基于中文知识图谱的电商领域问答系统

杜泽宇 杨 燕 贺 樑
（华东师范大学信息科学技术学院 上海 200062）

摘 要 随着知识图谱的迅速发展，面向知识图谱的中文领域问答系统已成为目前最新最热的研究方向之一，对于提高专业领域服务智能化程度具有较高的意义和价值。针对中文口语语义表达多样化、不符合语法规范以及电商领域特殊性问题，提出一套流式的中文知识图谱自动问答系统 CEQA，能够较好地完成电商领域商品咨询以及统计推理等复杂问题，特别是有效地提升了中英文混合商品名称识别、语义链接以及复杂问句的依存分析等方面的性能。实验结果表明，该系统在电商领域问答应用中具有较高的准确率和实用价值。

关键词 自动问答 知识图谱 语义网 本体

中图分类号 TP3 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2017.05.027

QUESTION ANSWERING SYSTEM OF ELECTRIC BUSINESS FIELD BASED ON CHINESE KNOWLEDGE MAP

Du Zeyu Yang Yan He Liang
(*School of Information Science Technology, East China Normal University, Shanghai 200062, China*)

Abstract With the rapid development of knowledge map, the Chinese domain question answering system for knowledge map has become one of the newest and hotest research directions at present, and it is of great significance and value to improve the intelligence level of professional field. In this paper, a set of streaming Chinese knowledge map automatic question answering system (CEQA) is proposed for the diversification of Chinese spoken language semantic expression, grammatical specification and the particularity of electricity business domain. It can accomplish the complex problem of commodity consultation and statistical reasoning in the field of electric business, especially the improvement of the interdependence between Chinese and English mixed commodity name recognition, semantic link and complex question. The experimental results show that the system has high accuracy and practical value in the application of question and answer.

Keywords Question answering Knowledge map Semantic Web Ontolog

0 引 言

知识图谱最早起源于 Google 的 Knowledge Graph，它本质上是一种语义网络，其结点代表实体或者概念，边代表实体/概念之间的各种语义关系。随着结构化数据源的剧增，互联网正在从大量互相链接的网页向包含大量描述各种实体和实体之间丰富关系的语义网演进。如今已经有很多著名的知识图谱知识库，如 DBpedia、Freebase、Yogo、百度知心、知立方等。知识图谱对搜索引擎提供语义层面上的支持，用户通过关键词搜索模式已经很难满足用户的需求[1]。用户更希望通过自然语言查询，直接得到所需的答案，智能问答系统正在成为新一代信息检索技术发展的必然趋势。

知识图谱构建是自底向上数据驱动型，相对于本体而言，数据语义表达灵活，实体覆盖率更高，语义关系也更加全面。现有的知识图谱的标准数据通常是由 RDF 三元组数据存储形式构成，即：<主语，谓语，宾语>，还有一些加入本体信息结构的 OWL 数据，其中包含本体的基本概念，例如类（Class）、属性（Proper-

收稿日期:2016-02-21。国家高技术研究发展计划项目（2015BAH12F01-04）；上海市科委重点项目（14511107000）。杜泽宇，硕士生，主研领域：自然语言处理。杨燕，博士生。贺樑，教授。

ty)、实例(Individual)等。庞大知识图谱不仅包含事实类知识,还有丰富的语义知识为自然语言理解、知识推理和计算等方面提供强有力的支持。

基于知识图谱的问答系统有两大核心问题,前端语义理解和后端知识图谱构建。通用的问答流程是将自然语言翻译成结构化的查询语言,比如 SQL[2]、SPARQL[3-5],以及其他的语言[6-8]查询知识图谱中的实体和关系。基于知识图谱的自动问答系统能够支持推理等更多复杂的问题,如包含逻辑判断的问句,如电商中"与 iphone5s 相同尺寸的手机有哪些?"等这类问句。近年来,IBM 的 Waston、Google Now 和 Siri 等都应用了知识图谱相关技术,目前,我国电商行业发展迅速,用户对于商品的咨询量较大,自动问答系统可以部分缓解人工客服压力,做到 24×7 在线服务,并且容易结合用户信息扩展为对用户提供个性化智能服务,例如京东的 JIMI 机器人可以提供基本查询和聊天等服务。

