基于知识图谱的广西文化旅游问答系统研究与实现*

何国对[1],黄容鑫[1],黄伟刚[1],李　航[1],覃　晓[1]**,元昌安[2],施　宇[1],廖兆琪[1]
(1.南宁师范大学计算机与信息工程学院,八桂学者创新团队实验室,广西南宁　530000;2.广西科学院,广西南宁　530007)

摘要:当前的旅游咨询服务还只是为用户提供自主网络搜索返回的碎片化信息,尚未能将地方特色文化智能反馈给用户。针对此实际情况,本研究基于广西民族文化旅游知识图谱,对广西民族文化旅游问答系统的关键技术加以研究,并设计相应的问答系统,在解决实际需求的同时,尝试提高用户咨询体验满意度。根据问答系统(Question Answering System,QA)结构,本研究设计并实现了基于BERT的命名实体识别模块(BERT based Entity_identification Model,BEiM),基于模版的关系抽取模块(Template based Relationship_extraction Module,TReM)和基于知识图谱的匹配推理模块(Knowledge Graph based Matching Module,KGMM)。在上述关键技术基础上,实现了广西文化旅游问答系统,并给出相关实验测试和应用效果。本研究构建的知识问答系统能够帮助游客高效地找到当地旅游的相关知识,提高游客自助服务的效率。对于人工智能助力广西旅游业的发展而言,本研究无疑是一项具有重要意义的工作。

关键词:知识图谱　问答系统　深度学习　自然语言处理　命名实体识别　关系抽取

中图分类号:TP31　　**文献标识码**:A　　**文章编号**:1005-9164(2020)06-0609-07
DOI:10.13656/j.cnki.gxkx.20210119.006

0　引言

问答系统(Question Answering System,QA)是人工智能和自然语言处理领域中一个倍受关注并具有广泛发展前景的研究方向[1],它能用准确、简洁的自然语言回答用户用自然语言提出的问题。旅游是问答系统的一个重要应用场景。一个完善的旅游知识问答系统,能够帮助人们在旅游前、旅游中,通过询问快速获得旅游资讯、了解旅游目的地的文化和特色旅游资源,对游客、管理部门和商家而言,都具有重要的应用价值。然而,当前针对旅游行业的知识咨询现状却难以令人满意:一方面游客对旅游地的文化知识和旅游资讯的咨询需求不断增大[2];另一方面,由于各个地方在旅游知识自动问答系统的建设方面投入不足,当前的咨询服务还停留在依靠用户独自在网上搜索碎片化信息阶段,远远不能满足用户的需求。

知识图谱技术[3]能够把大量不同种类的信息链接在一起,使其形成一个关系网络,为人们提供从“关

*国家自然科学基金项目(61962006),广西研究生教育创新计划项目(YCSW2019182)和广西创新驱动重大项目(AA18118047)资助。
【作者简介】
何国对(1993—),男,在读硕士研究生,主要从事自然语言处理和知识图谱研究。
【**通信作者】
覃　晓(1973—),女,副教授,主要从事人工智能和图像处理研究,E-mail:7670172@qq.com。
【引用本文】
何国对,黄容鑫,黄伟刚,等.基于知识图谱的广西文化旅游问答系统研究与实现[J].广西科学,2020,27(6):609-615.
HE G D,HUANG R X,HUANG W G,et al.Research and Implementation of Guangxi Cultural Tourism Question Answering System based on Knowledge Graph [J].Guangxi Sciences,2020,27(6):609-615.

系”角度分析问题的能力。当前,基于知识图谱的问答系统已经在油茶产业[4]、苹果种植销售产业[5]、水利信息管理[6]等方面得到充分研究和应用,而知识图谱所具备的推理功能,更是让其在新冠肺炎智能辅助问诊系统[7]、军事装备知识问答系统[8]、中医药知识问答与辅助开药系统[9]中表现出令人惊喜的效果。

大部分领域构建知识图谱和问答系统是希望通过构建知识库来提高领域知识的检索效率,同时辅助推理、决策等行为,这也是各个领域智能化的基本需求。知识图谱技术是领域智能化的一条路径,旅游领域的智能化无疑也需要借助知识图谱来实现。旅游领域的知识图谱可以用来辅助各种复杂的旅游应用分析,同时也可以对用户进行个性化的路线推荐[10,11]。广西是一个多民族聚居的地区,壮、汉、苗、瑶、侗等多个民族都在此地居住[12],各个民族都有着悠久的历史和灿烂的文化,在语言、社交、婚姻、服装、饮食、建筑等文化上各具特色,又相互交融。利用广西民族文化知识图谱,构建一个知识问答系统,对人工智能助力广西旅游业的发展而言,无疑是一项具有重要意义的工作。

