



基于多特征实体消歧的中文知识图谱问答

张鹏举, 贾永辉, 陈文亮

(苏州大学 计算机科学与技术学院, 江苏 苏州 215006)

摘要: 问答系统应用于人工智能、自然语言处理和信息检索领域获得了较好的效果,知识图谱问答(KBQA)作为其中的重要组成部分,是一项极具挑战性的自然语言处理任务。然而,目前常见的中文KBQA系统对于实体链接的实体消歧部分并没有给出很好的解决方法。提出一种基于多特征实体消歧的中文KBQA系统,通过结合实体自身的知名度特征、问句与实体关系的语义相似度特征、问句与实体的字符相似度特征和语义相似度特征,构建多特征实体消歧模型,提高实体链接准确率,为系统的问句分类和最优路径选取提供更准确的主题实体,从而提升系统性能。实验结果表明,该系统在CCKS2019-CKBQA评测数据的验证集上平均F1值为72.08%,其中采用多特征消歧模型的实体链接准确率达到90.84%,较使用知名度消歧模型和评测大赛第1名分别提升6.35和0.11个百分点。

关键词: 实体链接;实体消歧;主题实体;知识图谱问答;问答系统;问句分类;最优路径选取

开放科学(资源服务)标志码(OSID):



中文引用格式:张鹏举,贾永辉,陈文亮.基于多特征实体消歧的中文知识图谱问答[J].计算机工程,2022,48(2):47-54.

英文引用格式:ZHANG P J, JIA Y H, CHEN W L. Chinese knowledge based question answering based on multi-feature entity disambiguation[J]. Computer Engineering, 2022, 48(2): 47-54.

Chinese Knowledge Based Question Answering Based on Multi-feature Entity Disambiguation

ZHANG Pengju, JIA Yonghui, CHEN Wenliang

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

[Abstract] The application of question answering system to the fields of artificial intelligence, natural language processing and information retrieval has got excellent results. Knowledge Based Question Answering (KBQA) is an important part of question answering, and is a challenging natural language processing task. The commonly used Chinese KBQA systems do not provide a satisfying entity disambiguation solution for entity linking. To address the problem, this paper proposes a Chinese KBQA system based on multi-feature entity disambiguation. It jointly utilizes the entity's own popularity features, semantic similarity features of question and entity relations, character similarity features of question and entity, and semantic similarity features of question and entity, so as to implement entity disambiguation and improve entity linking. On this basis, the proposed system can provide more accurate subject entities for the question classification part and the optimal path selection part of the system to improve system performance. The experimental results show that the average F1 value of the proposed system on the verification set of CCKS2019-CKBQA evaluation data reaches 72.08%. Its entity linking module based on the multi-feature disambiguation model displays an accuracy of 90.84%, which is 6.35 percentage points higher than the module based on the popularity disambiguation model and 0.11 percentage points higher than the top 1 in CCKS2019-CKBQA evaluation competition.

[Key words] entity linking; entity disambiguation; subject entity; Knowledge Based Question Answering (KBQA); question answering system; question classification; optimal path selection

DOI: 10.19678/j.issn.1000-3428.0060438

0 概述

随着互联网信息资源激增,传统的搜索引擎无论从效率还是准确率上,都难以满足用户精准搜索

信息的需求。因此,问答系统被提出并迅速发展,其应用于人工智能、自然语言处理和信息检索领域获得了较好的效果,是目前具有较大发展前景的研究热点^[1]。而在问答系统中,知识图谱问答(Knowledge

基金项目:国家自然科学基金(61876115)。

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

收稿日期:2020-12-30 修回日期:2021-02-08 E-mail: pjzhang_stu@qq.com

Based Question Answering, KBQA) 是重要组成部分。

知识图谱问答系统的相关研究备受瞩目,与知识图谱的快速发展有密切关系。知识图谱由谷歌于2012年5月17日提出,其初衷是为了提高搜索引擎性能,改善用户的搜索质量以及搜索体验。知识图谱^[2]旨在描述真实世界中存在的各种实体或概念及其关系,构成一张巨大的语义网络图。在知识图谱中,节点代表实体或概念,边则表示属性或关系。目前知识图谱使用较为广泛的存储框架为资源描述框架(Resource Description Framework, RDF),表示形式一般用 SPO (Subject-Predicate-Object) 三元组表示,即“主语-谓语-宾语”。其中,“主语”一般为实体,“谓语”一般为关系或者属性,“宾语”一般为实体或者属性值。整个三元组表征了实体与实体之间的信息以及实体与自身属性之间的信息。

KBQA 系统的工作流程包含多个步骤。首先对于不同类别的问题需要进行分类处理。例如对问句“球星姚明的妻子是谁?”(涉及1个三元组)与“球星姚明妻子的星座是什么?”(涉及2个三元组)属于2种不同类型的问句,需要进行分类处理。其次要进行实体链接,即对问句进行实体识别与实体消歧。在实体链接中先要识别出问句中对应的主题实体提及,再从实体提及对应的所有候选实体中确定问句对应的唯一正确实体,最终完成实体链接。例如从问句中识别出“姚明”并链接到知识库中的实体节点“<姚明_(中联公司董事长兼总经理)>”。接着要对问句进行关系抽取,得到关系“妻子”“星座”,完成主题实体对应的关系抽取。最后在获得主题实体及其对应的关系后,进行三元组搜索“<姚明_(中联公司董事长兼总经理)>---妻子---叶莉_(中国著名篮球运动员)---星座---天蝎座_(占星学)”,得到“天蝎座_(占星学)”作为答案^[3],完成最终的问答。

