doi:10.11835/j.issn.1000.582X.2024.08.006

一种融合文本与知识图谱的问答系统模型

张佳豪,黄 勃,王晨明,曾国辉,刘 瑾
(上海工程技术大学 电子电气工程学院,上海 201620)

摘要:知识图谱是实现开放领域问答的关键技术之一,开放领域问答任务往往需要足够多的知识信息,而知识图谱的不完备性成为制约问答系统性能的重要因素。利用外部非结构化的文本与基于知识图谱的结构化知识相结合填补缺失信息时,检索外部文本的准确性和效率尤为关键,选取与问题相关度较高的文本可提升系统性能。相反,选取与问题相关性较弱的文本将引入知识噪声,降低问答任务的准确性。因此,设计了一种融合文本与知识图谱的问答系统模型,其中的文本检索器可充分挖掘问题和文本的语义信息,提高检索质量和查询子图的准确性;知识融合器将文本和知识库中的知识结合构建知识的融合表征。实验结果表明,相较对比模型,该模型在性能上存在一定优势。

关键词:问答系统;知识图谱;外部知识;文本检索;融合表征
中图分类号:TP183;TP391.1 **文献标志码:**A **文章编号:**1000-582X(2024)08-055-10

A question answering system model integrating text and knowledge graph

ZHANG Jiahao, *HUANG Bo*, *WANG Chenming*, *ZENG Guohui*, *LIU Jin*
(School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, P. R. China)

Abstract: Knowledge graph is one of the key technologies to realize question answering in open domain. Open domain question answering tasks often require enough knowledge information, and the incompleteness of knowledge graph becomes an important factor restricting the performance of question answering system. When combining external unstructured text with structured knowledge based on knowledge graphs to fill in missing information, the accuracy and efficiency of retrieving external texts are particularly critical, and selecting texts that are highly relevant to the problem can improve system performance. Conversely, selecting texts that are less relevant to the question will introduce knowledge noise, thereby reducing the accuracy of question answering tasks. Therefore, this paper designs a question answering system model that integrates text and knowledge graph, in which the text retriever can fully mine the semantic information of questions and texts to improve the quality of retrieval and the accuracy of query subgraphs. The knowledge mixer can combine knowledge from text and knowledge bases to build fusion representations of knowledge. The experimental results show that the proposed

收稿日期:2022-01-22
基金项目:科技创新 2030“新一代人工智能”重大项目(2020AAA0109300);国家自然科学基金青年项目(61802251)。
Supported by the Scientific and Technological Innovation 2030 Major Project of New Generation Artificial Intelligence (2020AAA0109300), and National Natural Science Foundation of China (61802251).
作者简介:张佳豪(1997—),男,硕士研究生,主要从事自然语言处理方向研究,(E-mail) 996765448@qq.com。
通信作者:黄勃(1985—),男,副教授,硕士生导师,主要从事人工智能方向研究,(E-mail) huangbosues@sues.edu.cn。

model has certain advantages in performance compared with the comparison models.

Keywords: question answering system; knowledge graph; external knowledge; text retrieval; fusion representation

开放领域问答[1-2]需要找到使用自然语言所描述问题的对应答案。当前开放领域的问答系统往往需要覆盖面足够广的知识库作支撑,而当今知识图谱的规模尚不足以作为开放领域问答系统的唯一知识源,其不完备性限制了问答系统性能。随着互联网发展,各类百科网站记载了越来越多领域的知识,以非结构化文本的形式呈现。陈丹琦等[3]首次将维基百科文本语料库引入开放领域问答。一方面,其拥有的知识量大、覆盖面广,并且规模日益增长;另一方面,其语言满足专业性和规范性,有利于转化为计算机易于存储的结构化形式。因此,文本语料库可作为不完备知识库的外部信息补充,与知识库相结合作为开放领域问答系统的知识源。图1显示了为回答无法直接从知识库中找到答案的问题需要结合非结构化文本信息的案例。