1 广西民族文化知识图谱

本研究构建的广西文化旅游问答系统,基于南宁师范大学“全域数字文化旅游智能服务技术研发及应用”项目团队所构建的广西民族文化知识图谱,知识图谱的部分内容如图 1 所示。该知识图谱的实体包含具有广西民族文化色彩的旅游景点、民族、民族文化、民族节假婚庆、民族服饰等概念,每一类概念下存储多个实体及实体关系。以民族服饰为例,其中包含广西各民族的服装、头饰、鞋子、帽子等实体,实体关系包含民族、支系、历史、服装部件、服装特点、穿着人群、相关人物、相关传说等。

广西民族文化知识图谱实体与实体关系以三元组(ei,rs,ej)的形式存储于 neo4j 数据库中,其中,ei,ej 分别表示实体 i 和实体 j,rs 表示实体 i 和实体 j 之间的关系。例如:三元组(香粽,食材,糯米)中,香粽和糯米为两个实体,食材为关系,表示糯米为香粽的食材。知识图谱中实体、实体关系均以单词方式存储,其中表示关系的单词称为关系词。

广西民族文化知识图谱建立了广西旅游景点、民族和民族文化的关联性,将三者之间的关联关系可视化,实现知识的关联与挖掘,使得这些民族旅游文化不再是数据孤岛,为基于广西民族文化知识图谱的问答系统的构建奠定了数据和技术基础。

