第47卷 第2期 计 算 机 工 程 2021年2月
Vol.47 No.2 Computer Engineering February 2021

·人工智能与模式识别· 文章编号：1000-3428(2021)02-0103-08 文献标志码：A 中图分类号：TP391

基于多标签策略的中文知识图谱问答系统研究

朱宗奎，张鹏举，贾永辉，陈文亮，张 民

（苏州大学 计算机科学与技术学院，江苏 苏州 215006）

摘 要：现有多数中文知识图谱问答(CKBQA)系统侧重于回答单个三元组查询的简单问题，而不能有效解决涉及多个实体和关系的复杂问题。提出一种基于多标签策略进行答案搜索的CKBQA系统，该系统主要包括问题处理和答案搜索2个部分。在问题处理部分，结合预训练语言模型构建新的模型框架，对问题进行实体提及识别、实体链接和关系抽取处理，通过设置3种分类标签将问题划分为简单问题、链式问题和多实体问题。在答案搜索部分，对上述3种分类问题分别给出不同的解决方法。实验结果表明，该系统在CCKS2019-CKBQA评测数据验证集上的平均F1值可达66.76%。

关键词：知识图谱；问答系统；分类；多标签策略；实体；关系

开放科学(资源服务)标志码(OSID)：

中文引用格式：朱宗奎，张鹏举，贾永辉，等.基于多标签策略的中文知识图谱问答系统研究[J].计算机工程，2021，47(2)：103-110，117.

英文引用格式：ZHU Zongkui，ZHANG Pengju，JIA Yonghui，et al.Chinese knowledge base question answering system based on multi-label strategy[J].Computer Engineering，2021，47(2)：103-110，117.

Study of Chinese Knowledge Base Question Answering System Based on Multi-Label Strategy

ZHU Zongkui，ZHANG Pengju，JIA Yonghui，CHEN Wenliang，ZHANG Min

（School of Computer Science and Technology，Soochow University，Suzhou，Jiangsu 215006，China）

【Abstract】Most of the existing Chinese Knowledge Base Question Answering（CKBQA）system focus on simple questions that need a single triplet query，but cannot solve complex questions involving multiple entities and relations. To address the problem，this paper proposes a CKBQA system for answer search based on multi-label strategy. The system mainly consists of two parts：question processing and answer search.In the question processing part，a new model framework is constructed based on the pre-trained language model to perform entity mention recognition，entity linking and relation extraction for the questions. By setting three classification labels，the questions are divided into simple questions，chain questions and multi-entity questions.In the answer search part，different processing methods are given for the above three kinds of classification questions. The experimental results show that the average F1 value of the proposed system reaches 66.76% on the validation set of evaluation data，CCKS2019-CKBQA.

【Key words】knowledge base；question answering system；classification；multi-label strategy；entity；relation

DOI：10. 19678/j. issn. 1000-3428. 0056763

0 概述

随着人机交互技术的快速发展，传统的搜索引擎已无法满足用户对信息获取的多样化需求，于是问答系统应运而生，并逐渐成为人工智能(AI)、自然语言处理(NLP)和信息检索(IR)领域中的一个研究热点，具有广阔的应用前景[1]。与传统的搜索引擎不同，问答系统可以更快速、更准确地向用户直接反馈所需的信息或答案，而非返回大量与用户查询相关的网页列表[2]。

根据答案来源的不同，问答系统可以分为基于结构化数据的问答系统(比如知识图谱问答)、基于

基金项目：国家自然科学基金(61936010)。

作者简介：朱宗奎(1994—)，男，硕士研究生，主研方向为自然语言处理；张鹏举、贾永辉，硕士研究生；陈文亮，教授、博士；张 民，教授、博士、博士生导师。

收稿日期：2019-12-02 **修回日期：**2020-02-05 **E-mail：**sudazzk@qq.com

文本的问答系统(比如机器阅读理解)以及基于问答对的问答系统(比如常见问题(FAQ)问答系统)[3]。基于中文知识图谱的问答(CKBQA)系统输入一个中文自然语言问题,问答系统从给定知识库中选择若干实体或属性值作为该问题的答案,问题均为客观事实型,不包含任何主观因素。目前,已有很多大规模的高质量知识图谱被提出,英文的包括Freebase[4]、YAGO[5]和DBpedia[6]等,中文的有百度知心、知立方、Zhishi.me[7]和XLore[8]等,这些知识大多来源于维基百科、百度百科等网站。现有知识图谱的标准数据存储形式一般是由资源描述框架(RDF)三元组组成,即<主语,谓语,宾语>或<头实体,关系,尾实体>,主要包括实体的基本属性、类型、提及信息以及实体与实体之间的语义关系等。知识图谱具有结构化的特点,已逐渐成为开放领域问答系统的重要资源,引起了研究人员的广泛关注。