国内外在语义网相关问答系统方面已经有了很长时间的研究。AquaLog[9]是较早基于多样化语义网资源进行自动问答的系统,其主要特点在于融合了消岐与排序的技术,可以处理多个语义网资源混合情况下的问答。其瓶颈在于无法处理类似于 < Counting ,how many , higher than > 等需要统计的复杂问题。ORAKEL[10]和 Pythia[11]是基于本体理论的语义网自动问答系统,本体的表达方法可以用于推理并解决复杂的语义问题。这类系统需要构建领域内的知识库词典,而不需进行实体的链接。虽然有较高的准确性,但人工构建的覆盖率和代价都过高。也有系统提出使用传统的语法解析方法,通过依存句法分析来进行初步的语义块提取。这类方法回答问题的准确度可以保证,但对于口语类型的短文本,单纯使用依存句法分析的结果,效果并不理想。TBSL[3]提出了基于模板的自动问答方法,是目前效果较好的方法,但生成的模板固定化,为了能够覆盖全部可能的问题,TBSL 往往会生成过多的候选项,使得系统性能下降。

目前大部分性能优秀的系统和研究都基于英文,因此在中文方面存在很多挑战:① 口语表达多样化,用户的表达往往无法在知识库中进行识别。②不符合语法,对于语法复杂的问句进行依存关系分析时存在大量语义提取错误的问题。③领域特殊性,例如,实体名称可能包含品牌型号等中英文混杂情况,如果用通用分词软件无法做到正确的实体识别。

本文在 TBSL 算法的基础上,针对中文特定领域内的知识库进行优化,提出了一套流式的中文知识图谱自动问答系统 CEQA,能够较好地完成商品咨询以及统计推理等复杂问题。针对商品名称特征,提出了混合词典的 CRF 方法,对该领域特殊实体识别有较好的效果;针对依存分析对于复杂问句三元组提取存在噪声的问题,本文在哈工大 LTP 语义依存分析 SDP(Semantic Dependency Parsing)[12]的基础上,提出了从三元组类别识别,到 SDP 依赖缩减,语义槽提取等一套算法框架,提高了语义三元组提取的准确率;为了解决自然语言翻译成 SPARQL 查询中自然语言多样性表达的问题,本文提出利用 Word2Vec[13]进行词与词直接的语义相似性计算,不需要标注大量数据,在电商领域的语义链接问题上取得了较好的效果。识图谱的自动问答系统已成为最新最热的研究范畴。

1 相关工作

基于知识图谱问答系统解决核心问题的方法主要有三类:基于模式的问答系统、基于统计学习的语义提取技术和基于依赖树的语义提取技术。基于模式的问答系统根据模板和规则最早的系统采用了基于模式匹配的语义提取方法,找到符合规则的问句,利用制定好的模板进行转换。如:找到一句话中含有(首都,国家)这一对关键词,则认为该句的问题是询问国家的首都。TBSL 系统第一步根据依赖关系、词性关系等生成基本的三元组,继而采用构建 SPARQL 解析器来生成查询模板。使用更多的信息提取三元组的准确率要高于直接使用依赖关系来构建查询。基于统计学习的语义提取技术主要是机器学习的思路,直接针对这种图结构与关系数据进行学习,包括 ILP 归纳逻辑编程和 SRL 统计关系学习[14]以及最近的一些研究,如:利用 SVM 进行语义在线学习[14-16]。推理一直是使用语义网的焦点,基于统计的方法虽然可以一定程度使用语义网的资源进行计算,但也会失去语义网结构中最重要的本体以及支持推理的特性。由于语义网结构数据大量涌现,在很多情况下基于统计的机器学习技术非常有效,大量的自动问答系统都应用了基于统计的基本思想。基于依赖树的语义提取技术,利用语法树进行语义提取非常符合语义网本身的链接结构,很多方法都依赖于一定的语法解析器。

另一些系统如 FREyA[17],在 QuestID[18]的基础上加入了用户模型,利用用户反馈信息提升领域词典映射的准确度。而 RTV[19]混合了一般基于字典的方法和统计机器学习的方法,将隐马尔科夫模型加入三元组映射中,相似的系统还有 Ngonga[20]。这些系统虽然在模型上有一定的优化,但都是针对英语系的知识库和语法规律进行。中文领域也有一些基于语义网的研究,最早在文献[21]的研究中提出了基于本体的自动

问答算法,回答了几种特殊的问题,但模板适用性有一定限制。最新的中文领域的文章[22]对问题进行了分类和细致的处理,但需要大量的问题库。本文在已有研究成果的基础上,提出了面向电商领域的中文知识图谱问答系统(CEQA)。