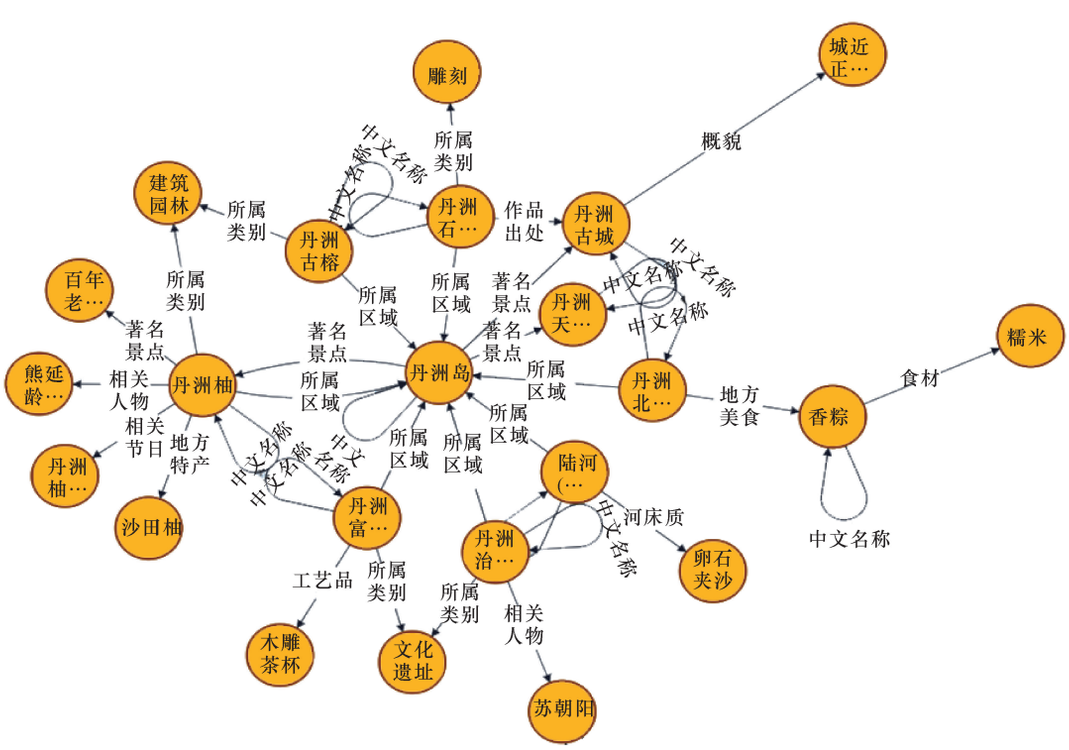

图 1 广西民族文化知识部分数据可视化结果

Fig. 1 Visualization results of some date of Guangxi ethnic cultural knowledge

2 文化旅游问答系统

基于广西民族文化知识图谱的问答系统,主要有 3 个核心模块:(1)基于 BERT 的命名实体识别模块(BERT based Entity_identification Model,BEiM),(2)基于模版的关系抽取模块(Template based Relationship_extraction Module,TReM),(3)基于知识图谱的匹配推理模块(Knowledge Graph based Matching Module,KGMM)。BEiM 模块主要功能:对给定对询问语句,识别出其中对实体词,帮助问答系统理解询问主体(即确定询问的范围)。TReM 模块通过对询问句的关系抽取,完成对询问句的语义解析,帮助问答系统理解询问的语义(即确定询问的内容)。KGMM 模块则是在广西民族文化知识图谱之上,构建查询匹配语句,完成针对询问句的匹配推理解答。

以在文化旅游问答系统询问"你知道丹洲书院在哪里吗?"为例。对于问句"你知道丹洲书院在哪里吗?",问答系统首先通过命名实体识别模块 BEiM 确定问句中所提及的实体"丹洲书院",然后通过关系抽取模块 TReM 确定问句所问的意图,既确定问句所涉及的关系"具体位置",最后通过匹配推理模块 KGMM 将"丹洲书院"与"具体位置"映射为知识图谱的结构化查询,并在知识图谱中最终确定问句的目标实体为"丹洲古镇"(图 2)。

以下详细介绍 3 个主要功能模块。

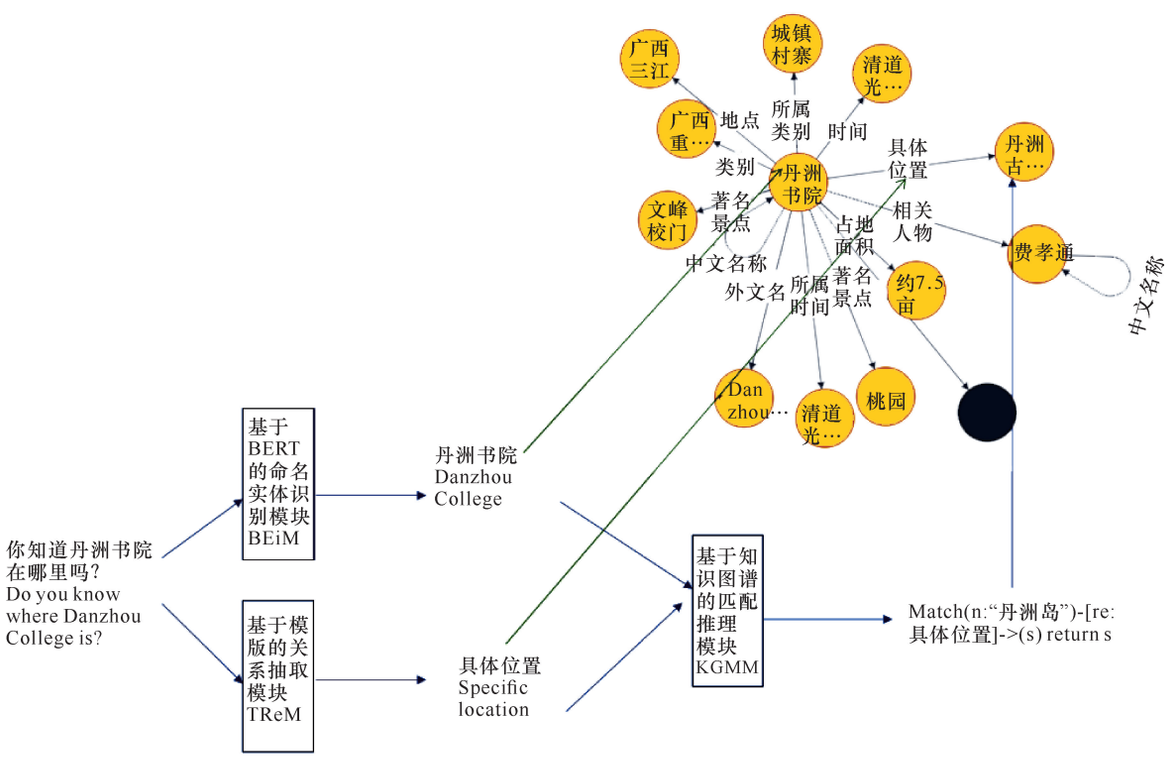

图 2 基于知识图谱的问答系统执行流程

Fig.2 Execution flow of question answering system based on knowledge graph

2.1 基于 BERT 的命名实体识别模块(BERT based Entity_identification Model,BEiM)