对于知识图谱问答系统,实体链接是至关重要的,只有确定了主题实体,才能根据实体对应的关系、属性三元组进行推理、判断,从而得到最终的答案。一旦实体链接出错,问答系统的后续工作就没有任何意义。实体链接一般分为主题实体识别和实体消歧2个步骤。实体识别模块可以采用序列标注模型和规则匹配结合的方法来进行,相对而言比较容易,并且还能取得较好的效果。然而,实体消歧较难取得很好的效果,这是因为单单从实体本身的信息来看,并不能完全确定问句对应的最优实体。例如实体提及“姚明”对应知识库中就有2个实体:“<姚明_(中联公司董事长兼总经理)>”和“<姚明_(陕西省城固县盐务局副局长)>”,而对于问句“姚明的职业生涯最高得分是多少?”,要进行最优实体的选取就难以下手。因此,实体消歧部分是实体链接的关键。

近年来,在很多大会评测比赛中都有单独的实体

链接任务。然而,知识图谱问答中的实体链接与这些发展成熟的实体链接却有所差别。因为在正常的实体链接任务中,会给出实体对应的描述文段,进而从文段中抽取重要的信息进行实体消歧,但是在知识图谱问答中,并没有实体对应的描述文档来帮助进行实体消歧,而只能借助实体对应的三元组信息。针对这一问题,本文构建一个多特征实体消歧模型,通过考虑实体知名度特征和问句与实体的多方面特征优化实体消歧过程,并在此基础上构建一个完整的知识图谱问答系统。

1 相关工作

1.1 实体链接

实体链接^[4]是指将文档中出现的文本片段(即实体提及)链向特定知识库中相应条目的过程,也被称作命名实体链接,其采用的知识库一般为较全面、较具体的知识库,如 TAP、维基百科等。

实体链接包含实体识别和实体消歧2项关键技术。实体识别旨在从文档中识别出可能链向知识库中特定条目的实体提及,也被称作命名实体识别。由于自然语言中普遍存在一词多义和别名现象,通过所识别的实体提及在多数情况下并不能唯一确定其所指向的实体,因此需要利用实体消歧技术,根据给定实体提及所在上下文,确定其所指向的实体。目前实体消歧大多采用分类方法、机器学习排序方法、基于图的方法、模型集成方法等。

对于中文实体链接任务,主要以中国计算机学会(CCF)或者中国中文信息学会举办的大会比赛评测任务为主,常见的有中国计算机学会国际自然语言处理与中文计算会议(NLPCC)与全国知识图谱与语义计算大会(CCKS)的实体链接评测任务,各参赛队伍使用机器学习排序方法居多。

1.2 知识图谱问答

知识图谱问答(KBQA)是一个具有吸引力和挑战性的任务,其最早伴随 Freebase^[5]、DBpedia^[6]、YAGO^[7]等大型知识库的出现而出现在人们视野中。简而言之,KBQA 任务定义为:以客观事实为基础,将自然语言问题作为输入、知识图谱中的实体或者属性值作为输出的一个综合性较高的任务。总体上 KBQA 方法分为两大类,一类是基于信息检索的方法,另一类是基于语义解析的方法。

基于信息检索的 KBQA 方法主要是通过构建不同的排序模型对检索出的候选答案信息进行排序,得到最优候选答案来完成 KBQA。Bordes 等^[8]提出先采用语义词向量嵌入的方法来表示问句和答案信息,再通过编码计算其相似度来进行知识图谱问答。此后,随着神经网络的兴起,越来越多的研究者采用神经网络模型(例如卷积神经网络、循环神经网络

络等)编码问句和答案来计算相似度,并且获得了不错的效果^[9-11]。

基于语义解析的KBQA方法^[12-14]相对比较传统,其通过对问句进行语义解析,得到对应的结构化查询图或者逻辑表达式,然后转化为结构化的查询语言(例如SPARQL)查询知识库得到最终答案。但是由于语义解析需要复杂的推理过程以及大量的手工规则特征,因此基于语义解析的方法实现起来颇有难度。

KBQA领域的研究最早是由国外KBQA研究者引领,并且他们提出的各种系统研究在Simple Questions数据集和Webquestions数据集上取得了不错的效果,在工业界也有很成熟的系统“start”。反观中文KBQA起步较晚,目前也主要是以NLPCCC和CCKS这2个公开会议所举办的评测任务比赛为主,同时大部分参与评测的队伍采用的是基于信息检索的方法。