基于知识图谱的问答系统包含了多个NLP任务,其在理解和回答问题的过程中需要进行实体识别、关系抽取和语义解析等不同的子任务,再通过SQL、SPARQL等查询语言对知识库进行搜索和推理以得到最终的答案[9]。例如,问题q1:“《湖上草》是谁的诗?”是一个简单问题,首先需要从问句中识别出主题实体的提及“湖上草”,再根据提及进行实体链接,确定主题实体为“<湖上草>”,然后从实体的所有候选关系中选出与问句表述最为相近的关系“<主要作品>”,最后利用SPARQL语言“select ? x where {? x<主要作品> <湖上草>.}”,查询答案为“<柳如是_(明末“秦淮八艳”之一)>”,只需要一个三元组知识即可完成;问题q2:“《根鸟》的作者是哪个民族的人”是一个复杂问题,解决方法类似于问题q1,但是需要2个三元组,先得到主题实体“<根鸟>”的“<作者>”,再查到其“<作者>”的“<民族>”为“<汉族>”,SPARQL语言为“select? x where{<根鸟> <作者>? y.? y<民族>? x.}”。

目前,中文知识图谱问答系统大多侧重于回答简单问题,但在实际应用中,很多用户提出的问题单靠一个三元组查询是无法解决的,许多复杂问题涉及多个实体与语义关系。因此,需要针对中文不同类型的问句设计不同的解决方案。虽然近年来有很多新模型和系统被提出以用于知识图谱问答,但大多基于英文语料,针对中文问题时仍存在局限性。中文知识图谱问答系统起步较晚,前期工作以简单问题为主,缺乏大规模公开的标注语料,且中文语言表达形式多样,相比英文更复杂,难以准确理解语义,同时中文分词技术存在领域特殊性和中英文混杂等情况[10]。

近年来,深度学习技术在NLP领域得到广泛应用,基于语言建模的神经网络模型也逐渐成为研究热点,比如ELMo[11]、BERT[12]等。本文将语言模型和中文知识图谱问答系统相结合,构建一种基于多标签策略的中文知识图谱问答系统。利用机器学习方法和预训练语言模型构建针对实体提及识别、实体链接和关系抽取3个任务的模型框架,通过设置不同的分类标签将中文问句划分成简单问题、链式问题和多实体问题3类,并提出处理链式问题和多实体问题的解决方法。

1 相关工作

在NLP领域,基于知识图谱的问答系统已经得到广泛研究。早在20世纪60年代,就有学者针对特定领域内小规模的知识库进行研究,以解决一些具体的专业问题。此后,研究方向逐渐从特定领域转向开放领域,从简单问题转向复杂问题。目前,英文语料主流的研究方法可以分为语义分析和信息检索2种。

早期多数知识图谱问答采用传统基于语义分析的方法[13-15],通过构建一个语义解析器,将自然语言问句映射成一种语义表示、逻辑表达式或查询图[16],然后基于知识库查询得到最终答案。虽然上述方法可以对问句进行深入解释,但由于推理的复杂性较高,需要特定领域语法、细粒度的标注数据以及手工设计规则和特征,使得这些方法难以进行大规模的训练,而且可移植性较差。

基于信息检索的方法[17-18]主要通过构建不同的排序模型检索出一组候选答案,通过分析进行排序从而完成知识图谱问答任务。BORDES等人[19]使用基于向量嵌入的方法编码问句和答案,计算两者之间的语义相似度并进行排序,随后又提出子图向量[20]、记忆网络[21]等方法。近年来,有很多先进的神经网络模型被提出以用于编码句子[22-24],包括卷积网络和循环网络等,这些网络只需简单地查询知识库而无需额外的语法知识和词典,并且能够隐式地完成候选答案的搜索和排序功能。

相较于英文,中文知识图谱问答系统的研究起步较晚,主要以中国计算机学会(CCF)国际自然语言处理与中文计算会议(NLPCC)、全国知识图谱与语义计算大会(CCKS)2个公开的评测任务为主。NLPCC 2015年—2018年的评测数据基本都是简单问题[25-27],而CCKS 2018年—2019年包含了简单问题和复杂问题2种[28-29],它们均使用基于信息检索的方法,针对问题和答案的语义相似度计算建立了不同的度量模型。

2　中文知识图谱问答系统

给定一个中文自然语言问句Q，CKBQA系统的目标是从一个中文知识图谱知识库KB中抽取答案A。本文提出的中文知识图谱问答系统流程如图1所示，其包括问句处理和答案搜索2个主要模块，其中，问句处理模块涉及分类模型、实体提及识别和实体链接模型，答案搜索模块涉及统一单跳问题搜索、链式问题搜索和多实体问题搜索3个部分。图1中的虚线部分表示3个搜索过程在知识图谱中完成。