BERT[13]是一种预训练语言表示的语言表征模型,它是谷歌公司在大量文本语料上训练出来的通用的"语言理解"模型。BERT 模型的核心功能是对输入的自然语言语料进行分析。在分析基础上将文本中各个字或词的一维词向量作为输入,经过一系列复杂的转换后,最终输出每个词的一维词向量表示,即 BERT 会对句子中的每个词作处理,并得到每个词最终的语义表示。对 BERT 的输出层进行微调,可以使其适应不同的文本分析需求,因此能够灵活应用于问答任务和语言推理,无需针对具体任务做大幅度架构修改。

基于 BERT 的强大功能,本研究设计并实现了基于 BERT 命名实体识别模型 BEiM。对于询问句 S,假设经过 BERT 模型处理后,得到 S 的字符集合为 $(S_1,S_2,\cdots,S_m)$,字符集合中的任意一个 S_i 代表输入的字符 i 的词向量。词的类别按命名实体识别

约定,分为 B-PER、I-ORG、E-PER、O 4 类,其中 B-PER 表示字符处在实体字符边界的开始,I-ORG 表示字符处在实体的中间,E-ORG 表示字符处在实体的结束位置,字符 O 表示不属于实体的无关字符,BEiM 模型可描述如下:

$$BEiM(S)=MLP(BERT(S))=MLP(S_1,S_2,\cdots,S_m)=\{p(S_1),p(S_2),\cdots,p(S_m)\},$$

其中,MLP 为对 BERT 模型的输出作简单全连接的操作,$p(S_i)$ 为对字符 S_i 类别的预测。

$p(S_i)=c_j$,且 $c_j \in$ { B-PER,I-ORG,E-PER,O},$j=1,2,3,4$。

从模型描述可知,BEiM 分两个阶段对输入的询问句 S 进行处理。第一阶段,使用 BERT 对输入的询问句 S 的每个字符进行 embedding,得到每个字符的向量表示,并将每个字符的 embedding 输入到 Transformer block(Trm)中,Trm 会计算句子中所有词对当前输入词的贡献,再根据得到的信息对当前输入词进行编码,获得询问句词向量($S_1,S_2,\cdots,S_m$)。第二阶段,采用 MLP 对词向量的类别进行预测,对获得的词向量($S_1,S_2,\cdots,S_m$)作全连接操作并进行多层感知机权重的调整。图 3 给出 BEiM 的架构,并说明询问句"丹洲岛上有什么"的处理过程。

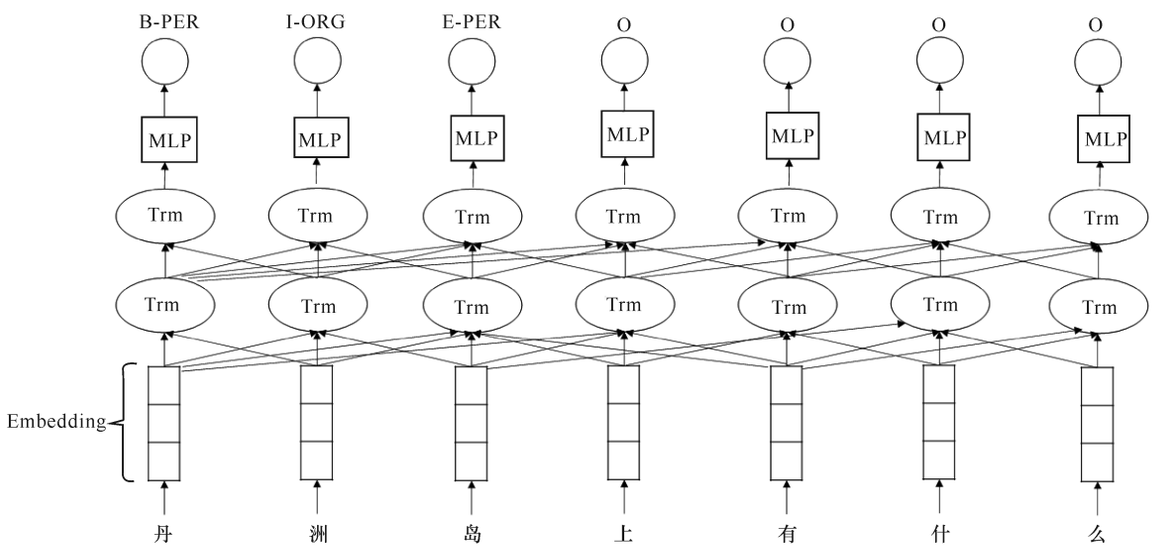

图 3　BEiM 架构图

Fig. 3　BEiM architecture diagram

2.2　基于模版的关系抽取模块(Template based Relationship_extraction Module,TReM)

