文章编号：1003-0077(2020)11-0104-09

基于知识图谱的在线商品问答研究

王思宇[1]，邱江涛[1]，洪川洋[1]，江 岭[2]

(1. 西南财经大学 经济信息工程学院，四川 成都 611130；
2. 成都晓多科技有限公司，四川 成都 610041)

摘 要：现阶段，针对商品的自动问答主要由意图识别和答案配置来实现，但问题答案的配置依赖人工且工作量巨大，容易造成答案质量不高。随着知识图谱技术的出现和发展，基于知识图谱的自动问答逐渐成为研究热点。目前，基于知识图谱的商品自动问答主要是通过规则解析的方法将文本形式问题解析为知识图谱查询语句来实现。虽然减少了人工配置工作，但其问答效果受限于规则的质量和数量，很难达到理想的效果。针对上述问题，该文提出一种基于知识图谱和规则推理的在线商品自动问答系统。主要贡献包括：①构建一个基于 LSTM 的属性注意力网络 SiameseATT(Siamese attention network)用于属性选择；②引入了本体推理规则，通过规则推理使得知识图谱能动态生成大量三元组，使得同样数据下可以回答更多问题。在 NIPCC-ICCPOL 2016 KBQA 数据集上的实验显示，该系统具有很好的性能。相比一些更复杂的模型，该问答系统更适合电商的应用场景。

关键词：问答系统；知识图谱；注意力机制；规则推理

中图分类号：TP391 **文献标识码**：A

Online Commodity KBQA Based on Knowledge Graph

WANG Siyu[1], QIU Jiangtao[1], HONG Chuanyang[1], JIANG Ling[2]

(1. School of Information Engineering, Southwestern University of Finance and Economics, Chengdu, Sichuan 611130, China;
2. Chengdu XiaoDuo Technology Co. Ltd., Chengdu, Sichuan 610041)

Abstract: In general, Question Answering System (QAS) for the commodity is mainly built via the intention identification and answer configuration. However, the configuration of answers of questions depends on manual labor, which easily results in poor quality of answers. With the introduction and development of Knowledge Graph (KG) technology, the KG-based QAS has gradually become a hot research topic. At present, the KG-based QAS for commodity is mainly implemented by employing rules to transform questions to queries in the KG. Although the manual configuration work is reduced, the performance of QAS is limited by the quality and quantity of the rules. In order to solve above problems, this paper proposes a question answering method for online commodities based on KG and rule reasoning. The main contributions include: (1) we built an LSTM-based property attention network named SiameseATT(Siamese Attention Network) for attribute selection; (2) we employed KG to infer rules, consequently generate a large number of triples to respond more questions. Finally, experiments on the NLPCC-ICCPOL 2016 dataset show that the model obtains good performance. Our QAS is more suitable for e-commerce applications.

Keywords: question answering system; knowledge graph; attention mechanism; rule-based reasoning

0 引言

近年来，随着互联网的发展，网上购物逐渐成为一种非常受欢迎的消费方式。网络购物时客户常常会对商品提出一系列的咨询问题，如“这款电视机带蓝牙功能吗？”“这款电视机是什么操作系统？”等等。由人工回答这些重复的问题非常耗时耗力，且客服

收稿日期：2019-12-02 **定稿日期**：2020-02-26

基金项目：国家自然科学基金(71571145)

人员知识的缺乏或者疏忽也经常会导致这些问题的错误回答。鉴于此，商家希望自动问答系统能够减少工作成本，提高问题的回答准确率。因此在线商品的自动问答系统具有很高的应用价值。

目前的在线商品自动问答主要通过意图识别和人工答案配置来实现。意图识别是一个分类问题，可以训练一个文本分类器来识别“问题”所表达的意图，再通过为每个意图配置固定答案来实现自动问答。这种自动问答方法较为灵活多样，可定制性强，但传统方法中，问题答案的配置依赖于人工且工作量巨大，因此，导致其提供的答案质量不高。

近年来，随着知识图谱技术的发展，基于知识图谱的智能问答成为自动问答中一项重要且充满挑战的研究工作。其核心是识别问题中的实体和属性，在知识图谱上检索实体属性对应的属性值，然后将它们作为答案返回给用户。由于知识图谱可以从结构化、半结构化信息和文本中自动构建，因此可大大减少答案配置的时间。

目前对基于知识图谱的问答系统的研究主要包括：①通过语法解析将问题转换为知识图谱相应的查询语句[1]，实现基于知识图谱的在线商品问答。其优点在于不用配置答案且具有很高的回答准确率，但其缺点为他们通过规则将用户问题转换为图谱查询语句，因此问答效果又受限于规则的质量和数量。②通过深度神经网络模型进行属性选择来实现知识图谱的问答[2]。其优点是，即使问题呈现多样性，深度神经网络也可以识别，且不会受到规则的限制，但其缺点是模型的性能仍有待提高。