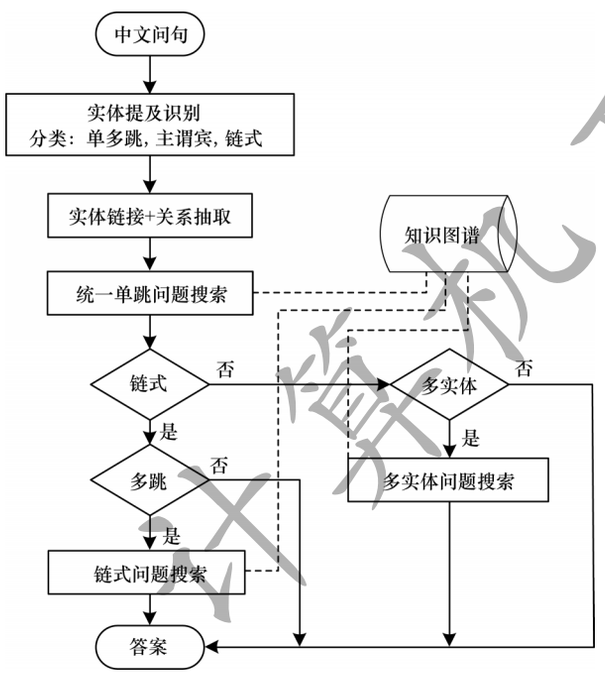

图1　中文知识图谱问答系统流程

Fig.1　Procedure of Chinese knowledge base question answering system

2.1　BERT模型

BERT（Bidirectional Encoder Representations from Transformers）模型结构如图2所示，其为一个多层双向的语言模型，模型输入由词向量、位置向量和分段向量共同组成。另外，句子的头部和尾部分别有2个特殊的标记符号[CLS]和[SEP]，用以区分不同的句子。模型输出是每个字经过M层编码器后对应的融合上下文信息的语义表示。假定一个中文自然语言问句的输入序列为$\boldsymbol{X}=(x_1,x_2,\cdots,x_n)$，经过文本分词器处理后为$\boldsymbol{S}=([\mathrm{CLS}],x_1,x_2,\cdots,x_n,[\mathrm{SEP}])$，再经过M层编码器后的输出序列为$\boldsymbol{H}=(h_0,h_1,\cdots,h_n,h_{n+1})$。预训练后的BERT模型提供了一个强大的上下文相关的句子特征表示，再通过微调后可以用于各种目标任务，包括单句分类、句子对分类和序列标注等。

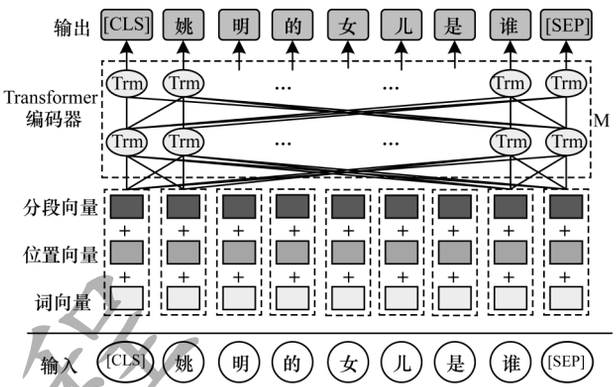

图2　BERT模型结构

Fig.2　Structure of BERT model

2.2　实体提及识别

实体提及识别指给定一个问句，从中识别出主题实体的提及。本文将实体提及识别当作一个序列标注任务，采用神经网络模型进行识别。首先，根据训练语料的SPARQL语句查找主题实体的提及；然后，构建序列标注所用的数据，训练一个提及识别模型。例如，一个问句“电影《怦然心动》的主要演员？”，从其SPQRQL语句“select？x where｛<怦然心动_（美国2010年罗伯·莱纳执导电影）><主演>？x.｝”中可知主题实体为“<怦然心动_（美国2010年罗伯·莱纳执导电影）>”，然后查询实体提及三元组知识，得到该实体的可能提及有“怦然心动”“FLIPPED”“冒失”等。根据最大长度优先匹配规则，标记出该问句的提及为“怦然心动”，设置标签为BIII，非提及部分标签设为O。如果匹配失败，则舍弃该问句，不进行标注。

本文将BERT语言模型和双向长短期记忆（BiLSTM）网络[30]相结合，输入到条件随机场（CRF）[31]模型中，构建一种BERT-BiLSTM-CRF模型，以预测每个字符的标签。首先，通过BERT语言模型得到问句中每个字符的深度上下文表示；然后，使用BiLSTM网络获取每个字符左侧和右侧的前后语义关系；最后，借助CRF模型确保预测的结果是合法的标签。上述过程的具体计算如式（1）、式（2）所示：

$$\boldsymbol{T}=\mathrm{BiLSTM}(\boldsymbol{H}) \tag{1}$$

$$\boldsymbol{Z}=\mathrm{CRF}(\boldsymbol{T}) \tag{2}$$

其中，$\boldsymbol{T}\in\mathbb{R}^{(n+2)\times 2D}$表示编码后的句子经过BiLSTM模型后的输出，$\boldsymbol{Z}\in\mathbb{R}^{1\times(n+2)}$表示CRF模型预测的标签，$D$表示BERT模型输出的隐藏层维度。