国内外有一些研究者设计了结合外部文本知识的知识图谱开放领域问答系统,虽然取得一些效果,但仍存在问题,导致效果未达预期。其中包括:1)检索文本的方法未曾涉及句子的语义信息,导致检索到的文本相关性被限制,无法充分挖掘文本中所蕴含与问句有关的信息,影响最终答案的准确性。例如Sun和Xiong等[4-5]利用词频信息检索文本,未涉及语义信息;2)知识图谱节点的表征未考虑差异化、邻接节点及边对其的重要性,使节点过于孤立,难以准确定位目标答案。例如,图卷积神经网络(graph convolution network, GCN)[6-8]在对查询子图节点的邻接节点进行卷积操作时使用相同的权重。

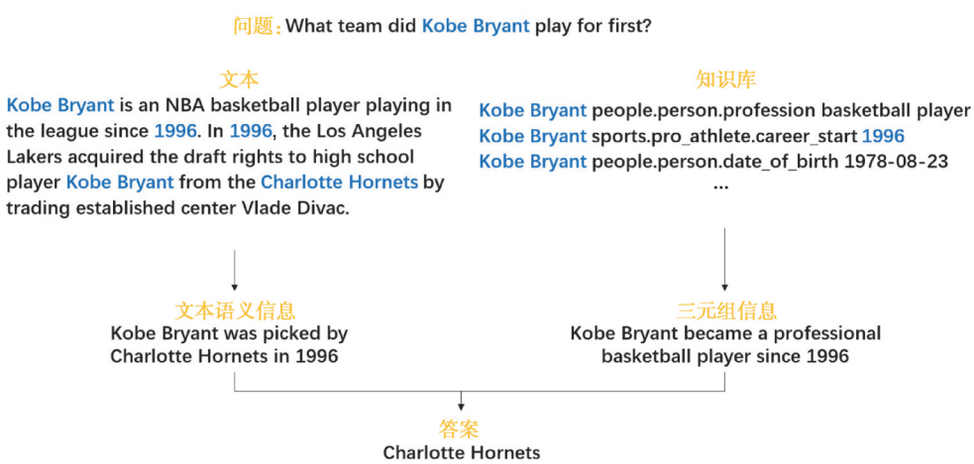


图1 1个来自WebquestionsSP数据集的问题示例

Fig. 1 An example of a question from the WebquestionsSP dataset

针对上述现有问答系统存在的2个问题,本文提出了新的融合文本与知识图谱的问答系统模型。针对问题1),模型设计基于语义信息进行精确检索的文本检索器,从问句的语义信息角度出发,在大型文本语料库中准确检索所有与问题相关的文本,为不完整的知识图谱提供推理依据和实体背景信息。针对问题2),本模型采用图注意力网络(graph attention network,GAT)[9-11]实现实体的表征和推理,用于计算注意力系数的得分函数,为邻接实体赋予不同的权重值,使实体的表征过程充分考虑到邻接实体的不同作用;最后,设计用于结合文本信息与知识库信息的知识融合器,采用早期融合策略,建立文本信息与知识图谱信息相结合的融合表征,得到实体为正确答案的概率。在2个公共数据集进行的实验结果表明,知识图谱不完整的情况下,本模型检索器检索到的文本对问答效果的提升帮助较大。在仅使用知识图谱作为数据源的情况下,本模型采用的图注意力网络相较于图卷积网络的模型存在一定优势。

1　任务定义

开放领域问答任务基于自然语言问句 $Q=\{q_1,q_2,\cdots,q_{U_Q}\}$，其中，$U_Q$ 表示问句 Q 中 token 的数量值，利用 1 个基于三元组的知识库 $K=\{(e_h,r,e_t)\}$，其中，e_h 和 e_t 分别表示头尾实体，r 表示头尾实体之间的关系，以及 1 个包含丰富外部信息的文本库 $D=\{d_1,d_2,\cdots,d_{U_D}\}$ 作为知识源，其中，U_D 表示文本库 D 中文本的数量值，根据问句 Q 中的中心实体集合 E_c，利用个性化 PageRank[9]算法，根据问句中的中心实体集合构建查询子图 G，包含与问句最相关的实体和关系，确保最终获得的答案具有高召回率。对文本库 D 中的文本利用 Facebook 公司开发的开源框架 FAISS 添加索引，依据问句 Q 进行语义相关性计算，筛选得到最相关文本集合。文本库 D 中的实体同样需要与知识库进行实体链接，完成知识表征操作。最终，由问句 Q、文本 D、知识图谱 K 联合构建知识表征，从根据问句 Q 的中心实体集合 E_c 构建的查询子图 G 中抽取答案。