基于知识图谱的问答系统的另一个问题是，候选属性是与实体直接相连的一阶属性。但许多时候用户的问题并非只是针对这种直接相连的一阶属性，我们更希望能够通过推理得到更高阶的属性。以图 1 为例，提交给系统一个问题“这个电视可以安装软件吗?”。因包含实体“TCL60F60 数字电视”的三元组中，不存在关系“安装软件”，故问答系统对这样的问题无法解答。

针对上述问题，本文提出一种基于知识图谱和规则推理的在线商品自动问答系统，其主要贡献包括：①开发了一个基于长短时记忆网络(long short-term memory，LSTM)的孪生属性注意力网络，称为 SiameseATT(Siamese attention network)。该网络通过共享权重的 LSTM 分别提取问题与属性

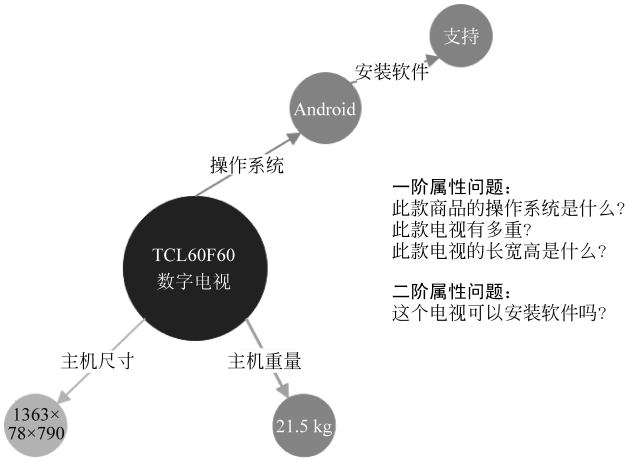

图 1　知识图谱问答示意图

的特征，在提取问题特征时引入了属性注意力机制。②为了提高知识图谱回答的覆盖率、复杂性和灵活性，系统采用基于 OWL 规则推理的方法，在知识图谱上推理得到更多的知识，提高了问答系统的解答能力。

本文在 NLPCC-ICCPOL 2016 问答数据集上的实验证明：①基于孪生属性注意力网络的问答系统能比多数模型展现出更好的性能；②通过规则推理使得问答系统可以获得更高的 F_1 值。相比其他系统，我们的问答系统更适合电子商务应用场景。最后本文构建了一个原型系统，在原型系统中展现了规则推理和知识图谱三元组的动态生成，以及基于推理的在线商品问答系统。

1　相关工作

在线商品自动问答主要有基于知识图谱和答案配置两种方法。答案配置的方法，主要是通过意图识别和配置答案实现。用户意图识别是指自动问答系统能够根据用户提问的直接或者间接的信息来判断用户的真实意图，主要包括基于规则的判断[3]和基于机器学习方法[4]的识别。意图识别可以看作是一个分类问题。

知识图谱以结构化和关联化的方式存储知识，可以作为问答系统高质量的数据源。曹明宇等[5]利用医疗领域数据构建知识图谱，在此基础上实现流水线式的问答系统。过程是，先识别问题中的实体，再结合 TF-IDF 和词向量生成句子向量，匹配最相似的问题模板，根据模板的语义及问题中的实体到

知识图谱中检索答案。该方法实现了基于知识图谱的智能问答在医疗领域的应用。在电商领域,杜泽宇等[6]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等问题,提出一套流式的中文知识图谱自动问答系统。目前知识图谱在各个领域的应用层出不穷,基于知识图谱的问答系统的研究也在不断发展。

当今主流的基于知识图谱的自动问答系统采用的方法可以分为两类:基于语义分析的方法和基于信息检索的方法。例如,Berant 等[7]基于语义分析的方法首先将自然语言形式的问题转换为某种特定类型的逻辑表达形式;Zettlemoyer 等[7]提出了 Lambda 表达式,这类逻辑表达形式通常也适用于在知识库中进行查询,进而找出问题的答案。语义分析的方法简单且不需要训练集,但由于其使用规则模板,因此在应用上存在一定的限制。

基于信息检索的方法首先从知识库中检索一系列候选答案,然后对问题和候选答案进行特征抽取,通过计算它们之间的匹配评分来选择最终答案。该方法的核心在于抽取有效特征。Yao 等[8]提出了使用问题特征与 Freebase 知识图谱中的特征相结合的信息抽取方法,在 QA 语料库上获得了当时最好的结果。近年来随着人工智能的发展,深度神经网络的方法也开始应用于信息检索,并取得了相比于语义分析更好的实验结果。Bordes 等[9]提出了一个基于知识库 Freebase 的方法,根据问题中的主题词在知识库中确定候选答案;同时构建一个模型来学习问题和候选答案的词嵌入;最后通过这些词嵌入来计算问题和候选答案的相关度,据此选出正确答案。在不使用词表、规则、句法和依存树解析等条件下,该方法超越了当时最好的结果。紧接着 Dong 等[10]使用了卷积神经网络的一种变体从答案路径(answer path)、答案上下文信息(answer context)、答案类型(answer type)对问题和答案的分布式表示进行学习,使得该分布式表示相比之前的向量建模方法能够包含更多的有效特征。Hao 等[11]针对前人工作中存在没有充分考虑候选答案的相关信息的问题,提出了一个使用交叉注意力机制的神经网络模型来针对候选答案的不同方面信息。Qu 等[12]提出了基于相似矩阵的卷积神经网络,利用 RNN 和 CNN 的优势捕捉全局的层次信息。通过 RNN 的顺序建模性质来捕捉语义级别的关联,同时加入注意力机制跟踪实体和关系。再利用 CNN 建模数据之间的空间相关性,提取文字级别的语义信息。最后再汇总两种类型的语义信息。