BERT-BiLSTM-CRF模型结构如图3所示。

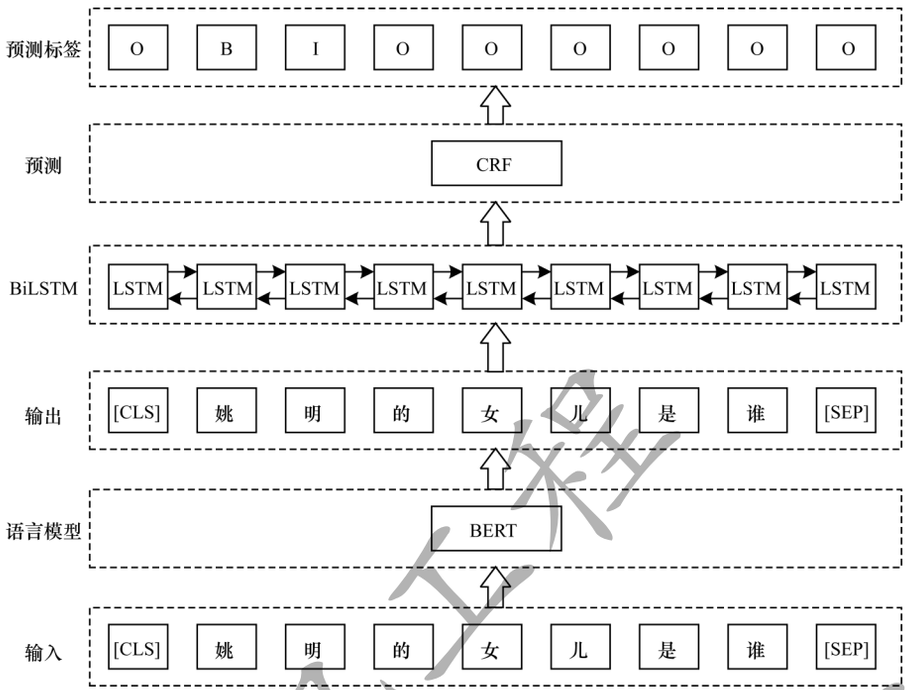

图3 BERT-BiLSTM-CRF模型结构

Fig.3 Structure of BERT-BiLSTM-CRF model

2.3 分类模型

在实际应用场景中,用户提出的问题往往不局限于简单问题,很多包含了复杂的多跳问题。因此,本文将问题划分成单跳问题和多跳问题2类,其中,单跳问题再分为主、谓、宾3个位置的答案查询,多跳问题可以分成链式问题和多实体问题2种。

2.3.1 单多跳分类

单跳问题(简单问题)指问句对应单个三元组查询,而多跳问题(复杂问题)指问句对应多个三元组查询。表1所示为2种类型的问句示例。由于训练数据提供了每个问句的SPARQL查询语句,根据大括号中字段的数量,将训练数据切分成单跳数据(数量=3)和多跳数据(数量>3),单跳标签设为0,多跳标签设为1,然后基于BERT模型训练一个二分类模型。对于单句子分类任务,文献[12]给出了BERT的基本分类框架,即将模型最后一层的第一个标记[CLS]的输出直接作为整个句子的融合表示,然后经过一个多层感知器进行分类,其模型结构如图4所示,最后一步的计算如式(3)所示:

$$y=\text{softmax}(h_0\boldsymbol{W}^{\mathrm{T}}+b) \tag{3}$$

其中,softmax表示激活函数,其计算每个类别的概率分布,$\boldsymbol{W}\in\mathbb{R}^{K\times D}$是隐藏层的权重,$b\in\mathbb{R}^{1\times K}$是偏置,$K$表示类别个数。

表1 单多跳分类示例

Table 1 Examples of single-multi hop classification

标签	例句	SPARQL
0	莫妮卡·贝鲁奇的代表作?	select ?x where { <莫妮卡·贝鲁奇> <代表作品> ?x. }
1	《红豆》的演唱者出生在?	select ?y where { ?x <代表作品> <红豆_(王菲演唱歌曲)>. ?x <出生地> ?y. }

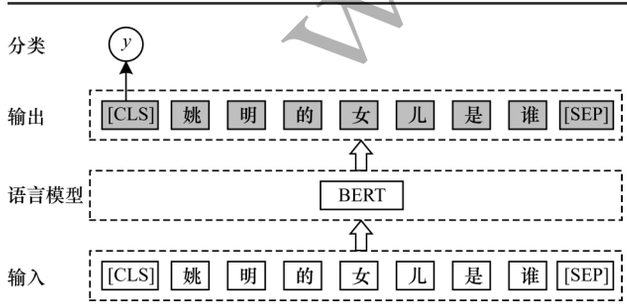

图4 BERT分类模型结构

Fig.4 Structure of BERT classification model

2.3.2 主谓宾分类

主谓宾分类指单跳问句的答案对应于三元组中的主语、谓语或宾语中的一个。当已知一个问句的主题实体时,无法知道该实体对应于知识库三元组中的主语位置还是宾语位置,因此,本文将单跳问题划分成主谓宾3类来查找答案。根据单跳问题的SPARQL语句三元组中问号的所在位置,将单跳问题的数据划分成3类,标签分别设为0、1、2,数据样例如表2所示,然后训练一个三分类模型,模型结构如图4所示。

表2　主谓宾分类示例

Table 2　Examples of subject-predicate-object classification

标签	例句	SPARQL
0	《悼李夫人赋》是谁的作品？	select ?x where { ?x <代表作品> <悼李夫人赋>. }
1	林徽因和梁思成是什么关系？	select ?x where { <林徽因_(中国建筑师、诗人、作家)> ?x <梁思成>. }
2	天津大学的现任校长是谁？	select ?x where { <天津大学> <现任校长> ?x . }

2.3.3　链式分类

链式问题指问句涉及多个三元组查询，并且三元组之间呈递进关系，这类复杂问题的问句中均包含多个关系属性。根据SPARQL语句中三元组是否呈递进关系，可以将所有数据切分成链式问题和非链式问题，因为单跳问题也可能存在问句中有多个实体的情况，所以没有直接将多跳问题划分成链式问题和多实体问题。在此基础上，训练一个二分类模型，模型结构如图4所示。表3所示为2种类型问句的链式分类示例。