2 实体链接模型

知识图谱问答的实体链接模型分为实体识别和实体消歧2个部分,后者依靠前者所识别出来的实体提及对应的知识库实体进行消歧,通过将实体链接到知识库中完成实体链接。因此,实体识别部分必须达到很高的性能,才能够降低错误传播,防止实体消歧时对错误的实体提及对应的知识库实体进行消歧。对于实体识别模型,本文采用性能较好的BERT(Bidirectional Encoder Representations from Transformers)预训练模型作为基础模型。而在实体消歧部分,由于基于KBQA的实体链接任务并没有实体的描述文段,因此只能借助知识库和问句信息来进行消歧。本文采用多特征集成模型来进行实体消歧。

2.1 BERT预训练模型

BERT预训练语言模型^[15]是一个基于Transformer神经网络单元的双向语言模型,其结构如图1所示。由于Transformer是基于能够彻底捕捉语句中每个词之间时序信息的self-attention编码器,因此在句子级别的任务上,BERT能够实现真正意义上的前向、后向的双向信息传递,从而获得更高的性能和更好的效果。整个模型的输入由词向量输入、位置向量输入、句子分段向量输入3个部分构成。整个句子的首部和尾部分别有特殊的标记[CLS]和[SEP],这2个标记用来区别不同的2个句子。模型的输出是[CLS]、[SEP]以及每个词经过Transformer编码器得到的语义编码向量。给定一个自然语言句子的输入序列 $Q=(q_1, q_2, \dots, q_n)$,经过BERT的预处理和向量化后得到其对应的句子输入向量 $E=[CLS, E_1, E_2, \dots, E_n, SEP]$,再经过12层或者24层编码器得到最后的输出向量 $H=(H_0, H_1, \dots, H_n)$ 。经过预训练的BERT模型包含很

强的上下文关联语义特征,再经过微调即可用于分类、序列标注、阅读理解等多个任务上。由于BERT能够得到较好的训练效果,因此其在工业界被广泛应用。

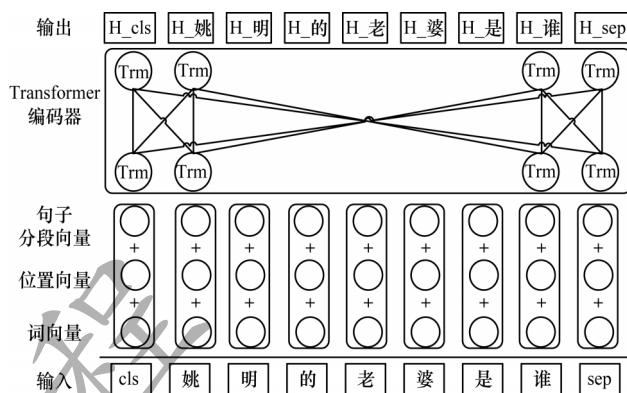


图1 BERT模型结构

Fig.1 Structure of BERT model

2.2 问句与路径语义相似度模型

问句与路径语义相似度模型指的是在完成实体链接后,确定该实体与问句语义最相关的关系所使用的模型。例如对于问句“球星姚明的老婆的星座是什么?”,完成实体链接得到主题实体“<姚明_(中联公司董事长兼总经理)>”,需要确定该实体对应的最优关系“妻子”和第2个三元组的最优关系“星座”,这里没有采用关系抽取的方法来进行,而是结合BERT预训练模型擅长处理句子级任务的特点,构建以主题实体为核心的三元组候选路径与问句组成句子对“球星姚明老婆的星座是什么?[SEP]<姚明_(中联公司董事长兼总经理)>---<妻子>---<星座>---<PAD>”,其中“<PAD>”表示将实体泛化所用的特殊标签。采用相似度模型选出最优路径,进而选出最优关系。这样处理的原因如下:

1) 能够避免错误传递。因为进行关系抽取时只抽1个关系,那么对于第2个关系来说,第1个关系识别错误就会造成错误传递。

2) 三元组顺序不同。有些问句格式对应的三元组内实体顺序是反向三元组。例如问句“万岛之国指的是哪个国家?”,其在知识图谱中对应的三元组是“<挪威>---<别称>---<万岛之国>”,然而“<挪威>”是该问句的答案,那么正确的标注就是“<PAD>---<别称>---<万岛之国>”,这时三元组的顺序已经反向,那么再使用关系抽取识别出“<别称>”在知识图谱中进行查找时,由于顺序的问题,必然得不出答案。

3) BERT模型的特点。本文使用的基础模型是BERT,而在BERT模型训练中的第2个任务是句子下一句的预测,这个任务就是为了更好地理解 and 处理2个句子中的信息。因此,BERT更擅长处理基于句子与句子的任务。本文将候选三元组组成一个短句,结合问句形成一个句子与句子之间的语义相似度计算任务,例如问句“万岛之国指的是哪个国家”

与三元组路径“<PAD>--<别称>--<万岛之国>”的语义相似度计算任务,这样能够更契合BERT模型的训练和预测,从而使相似度模型获得较好的效果。

在构建模型的训练语料时,将问句作为“SEN1”,将答案路径作为“SEN2”构建句子对,再把含有正确答案路径的句子对标注为“1”,错误的标注为“0”,并且以正、负例比为1:10进行构建。得到训练语料后,通过微调BERT分类模型进行训练得到问句与路径相似度模型,如图2所示。在最终预测时取出模型最后一层隐层,经过分类层得到标签为“1”的各条答案路径的向量,再通过softmax得到每个问句的得分,选取top1作为最优答案路径,完成问句与最优答案路径的选取。