2　模　型

本模型核心部分主要由 2 个模块构成，分别为基于语义信息进行精确检索的文本检索器、结合文本信息与知识库信息的知识融合器，总体框架如图 2 所示。

该模型从问句出发，围绕问句的中心实体构建查询子图，利用图注意力网络对查询子图 G 中的实体进行嵌入表征。同时，利用文本的语义信息检索，结合文本所链接的实体信息对文本进行表征。最后，将问句信息、文本信息、知识图谱的实体信息进行融合，得到最终知识表征，计算某实体为正确答案的概率。

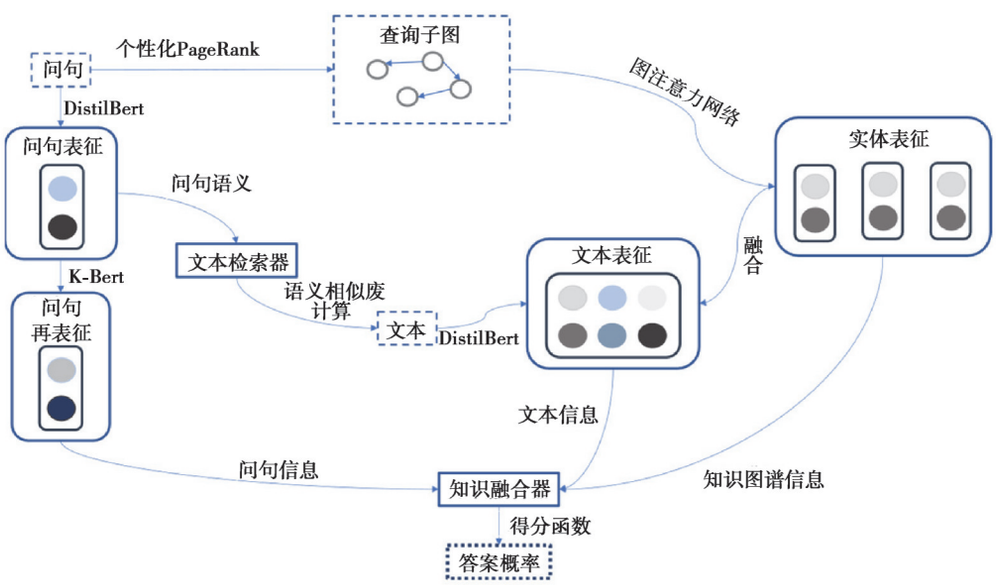


图 2　模型总体框架图

Fig. 2　Framework of the model

2.1　文本检索器

2.1.1　问句编码和文本编码

将自然语言问句和维基百科文本编码为向量形式，为文本检索打好基础。为充分挖掘问句和文本中蕴含的语义信息，采用 DistilBert[12]作为预训练模型，为自然语言句子提供基于语义的嵌入表示，该嵌入为句子级别的嵌入。与此同时，该模型为大型预训练模型 Bert 经知识蒸馏得到的轻量级模型，其参数量仅为 Bert 的 40%，且性能几乎保持相同，可在为大规模文本语料编码时节省时间。研究采用独立 DistilBert 编码器对问句和文本分别进行编码，以问句 Q 为例，编码过程如式(1)所示

$$\overrightarrow{\boldsymbol{h}_q}=\mathrm{DistilBert}(Q), \tag{1}$$

设编码后问句表征为$\overrightarrow{\boldsymbol{h}_q}$。

2.1.2 文本检索

利用问句和文本句子级别的嵌入表示进行相关性匹配，采用高效且运算简单的点积作为计算问句与文本语料库中文本之间的匹配得分函数。如式(2)所示。

$$\mathrm{Score}\left(q,d\right)=\overrightarrow{\boldsymbol{h}_q^{\mathrm{T}}}\cdot\overrightarrow{\boldsymbol{h}_d}, \tag{2}$$