在中文领域,随着 2016 年 NLPCC-ICCPOL KBQA 任务提供的知识库的发布,中文问答系统的研究取得很大的发展。Lai 等[13]通过简单的词向量相似度运算,结合细粒度的分词进行属性映射,同时结合多种人工构造的规则和特征,取得了该任务当年最好的效果。周博通等[14]利用别名词典获取候选实体,并通过 LSTM 语言模型结合简单的文本特征进行打分,进行命名实体识别。然后结合两种不同的注意力机制使用双向 LSTM 模型进行属性映射,最后综合前两步的结果进行实体消歧和答案选择,取得了较好的效果。最近,Liu 等[15]将预训练的 BERT 用于知识图谱问答,在该数据集上获得了目前最好的成绩。

2 基于孪生注意力网络的中文知识图谱问答系统

知识图谱是一个<实体,关系,实体>的三元组的集合。在本文的电商知识图谱中,将属性视为"关系",因此电商知识图谱中拥有大量<实体,属性,属性值>的三元组。例如,图 1 中的一个三元组<TCL60F60 数字电视,操作系统,Android>。

本文开发了一个知识图谱问答系统。该系统的工作流程如图 2 所示,主要包括以下几个步骤:①主题实体获取:从用户的问题中获得命名实体,将命名实体链接到知识图谱中获得主题实体 e。所谓主题实体是问句中的询问对象(命名实体)所对应的知识图谱中的实体。图 2 中,问题"TCL60F60 这款有多大?"中的主题实体为"TCL60F60";②检索候选属性:在知识图谱上检索包含实体 e 的三元组,得到候选属性集合 A;③属性选择:根据问题 q(一段文本)从候选属性集合中挑选属性是一个排序问题。问题 q 和一个属性 $p \in A$ 通过共享权重双向 LSTM 网络编码后分别得到两个向量,计算两个向量的余弦相似度作为语义相似评分,选择评分最高的属性 $p_{\max}$。④生成答案:在知识图谱上检索三元组<e,$p_{\max}$,v>,返回三元组包含的属性值 v 作为问题的答案。

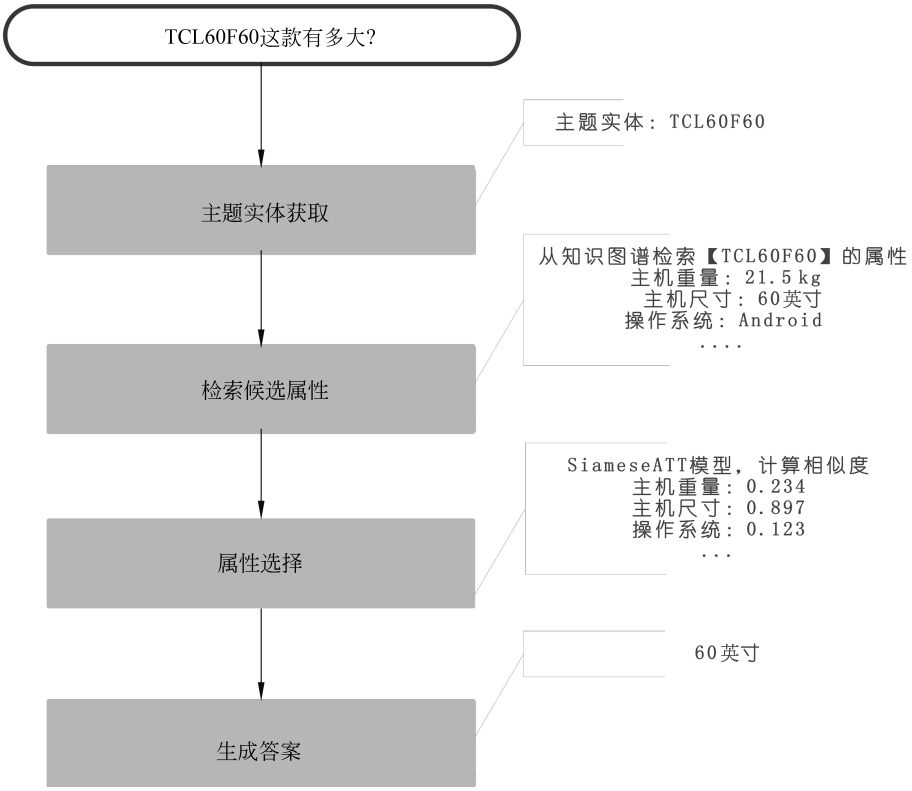

图 2　基于知识图谱的在线商品自动问答流程

知识图谱问答的第一步是识别问题中的主题实体，该任务与命名实体识别任务类似，因此本文采用由 Bi-LSTM 和 CRF 模型组成的 Bi-LSTM-CRF 模型[16]来获得命名实体。接着，将命名实体链接到知识图谱获得主题实体。

