsilon W}^{m}\mathrm{Rel}(G,w_i)*\mathrm{tf\_idf}(w_i),$$

其中关系 l 有 m 个相关度比较高的词，W 是与 l 相关较高的词的集合。$\mathrm{Rel}(G,w_i)$ 的定义如下:

$$\mathrm{Rel}(G,w_i)=\begin{cases}1, G=w_i,\\ 0, \mathrm{G}\neq w_i。\end{cases}$$

最后，我们规定 G 映射到的关系满足如下的条件:

$$\mathrm{Match}(G)=\arg\max_{l}\mathrm{Rel}(G,L)。$$

5 生成 SPARQL 查询

在进行实体消歧和谓词消歧后，查询语义图中的每一个实体都会有一个候选集，每一个谓词都代表知识库中的一个关系。根据前面的介绍可知，查询语义图的每一条边都代表一个三元组，所以我们直接将查询语义图每条边所代表的三元组组合起来，可以得到 SPARQL 查询语句。比如对于图 1 的查询语义图, SPARQL 语句如下:

```
select ?y where{
刘德华　妻子?x
?x　籍贯?y
}
```

最终，可以利用 RDF 搜索引擎 g-store[6]进行查询得到最终结果。

6 实验与分析

百度知道是由全球最大的中文搜索引擎百度自主研发、基于搜索的互动式知识问答分享平台。经过统计，百度知道已经解决了 230 万个问题。由于中文并没有类似 QALD 这样的测试集，所以为了测试中文的问句理解，我们从百度知道中抽取出 42 个问题。这些问题按照查询实体的类别共分为人物、地点、组织机构三大类，其中每一类有 14 个问题。表 2 列出了部分问题。这些问题所涉及到的实体和相关关系都可以链接到知识库。我们手动地为这些问题写出相应的 SPARQL 查询，并用 g-store 进行查询，由于知识库的覆盖面问题，只有 22 个问题可以得到结果。在判断生成的 SPARQL 语句和谓词消歧是否正确时，需要人工测评，共有 3 人参与，3 人都熟悉 SPARQL 和底层知识库，当有

表 2　问题列表

Table 2　Question List

序号	类别	问题描述	序号	类别	问题描述	序号	类别	问题描述
1	人物	张杰是哪里人?	15	地点	北京有哪些地标性建筑?	29	组织	国民党是谁建立的?
2	人物	张杰的第一张专辑是什	16	地点	北京的人口有多少?	30	组织	中国共产党现任总书记是谁?
3	人物	张杰的生日是啥时候?	17	地点	北京市的现任市长是谁?	31	组织	九三学社的加入条件是什么?
4	人物	张杰的老婆是谁?	18	地点	北京市有哪些知名企业?	32	组织	国民党的政治理念是什么?
5	人物	刘德华的原名叫什么?	19	地点	北京市有哪些机场?	33	组织	共产党创建时间是哪一年?
6	人物	刘德华的爸爸是谁?	20	地点	北京市有哪些行政区?	34	组织	中国铁道部部长是谁?
7	人物	刘德华的女儿演过什么电	21	地点	朝阳区的邮政编码是多	35	组织	中国工会的会长是谁?
8	人物	梁朝伟演过什么电影?	22	地点	北京有哪些特产?	36	组织	国际红十字会的英文缩写是什
9	人物	刘德华出生在哪里?	23	地点	北京的著名景点有哪些?	37	组织	国民党什么时候成立的?
10	人物	梁朝伟的女朋友有哪些?	24	地点	北京的面积有多大?	38	组织	发改委的办公驻地在哪里?
11	人物	梁朝伟的身高有多少?	25	地点	北京有哪些地标?	39	组织	发改委的领导人是谁?
12	人物	梁朝伟写过那些书?	26	地点	上海的主要街道有哪些?	40	组织	发改委的全称是什么?
13	人物	梁朝伟高中毕业于哪里?	27	地点	上海的电话区号是多少?	41	组织	中国农业部的网站是什么?
14	人物	梁朝伟出生在哪里?	28	地点	上海的名人有哪些?	42	组织	中国农业部的职能是什么?