式中：$\overrightarrow{\boldsymbol{h}_q}$为利用DistilBert得到的问句表征；$\overrightarrow{\boldsymbol{h}_d}$为利用DistilBert得到的某一文本语句表征。

由于传统SQL查询的方式在相关性检索方面效率低下，因此，选择Facebook公司开发的FAISS，它是为稠密向量提供高效相似度搜索和聚类的框架。利用FAISS框架对文本语料库中的所有文本做索引。依据得分函数的计算结果筛选前20个句子作为最相关的文本，成为后续知识融合的重要组成部分。

2.2 知识融合器

该部分需要将某个问题的3部分分别进行表征并融合，包括问句、相关文本和查询子图中的实体。

2.2.1 查询子图

不同于传统构建查询子图的方法，本文并非固定多跳范围，选择采用个性化PageRank算法动态构建查询子图，确保查询子图中的实体仅包括问句的中心实体及相关实体，不涉及其他实体。初始化时，从问句的中心实体及邻接实体出发，定义PageRank得分，并为其赋得分初始值。若该实体为中心实体e_c，则其得分初始值为中心实体个数的倒数，表示为$1/U_c$(其中，U_c表示中心实体的个数)，否则为0，初始化如式(3)所示

$$PR^{(0)}=\begin{cases}1/U_c & \text{if } e_c\in E_c,\\ 0 & \text{o.w},\end{cases} \tag{3}$$

式中，E_c表示问句的中心实体集合。

式(1)使用DistilBert预训练模型得到查询子图中所有关系的表征$\vec{\boldsymbol{r}}$和问句表征$\overrightarrow{\boldsymbol{h}_q}$，通过计算得到二者的匹配得分作为式(4)的前项。此外，定义控制函数I_r，其取值作为式(4)的后项。当关系r的2端中至少有1端连接问句的中心实体时，控制函数I_r取值为1，否则取值为0，匹配得分函数如式(4)所示

$$S_{r,q}=\overrightarrow{\boldsymbol{r}^{\mathrm{T}}}\cdot\overrightarrow{\boldsymbol{h}_q}+I_r。 \tag{4}$$

在更新过程中，与问句相关性较强的关系所连接的实体具有较高权重，实体PageRank得分函数的更新过程如式(5)所示

$$PR^{(l+1)}=\lambda_P PR^{(l)}+\left(1-\lambda_P\right)\sum_{e_i\in E_N}S_{r_i,q}\cdot PR_{e_i}^{(l)}。 \tag{5}$$

该得分函数由2部分组成，前项为某实体上1轮更新所得的PageRank得分，后项为某实体邻接实体上1轮更新所得的PageRank得分的加权之和。2者通过1个取值范围为0~1的平衡因子λ_P相结合。E_N表示当前实体的邻接实体集合。当实体的PageRank得分收敛后，选取其值大于0.005的实体用以构建查询子图。

2.2.2 问句再表征

为构建节点的融合表征，需将自然语言问句信息与三元组信息相融合。利用式(1)得到的问句表征$\overrightarrow{\boldsymbol{h}_q}$仅能包含句子的语义信息，无法涵盖到知识图谱中的实体信息，不利于构建融合表征。由北京大学推出的K-Bert模型[13]知识层可从知识图谱中查询自然语言问句中涉及的所有三元组，将其注入到问句中，形成句子树。由K-Bert模型构造的句子树如图3所示。将句子树转化为涵盖知识图谱结构信息的新问句表征$\vec{\boldsymbol{q}}$，如式(6)所示

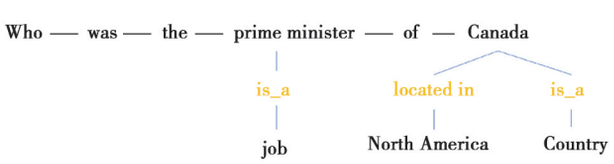


图3 问句的句子树

Fig. 3 Sentence tree of a question

$$\vec{\boldsymbol{q}}=K-\mathrm{Bert}\left(\vec{\boldsymbol{h}}_q\right)。\tag{6}$$