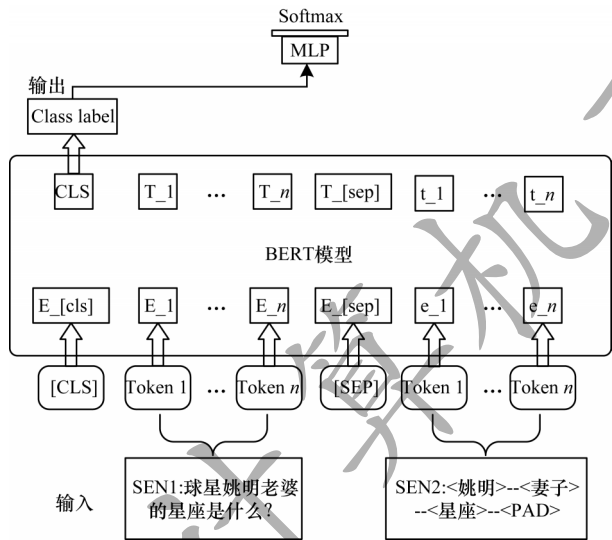


图2 问句与答案路径相似度模型结构

Fig.2 Structure of similarity model of question and answer path

2.3 实体识别

实体识别指的是从问句中识别出主题实体提及,例如从问句“姚明的老婆是谁?”中识别出“姚明”这个主题实体提及。本文采用序列标注模型作为实体识别的基础模型,使用CKKS2019-CKBQA数据集含有SPARQL标注语料的训练数据集。面对含有标注的问句“姚明的老婆是谁?”,其对应的SPARQL语句为“select ? x where {<姚明_(中联公司董事长兼总经理)> <妻子> ? x}”,从中对实体“<姚明_(中联公司董事长兼总经理)>”进行泛化处理,得到“姚明”作为句子对应的实体提及。然后根据序列标注模型的数据处理方法,将问句中“姚明”对应的位置标记为“BI”,把其他非提及部分标记为“O”,依照“BIO”标记进行序列标注模型训练。

本文将BERT语言模型和条件随机场(Conditional Radom Field,CRF)^[16]相结合训练,并预测每个字符对应的标签,如图3所示。首先通过BERT语言模型得到富含语义信息的每个词的上下文表示,然后通过CRF

模型预测标签序列的正确性。在完成模型训练后,根据用户问句进行实体识别,得到问句对应的实体提及。

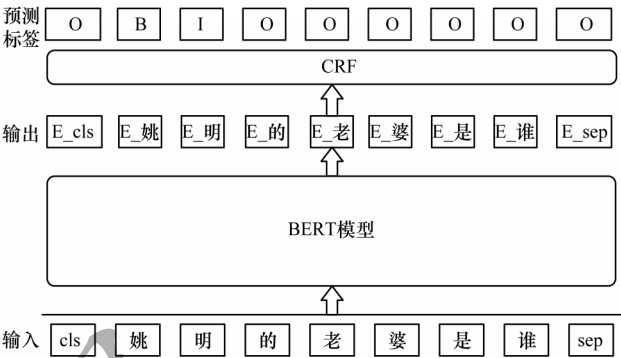


图3 实体识别模型结构

Fig.3 Structure of entity recognition model

2.4 实体消歧

完成实体识别后得到主题实体提及,例如“姚明”,但是在知识库中,“姚明”这一个提及在知识库中可能对应2个实体:“<姚明_(中联公司董事长兼总经理)>”和“<姚明_(陕西省城固县盐务局副局长)>”,那么对于问句“球星姚明的老婆是谁?”,其主题实体显然是前者,因此,最终通过实体消歧得到的实体为“<姚明_(中联公司董事长兼总经理)>”。

研究者通常使用基于实体知名度的方法来进行实体消歧。实体知名度指的是该实体在知识图谱中对应的知名程度(热度)得分。对于问句“球星姚明的老婆是谁?”,采用知名度方法进行实体消歧得到的最终结果就是正确实体“<姚明_(中联公司董事长兼总经理)>”,但是对于问句“姚明副局的执政理念是什么?”,如果采用知名度得分的话显然是错误的。因此,本文提出一个基于多特征的实体消歧模型,即结合知名度特征、问句与实体关系的语义相似度特征、问句与实体的字符相似度特征、问句与实体的语义相似度特征这4个特征的语义模型,如图4所示。