本文的工作重点在于“属性选择”。首先开发了 SiameseATT 模型来进行候选属性选择，详细内容见 2.1 节。另外，我们依据事先定义的规则，或者由专家制定的生成规则对知识图谱进行推理，进而得到满足条件实体的更多属性。详细内容见 2.2 节。

2.1　基于 SiameseATT 网络的属性选择方法

从属性集合 A 中挑选一个当前问题涉及的属性称为属性选择。属性选择的难点在于问题中属性描述可能与知识图谱中的属性名称存在较大的差异，例如问题“这款电视的长宽高是多少”。该问题涉及了属性“商品主机尺寸”，但没有直接询问，而是用同义的“长宽高”代替。因此，如何设计一个模型有效地将问题与属性进行匹配是本文的研究重点之一。为了更好地进行匹配，我们设计了孪生属性注意力网络(SiameseATT)。该网络借鉴了人脸识别中两张人脸相似度的计算的 Siamese 网络的思想，通过两个共享权重的双向 LSTM 网络分别对问题 q 和属性 p 进行编码。在对问题 q 进行编码时，我们引入了基于属性 p 的注意力机制。编码后得到问题和属性的语义向量 $\boldsymbol{S}_p$ 和 $\boldsymbol{S}_q$，随后该模型计算问题 q 与所有属性 $p \in A$ 的余弦相似度评分，最后挑选出评分最高的属性。

Siamese 属性选择网络架构图参见图 3。具体工作步骤如下：

(1) 属性编码。将属性看作是长度为 m 的字的序列，使用预先训练的字向量得到属性的字向量序列 $p=(\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_m)$。将 p 送入一个双向 LSTM 网络。双向 LSTM 模型的时间步 t 有两个隐状态，如式(1)所示。

$$\begin{aligned}\overleftarrow{h_t}&=\overleftarrow{\mathrm{LSTM}}(\boldsymbol{c}_t,\overleftarrow{h_{t-1}})\\ \overrightarrow{h_t}&=\overrightarrow{\mathrm{LSTM}}(\boldsymbol{c}_t,\overrightarrow{h_{t-1}})\end{aligned}\tag{1}$$

其中，$c_t \in p$ 是一个字向量。设$\overleftarrow{h_t}$和$\overrightarrow{h_t}$的向量长度为 $u/2$，将两个隐状态拼接可以得到时间步 t 的输出 $\boldsymbol{h}_t$，$\boldsymbol{h}_t$ 是一个长度为 u 的向量。将 m 个时间步进行平均池化操作得到向量 $\boldsymbol{S}_p$，即属性编码得到的向量。

(2) 基于属性注意力的问题编码。我们将输入的问题也看作是长度为 n 的字序列。使用预训练的中文字向量对输入进行表示，得到一个问题的字向量序列 $\boldsymbol{q}=(\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_n)$。将 q 送入一个双向 LSTM 模型，每个时间步产生两个隐状态。与上面属