2.2.3　实体表征

知识图谱信息的实体表达是问答系统的知识源之一。由于知识图谱可被看作异构网络图,因此,图神经网络[14]可用于知识图谱表征学习。为充分利用邻接实体信息获取自身的嵌入表示,使用图注意力网络对查询子图中的实体进行表征,得到实体的向量化表示。此过程需计算每个邻接实体对自身实体的注意力得分。本文利用计算实体 PageRank 得分时定义的“问句-关系”匹配得分 $S_{r,q}$ 作为注意力得分的组成部分之一。定义另 1 个控制函数 I_e,若该中心实体的邻接实体 e_i 也为问句的中心实体,则控制函数 I_e 取值为 1,此时这 2 个实体对问句的相关性较强,否则取值为 0,表示这 2 个实体对问句的相关性较弱。最后,将计算结果用 softmax 函数进行归一化,邻接实体 e_i 的注意力得分计算方式如式(7)所示。

$$S_{r_i,e_i}=\mathrm{softmax}\left(S_{r_i,q}+I_{e_i}\right)。\tag{7}$$

实体表征的更新过程如式(8)所示

$$\vec{\boldsymbol{e}}_c'=\lambda_{e_1}\vec{\boldsymbol{e}}_c+\lambda_{e_2}\sigma\left(\sum_{e_i\in E_N}S_{r_i,e_i}\boldsymbol{W}\left[\vec{\boldsymbol{e}}_c\|\vec{\boldsymbol{e}}_i\|\vec{\boldsymbol{r}}_i\right]\right),\tag{8}$$

式中:E_N 表示当前实体的邻接实体集合;符号 $\|$ 表示向量拼接操作;$\sigma(\cdot)$ 为 sigmoid 激活函数。式(8)由 2 部分组成,前项为该实体上 1 轮更新的实体表征,后项利用图注意力网络根据邻接实体的注意力得分进行聚合计算,得到实体表征,将 $\sigma\left(\sum_{e_i\in E_N}S_{r_i,e_i}\boldsymbol{W}\left[\vec{\boldsymbol{e}}_c\|\vec{\boldsymbol{e}}_i\|\vec{\boldsymbol{r}}_i\right]\right)$ 记为 $\vec{\boldsymbol{e}}_N$。为综合二者,设定取值范围为 0~1 的平衡因子 λ_{e_1} 和 λ_{e_2}。利用问句表征 $\vec{\boldsymbol{q}}$ 分别与实体 $\vec{\boldsymbol{e}}_c$ 和由邻接实体得到的实体表征 $\vec{\boldsymbol{e}}_N$ 做匹配计算,将计算结果作为系数平衡两者的权重,将当前实体与邻接实体相结合。平衡因子 λ_{e_1} 和 λ_{e_2} 的计算过程如式(9)所示

$$\begin{cases}\lambda_{e_1}=\mathrm{soft\,max}\left(\vec{\boldsymbol{q}}^{\mathrm{T}}\cdot\vec{\boldsymbol{e}}_c\right),\\ \lambda_{e_2}=\mathrm{soft\,max}\left(\vec{\boldsymbol{q}}^{\mathrm{T}}\cdot\vec{\boldsymbol{e}}_N\right)。\end{cases}\tag{9}$$

2.2.4　文本再表征

文本同样是问答系统的知识源之一,需要将知识图谱的实体信息与文本语义信息相融合,为后续知识融合提供便利。将利用式(1)得到的文本表征 $\vec{\boldsymbol{h}}_d$ 作为初始值进行实体链接,建立了文本与知识图谱实体的对应关系,将这些实体作为文本最终表征的重要组成部分。从文本表征的第 2 轮开始,用 $\vec{\boldsymbol{d}}$ 代替初始值 $\vec{\boldsymbol{h}}_d$。文本最终表征如式(10)所示

$$\vec{\boldsymbol{d}}'=\lambda_{d_1}\vec{\boldsymbol{d}}+\lambda_{d_2}\vec{\boldsymbol{e}}_d\ 。\tag{10}$$

式中的 $\vec{\boldsymbol{e}}_d$ 表示与文本 d 相链接的实体表征,可通过实体表征式(8)获得。若文本 D 链接多个实体,则对其求平均值,其表达式如式(11)所示

$$\vec{\boldsymbol{e}}_d=\frac{\sum_{e_j\in E_d}\vec{\boldsymbol{e}}_j}{U_d}。\tag{11}$$