过半数的人赞成某种结果时, 我们就认为该结果是正确的。本文从 3 个方面来评测我们的方法。

1)回答问题的准确率。对于可以用知识库回答的 22 个问题, 我们利用本文的方法生成 SPARQL 查询, 并利用 g-store[6]系统得到答案。经过统计, 共回答正确 10 个问题, 正确率为 45%。

2)生成 SPARQL 查询的准确率。对于实验数据集中的 42 个问题, 我们将手动生成的 SPARQL 查询与利用本文方法生成的 SPARQL 查询进行比较。实验采用两种比较方法。

①自动评价。只有生成的 SPARQL 查询与手动生成的 SPARQL 查询字符串完全相同, 才认为生成的 SAPRQL 查询是正确的。经过统计, 正确率为 36%。

②人工评价。人工地判断生成的 SPARQL 查询与手动生成的 SPARQL 查询是否在语义上相同。如果相同, 则认为生成的 SPARQL 查询是正确的。经过统计, 正确率为 48%。下面是部分实验结果。

问题 1 张杰的生日是啥时候?
SPARQL 查询
select ?*x* where{
张杰出生年月?*x*
}
结果 11982 年 12 月 20 日。
问题 2 成龙的儿子演过什么电影?
SPARQL 查询
select ?*y* where{
成龙儿子 ?*x*
?*x*　主演 ?*y*
}
结果 2《千机变》。

3)谓词消歧的准确率。对于本文计算框架中的基于语料库的谓词消歧部分也进行手工评价。特别地, 我们还实现了一个基于 HowNet[17]的消歧方法。在 HowNet 中, 词用概念来描述, 一个词可以表达为几个概念。对于两个汉语词语 W_1 和 W_2, 如果 W_1 有 n 个概念: $S_{11}...S_{1n}$, W_2 有 m 个概念: $S_{21}...S_{2m}$, 那么规定 W_1 和 W_2 的相似度为各个概念的相似度的值的最大值:

$$\mathrm{Sim}(W_1,W_2)=\max_{i=1..n\ j=1..m}\mathrm{Sim}(S_{1i},S_{2j})。$$

同时, 我们规定概念相似度的计算公式如下: $\mathrm{Sim}(p,q)=\frac{1}{d}$, 其中 p 和 q 代表两种概念, d 是 p 和 q 在概念层次体系中的路径长度, 为一个正整数。通过计算谓词与所有关系的相似度, 将谓词映射到最相似的关系, 然后手动判断谓词映射是否正确。经过统计, 共有 30 个谓词可以映射到知识库, 谓词映射正确率为 10%。而利用本文中的方法计算谓词与关系的相似度, 经过统计, 准确率为 36%。

通过实验可以发现, 本文方法产生的 SPARQL 查询通过人工语义比较确定的准确率会更高。这是因为我们使用的知识库是从互动百科自动构建的, 其关系体系并不完美, 存在一些关系可以被映射到多个关系, 例如, “出生于”可以映射到关系“籍贯”和“出生地”。实验还表明, 在谓词消歧时, 利用本文的方法可以获得更高的准确率。通过错误分析, 我们发现 80%的 SPARQL 语句错误是由谓词消歧造成, 主要原因是, 由于部分属性所能收集到的语料较少, 弱监督假设引入的噪声等问题, 我们收集到的词与关系相关度并不一定都正确。表 3 列出收集到的与部分关系相关度较高的词。其中带下划线的词其实与关系相关度并不高。例如, 对于“导演”这个关系, “金庸”被计算出相关度比较高, 出现这种问题的原因可能有两个: 1) 对于“导演”这个关系, 收集到的实体对少, 而且描述“导演”的句子也少。从而收集到一些与“导演”相关度不高的词; 2) 对于“导演”这个关系, 在收集到的实体对中, 有关“金庸”的实体对比例大, 从而计算出“金庸”与“导演”相关度较高。另外, “武侠小说”、“中篇小说”和“长篇小说”被认为与关系“作者”相关度较高也是由于上述的两个原因。