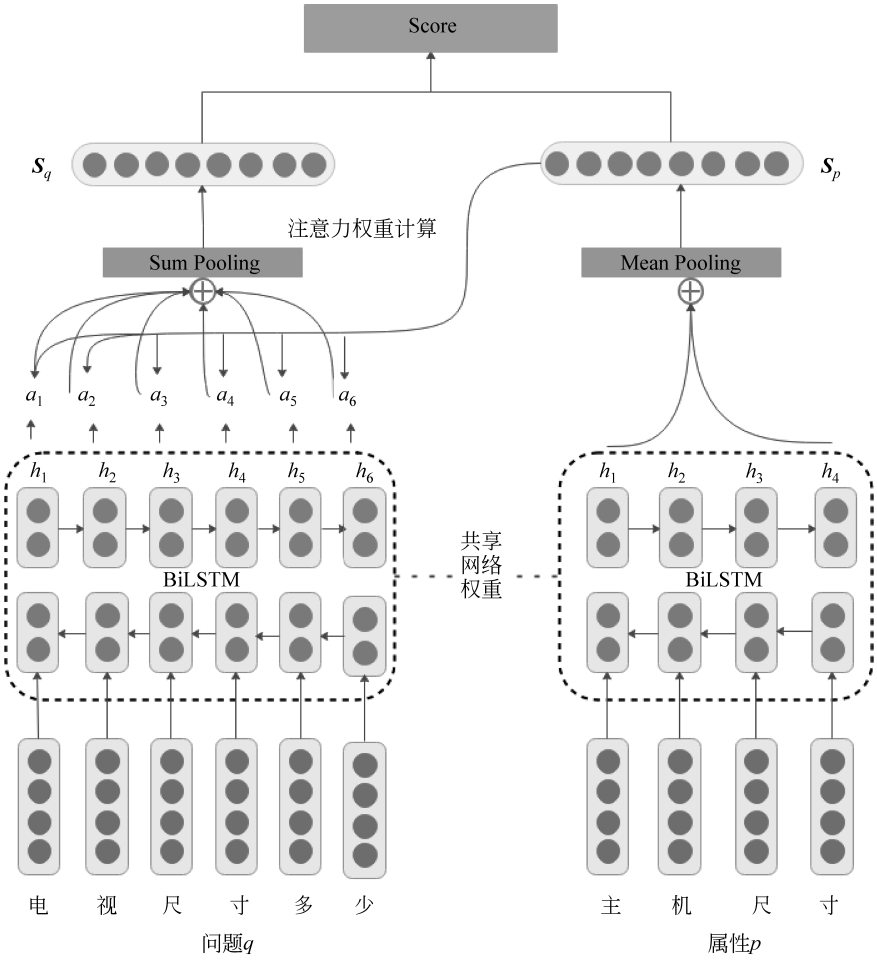

图 3 Siamese 属性选择网络架构图

性编码的步骤相同,我们把得到的两个隐状态拼接为一个长度为 u 的向量。设 $\boldsymbol{h}_t$ 是时间步 t 的最后输出。在对问题编码时,我们设计了一个注意力机制,它根据属性编码得到的向量 $\boldsymbol{S}_p$ 和问题的时间步 t 的输出 $\boldsymbol{h}_i$ 来计算权重值 α_i。见式(3)和式(4)。双向 LSTM 每个时间步 h_i 与权重 α_i 的乘积加权求和得到向量 $\boldsymbol{S}_q$,即问题编码得到的向量,如式(2)所示。

$$\boldsymbol{S}_q = \sum_{i=1}^{n} a_i h_i \tag{2}$$

$$\alpha_i = \frac{\exp(e_i)}{\sum_{k=1}^{n} \exp(e_k)} \tag{3}$$

$$e_i = \boldsymbol{h}_i^{\mathrm{T}} \boldsymbol{S}_p \tag{4}$$

(3) 相似度计算。在得到属性编码向量 $\boldsymbol{S}_p$ 和问题编码向量 $\boldsymbol{S}_q$ 后,我们需要计算这两个向量的相似度,以衡量问题是否涉及此属性。相似度的计算采用余弦相似度,如式(5)所示。

$$S(\boldsymbol{S}_q, \boldsymbol{S}_p) = \frac{\boldsymbol{S}_q \cdot \boldsymbol{S}_p}{|\boldsymbol{S}_q||\boldsymbol{S}_p|} \tag{5}$$

(4) 模型训练。我们设计了一个合页损失函数(Hinge Loss),如式(6)所示。

$$L(\boldsymbol{S}_q, \boldsymbol{S}_p, \overline{\boldsymbol{S}_p}) = \max(0, [\gamma - S(\boldsymbol{S}_q, \boldsymbol{S}_p) + S(\boldsymbol{S}_q, \overline{\boldsymbol{S}_p})]) \tag{6}$$

其中,$S(\boldsymbol{S}_q, \boldsymbol{S}_p)$,$S(\boldsymbol{S}_q, \overline{\boldsymbol{S}_p})$表示问题和正例属性、问题和负例属性分别计算的得分。我们采用负采样的方法从属性集合中随机选择除正确属性外的其他属性建立负例集合。其中超参数 γ 是一个正的实数值。被编码后的属性和问题在计算相似度时,正例和负例之间应该有个间隔,γ 设置了这个间隔。

2.2 规则推理

通过属性选择模型(SiameseATT)已经可以将问题中的描述正确地链接到知识图谱中的主题实体属性上了。如图 1 所示的问题“此款商品的操作系统是什么?”,其主题实体候选属性包括“主机重量”“主机尺寸”“操作系统”。SiameseATT 计算问题分别对 3 个属性的相似度评分,选择评分最高的属性“操作系统”的值作为答案。但对于问题 “这个电视可以安装

软件吗?”主题实体没有属性描述“是否可以安装软件”,因此不能正确回答这样的问题。这就是我们前面提及的“当前的模型可以回答一阶属性的问题,但不能回答高阶属性的问题”的情况。为了解决该问题,本文通过规则推理让知识图谱自动生成属性来扩充候选属性集合。此方法充分利用了知识图谱的推理功能,动态生成三元组,使得问答系统在不更新知识图谱数据的情况下回答更高阶的问题。推理过程如下。