该式表示将所有实体的表征求和,并求平均值作为文本链接实体的表征。其中:E_d 表示文本链接的实体集合;U_d 表示集合中实体的个数。

式(10)由 2 部分组成,前项表示该文本上 1 轮更新的表征,后项表示该文本所链接实体的表征,为综合二者,设定取值范围为 0~1 的平衡因子 λ_{d_1} 和 λ_{d_2}。利用问句表征 $\vec{\boldsymbol{q}}$ 分别与文本表征 $\vec{\boldsymbol{d}}$ 和链接的实体表征 $\vec{\boldsymbol{e}}_d$ 作匹配计算,得到系数平衡 2 者权重,将文本与实体相结合。平衡因子 λ_d 的计算过程如式(12)所示

$$\begin{cases}\lambda_{d_1}=\mathrm{soft\,max}\left(\vec{\boldsymbol{q}}^{\mathrm{T}}\cdot\vec{\boldsymbol{d}}\right),\\ \lambda_{d_2}=\mathrm{soft\,max}\left(\vec{\boldsymbol{q}}^{\mathrm{T}}\cdot\vec{\boldsymbol{e}}_d\right),\end{cases}\tag{12}$$

对于与问句Q相关的所有文本表征$\vec{d}$,取对应位置元素的均值作为最终的$\vec{d}'$。

2.2.5 融合表征

该部分将此前获得的问句表征,以及由问句构建查询子图中的实体自身的表征、邻接实体表征、文本表征进行融合,得到知识的最终表征。其中邻接实体表征$\vec{e_N}$的计算过程如式(13)所示

$$\vec{e_N}=\frac{\sum_{e_i\in E_N}\vec{e_i}}{U_N}, \tag{13}$$

式中:某实体的邻接实体集合为E_N;U_N表示集合中实体的个数。这里邻接实体表征取所有邻接实体表征的平均值。最终的知识融合表征$\vec{k_e}$如式(14)所示。

$$\vec{k_e}=\vec{q}^{\mathrm{T}}\left[\vec{e_c'}\|\vec{d}'\|\vec{e_N}\right], \tag{14}$$

式中,符号‖表示向量拼接操作。最后,通过式(15)得到实体为正确答案的概率

$$S_{\text{final}}=\text{softmax}\left(\vec{k_e}\right), \tag{15}$$

式中,softmax(·)为激活函数,将实体为正确答案的概率映射到0~1之间。

3 实 验

3.1 数据集

本实验使用WikiMovies-10K和WebQuestionsSP作为数据集[15]。

3.1.1 WikiMovies-10K

该数据集由Miller等[16]于2016年引入,包含来自WikiMovies数据集的10 K个电影领域问答数据,使用Wikipedia的子集(电影领域文章的标题和内容)作为知识库和文本语料库。

3.1.2 WebQuestionsSP

WebQuestionsSP是WebQuestions包含SPARQL标注的升级版本,包含4 737个基于Freebase[17]实体的自然语言问句,使用Freebase作为知识库。WikiMovies-10K数据集和WebQuestionsSP数据集的基本信息统计如表1所示。

表1 数据集基本信息统计

Table 1 Dataset basic information statistics

数据集名称	数据规模 train / dev / test	实体个数	关系类型数	文本数
WikiMovies-10K	10 000 / 9 999 / 9 951	43 235	9	79 728
WebquestionsSP	2 848 /250 /1 639	528 617	513	235 567

3.2 实验设置

为验证本文提出模型的有效性,实验在Python3.6、CUDA11.1环境下进行,基于PyTorch框架编写代码。所使用计算机配置环境的硬件参数为:处理器AMD R5-2600X、内存16 G、显卡NVIDIA GeForce GTX 1080Ti(显卡芯片内存容量为11 G)。学习率α=0.001,epoch=100,PageRank的平衡因子λ_p设置为0.6。

3.3 评价指标

实验采用Hit@1和$F1$分数来评估不同模型的性能效果,其中,Hit@1表示模型预测最佳答案的准确性。$F1$分数同时考虑精确率和召回率,2者同时达到最高,取得平衡。$F1$分数的计算方法如式(16)所示

$$F1=\frac{2*\text{precision}*\text{recall}}{\text{precision}+\text{recall}}, \tag{16}$$