2　系统架构

2.1　系统结构

CE-QA 方法是一套针对特定领域的算法框架,重点解决将中文自然语言转换为 SPARQL 查询的问题。本文特别针对电商领域进行了实验,在准确率和算法运行效率方面与其他方法进行了对比,取得了较好的效果。整体算法框架如图 1 所示。

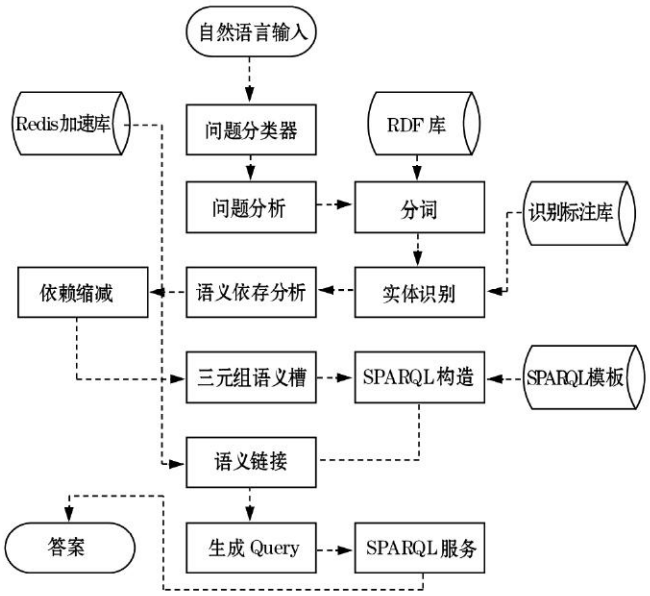

图 1　CEQA 系统结构

(1) 自然语言问题输入:输入电商领域与商品查询相关的问题,例如,夏普支持翻盖的手机有哪些?

(2) 问题分类:对于输入的自然语言,进行问题的分类。本文采用基于 SVM 算法分类。

(3) 问题分析:主要完成分词、词性标注、实体识别和实体消歧工作。本文基于 LTP 的分词包之后,如,诺基亚 8200 被切分成 < 诺基亚,8200 >, 斯黛尔塑颜腮红被切分成 < 斯黛尔,塑颜,腮红 >,另外,苹果在电商领域中为品牌词,而不是水果。所以需要针对电商领域的数据库构建词典并训练其特定的实体识别器。在得到分词序列和体序列之后,本文依据 SDP 的初步依赖结果进行缩减,提出了 SDP-Reduce 的方法,缩减了复杂的依赖关系。

(4) 语义槽提取:语义槽是代表自然语言的三元组集合,是表达问句语义的基本组成,其中的槽代表待链接的自然语言描述,由 3 个部分构成:一个变量、一个可能的 URL(类别:class,属性:property,实体:resource)、语义块(词或词组)。本模块主要完成类型判别,例如夏普 = resource,翻盖 = property,手机 = class,以及变量提取,< ? x,resource,夏普 > < ? y,property 翻盖 >, < ? z,class,手机 >。本文提出了粗分类的方式,先简单地将依赖缩减后的语义块分别映射到资源、属性、和类别上,这里简化 RDF 的类别仅分为 3 类,保证粗分类的准确度。

(5) SPARQL 抽取:主要完成构造 SPARQL 模板工作。例如,Select? x WHERE {? x? p? y;? x rdf:type? z}。

(6) 语义链接:主要解决语义槽中的待链接自然语言表达分别链接到 < 类别,资源,实体 > 对应的知识图谱中的 URL 上。例如,< res:夏普,resource,夏普 >, < db:翻盖,property,翻盖 >, < db:手机,type,手机 >。其中,res: 代表命名空间。http://ica.ecnu.com/resource 的缩写,后文均以缩写形式表示。

(7) SPARQL 查询生成:查询生成模块以及问题类别,以及连接完成的实体,构造标准的 SPARQL 查询。

```
PREFIX db: < http://ica.ecnu.com/spus >
PREFIX res: < http://ica.ecnu.com/resource >
SELECT DISTINCT ? x WHERE ({
  ? x ? p res:夏普.
  ? x rdf:type db:手机.
  ? x db:翻盖 ? z
}
```