每一个存在于知识图谱的三元组都可以表示为 $r(h,t)$,其中 h,t 分别代表知识图谱中的两个实体,r 代表两个实体的关系。逻辑规则形式如式(7)所示,如果 x_1,x_2 存在 r_1 关系并且 x_2,x_3 存在 r_2 关系,那么 x_1,x_3 就一定存在 r_3 关系。

$$r_1(x_1,x_2) \wedge r_2(x_2,x_3) \rightarrow r_3(x_1,x_3) \quad (7)$$

如图 1 所示的例子,我们给定一个规则:如果一个实体 x 的操作系统能够安装软件则此实体 x 可以安装软件。式(8)描述了规则的逻辑形式。

操作系统(x,linux) ∧ 安装软件(linux,true)
→ 安装软件(x,true)　　(8)

将此规则输入推理机中进行推理,知识图谱会生成新的关系(属性),当再次查询实体 x 时,“安装软件”属性会被添加到候选属性集合。

3　实验

3.1　数据准备

本文实验使用了 NLPCC-ICCPOL 2016 KBQA 数据集(表 1)①。该数据集是目前最大的公开中文知识图谱问答数据集,包含大约 4 300 万个三元组和 600 万个实体[17]。该知识图谱的三元组大部分来自百度百科,训练数据示例如表 2 所示。

表 1　NLPCC-ICCPOL 2016 KBQA 数据集描述

数据集	训练数据	测试数据
NLPCC-ICCPOL 2016 KBQA	14 609	9 870

表 2　NLPCC-ICCPOL 2016 KBQA 数据集问题(属性示例)

属性	问题
作者	《机械设计基础》这本书的作者是谁
国籍	安德烈是哪个国家的人呢
连载网站	大家知道《夏想》是在哪个网站连载的吗
总部地点	万达的总部在哪

3.2　实验设置

SiameseATT 模型的设置见表 3。其中预训练字向量采用 Li 等[18]所提供的百度百科数据集训练字向量②。实验使用 Ubuntu 18.04 操作系统和 GTX 1080ti 显卡用于训练,采用 Pytorch 1.1.0 实现模型。

表 3　实验参数设置

参数名	属性选择(SiameseATT)
LSTM 隐层维度	300
预训练字向量维度	300
学习率	0.0001
γ	0.2
Batch size	16
优化器	Adam
训练 Epoch	20
负采样	1∶20

3.3　实验结果

在自动问答的实验中,本文和以下几种基准方法进行了比较:

(1) **CCNU**[19]:将基于卷积神经网络和循环神经网络的 DSSM(deep structured semantic models)用于问题与候选三元组匹配。

(2) **UCAS**[20]:提出一种多粒度特征表示的属性选择模型。该模型采用字符级别以及词级别分别对问句和属性进行嵌入表示并通过编码器进行编码,用于属性选择。

(3) **BiLstm_AC12_Overlap**[14]:首先采用别名词典结合 LSTM 语言模型进行命名实体识别,然后使用双向 LSTM 模型结合两种不同的注意力机制进行属性映射,最后综合前两步的结果进行实体消歧和答案选择。

(4) **NUDT**[21]:提出一种混合深度学习模型和信息抽取的方法用于问答系统。

(5) **CCNU-TE**[22]:提出一种主题强化(topic enhanced)的 DSSM 用于基于知识图谱的问答系统。

(6) **PKU**[13]:采用 SPE(subject predicate extraction)算法能自动抽取问句中的主题和属性,并

① http://tcci.ccf.org.cn/conference/2016/pages/page05_evadata.html

② https://github.com/Embedding/Chinese-Word-Vectors

将之翻译为知识图谱查询语句。

(7) **BUPT**[15]：将预训练的 BERT 用于知识图谱问答，在 NLPCC-ICCPOL 2016 数据集上获得了目前为止最好的成绩。

最后各方法在 NLPCC-ICCPOL 2016 KBQA 自动问答的评测结果如表 4 所示。SiameseATT 模型在只使用 LSTM 和属性注意力的情况下获得了 81.81% F_1 值，略高于模型 CCNU、UCAS、BiLSTM_AC12_Overlap 和 NUDT，另外从实验中也可以看到加入属性注意力后问答系统最终 F_1 值有 0.25% 的提升。但我们的模型的 F_1 值与 CCNU-TE、PKU、BUPT 相比略低，这些模型都使用较为复杂的网络或人工规则获得了较好的效果。SiameseATT 主要用于商品问答，在商品问答中每个店铺所出售的商品不同且比较单一，因此每个店铺都需要一个单独的模型，如果结合规则来进行问答，构建规则的工作量巨大。同样地，由于店铺规模较小并且商品单一，每个店铺也没有大量数据来训练较为复杂的模型，因此相比 CCNU、PKU、BUPT 等模型，SiameseATT 更适合电商领域的商品问答。