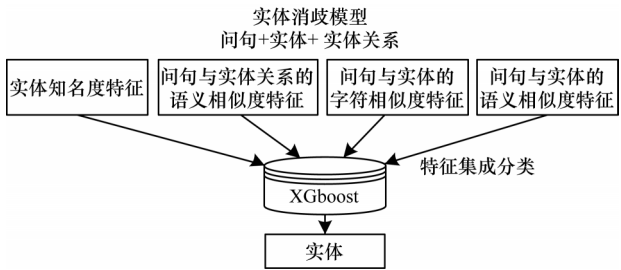


图4 多特征实体消歧模型结构

Fig.4 Structure of multi-feature entity disambiguation model

1) 知名度特征

采用知名度特征应获取实体对应的知名度(热度)。在开放领域的问句中,人们一般所问实体的知名度比重都比较高,因此,对于开放领域的知识图谱问答,实体的知名度是必要的。对于较为完备的知识图谱,都会有一个实体对应的知名度得分排序表,

这样就可以根据这个知名度排序表得到对应的排序特征。例如“<姚明_(中联公司董事长兼总经理)>”和“<姚明_(陕西省城固县盐务局副局长)>”对应排名分别为1和5。

2) 问句与实体关系的语义相似度特征

在实体消歧时,如何利用好实体的上下文十分重要,而在问句中获取上下文信息最好的方法就是找到实体信息中与问句关联的关系(属性)或者实体解释。但是对于知识图谱问答而言,知识图谱中并没有每个实体对应的具体描述文段,所以,只能采用知识图谱中每个实体对应的关系或者属性来进行消歧。首先要选出实体所有关系中,与问句关联度最高的关系(属性)。本文通过上文所提到的问句与路径的语义相似度模型来进行最优关系的预测,构建每个实体对应的三元组路径,然后分别通过语义相似度模型得到最后一层分类层的语义向量,选取正确标签上的每条路径对应特征向量,将其作为每条路径得分。从所有路径中选取得出得分为top1的路径作为实体对应的最优路径得分,即为最优关系的得分。例如,对于问句“球星姚明的老婆是谁?”的候选实体“<姚明_(中联公司董事长兼总经理)>”的所有关系,关系“妻子”的相似度模型得分(0.97)最高,因此,就将该得分作为“<姚明_(中联公司董事长兼总经理)>”的实体关系与问句的语义相似度特征。

3) 实体与问句的字符相似度特征

采用问句与实体的字符相似度特征,是因为对于一些含有实体别名的问句,必须依靠实体的字符相似度来进行消歧处理。例如对于问句“小说中风清扬的徒弟是谁?”,“风清扬”在知识图谱中链接到“马云”“风清扬”,然而在知识图谱中,“马云”的知名度得分大于“风清扬”,另外根据问句与实体的最优关系语义相似度特征来看,2个实体都有“徒弟”这个最优关系,在这种情况下,就会选择错误的实体“马云”。为避免这种错误的情况,就需要问句与实体的字符相似度特征来进行辅助修正。

4) 实体与问句的语义相似度特征

就实体本身而言,其在问句中就包含有对应的问句语义信息,因此,本文计算问句与实体的语义相似度作为问句与实体的语义特征。这里指的语义特征是将问句与实体共同映射到一个向量空间,分别将问句和实体向量化来计算语义相似度实现的。本文同样使用之前提到的问句与路径的语义相似度模型来实现同一个向量空间的映射,但与之不同的是将问句对应的路径改为“<实体><pad><pad>”的形式,用来代替实体部分的输入,问句部分输入保持不变。最终得到模型的分层输出向量,取标签为“1”的位置上的所有实体对应的向量得分,分别作为每个候选实体与问句的语义相似度特征。

通过上述4个特征能够获得较为完整的关于问句的实体信息。然后通过性能较好的特征拟合模型

XGBOOST^[17]对4个特征进行拟合训练。在预测时,本文采用二分类方法对每个实体进行得分计算(标签为正确标签的概率得分),选择得分top1的实体作为最终实体消歧得到的实体,同时也作为实体链接得到的最终实体。

3 中文知识图谱问答系统

本文设计了一个基于多特征实体消歧的中文知识图谱问答系统,如图5所示。整个系统主要分为3个模块:问句预处理模块,问句实体链接模块,最优答案路径筛选模块。问句处理模块包括对问句的预处理、问句分类。问句实体链接模块包括实体识别和实体链接2个部分。最优答案路径筛选模块包括对规则问句的路径选取、对单跳问句答案路径的选取、对两跳链式问句的答案路径选取以及对单、多跳难以分类问句的答案路径选取。上文已经阐述了实体链接模块的工作流程,本节将介绍不同类型问句的分类处理和最优答案路径的选取。

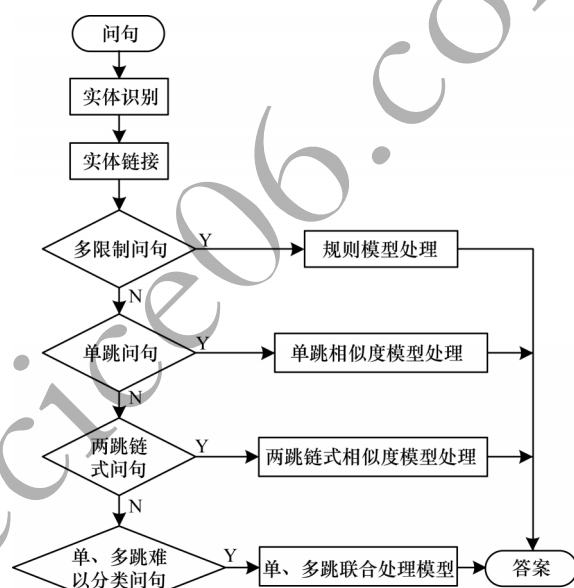


图5 基于多特征实体消歧的中文KBQA系统工作流程

Fig.5 Workflow of Chinese KBQA system based on multi-feature entity disambiguation