式中:precision为精度,表示被分为正确答案的示例中实际为正确答案的比例;recall为召回率,是覆盖面的度

量,表示系统预测答案覆盖正确答案的程度。

3.4　对比实验

本实验将提出的模型与对比模型分别在 WikiMovies-10K 数据集和 WebQuestionsSP 数据集上进行对比。为验证本模型中检索器对问答效果的提升,分别进行了仅采用知识图谱作为数据源和同时采用文本加知识图谱作为数据源的实验,实验结果如表 2 和表 3 及图 4 和图 5 所示。

为验证本模型检索的相关文本知识对补充不完整知识图谱的作用,以及本模型与其他问答系统模型对不同完整性程度知识图谱是否能保持相对稳定的性能表现,本实验在上述 2 个数据中分别构造完整度不同的 3 个数据集,将知识图谱三元组的数量降至原始数据的 10%、30% 和 50%,模拟知识图谱中存在不同程度完整性的情况。

3.4.1　对比模型

KVMemNet 是 Miller 等[16]提出的端到端记忆网络,它将知识图谱三元组和文本视为记忆单元,并封装成键值对的形式,KV-EF 是该模型基于早期融合策略的版本。GraftNet 是 Sun 等[4]提出的基于图卷积网络的问答系统模型,该模型将文本视为知识图谱中的特殊类型节点,利用图卷积网络聚合信息。GN-LF 和 GN-EF 分别是 GraftNet 采用晚期融合和早期融合策略的版本。SG-KA 是 Xiong 等[5]提出的问答系统模型,该模型的 Knowledge-Aware Text Reader 模块利用知识库信息从文本中找出正确答案。PullNet 是 Sun 等[18]提出的问答系统模型,该模型可自主学习如何检索与回答问题相关的子图,并以迭代的方式构建子图。

表 2　提出模型与对比模型在 WikiMovies-10K 数据集下的结果对比

Table 2　Comparison of the results of the proposed model and comparison models under the WikiMovies-10K dataset

模型	10%		30%		50%	
	KB	KB+Text	KB	KB+Text	KB	KB+Text
KV-EF	15.8 / 9.8	53.6 / 44.0	44.7 / 30.4	60.6 / 48.1	63.8 / 46.4	75.3 / 59.1
SG-KA	19.1 / 13.4	49.4 / 37.8	47.5 / **37.4**	71.7 / 53.4	66.5 / 53.9	80.6 / 66.7
GN-LF	19.7 / 17.3	74.5 / 65.4	**48.4** / 37.1	78.7 / 68.5	67.7 / 58.1	83.3 / 74.2
GN-EF	19.7 / 17.3	75.4 / **66.3**	**48.4** / 37.1	82.6 / 71.3	67.7 / 58.1	87.6 / 76.2
PullNet	—	—	—	—	65.1 / —	92.4 / —
Ours	**20.3** / **17.5**	**77.6** / 65.8	48.1 / 37.2	**83.3** / **74.8**	**68.4** / **60.2**	**93.1** /**78.6**

注:加黑数据表示特定条件下,所有模型中的最佳实验值。

表 3　提出模型与对比模型在 WebQuestionsSP 数据集下的结果对比

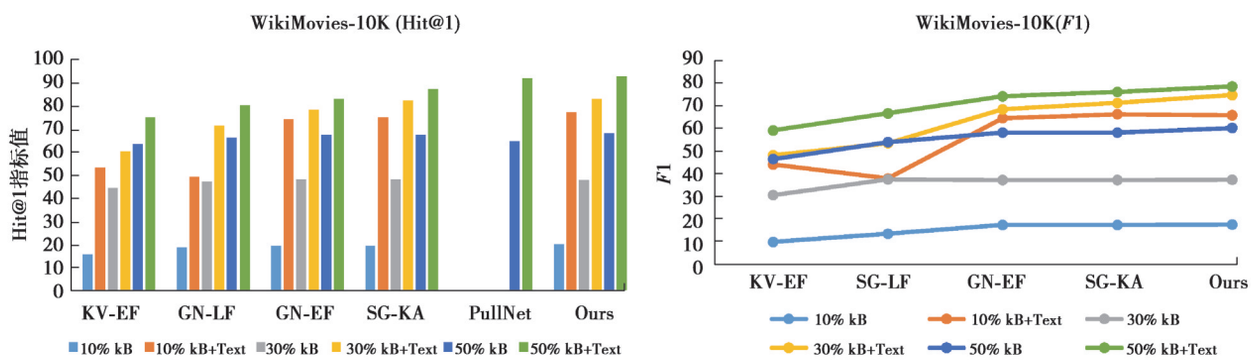
Table 3　Comparison of the results of the proposed model and comparison models under the WebQuestionsSP dataset