2.2　问题分类器

将知识图谱中的实体概念和属性等词加入领域词库,同时初始化分词器,完成领域分词器的构建。针对百度知道抓取获得的用户问题分词后的结果,根据抓取的集进行抽取标注,共定义 8 类问题类别,见表 1 所示。

表 1　问题类别

问题标识	类别名称(问题比例)	例子
TYPE_UNKNOWN	其他问题(12%)	…
TYPE_COUNT	计数类别	红色的三星手机有多少种
TYPE_MAX	最大值问题	最贵的手机是哪一种?
TYPE_MIN	最小值问题	最薄的手机是哪一种?
TYPE_NUMERIC	数值类问题	Iphone5s 的屏幕像素是多少?
TYPE_BOOL	是否类问题(35%)	夏普 x280 是否支持翻盖?
TYPE_FACT	事实类问题	联想 Y450 的屏幕大小是多少?
TYPE_LIST	列表类问题	联想有哪些电脑?

对于输入的自然语言，首先进行问题的分类，根据问题类别的关键词（“能”、“吗”、“有”、“可以”、“哪些”、“多大”等词）构造出问句类别向量，问题分类大多是从统计学的角度进行分类。由于本文初步问题分类类别少，特征突出，所以本文基于 LibSVM[23] 进行多分类器的训练。

2.3 实体识别与消歧

传统的实体识别包括人名、机构名等命名实体识别，主流的算法是基于条件随机场（CRF）的命名实体识别算法。而电商领域的实体不同于传统的命名实体，其主要包括品牌名（BrandName）、型号名（SerialName），单品名（TrunkName），并且电商领域内的实体往往由中英文混搭、长实体等多种不同的形式构成。例如，商品标题为：＜娇韵诗（clarins）花样年华纤柔美腹霜 200 ml＞，其中单品名为（花样年华纤柔美腹霜）；或者 ＜三星 Galaxy Note4 ＞ 型号名为（Note4）。由于 CRF 没有 HMM 那样严格的独立性假设条件，因而可以容纳任意的上下文信息，特征设计灵活。针对电商领域特点，本系统在 CRF ++ [24] 的基础上混合 n-gram 特征模板混合词型和单品特征在商品标题数据训练领域内的实体识别模型。线性 CRF 主要目标函数如式（1）所示，其中 t_k 和 s_l 是特征函数，而 $\lambda_l\mu_l$ 分别是它们对应的参数。这里特征函数 t_k 和 s_l 取 0 或 1，当满足条件时为 0，不满足时为 1，对于上述特征模板实际上会转化为 0 或 1 的特征向量。部分特征模板如表 2 所示。

$$Z(x)=\sum_{y}\left(\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)\right) \quad (1)$$

表 2 类目下特征模板

问题标识	类别名称	例子
末尾是否颜色词	单品模板	谜尚（MISSHA）斯黛尔塑颜腮红 5g 橙红
与类目名有交集的产品词	单品模板	雨靴类目 -＞（产品：靴子，长靴，短靴，等）
长度大于 K_1 的连续数字	型号模板	K_1 是参数，诺基亚 8200
字母数字下划线的连续串长度大于 K_2	型号模板	K_1 是参数，三星 s4
字母数字下划线的连续串距离品牌的距离 dis	型号模板	三星 s4 dis = 1
%x [-1,0]	N - gram 模板	前一个词
%x [0,0]	N - gram 模板	本身的词
%x [1,0]	N - gram 模板	后一个词

2.4 序列词性依赖标注

本文基于哈工大 LTP 工具进行词性标注，获得标注好的词序列。中文领域不同于英文的语法结构，传统的依存句法分析关注实词和实词之间的介词关系，而针对问答，则更关心有语义关系的词。这里我们结合 LTP 的语义依存分析 SDP（Semantic Dependency Parsing），替代了传统的依存语法 DP（Dependency Parsing）。虽然 SDP 能够部分有效地提取语义相关的词汇关系，但用于特定领域的问句时存在两个问题，一是 SDP 的训练和效果依赖于语料，并不能广泛适用于特定领域；二是 SDP 的依赖过于复杂，同时针对一些较短的语句不能很好提取。本文在 SDP 的语义依存序列和领域内实体序列的基础上提出了依赖缩减算法。