3.1 多限制问句

多限制问句指的是在一些特定的领域,含有很强的逻辑性或者规则的问句。将问句中对应的2个实体三元组的尾实体或者属性值有交集的问句作为多限制问句类型1,将问句中对应的2个实体三元组的头实体有交集的问句作为多限制问句类型2。这2类问句是CKKS2019-CKBQA评测任务数据集中逻辑性较强的问句。按照这样的逻辑性和规则,将问句分为多限制问句和非多限制问句。但在实际应用场景中还有更多种类的问句,但目前就实验数据,对于多限制类型问句,本文就只分为这2个类型。

多限制问句类型 1:“陈奕迅和王菲共同演唱了那首歌曲”。

多限制问句类型 2:“哈佛大学出了哪些物理学家?”。

由于多限制问句具有强逻辑性的特点,因此解决这类问句也变得十分清晰。在完成实体链接后,得到问句对应的 2 个实体,按照这 2 种类型问句的三元组特点,分别进行对应的三元组的规则性查找就能完成问答。

3.2 非多限制问句

在非多限制问句中,根据数据集的特点,本文将问句细分为单跳问句、两跳链式问句和单、多跳难以分类的问句。虽然分为 3 种问句类型,但处理方法基本一致。

3.2.1 单跳问句

单跳和多跳问句的定义为:只涉及一个三元组的问句称作单跳问句,涉及 2 个及以上三元组的问句称作多跳问句,如表 1 所示。因此,可以将这类问题当作二分类问题来处理。由于本文采用的是 CCKS2019-CKBQA 评测任务中的数据,每个问句都含有对应 SPARQL 结构化查询语句的标注数据,所

以按照标注语句来构建二分类模型的训练数据,将含有一个三元组打上标签“0”作为单跳问句,含有 2 个或 2 个以上三元组打上标签“1”作为多跳问句。最后利用 BERT 预训练语言模型进行模型的二分类微调训练。在预测时,采用模型的最后一层隐层输出中的[CLS]向量作为整个句子的语义分类向量,将其经过一个多层感知机(Multi-Layer Perceptron, MLP)分类(如图 6 所示),得到最终的分类结果,其中,标签为“1”表示多跳问句,标签为“0”表示单跳问句。除使用 BERT 二分类模型进行预测外,还需要结合实体链接的结果进行修正。从实体链接结果中找出只包含一个实体的问句,将这些问句与 BERT 二分类模型预测的单跳问句进行求交集处理,完成单跳问句的分类,其他问句作为多跳问句。完成问句分类与实体链接后,按照上文问句与路径语义相似度模型,根据问句对应的唯一主题实体,获得主题实体对应的候选答案路径。最后再通过问句与路径语义相似度模型得到最优答案路径,确定答案三元组,根据三元组检索答案完成单跳问句的问答。

表 1 单、多跳问句示例

问句类型	示例	SPARQL 语句
单跳问句	姚明的妻子是谁?	Select ?x where {<姚明><妻子> ?x
多跳问句	姚明妻子的星座是什么?	where {<姚明><妻子> ?x.?x <星座> ?y}

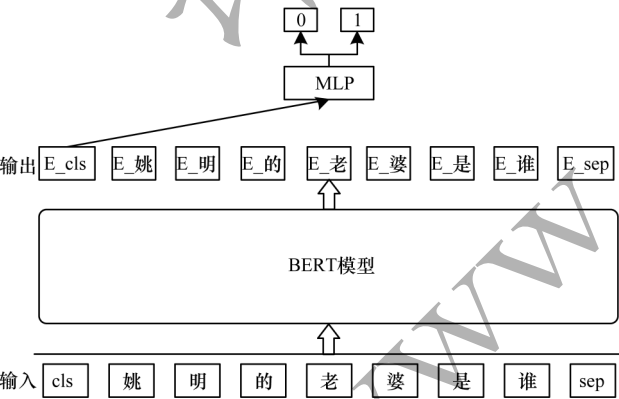


图 6 BERT 模型问句二分类示例

Fig.6 Example of two-classification of question by BERT model

3.2.2 两跳链式问句与难以分类问句

两跳链式问句指的是那些包含 2 个顺序排列的三元组对应的问句(上一个三元组的尾实体是下一个三元组的头实体),例如 SPARQL 语句为“select ? x where {<姚明_(中职联公司董事长兼总经理)> <妻子> ? y.? y <星座> ? x.}”的问句“姚明妻子的星座是什么?”。同构建单跳问句分类数据集一样,符合链式问句的 SPARQL 的问句标记为“0”,不符合的标记为“1”。同上文提到