关系抽取是问答系统中帮助系统理解询问句语义的环节,只有理解并获取了询问句中的语义关系,才能把该关系映射到知识图谱中,最终获取答案。

为方便说明 TReM 构建方法,下面先对相关概念进行定义和描述。

设在领域知识图谱中,共定义了 n 个关系,则可将关系集记为 $R=\{r1,r2,\cdots,rn\}$。其中 $fr(wr_i)=r_j$ 表示从关系词 wr_i 到关系 r_j 的映射。假设知识图谱中存储的关系词数目为 m,则 m 个关系词构成集合 WR。

定义 1(关系词集合)　由广西民族文化知识图谱中所有关系词构成的集合,记为 WR:

$$WR=\{wr_i \mid fr(wr_i)=r_j, i \in [1,m], j \in [1,n], m>n\}。$$

由知识图谱的特性可知,在知识图谱中,关系词的数目 m 往往大于关系数目 $n(m>n)$,因此多个关系词将会被映射到同一个关系。本文将 fr 函数定义为关系模板,它能将具有同一种语义关系的关系词映射为同一个关系。由 n 个关系模版构成的集合定义为关系模版集。

定义 2(关系模版和关系模版集)　关系模版是指被映射为同一个关系的关系词向量,关系模版集是由所有关系模版构成的集合,记为 WRS:

$$WRS=\{WRS^j \mid j \in [1,n]\}。$$

则第 j 个关系模版记为 WRS^j:

$$WRS^j=\{(wr^j_{s_1},wr^j_{s_2},\cdots,wr^j_{s_i},\cdots,wr^j_{s_n}) \mid$$

$wr^{j}_{s_i} \in WR$ 且 $fr(wr^{j}_{s_i})=r_j, s_i \in [1,m]\}$,
即 WRS 是由 n 个不定长的关系模版构成的集合。对于一个关系模版而言,该模版中的所有关系词,均映射到同一个关系中。

TReM 的具体实现方法:首先构建广西民族文化知识图谱关系词组集 WRS;然后调用分词函数 split(),获取询问句分词向量 W;最后,在关系词组集中对问句分词向量进行匹配检索,如果检索成功,则问句关系即可判定为匹配关系。TReM 算法描述如下:

Template based Relationship_extraction algorithm

input:query senten S,WRS

output:relation of words in S,记为 S_r

begin

(1) $W:(w_1,w_2,\cdots,w_m) \leftarrow$ split(S) //

(2)for i=1 to m do

(3) for j=1 to n do

(4) if w_i in $WRSj$:

(5) $S_r \leftarrow r_j$

(6) end for

(7)end for

(8)end

2.3 基于知识图谱的匹配推理模块(Knowledge Graph based Matching Module,KGMM)

KGMM 的主要功能是基于广西民族文化知识图谱,根据 BEiM 和 TReM 的输出结果,构造 Cypher 查询模板进行答案的查询。对于问答系统,只要确定查询实体的 E,然后再确定查询实体关联关系 r,便可构造。构造的查询语句为“MATCH (n:E)-[re:r]->(s) return s”,该语句通过确定问句的实体 E,并通过关系链路 r 确定答案 s,其中查询语句中 n:E 表示将实体名称 E 赋值给实体 n,re:r 表示将关系名称 r 赋值给关系链路 re。该查询语句会查询与实体 E 具有关系 r 的实体并返回。例如:对于问句“你知道丹洲书院在哪里吗?”,确定询问实体“丹洲书院”与实体关系“具体位置”,便可将查询语句构造为“MATCH (n:丹洲书院)-[re:具体位置]->(s) return s”,该语句会匹配与“丹洲书院”具有“具体位置”关系的实体并返回。