表 3　与关系相关度较高的词

Table3　Relations and the corresponding high relevance words

关系	相关度较高的词
毕业院校	毕业、考入、就读、大学
出生地	出生、生于、原名、祖籍
代表作品	作品、著作、演、参与
导演	执导、电视剧、作品、金庸
作者	武侠小说、中篇小说、长篇小说、代表作

7　总结

本文提出了查询语义图的概念, 并且利用图结构来表示自然语言问题的语义, 最终将其转化为结构化查询的方法。该方法从自然语言问句的句法结构入手, 提出了一套启发式识别实体与关系的方法, 并利用语料库建立了从实体到知识库的映射, 对谓

词进行消歧，进而转化为计算机可理解的结构化查询语言。实验中，我们从百度知道中抽取出了人物、地点、组织三类一共 42 个问题作为标准测试集。实验结果表明，本文提出的框架能够有效地将中文自然语言问句转换为结构化查询。

在实验中，通过错误分析，我们发现 80%的 SPARQL 语句错误是由谓词消歧造成。所以未来的工作将考虑如何提高收集的词与关系的相关度，增加谓词消歧的准确率。

参考文献

[1] Suchanek F M,Kasneci G,Weikum G. Yago: a core of semantic knowledge // WWW.Beijing, 2007: 697–706

[2] Hoffart J,Suchanek F M,Berberich K, et al. Yago2: exploring and querying world knowledge in time, space, context, and many languages // ACM. Scottsdale, Arizona, 2011: 229–232

[3] Auer S, Bizer C,Kobilarov G,et al. DBpedia: a nucleus for a web of open data // ISWC/ASWC. Busan, Korea, 2007:722–735

[4] Bollacker K D, Evans C,Paritosh P,et al. Freebase: a collaboratively created graph database for structuring human knowledge // SIGMOD. Vancouver, 2008:1247–1250

[5] Prudhommeaux, Seaborne A. SPARQL query language for RDF. SIGMOD.Vancouver, 2008:1217–1220

[6] ZouLei, Mo Jinghui, Chen Lei, et al. GStore: answering SPARQL queries via subgraph matching // VLDB.Seattle, 2011: 482–493

[7] Pedersen T, Patwardhan S, Michelizzi J. 2004.WordNet:Similarity:measuring the relatedness of concepts // HLT-NAACL. Montréal, Canada, 2012:28–31

[8] Nakashole N, Weikum G, Suchanek F. PATTY: a taxonomy of relational patterns with semantic types // Empirical Methods in Natural Language Learning. Jeju Island, Korea, 2012: 1135–1145

[9] Toutanova K, Manning C D. Enriching the knowledge sources used in a maximum entropy // EMNLP/VLC-2000. Kong, 2000: 63–70

[10] Toutanova K, Klein D, Manning C D, et al. Feature-rich part-of-speech tagging with a cyclic dependency network // HLT-NAACL.Edmonton, Canada,2003: 252–259

[11] Yahya M, Berberich K, Elbassuoni S, et al. Natural language questions for the web of data // Natural Language Processing and Natural Language Learning. Sydney, Australia,2012:379–390

[12] Unger C, Buhmann L, Lehmann J, et al. Template-based question answering over RDF data // WWW. Seoul, Korea,2012: 639–648

[13] Ferrucci D, Brown E, Chu-Carroll J, et al. Building watson: an overview of the DeepQA project // AAAI. San Jose, California,2004: 59–79

[14] Unger C,CimianoP. Pythia: compositional meaning construction for ontology-based question answering on the Semantic Web // NLDB. Salford, UK,2011: 153–160

[15] Damljanovic D, Agatonovic M, Cunningham H. FREyA: an interactive way of querying linked data using natural language // QALD-1, ESWC. Crete, Greece,2011:115–120

[16] Lopez V, Fernandez M, Stieler N, et al. PowerAqua: supporting users in querying and exploring the Semantic web content.Semantic Web Journal, 2001:1250–1256

[17] Dong Zhendong,Dong Qiang. HowNet and the computation of meaning // World Scientific Publishing Co Pte Ltd. Singapore, 2006:112–116

[18] Mintz M, Bills S, Snow R, et al. Distant supervision for relation extraction without labeled data //ACL. Singapore, 2009:1003–1011