表3　链式分类示例

Table 3　Examples of chain classification

标签	例句	SPARQL
0	莫妮卡·贝鲁奇的代表作？	select ?x where { <莫妮卡·贝鲁奇> <代表作品> ?x. }
1	纳兰性德的父亲担任过什么官职？	select ?y where { <纳兰性德> <父亲> ?x. ?x <主要职位> ?y. }
1	宗馥莉任董事长的公司的公司口号是？	select ?y where { ?x <董事长> <宗馥莉>. ?x <公司口号> ?y. }

2.3.4　关系抽取

关系抽取指已知给定问句的主题实体，查找实体的所有候选关系中与问句表达最相近的关系。在很多情况下，中文问句中的关系表述偏口语化，缺乏规范，与知识库中的表达不一致，无法直接通过字符对齐来实现关系抽取。本文基于BERT模型，设计一个问句和关系的语义相似度计算方法。例如，一个问句“里奥·梅西的生日是什么时候？”，从SPARQL语句得知主题实体为“<里奥·梅西_(阿根廷足球运动员)>”，但该实体有很多候选关系，包括“中文名”“外文名”“妻子”“出生日期”“所属运动队”等。本文构建一个相似度计算模型的训练数据，令正例的标签为1，5个负例的标签为0，使用训练好的模型计算问句和每个候选关系的相似度（分类为标签1的概率值），然后进行排序，选择相似度最高的关系来搜索最终答案。模型结构如图4所示，但不同的是输入序列为问句 $\boldsymbol{Q}=(x_1,x_2,\cdots,x_n)$ 和关系 $\boldsymbol{P}=(k_1,k_2,\cdots,k_m)$，然后经过BERT的中文文本分词器处理后的序列为 $\boldsymbol{S}=([\mathrm{CLS}],x_1,x_2,\cdots,x_n,[\mathrm{SEP}],k_1,k_2,\cdots,k_m,[\mathrm{SEP}])$。

2.4　实体链接

实体链接指将问句中识别出的主题实体提及链接到知识库中唯一的实体。因为识别出的提及不能直接链接到具体实体，很多存在一个提及对应多个实体的情况，而且受到模型性能的影响，识别出的提及会有边界错误，所以本文设计3类共10个特征来完成候选实体的排序任务。

2.4.1　提及特征

提及特征共包括以下3种特征：

1）S1，实体提及的初始分。提及识别模型抽取出的提及初始分S1=1，但其只能作为候选，因为很多情况下识别存在边界错误，此时需要对候选的左右字符进行扩展或删减，增加或减少1个字符扣0.1分，最多扩展5个字符，删减最少剩1个字符。

2）S2，实体提及的长度，表示实体对应的提及的字符个数。

3）S3，实体提及的长度占问句的长度比，即提及的字符个数占问句的字符个数的比例。

2.4.2　实体特征

实体特征共包括以下5种特征：

1）S4，实体对应的排名。知识图谱的实体提及三元组中包含了提及所对应的每个实体的具体排名，即优先级0，1，2，…。

2）S5，实体对应的排名的倒数，如果排名为0则设为1，否则为 $\frac{1}{\text{排名}}$。

3）S6，问句和实体的语义相似度，此处相似度度量通过关系相似度抽取模型实现。

4）S7，问句和实体后缀的语义相似度。实体后缀指实体知识三元组中实体名字括号中的部分，通过该信息可以完成实体消歧任务。

5）S8，问句和实体后缀的杰卡德系数，此处杰卡德系数指2个字符串的字符交集个数与并集个数的比值，其值越大，表明字符重叠度越高。

2.4.3　关系特征

关系特征共包括以下2种特征：

1）S9，问句和实体候选关系的最大语义相似度，

该相似度指实体的所有候选关系中与问句语义最相似的关系的相似度值。

2)S10,问句和实体候选关系的最大杰卡德系数,该系数指实体的所有候选关系中与问句字符最相似的关系的杰卡德系数值。

在训练数据的过程中,令正确实体的标签为1,其余候选实体标签为0,采用XGBoost模型[32]对上述特征进行拟合,完成二分类任务,然后在验证集和测试集上,使用训练好的模型对每个候选实体进行打分(分数即分类为标签1的概率值),选择排名第1的实体作为最终答案。

2.5 答案搜索

答案的搜索流程如图1所示,具体步骤如下:

1)先对问句进行分类,判断是否为单多跳、主谓宾或者链式,然后实现实体提及识别。

2)根据识别到的提及进行左右扩展或删减,搜索所有可能的候选实体,再根据一组特征,通过实体链接模型对候选实体进行打分排序,选择得分最高的实体。

3)根据问句的主谓宾标签搜索实体对应的所有关系,通过关系抽取模型计算它们与当前问句的语义相似度,取得分最高的关系,搜索知识库得到统一单跳问题的答案。

4)若问句是链式且为多跳问题,将第3步得到的答案作为主题实体再执行一遍第3步,得到多跳链式问题的答案。

5)若问句是非链式且识别到多个实体,对每个实体搜索数据库,查询对应的所有候选三元组,然后两两求交集得到多实体问题的答案。

图5所示为多实体问题搜索的2个例子,分别为“由黄渤和徐峥共同主演的电影有哪些?”和“清华大学出了哪些物理学家?”,两者都具有2个尾实体,前者是相同谓语,后者是不同谓语,通过计算2个实体三元组之间的交集可以得到问题的答案。