表 4 NLPCC 自动问答评测 F_1 值与其他模型对比

方法	F_1 值/%
NLPCC Baseline	52.48
Naive Search	47.39
UCAS	73.96
CCNU	79.57
BiLSTM_AC12_Overlap	81.06
NUDT	81.59
Siamese	81.56
SiameseATT	81.81
CCNU-TE	82.43
PKU	82.47
BUPT	84.12

4 原型系统

本研究构建了一个某品牌数字电视的知识图谱和一个基于该知识图谱的自动问答原型系统①。图 4 是该问答系统的界面。

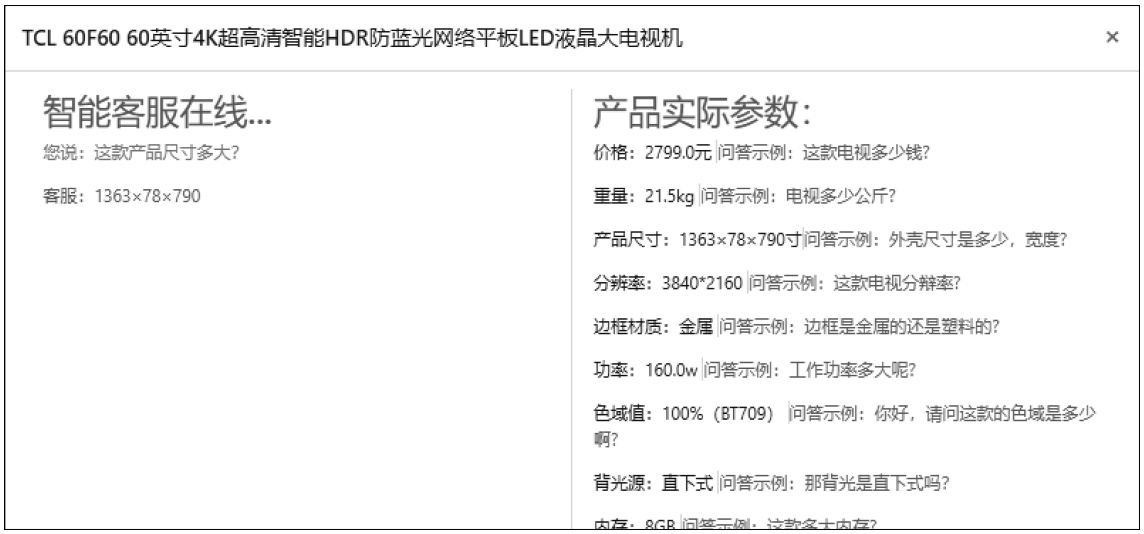

图 4 商品问答系统的界面

参考某电商平台在线问答页面，通过人工标记问题及其涉及的属性，构建了一个数据集。在训练属性选择模型(SiameseATT)时，采用留出法(hold-out)按照 80% 和 20% 的比例划分数据到训练集和测试两个部分。最终，包括训练数据 26 363 条，测试数据 6 503 条②。商品问答数据示例如表 5 所示。

表 5 商品问答数据示例

属性	问题样本示例
主机尺寸	24 英寸，外壳尺寸是多少，宽度
主机重量	电视多少公斤

① http://biswufe.cn/app

② https://pan.baidu.com/s/1LkNijeklPrG0Q_MZtPDtEg

续表

属性	问题样本示例
品牌	是什么牌子的电视
分辨率	这款电视分辨率
能效等级	这个能效等级三级是什么意思呢
上市时间	这一款是哪一年上的市
屏幕尺寸	699 这个是 32 寸吗
接口类型	有输入哪些接口呢

模型训练的超参数与表 3 一致，但属性较少，因此负采样比例为 1∶5。在进行了 5 轮迭代训练后结果如表 6 所示，可以看到在没有进行推理时 F_1 值为 94.69%。

表 6　商品问答平均 F_1 值

模型	F_1 值/%
Siamese	94.21
SiameseATT	94.69
Siamese 推理后	97.97
SiameseATT 推理后	**98.49**

表 7 列出了两个推理规则，可以加入原型系统中。从表 6 可以看出，加入推理规则后，问答系统的 F_1 值有显著的提升，达到 98.49%。

表 7　规则推理的属性

属性	问题样本示例	推理规则
是否投屏功能	还有支持投屏么？	Android(x, true)→ 投屏(x, true)
能否连接电脑	电视可以连电脑主机吗？	HDMI (x, numbers) ^ greaterThan(numbers, 0) →连接电脑(x, true)

5　结论

本文提出了基于知识图谱和规则推理的在线商品自动问答系统，其核心包括：①基于 LSTM 的 Siamese 属性注意力网络实现了候选属性选择算法，解决了一阶属性选择的问题。②在知识图谱上进行规则推理，解决了高阶属性选择的问题。实验证明，我们的问答系统获得了较好的性能，且适合电商的应用场景。同时，我们也开发了一个原型系统，用于展示工作。

虽然我们的问答系统效果达到了预期，但也存在一些不足，例如，在 SiameseATT 的训练中并没有结合知识图谱的信息。下一步可以考虑将词嵌入和知识图谱的嵌入[23]结合起来以提高属性选择的准确率。此外，还将研究是否可以动态地生成规则，而非人工给出规则。最后，问答系统给出的结果较为生硬，如何使得答案更为自然也是下一步的研究的方向。

参考文献

[1] Berant J, Chou A K, Frostig R, et al. Semantic parsing on freebase from question-answer pairs[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013: 1533-1544.