1）生成基于 SDP 初始化的依赖图。如图 2 所示：每个节点表示一个词，每条边表示它们的依赖关系。在下图的例子中，Agt 表示施事关系（如我送她一束花（我 ＜ -- 送）），Feat 表示描写关系，Cont 表示客事关系，在这个句子中相应询问的实体是 ＜中兴 C580＞，而其属性是 ＜网络类型＞，三句话表述相同的含义，由于缺少领域实体的支持，在 SDP 的描述下形成复杂的依赖结构。

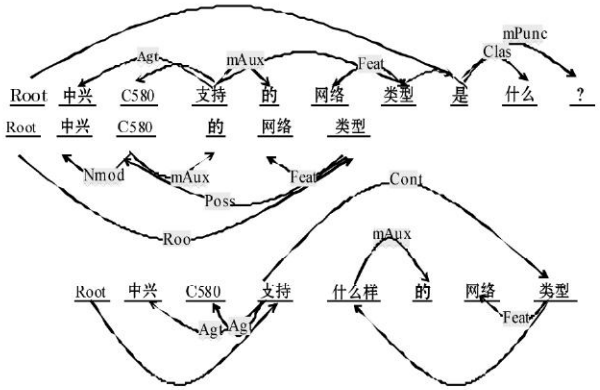

图 2 SDP 依赖初始化

2）利用 CRF 识别出的实体进行初步合并，如图 3 所示，合并品牌名和型号名为同一个产品词，并利用约简规则减少标签，我们对于每种依赖关系定义了四种基本操作：删除（OMT）、合并（MRG），反转（REV）和保持（REM）。OMT：表示删除该条关系，并分别删除两端节点词之间的链接。MRG：表示 A 词和 B 词之间的关系需要合并，合并后词保留在源节点中，并使用源节点的指向关系，如图 3 所示，要合并（类型 -＞ feat -＞ 网络）则保留类型的指向关系（网络类型 -＞ Feat - 什么样）。REV：表示反转关系。REM：表示关系保持，及保留这一条边。

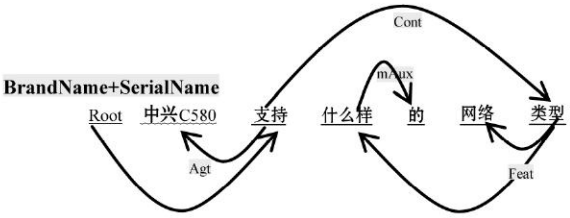

图 3　SDP 实体初步合并

我们得到带依赖序列的词序列,并定义依赖缩减规则。定义缩减操作为 F,对于任意语义依赖关系 s,一定能找到一种操作(OMT、MRG、REV、REM 中的一种)对该依赖进行操作。

F(Agt ,中兴 C580 [brandName + serialName],支持 [v]) = REM

F(mAux ,的 [u],什么样 [r]) = OMT

如图 4 所示,最后获得的实体序列为: < 中兴 c580 > 、< 支持 > 、< 网络类型 > 。基于依赖缩减的简化规则简化了复杂依赖关系,保留保护语义的实体块之间的依赖关系。

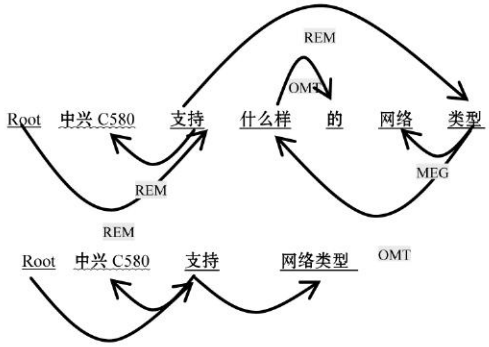

图 4　SDP 依赖缩减

2.5　SPARQL 提取

SPARQL 模板是在上述的标注序列中生成的,针对不同的问题类别采用不同的策略。

事实类问题:定义了基本的查询模板。对于类是完全正确的,使用?c 替换对应的类。使用的词汇信息(连词如“和”,关系代词如“什么”)和依存句法分析进行三元组的提取。

计数最值类问题:包含一个聚合函数问题的模板,使用了“聚合”属性,说明是否需要添加聚合语句。定义的“聚合”函数有“计数”、“过滤器”和“比较器”这三种,并定义了目标作为聚合函数的目标。例如,使用了两种类型的函数,分别用于回答计数类和比较类的问题:

1) COUNT: SELECT COUNT (DISTINCT? x) WHERE {? x? p? y.}

2) ORDER: SELECT DISTINCT? x WHERE {? x? p? y.} ORDER BY DESC(? x) OFFSET 0 LIMIT n

另外,对于一些词有特殊的功能类型。我们将这些话定义为“聚合指标”。例如,如果一个句子包含“多少”,则提取计数模板和提取技术指标。如果问题中包含一个“高级的”,我们认为这个问题需要过滤器。如果问题包含比较,推断这个问题需要一个比较器操作。我们使用“聚合指标”来检测这些类型的操作的目标和常数(如价格多于 2 000 的手机,会直接使用 2 000 和手机的依赖关系,Quan (手机,2 000)。针对不同问题类别,制定相应的语义模板,再结合之前获得实体序列获得初步的 SPARQL 表达式。

2.6　语义链接

生成可以执行的 SPARQL 后,还有复杂的链接问题需要处理。由于上文得到的函数式还包括自然语言,语义网中的表达则是以 URL 为单位的,本文提出先粗分类再利用 Word2Vec 混合词典链接的模型。

主要流程如图 5 所示:

1) 构建字典,直接从 RDF 中建立到名词短语识别资源/类 URI 的链接关系词典,并基于 Redis(一个开源的 Key-Value 存储系统)进行数据缓存。

2) 自然语言表达首先根据如下公式进行粗分类。

$$Score_{\text{class}} W_{(i)} = W_{\text{SDP}}(W_{(i)}) + W_{\text{depm}}(W_{(i)}) + W_{\text{hdpt}}(W_{(i)}) \quad (2)$$

其中 W_i 表示原问句中的一个词,$W_{\text{SDP}}(W_{(i)})$ 表示当这个词的词性得分,$W_{\text{depm}}(W_{(i)})$ 表示这个词的依赖得分,$W_{\text{hdpt}}(W_{(i)})$ 代表这个词是否在 RDF 词典中存在这个类别(存在为 1,不存在为 0)。利用 SDP 的依赖结果标注一部分数据进行训练,最终获得 W_{SDP}、W_{depm} 两个参数。

3) 利用粗分类的结果,分别对每个自然语言的表达构建候选项集合。

4) 计算相似度,本文并没有采用 WordNet 或者同义词网络,而是采用 Word2Vec 寻找相似词(找到 花花公子 = Playboy simi = 0.89)。Word2vec 是 Google 的词向量化工具,使用深度学习的思想,可以计算词与词之间的相似性,其主要假设是相似的单词拥有相似的语境。主要目标函数如下所示:单词 w 用长度为 d 的列向量表示,条件概率 $p_{(c|w)}$ 表示当 w 出现时,某一语境 c 出现的概率,θ 表示模型参数,D 表示所有单词 w 和它的语境 $C_{(w)}$ 构成的组合的集合。

$$\underset{\theta}{\operatorname{argmax}} \prod_{(w,c) \in D} p(c \mid w;\theta) \quad (3)$$

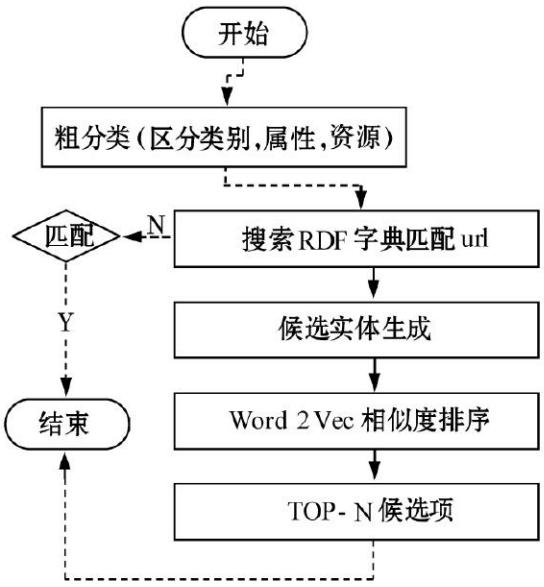

图5 语义链接

5）对于直接命中的进行构建查询，对于未命中的自然语言表达，取满足阈值 T 的 TopN 的相似词进行查询，直到有查询结果为止。

3 实验与结论

3.1 数据准备

实验主要对比该系统在实际电商领域问答数据中的问答准确度，我们参考国际标准比赛 QALD 的评测方法，利用准确率，召回率，和 F 值三个指标综合衡量系统的准确程度，计算方法见式(4)。P 表示准确率，R 表示召回率，$Pnum$ 是准确答案的数目，$ResNum$ 是对所有问题系统给出的答案数据，$AucNum$ 是实际的答案数目。

$$P=\frac{Pnum}{ResNum} \quad R=\frac{Pnum}{AucNum} \quad F=\frac{P\times R\times 2}{P+R} \qquad (4)$$