的一样,使用 BERT 二分类模型,进行微调得到最终的链式问句分类模型。同时也采用实体链接的结果进行修正,得到最终的两跳链式问句,剩下的问句作为单、多跳难以分类问句。完成问句分类后,根据实体链接得到主题实体,构建候选答案路径,通过问句与路径语义相似度模型选出最优路径,完成答案两跳三元组确定,最终完成两跳链式问句的问答。这里需要说明的是,两跳链式问句采用的相似度模型与处理单跳问句是不同的模型,其根本的区别在于采用了不同的数据集进行训练,但预测方式基本一致。

在完成链式问句问答后,就只剩下难以分类问句的处理。由于这部分问句的数量只占总问句的 9.21%,数据量较小,因此对于这一部分问句,将使用单、多跳联合处理模型进行语义相似度匹配统一处理,不再进行细分。联合处理模型是通过包含单、多跳数据集训练得到的模型,可以处理单、多跳混合问句最优答案路径的选取。同样,在得到实体链接产生的主题实体后,候选路径的生成也是单、多跳路径同时生成的。最终通过相似度模型得到最优的候选答案路径作为最终的答案三元组,完成问答。

4 实验

4.1 实验数据

本文使用 CCKS2019-CKBQA 公开的评测数据集进行实验,其中包括3份数据集和1份知识图谱。评测数据由北京大学和恒生电子股份有限公司人工构建与标注,包括2 298条训练集、766条验证集和766条测试集。在问答数据集中,验证集和测试集分别是比赛初赛和复赛所用的数据集。知识图谱使用的是北京大学构建的知识图谱 PKUBASE,由41 009 141条实体三元组、13 930 117条实体提及三元组和25 182 627条实体类型三元组构成。在实验过程中,虽然 CCKS2019-CKBQA 数据集既包含简单问句又包含复杂问句^[18-19],但是数量较少,因此,使用 NLPCC2016-KBQA 的数据集^[20]作为额外的训练集训练模型(CCKS2019-CKBQA 评测比赛允许使用额外的公开数据集。参赛队伍同样使用了 NLPCC2016-KBQA 的数据集作为训练集进行模型训练)。

4.2 实验设置

本文使用的 BERT 预训练模型为基于 PyTorch 深度学习框架的 BERT-Base-Chinese 模型,其中共有12层编码器,隐层输出维度为768,中文最大句长设置为55。模型采用 Adam 优化器进行参数的更新和微调,初始学习率设置为 $5e-5$,采用大小为101的批量训练方法,dropout 设置为0.1,最大迭代次数为100次,设置每训练2轮进行开发集的验证。整个实验分为2个部分:

1) 使用知名度实体消歧模型的实体链接与使用多特征实体消歧的实体链接的对比实验。数据集为含有标注数据的766条测试集中的实体标注数据,该部分的实验指标为所有问句对应的实体链接的准确率 P 。设置所有问句个数为 N_a ,所有问句中实体链接正确的问句个数为 N_c ,则 P 计算公式如下:

$$P = \frac{N_c}{N_a} \quad (1)$$

2) 验证使用多特征的实体链接对 KBQA 系统性能提升的实验。采用766条测试集中的答案标注数据进行系统最终的性能实验。KBQA 系统部分评价指标为平均 F1 值 F_1^{Avg} 。设置问题集合为 Q , a_i 表示系统所给出的第 i 个问题的答案集, a_i^f 表示第 i 个问题的标准答案集, P_i 为第 i 个问题的答案准确率(如式(2)所示), R_i 为第 i 个问题的答案召回率(如式(3)所示),则 F_1^{Avg} 计算公式如式(4)所示:

$$P_i = \frac{|a_i \cap a_i^f|}{a_i} \quad (2)$$

$$R_i = \frac{|a_i \cap a_i^f|}{a_i^f} \quad (3)$$

$$F_1^{Avg} = \frac{1}{Q} \sum_{i=1}^{|Q|} \frac{2P_i R_i}{P_i + R_i} \quad (4)$$

4.3 实验结果与分析

表2展示了采用不同实体消歧模型的实体链接性能比较以及与评测比赛第1名的实体链接性能比较。从表中可以看出,采用多特征实体消歧模型的实体链接后,在实体链接模块,相比于采用知名度实体消歧的实体链接性能提升了6.35个百分点,同时相比于第1名的实体链接模型高出0.11个百分点,表明本文提出的采用多特征实体消歧模型能够很好地结合上下文信息和实体本身的信息并取得不错的性能。

表2 实体链接性能比较

实体链接模型	准确率
评测比赛第1名	90.73
知名度消歧模型	84.49
多特征消歧模型	90.84

表3展示了本文提出的系统与采用这个数据集进行评测比赛的前3名的系统的性能差异。根据平均 F1 值的比较,本文系统性能仅次于第2名,但是第1名和第2名分别在候选答案路径模块采用了特征集成与模型融合的方法,本文则是仅采用一个特征(模型)来进行候选答案路径的选取。另外,从单特征角度来看,本文提出的系统性能已经优于第1名单特征的系统性能(69.02%)^[19],因此表明本文构建的知识图谱问答系统已取得不错的效果。