[2] Cui W, Xiao Y, Wang H, et al. KBQA: Learning question answering over QA corpora and knowledge bases[J/OL].arXiv preprint arXiv: 1903.02419, 2019.

[3] Madabushi H T, Lee M. Highaccuracy rule-based question classification using question syntax and semantics[C]//Proceedings of the 26th International Conference on Computational Linguistics, 2016: 1220-1230.

[4] Amin, Muhammad Zain, Noman Nadeem, et al. Convolutional neural network: Text classification model for open domain question answering system[J/OL]. arXiv preprint arXiv: 1809.02479, 2018.

[5] 曹明宇，李青青，杨志豪，等. 基于知识图谱的原发性肝癌知识问答系统[J]. 中文信息学报，2019，33(6)：88-93.

[6] 杜泽宇，杨燕，贺樑. 基于中文知识图谱的电商领域问答系统[J]. 计算机应用与软件，2017，34(5)：153-159.

[7] Zettlemoyer L, Collins M. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars[J/OL].arXiv preprint arXiv: 1207.1420, 2012.

[8] Yao X, Van Durme B. Information extraction over structured data: Question answering with freebase[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014: 956-966.

[9] Bordes A, Chopra S, Weston J, et al. Questionanswering with subgraph embeddings[J]. ArXiv prepine arXiv: 1406.3676, 2014.

[10] Dong L, Wei F, Zhou M, et al. Question answering over freebase with multi-column convolutional neural networks[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics

and the 7th International Joint Conference on Natural Language Processing, 2015: 260-269.

[11] Hao Y, Zhang Y, Liu K, et al. An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge [C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017: 221-231.

[12] Qu Y, Liu J, Kang L, et al. Questionanswering over freebase via attentive rnn with similarity matrix based cnn[J/OL]. arXiv preprintarXiv: 1804.03317, 2018.

[13] Lai Y, Lin Y, Chen J, et al. Open domain question answering system based on knowledge base[J]. Natural Language Understanding and Intelligent Applications. Springer, Cham, 2016: 722-733.

[14] 周博通, 孙承杰, 林磊, 等. 基于 LSTM 的大规模知识库自动问答[J]. 北京大学学报(自然科学版), 2018, 54(2): 286-292.

[15] Liu A, Huang Z, Lu H, et al. BB-KBQA: bert-based knowledge base question answering[C]//Proceedings of the 18th China National Conference on Chinese Computational Linguistics. Springer, Cham, 2019: 81-92.

[16] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition [C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016: 260-270.

[17] Duan N. Overview of the NLPCC-ICCPOL 2016 shared task: Open domain Chinese question answering [G]//LNCS10102: Natural Language Understanding and Intelligent Applications. Berlin: Springer, 2016: 942-948.

[18] Li S, Zhao Z, Hu R, et al. Analogical reasoning on Chinese morphological and semantic relations[J/OL]. arXiv preprint arXiv: 1805.06504, 2018.

[19] Xie Z, Zeng Z, Zhou G, et al. Knowledge base question answering based on deep learning models[G]// LNCS10102: Natural Language Understanding and Intelligent Applications. Berlin: Springer, 2016: 300-311.

[20] 申存, 黄廷磊, 梁霄. 基于多粒度特征表示的知识图谱问答[J]. 计算机与现代化, 2018(9): 5-10.

[21] Yang F, Gan L, Li A, et al. Combining deep learning with information retrieval for question answering [G]//LNCS10102: Natural Language Understanding and Intelligent Applications. Berlin: Springer, 2016: 917-925..

[22] Xie Z, Zeng Z, Zhou G, et al. Topic enhanced deep structured semantic models for knowledge base question answering[J]. Science China (Information Sciences), 2017,64(11): 28-42.

[23] Lin Y, Liu Z, Sun M, et al. Learning entity and relation embeddings for knowledge graph completion [C]//Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI'15), 2015: 2181-2187.

王思宇(1982—),博士研究生,主要研究领域为知识图谱、数据挖掘、自然语言处理。
E-mail: siyu_wang@smail.swufe.edu.cn

洪川洋(1994—),博士研究生,主要研究领域为知识图谱、自然语言处理。
E-mail: hongchuanyang@smail.swufe.edu.cn

邱江涛(1972—),通信作者,教授,博士生导师,主要研究领域为数据挖掘、商务智能、社会计算。
E-mail: qiujt_t@swufe.edu.cn