实验抓取了京东、苏宁等电商手机和电脑类的 SPU 数据整理成 RDF 资源，共有手机类(103 137 个三元组)，电脑类(508 123 个三元组)。并利用百度知道和新浪爱问的相关问题 100 个和手工整理不同类别问题共 200 个作为训练数据。

3.2 实验分析

实验主要比较了 3 种方法，最基础的是 TBSL(一个开源的问答系统，由 DBpedia 团队设计)算法的基础上加入了实体识别，CEQA-N-W2V 是 CEQA 框架下加入 Word2Vec 的实体链接的方法，后者取得了比以前更高的 F 值。此外，我们对后者是否添加依赖缩减算法进行对比，缩减图的方法取得了最高的 F 值。如表 3 所示。

表3 实验结果

算法	生成 SPARQL 数目	准确率	召回率	F 值
TBSL-NER	42	0.33	0.21	0.25
CEQA-N-W2V	56	0.35	0.28	0.31
CEQA-N-Reduce	62	0.51	0.31	0.38

在电商领域内，由于用户关注的实体词具有特殊性，构建友好的问答系统需要解决领域内的实体词识别，以及行业内的自然语言链接。从 TBSL-NER 实验中可以发现，在加入品牌名、型号名、单品名识别的基础上 F 值可以达到 0.25，而不使用实体识别的情况下无法正常解析用户问句，由此说明的 CEQA 中加入领域特征的实体识别已经初步具备回答问题的能力。进一步，我们对比了先粗分类，再利用 Word2Vec 做链接的 CEQA-N-W2V 方法，将类别、属性、资源分别进行链接，并在 Word2Vec 的基础上解决语义槽到图谱数据链接的问题。在实验中我们发现，如这样的句子“ThinkPad R4007445A46 的硬盘容量有多大?”，Thinkpad 本身并不在知识库中，而“联想”在知识库中，利用 word2vec 训练商品标题可以增加自然语言表达的丰富性。从实验中也可以看出其增加了 6% 的 F 值，并且准确率没有下降。虽然 word2vec 并不会比人工构建同义词库准确度高，但是人工构建的代价太大，在存储电商标题数据的时候，直接使用标题无监督训练出词向量更有优势。最后，加入了依赖缩减规则的 CEQA-N-Reduce 可以将一部分语义过于复杂的语句进行缩减，效果最好。在实验中我们发现，如“中兴 C580 支持什么样的网络类型”，“中兴 C580 的网络类型”均可以进行有效解析答案，实验结果符合预期缩减 SDP 提取三元组的设想，其最终增加了 7% 的 F 值。并由于依赖缩减，使得准确率有了 17% 的提升，说明在 SDP 基础上的依赖缩减对于电商领域的问答不仅可以回答更多的问题，而且更加准确。

现有实验说明 CEQA 的整套框架可以有效地在电商领域的知识图谱数据上提供问答服务，整套框架中各个模块都可以持续优化，而实际上链接算法、分类算法针对其他领域的问题方便设计特征进行替换，算法的可移植性也很好。

4 结 语

本文研发了基于中文知识图谱的电商领域自动问答系统，利用语义依存分析等自然语言处理技术，提出

缩减依赖算法提高问题的识别率,提取相应的语义槽,构建 SPARQL 查询。先进行粗分类,再结合 Word2Vec 完成了自然语言的链接,提高了 URL 的匹配的覆盖率。另外,我们结合特定领域的特征在实体识别部分加入了特定的实体识别,使得进一步使用 LTP 变为可能。然而本文提出的系统仍然有局限性,制定规则来确定标签是一项艰巨的任务,比如对于这样的问题:“给我所有的手机与手机的颜色。”这样的句子规则难以提取,同时 LTP 的精确度也有很大的影响。在未来的工作中,我们将重点放在 LTP 缩减的问题上,目前的缩减规则准确但是覆盖率不够,下一步将使用更多的机器学习算法提取。此外,我们还将研究在答案存在多个或没有答案时的推荐式展现策略。

参 考 文 献

[1] Lopez V, Unger C, Cimiano P, et al. Evaluating question answering over linked data [J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2013, 21: 3-13.