表3 不同 KBQA 系统在最终测试集上的平均 F1 值

KBQA 系统	平均 F1 值
评测比赛第1名(特征集成)	73.55
评测比赛第2名(模型融合)	73.08
评测比赛第3名(单系统)	70.45
本文-多特征实体消歧(单系统)	72.08
本文-知名度实体消歧(单系统)	70.22

表3还表明了采用多特征实体消歧的系统性能要优于采用知名度实体消歧的系统。通过比较可以得出,采用多特征实体消歧从系统层面上提升了1.86个百分点,印证了采用多特征实体消歧的实体链接是十分有效的。但同时也可以看出,虽然实体链接部分提升了6.35个百分点,但是整体系统却只提升了1.86个百分点,这表明想要提升 KBQA 整个系统的性能单单从实体链接部分提升是不够的,除了实体链接之外,分类模块、候选答案路径计算模块也需要进一步优化。

5 结束语

本文设计一个基于多特征实体消歧的中文知识图谱问答系统。从知名度、字符和语义层面综合确定一个实体提及对于问句的最优实体,提升问句中实体链接的实体消歧性能,同时提出一种更契合BERT预训练模型的问句与路径语义相似度模型,准确抽取出问句对应的关系、属性,并最终经问句具体分类确定中文知识图谱问答中用户提出的问句。本文在实体链接之后采用基于检索的方法构建知识图谱问答系统,这种方法容易造成误差传递。下一步研究将基于语义解析方法构建一个结合检索和语义解析的中文知识图谱问答系统,从而避免这一问题。

参考文献

- [1] 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.
MAO X L, LI X M. A survey on question and answering systems[J]. Journal of Frontiers of Computer Science and Technology, 2012, 6(3): 193-207. (in Chinese)
- [2] 孙建军. 链接分析:知识基础,研究主体,研究热点与前沿综述——基于科学知识图谱的途径[J]. 情报学报,2014,33(6):659-672.
SUN J J. Link analysis: knowledge base, research subjects, a review of research hotspots and frontiers[J]. Journal of Information Science, 2014, 33(6): 659-672. (in Chinese)
- [3] 王思宇,邱江涛,洪川洋,等. 基于知识图谱的在线商品问答研究[J]. 中文信息学报,2020,34(11):104-112.
WANG S Y, QIU J T, HONG C Y, et al. Online commodity KBQA based on knowledge graph[J]. Journal of Chinese Information Processing, 2020, 34(11): 104-112. (in Chinese)
- [4] 罗念,杨燕,贺晖. 命名实体链接技术研究综述[J]. 计算机应用与软件,2016,33(12):6-10.
LUO N, YANG Y, HE Y. A survey of named entity linking technology[J]. Journal of Frontiers of Computer Applications and Software, 2016, 33(12): 6-10. (in Chinese)
- [5] BOLLACKER K, EVANS C, PARITOSH P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]//Proceedings of ACM International Conference on Management of Data. New York, USA: ACM Press, 2008: 1247-1250.
- [6] AUER S, BIZER C, KOBILAROV G, et al. DBpedia: a nucleus for a Web of open data[M]//ABERER K, CHOI K S, NOY N. The semantic Web. Berlin, Germany: Springer, 2007: 722-735.
- [7] 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.
- [8] BORDES A, WESTON J, USUNIER N. Open question answering with weakly supervised embedding models[C]//Proceedings of Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Berlin, Germany: Springer, 2014: 165-180.
- [9] 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. [S. l.]: Association for Computational Linguistics, 2015: 260-269.
- [10] 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. [S. l.]: Association for Computational Linguistics, 2017: 221-231.
- [11] LAI Y, FENG Y, YU X, et al. Lattice CNNs for matching based Chinese question answering[EB/OL]. (2019-02-25) [2020-11-20]. <https://arxiv.org/pdf/1902.09087.pdf>.
- [12] 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. [S. l.]: Association for Computational Linguistics, 2013: 1533-1544.
- [13] REDDY S, LAPATA M, STEEDMAN M. Large-scale semantic parsing without question-answer pairs[J]. Transactions of the Association for Computational Linguistics, 2014, 2(1): 377-392.
- [14] LAN Y, JIANG J. Query graph generation for answering multi-hop complex questions from knowledge bases[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. [S. l.]: Association for Computational Linguistics, 2020: 1-5.
- [15] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. [S. l.]: Association for Computational Linguistics, 2019: 4171-4186.
- [16] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
- [17] CHEN T, GUESTRIN C. XGBoost: a scalable tree boosting system[C]//Proceedings of ACM International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2016: 785-794.
- [18] SUN Z, SONG L, YU J. A QA search algorithm based on the fusion integration of text similarity and graph computation[C]//Proceedings of the Evaluation Tasks at the China Conference on Knowledge Graph and Semantic Computing. Tianjin, China: [s. n.], 2018: 89-94.
- [19] LUO J C, YIN C X, WU X H, et al. Question answering system base Chinese knowledge based on hybrid semantic similarity[C]//Proceedings of the Evaluation Tasks at the China Conference on Knowledge Graph and Semantic Computing. Hangzhou, China: [s. n.], 2018: 1-5.
- [20] 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: 942-948.