至此,本研究介绍了基于知识图谱的广西民族文化问答系统的关键技术。系统的具体实现方法如下:首先构建广西民族文化知识图谱 KG 关系词组集 WRS;然后将问句 S 输入到 BEiM 模型,得到问句的询问实体 E,再将问句 S 输入到关系抽取模块 TReM,得到关系 r,进而将实体 E 和关系 r 输入到匹配推理模块 KGMM,得到查询语句 Q,最后基于语句 Q 在广西民族文化知识图谱上查询答案 t 并返回。广西民族文化问答系统模型算法(Knowledge question answering algorithm)描述如下:

Knowledge question answering algorithm

input:query senten S,WRS,KG

output:answer t of KG

(1) $E=BEiM(S)$

(2)R=TReM(S,WRS)

(3) $Q \leftarrow$ KGMM(E,R)= ‘MATCH (n:E)-[re:R]->(t) return t’

(4)match t based on Q from KG

(5)return t

3 测试实验

对问答系统而言,能否正确识别问句中所询问的实体是问答系统的关键,只有先确定了实体,才能通过关系链路把正确答案找出来,因此命名实体识别模型通常需要达到较高的精度才能投入到问答系统中使用。为此,先对基于 BERT 的命名实体识别模型(BEiM)进行训练测试,其中训练数据集以及验证集使用的是含有人名、地名、机构名、时间、景点、饮食等实体的标注序列数据,其数据为通用的中文命名实体识别数据集 Uniuersal data;同时,对广西民族文化知识图谱中的部分实体数据集进行测试。

3.1 命名实体识别测试

本研究使用准确率作为命名实体识别的评价指标。本研究使用 BEiM 预训练模型来训练中文命名实体识别数据集。BEim 对输入字特征进行微调,通过微调可以更好地学习与命名实体相关的特征。从 BEiM 模型对部分广西民族文化知识图谱实体进行命名实体识别的实验结果可知,BEiM 在 Universal data 验证集上的准确率达到了 99.30%,BEiM 模型对部分广西文化知识图谱上的命名实体识别精度达到 94.45%,完全能够胜任广西文化旅游问答系统的命名实体识别任务。

3.2 系统应用测试

如表 1 所示,用户分别询问了景点的票价、旅游地点的美食、景区内的主要景点、当地美食的主要食材等问题,问答系统都能很好地对用户的问句进行解

析,并通过广西民族旅游文化知识图谱进行检索,得到了准确的答案,并用精简的方式返回给用户,增强用户的体验感。

表 1　问答系统询问测试

Table 1　Query test to the question answering system

问句 Question	问答系统答案 Answers of question answering system	真实答案 Real answer
花山风景区的票价 Price of Huashan Scenic Spot	80 元 80 Yuan	80 元 80 Yuan
丹洲北门街的美食有什么 What are the delicacies in the North Gate Street of Danzhou	白切猪手、糯米饭、香粽、土鸡、秘制腊肉、清蒸鱼、芝麻粑 White cut pork knuckles, glutinous rice, fragrant zongzi, chicken, secret bacon, steamed fish, sesame cake	白切猪手、糯米饭、香粽、土鸡、秘制腊肉、清蒸鱼、芝麻粑 White cut pork knuckles, glutinous rice, fragrant zongzi, chicken, secret bacon, steamed fish, sesame cake
能告诉我花山岩画的主要景点有哪些吗 Can you tell me the main scenic spots of Huashan rock paintings	龙峡、高山、明江东岸花山岩画、将军崖花山岩画、花山岩壁画、珠山 Longxia mountain, Gaoshan mountain, Huashan rock paintings on the East Bank of Mingjiang River, Huashan rock paintings on Jiangjunya, Huashan rock mural, Zhushan mountain	龙峡、高山、明江东岸花山岩画、将军崖花山岩画、花山岩壁画、珠山 Longxia mountain, Gaoshan mountain, Huashan rock paintings on the East Bank of Mingjiang River, Huashan rock paintings on Jiangjunya, Huashan rock mural, Zhushan mountain
芝麻糍粑由什么制成 What is sesame Ciba made of	黑芝麻、糯米粉 Black sesame, glutinous rice flour	黑芝麻、糯米粉 Black sesame, glutinous rice flour

4　结束语

本文设计并实现了基于广西民族旅游文化知识图谱的智能问答系统,重点介绍了问答系统中命名实体识别和询问句关系抽取关键技术。在实现的智能问答系统上进行询问测试,其结果表明系统能够准确解析询问句子,从知识图谱中检索到准确答案,并以简洁的实体树形式展示。该智能问答系统不足之处在于:基于模版的命名实体识别方法需要构建大量的关系词组模版,才能保证将用户的问句映射为知识图谱中的关系,而手工构建关系词组模版将导致高额的系统构建代价。因此在下一步工作中,可以设计一套完整的映射模型,将用户问句自动映射到知识图谱中的关系列表中,从而提升系统的智能程度,降低系统构建代价。

参考文献

[1] 詹晨迪.基于知识库的自然语言问答方法研究[D].合肥:中国科学技术大学,2017.