[2] Popescu A M, Etzioni O, Kautz H. Towards a theory of natural language interfaces to databases [C] //Proceedings of the 8th International Conference on Intelligent User Interfaces. ACM, 2003: 149-157.

[3] Unger C, Bühmann L, Lehmann J, et al. Template-based question answering over RDF data [C] //Proceedings of the 21st International Conference on World Wide Web. Lyon, France: ACM, 2012: 639-648.

[4] Yahya M, Berberich K, Elbassuoni S, et al. Natural language questions for the web of data [C] //Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2012: 379-390.

[5] 郭磊. 基于领域本体中文自动问答系统相关技术的研究与实现 [D]. 上海:华东理工大学, 2013.

[6] Berant J, Chou A, Frostig R, et al. Semantic parsing on freebase from question-answer pairs [C] //Proceedings of EMNLP, 2013: 1533-1544.

[7] Berners-Lee T, Hendler J, Lassila O. The Semantic Web [J]. Scientific American, 2001, 284(5): 35-43.

[8] Fazzinga B, Lukasiewicz T. Semantic Search on the Web [J]. Semantic Web, 2010, 1(1-2): 89-96.

[9] Lopez V, Pasin M, Motta E. AquaLog: An ontology-portable question answering system for the semantic web [C] //Second European Semantic Web Conference, 2005: 546-562.

[10] Cimiano P, Haase P, Heizmann J. Porting natural language interfaces between domains: an experimental user study with the ORAKEL system [C] //Proceedings of the 12th International Conference on Intelligent User Interfaces. ACM, 2007: 180-189.

[11] Unger C, Cimiano P. Pythia: Compositional meaning construction for ontology-based question answering on the semantic web [C] //16th International Conference on Applications of Natural Language to Information Systems, 2011: 153-160.

[12] Che W, Li Z, Liu T. LTP: A Chinese Language Technology Platform [C] //Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations. ACM, 2010: 13-16.

[13] Goldberg Y, Levy O. word2vec Explained: deriving Mikolov et al. 's negative-sampling word-embedding method [DB]. arXiv preprint arXiv: 1402.3722.

[14] Muggleton S, Raedt L D. Inductive Logic Programming: Theory and Methods [J]. The Journal of Logic Programming, 1994, 19-20(S1): 629-679.

[15] Bordes A, Glorot X, Weston J, et al. A semantic matching energy function for learning with multi-relational data [J]. Machine Learning, 2014, 94(2): 233-259.

[16] Fanizzi N, d'Amato C, Esposito F. Learning with kernels in description logics [C] //18th International Conference on Inductive Logic Programming, 2008: 210-225.

[17] Damljanovic D, Agatonovic M, Cunningham H. FREyA: an interactive way of querying linked data using natural language [C] //Proceedings of the 1st Workshop on Question Answering over Linked Data lab (QALD-1), 2011: 125-138.

[18] Damljanovic D, Tablan V, Bontcheva K. A text-based query interface to OWL ontologies [C] //Proceedings of the 6th Language Resources and Evaluation Conference (LREC), 2008: 205-212.

[19] Giannone C, Bellomaria V, Basili R. A HMM-based approach to question answering against linked data [C] //Proceedings of the 3rd Workshop on Question Answering over Linked Data lab (QALD-3) at CLEF, 2013: 1-13.

[20] Shekarpour S, Ngomo A C N, Auer S. Question answering on interlinked data [C] //Proceedings of the 22nd International Conference on World Wide Web (WWW). ACM, 2013: 1145-1156.

[21] 何海芸, 袁春风. 基于 Ontology 的领域知识构建技术综述 [J]. 计算机应用研究, 2005, 22(3): 14-18.

[22] 洪韵佳, 许鑫. 联合虚拟参考咨询系统知识库的发展现状与趋势 [J]. 现代图书情报技术, 2012(9): 2-9.

[23] 张巍, 陈俊杰. 信息熵方法及在中文问题分类中的应用 [J]. 计算机工程与应用, 2013, 49(10): 129-131,179.

[24] 唐旭日, 陈小荷, 张雪英. 中文文本的地名解析方法研究 [J]. 武汉大学学报信息科学版, 2010, 35(8): 930-935,982.