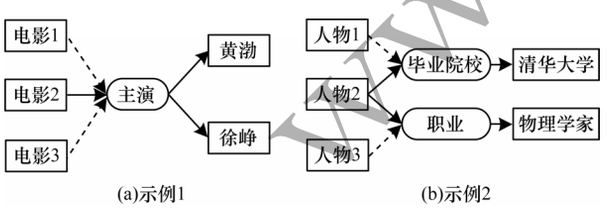

图5 多实体问题的搜索示例

Fig.5 Search examples for multi entity problems

3 实验结果与分析

3.1 实验数据

本文实验使用的数据来自CCKS2019-CKBQA公开评测数据,包括3份问答数据集和1份开放知识图谱。评测数据均由人工构建和标注,其中,北京大学计算机技术研究所提供了3/4的开放领域问答数据,恒生电子股份有限公司提供了1/4的金融领域问答数据。问答数据集包含2 298条训练集,766条验证集(初赛)和766条测试集(复赛)。开放知识图谱使用一个大型的中文知识图谱PKUBASE,该图谱包含41 009 141条实体知识三元组、13 930 117条实体提及三元组和25 182 627条实体类型三元组。另外,由于关系抽取模型的训练数据过少,本文实验额外增加了NLPCC2016-KBQA[33]公开评测数据。NLPCC2016-KBQA数据主要包含简单问题,而CCKS2019-CKBQA数据还包含很多的复杂问题,因此,本文选取CCKS2019-CKBQA数据作为实验数据。

3.2 实验设置

本文使用的BERT预训练模型为BERT-Base Chinese[12],其基于Tensorflow框架实现,有12层编码器,每一层隐状态的输出维度为768,中文问句的最大长度为60。模型的优化方式采用Adam算法对参数进行更新和微调,初始学习率均为2e-5。训练时采用批量训练的方法,批量大小为32。Dropout比率默认设置为0.1,最大迭代次数为100,训练时每50步保存一次模型并验证一次开发集。

实验结果的评价指标包括宏观准确率(P_{Macro})、宏观召回率(R_{Macro})和平均F1值($\text{F1}_{\text{Average}}$),评测结果最终排名以平均F1值为基准。设$Q$为所有问题集合,$A_i$为第$i$个问题给出的答案集合,$G_i$为第$i$个问题的标准答案集合,相关指标的计算如式(4)~式(6)所示:

$$P_{\text{Macro}}=\frac{1}{|Q|}\sum_{i=1}^{|Q|}P_i,\ P_i=\frac{|A_i\cap G_i|}{|A_i|} \tag{4}$$

$$R_{\text{Macro}}=\frac{1}{|Q|}\sum_{i=1}^{|Q|}R_i,\ R_i=\frac{|A_i\cap G_i|}{|G_i|} \tag{5}$$

$$\text{F1}_{\text{Average}}=\frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{2P_iR_i}{|P_i+R_i|} \tag{6}$$

3.3 实验结果

由于评测组织者只对验证集(初赛)公开了标准答案,因此本文相关实验只在验证集上进行测试并呈现基于本文方法的模型应用于测试集(复赛)上的结果,表4所示为评测前4名系统和本文方法的结果对比,其中,“评测第2名”是本文系统与其他系统融合的结果。从表4可以看出,本文方法略优于第4名系统。值得注意的是,评测前4名系统都采用模型融合的策略,本文提出单模型方法,在结构尽量简单的情况下也取得了较好的实验结果,从而验证了该系统的有效性。

表4　不同系统的性能比较

Table 4　Performance comparison of different systems　%

系统	初赛平均F1值	复赛平均F1值
评测第1名系统	70.69	73.55
评测第2名系统	70.41	73.08
评测第3名系统	67.23	70.45
评测第4名系统	66.39	67.68
本文系统	66.76	—

3.4　实验分析

表5所示为本文系统各个子模型的性能对比结果,从表5可以看出,实体提及识别模型的性能并不高,为了提高识别的召回率,本文对模型识别到的候选提及进行左右字符的扩展和删减,以增加候选实体的数量。单多跳分类模型的准确率只有89.13%,其余模型的准确率均在93%以上。表6所示为分类错误的具体样例,从表6可以看出,多跳问题实际上可以用单跳方法来解决,即别名提及可以通过实体链接得到其主题实体,而无需多余的三元组。

表5　不同子模型的性能比较

Table 5　Performance comparison of different sub-models　%

子模型	准确率
实体提及识别模型	87.09
单多跳分类模型	89.13
主谓宾分类模型	93.56
链式分类模型	94.23
关系抽取分类模型	94.80
实体链接模型	98.68

表6　多跳分类错误的示例

Table 6　Examples of multi-hop classification error

例句	SPARQL
大姚是哪里毕业的?	select ?y where { ?x <别名>"大姚". ?x <毕业院校> ?y . }
小马哥有哪些主要成就?	select ?y where { ?x <别名>"小马哥". ?x <主要成就> ?y. }