[2] 张苗荧."互联网+旅游"迎来更大发展机遇[N].中国旅游报,2020-06-09(003).

[3] 赵军,刘康,何世柱,等.知识图谱[M].北京:高等教育出版社,2018:7-10.

[4] 丁浩宸,王忠明.基于本体的油茶中文知识图谱构建与应用[J].世界林业研究,2020,33(4):50-55.

[5] 陈亚东,鲜国建,寇远涛,等.我国苹果产业知识图谱构建研究[J].中国农业资源与区划,2017,38(11):40-45.

[6] 张紫璇,陆佳民,姜笑,等.面向水利信息资源的智能问答系统构建与应用[J].计算机与现代化,2020(3):65-71.

[7] 浙江移动:基于医疗知识图谱的新冠肺炎智能辅助问诊系统[J].杭州科技,2020(1):62-63.

[8] 车金立,唐力伟,邓士杰,等.基于百科知识的军事装备知识图谱构建与应用[J].兵器装备工程学报,2019,40(1):148-153.

[9] 阮彤,孙程琳,王昊奋,等.中医药知识图谱构建与应用[J].医学信息学杂志,2016,37(4):8-13.

[10] 贾中浩,宾辰忠,古天龙,等.基于知识图谱和用户长短期偏好的个性化景点推荐[J/OL].智能系统学报:1-9.[2020-06-10].http://kns.cnki.net/kcms/detail/23.1538.TP.20190906.1314.002.html.

[11] 孙彦鹏.面向智能旅游服务机器人的个性化推荐算法研究[D].桂林:桂林电子科技大学,2019.

[12] 朱华丽.多民族地区乡村文化发展禀赋、问题及路径设计[J].湖南行政学院学报,2019(3):49-56.

[13] DEVLIN J,CHANG M W,LEE K,et al.BERT:Pre-training of deep bidirectional transformers for language und erstanding [Z/OL].[2020-06-10].Google AI Language,2018.https://nlp.stanford.edu/seminar/details/jdevlin.pdf.

Research and Implementation of Guangxi Cultural Tourism Question Answering System based on Knowledge Graph

HE Guodui[1], HUANG Rongxin[1], HUANG Weigang[1], LI Hang[1], QIN Xiao[1], YUAN Chang'an[2], SHI Yu[1], LIAO Zhaoqi[1]

(1. BAGUI Scholar Innovation Team Laboratory, School of Computer & Information Engineering, Nanning Normal University, Nanning, Guangxi, 530000, China; 2. Guangxi Academy of Sciences, Nanning, Guangxi, 530007, China)

Abstract: The current tourism consulting service is only to provide users with fragmented information returned by independent web search, and has not yet been able to feed back the local characteristic culture intelligently to users. In response to this actual situation, the key technologies of the Guangxi ethnic culture and tourism question answering system based on the knowledge map of Guangxi ethnic culture and tourism are studied in this article. And the corresponding question and answer system is designed to try to improve the satisfaction of users consultation experience while solving the actual needs. According to the structure of the question answering system, this article designs and implements a BERT-based named entity identification module (BERT based Entity_identification Model, BEiM), a template-based relationship extraction module (Template based Relationship_extraction Module, TReM) and a knowledge graph-based matching inference module (Knowledge Graph based Matching Module, KGMM). On the basis of the above key technologies, the Guangxi cultural tourism question answering system is implemented, and the relevant experiments and application effects are given. The knowledge question answering system constructed in this research can help tourists find the relevant knowledge of local tourism efficiently and improve the efficiency of tourists' self-service. This research is undoubtedly an important work for artificial intelligence to help the development of Guangxi tourism.

Key words: knowledge graph, question answering system, deep learning, natural language processing, named entity recognition, relationship extraction

责任编辑:米慧芝