本文问答系统考虑到子模型的性能,并未将中文问题单独划分成单跳和多跳来处理,而是对所有问题统一进行了一遍单跳搜索,从而提高系统性能。由于单跳问题也有可能含有多个实体,因此该系统以是否链式来判断问句是链式问题还是多实体问题。此外,部分问句被分类为链式问题但不是多跳问题,因此,本文对链式问题增加一层约束判断,以降低因为模型分类错误而带来的影响。

在表4评测第1名系统[29]中,实体提及部分并未采用序列标注模型来识别,而是通过构建词典进行字符串匹配和外加命名实体识别器的方法,提高实体识别的精度。在实体链接部分,本文所提方法只保留候选得分最高的唯一实体,而没有增加候选实体的数量,导致召回率降低。另外,评测第1名系统没有对中文问题进行分类,而是统一地使用基于路径相似度匹配的策略,相比于只用实体关系和问题进行匹配的策略,该策略在语义上更准确,也减少了错误传播。因此,本文在模型融合时加入实体路径和问题匹配方法。在未来的研究中,可以借鉴评测第1名系统的优点来改进本文模型的系统性能。

为验证不同答案搜索模块对本文系统的影响,分别对某个模块进行屏蔽后进行实验,结果如表7所示。从表7可以看出,不同搜索模块对系统整体性能都有较大影响。如果将所有问题都当成简单问题来解决,系统的F1值只有52.02%。相较于简单问题,本文所提系统针对复杂问题中的链式和多实体问题的F1值提高了14.74个百分点(66.76%-52.02%),验证了该系统将中文问题设置不同的标签进行答案搜索的策略具有有效性。

表7　不同模块设置下的系统性能对比

Table 7　Comparison of system performance under different module settings　%

模块设置	平均F1值
只有统一单跳搜索模块	52.02
屏蔽链式问题搜索模块	57.01
屏蔽多实体搜索模块	61.68
不屏蔽任何模块	66.76

4　结束语

本文提出一种基于多标签策略进行答案搜索的中文知识图谱问答系统。对问句设置不同的标签,以利用不同的模块来搜索问句答案并解决复杂问题中的链式和多实体问题。在实体提及识别部分,提出将预训练语言模型BERT和BiLSTM网络相结合的方法。在关系抽取部分,摒弃复杂的模型结构而直接基于BERT模型实现问句和候选关系的相似度计算。在实体链接部分,借助XGBoost模型设计不同的特征以提高系统性能。实验结果表明,该系统可以有效解决中文知识图谱问答中不同类型的简单、链式和多实体问题。

虽然本文利用多标签的方法取得了较好的效果,但也存在一个弊端,即通过不同的分类模型对问句设置多个标签,将存在一个错误传递的过程,系统整体性能会受到多个子模块性能的影响。因此,今后将研究并实现一种端到端的方法来完成中文知识图谱问答。此外,NL2SQL技术可以将用户的自然语句直接转为可执行的SQL语句,如何有效地将NL2SQL技术引入到中文知识图谱问答任务中也是下一步的研究方向。

参考文献

[1] ALLAM A,HAGGAG M. The question answering systems: a survey[EB/OL]. [2019-11-10]. http://www. aast. edu/pheed/staffadminview/pdf_retreive. php? url=19955_401_32_QA%20Survey%20Paper. pdf&stafftype=staffpdf.

[2] MISHRA A,JAIN S K. A survey on question answering systems with classification[J]. Journal of King Saud University-Computer and Information Sciences, 2016, 28(3):345-361.

[3] MAO Xianling,LI Xiaoming. A survey on question and answering systems[J]. Journal of Frontiers of Computer Science and Technology,2012,6(3):193-207. (in Chinese)
毛先领,李晓明. 问答系统研究综述[J]. 计算机科学与探索,2012,6(3):193-207.

[4] BOLLACKER K, EVANS C, PARITOSH P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]//Proceedings of 2008 ACM SIGMOD International Conference on Management of Data. New York,USA:ACM Press,2008:1247-1250.

[5] SUCHANEK F M, KASNECI G, WEIKUM G. Yago: a core of semantic knowledge[C]//Proceedings of the 16th International Conference on World Wide Web. New York,USA:ACM Press,2007:697-706.

[6] AUER S,BIZER C,KOBILAROV G,et al. DBpedia: a nucleus for a Web of open data[M]. Berlin,Germany: Springer,2007.

[7] NIU Xing,SUN Xinruo,WANG Haofen, et al. Zhishi. me-weaving Chinese linking open data[C]//Proceedings of International Semantic Web Conference. Berlin,Germany: Springer,2011:205-220.

[8] PAN J Z. Knowledge extraction from Chinese wiki encyclopedias[J]. Journal of Zhejiang University-Science C(Computers & Electronics),2012,13(4):268-280.

[9] DU Zeyu, YANG Yan, HE Liang. Question answering system of electric business field based on Chinese knowledge map[J]. Computer Applications and Software, 2017,34(5):159-165. (in Chinese)
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):159-165.

[10] SHI Yu,GU Tianlong,BIN Chenzhong,et al. Question and answer system of tourist attractions based on knowledge graph[J]. Journal of Guilin University of Electronic Technology,2018,38(4):42-48. (in Chinese)
时雨,古天龙,宾辰忠,等. 基于知识图谱的旅游景点问答系统[J]. 桂林电子科技大学学报,2018,38(4):42-48.

[11] PETERS M, NEUMANN M, IYYER M, et al. Deep contextualized word representations[C]//Proceedings of the 9th American Chapter of the Association for Computational Linguistics. New Orleans, USA: Association for Computational Linguistics,2018:2227-2237.

[12] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 10th American Chapter of the Association for Computational Linguistics. New Orleans, USA: Association for Computational Linguistics, 2019:4171-4186.

[13] BERANT J,CHOU A,FROSTIG R,et al. Semantic parsing on freebase from question-answer pairs[C]//Proceedings of 2013 Conference on Empirical Methods in Natural Language Processing. Washington D. C., USA: IEEE Press,2013:1533-1544.

[14] CAI Q,YATES A. Semantic parsing freebase:towards open-domain semantic parsing[C]//Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics. Washington D. C. ,USA:IEEE Press,2013: 328-338.

[15] REDDY S,LAPATA M,STEEDMAN M. Large-scale semantic parsing without question-answer pairs[J]. Transactions of the Association for Computational Linguistics,2014,2:377-392.

[16] TAU YIH W,WEI CHANG M,HE X D,et al. Semantic parsing via staged query graph generation: question answering with knowledge base[EB/OL]. [2019-11-10]. https://www. aclweb. org/anthology/P15-1128. pdf.

[17] 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. Washington D. C., USA: IEEE Press,2014:956-966.

[18] BAST H, HAUSSMANN E. More accurate question answering on freebase[C]//Proceedings of the 24th ACM International Conference on Information and Knowledge Management. New York,USA:ACM Press,2015:1431-1440.

[19] BORDES A, WESTON J, USUNIER N. Open question answering with weakly supervised embedding models[M]// DAELEMANS W,GOETHALS B,MORIK K. Machine learning and knowledge discovery in databases. Berlin, Germany:Springer,2014.

[20] BORDES A,CHOPRA S,WESTON J. Question answering with subgraph embeddings[EB/OL]. [2019-11-10]. https://arxiv. org/abs/1406. 3676.

[21] BORDES A,USUNIER N,CHOPRA S,et al. Large-scale simple question answering with memory networks[EB/OL]. [2019-11-10]. https://arxiv. org/abs/1506. 02075.

[22] DONG Li, WEI Furu, ZHOU Ming, 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. Stroudsburg, USA: Association for Computational Linguistics,2015:260-269.

[23] HAO Yanchao,ZHANG Yuanzhe,LIU Kang,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. Stroudsburg, USA: Association for Computational Linguistics,2017:221-231.

[24] LAI Yuxuan,FENG Yansong,YU Xiaohan,et al. Lattice CNNs for matching based Chinese question answering[C]// Proceedings of AAAI Conference on Artificial Intelligence. New York,USA:ACM Press,2019:6634-6641.

[25] LAI Yuxuan, LIN Yang, CHEN Jiahao, et al. Open domain question answering system based on knowledge base[EB/OL]. [2019-11-10]. http://tcci. ccf. org. cn/conference/2016/papers/319. pdf.

(下转第117页)

（上接第110页）

[26] LAI Yuxuan, JIA Yanyan, LIN Yang, et al. A Chinese question answering system for single-relation factoid questions[EB/OL]. [2019-11-10]. http://tcci.ccf.org.cn/conference/2017/papers/2003.pdf.

[27] ZHOU Botong, SUN Chengjie, LIN Lei, et al. LSTM based question answering for large scale knowledge base [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2018, 54(2): 286-292. (in Chinese)
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.

[28] SUN Zhaoyu, SONG Lei, YU Jiaming. A QA search algorithm based on the fusion integration of text similarity and graph computation[EB/OL]. [2019-11-10]. http://ceur-ws.org/Vol-2242/paper13.pdf.

[29] LUO Jinchang, YIN Cunxiang, WU Xiaohui, et al. Chinese knowledge base question answering system based on hybrid semantic similarity [EB/OL]. [2019-11-10]. https://conference.bj.bcebos.com/ccks2019/eval/webpage/pdfs/eval_paper_6_1.pdf. (in Chinese)
骆金昌，尹存祥，吴晓晖，等. 混合语义相似度的中文知识图谱问答系统[EB/OL]. [2019-11-10]. https://conference.bj.bcebos.com/ccks2019/eval/webpage/pdfs/eval_paper_6_1.pdf.

[30] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[31] LAFFERTY J, MCCALLUM A, PEREIRA F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data[EB/OL]. [2019-11-10]. https://repository.upenn.edu/cgi/viewcontent.cgi?article=1162&context=cis_papers.

[32] CHEN T, GUESTRIN C. XGBoost: a scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2016: 785-794.

[33] DUAN N. Overview of the NLPCC-ICCPOL 2016 shared task: open domain Chinese question answering [M]// LIN C Y, XUE N W, ZHAO D Y, et al. Natural language understanding and intelligent applications. Berlin, Germany: Springer, 2016.

编辑　吴云芳