模型	10%		30%		50%	
	KB	KB+Text	KB	KB+Text	KB	KB+Text
KV-EF	12.5 / 4.3	24.6 / 14.4	25.8 / 13.8	27.0 / 17.7	33.3 / 21.3	32.5 / 23.6
GN-LF	15.5 / 6.5	29.8 / 17.0	34.9 / 20.4	39.1 / 25.9	47.7 / 34.3	46.2 / 35.6
GN-EF	15.5 / 6.5	31.5 / 17.7	34.9 / 20.4	40.7 / 25.2	47.7 / 34.3	49.9 / 34.7
SG-KA	**17.1** / 7.0	33.6 / 18.9	35.9 / 20.2	42.6 / 27.1	49.2 / 33.5	52.7 / 36.1
PullNet	—	—	—	—	**50.3** / —	51.9 / —
Ours	17.0 / **8.6**	**35.1** / **20.4**	**36.1** / **20.6**	**43.2** / **27.8**	49.6 / **35.2**	**53.9** / **37.6**

注:加黑数据表示特定条件下,所有模型中的最佳实验值。

3.4.2 实验结果分析

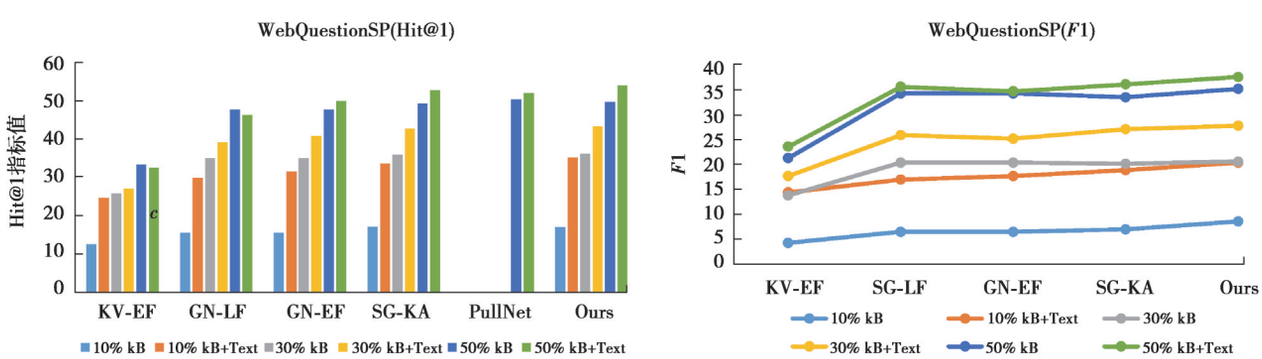
表 2 和表 3 分别展示所提出的模型与对比模型在 WikiMovies-10K 数据集和 WebQuestionSP 数据集下的实验结果。图 4 和图 5 以柱状图和折线图形式直观展示了本模型与对比模型在 2 种数据集下对 2 种评价指标的实验结果。从表 2 和表 3 及图 4 和图 5 中的实验结果数据看出，所提出的模型相较对比模型在公共数据集 WikiMovies-10K 和 WebQuestionsSP 的表现更好，特别是在采用“KB+Text”的方式时提升尤为明显，由此验证本模型的检索器深层次考虑自然语言句子的语义信息，对文本检索与匹配精确度有一定提升，在知识图谱完整性较低时，为图注意力网络提供推理依据和实体背景知识，提升问答系统性能。而对于其他模型，本模型的文本检索器并非只考虑句子中关键词的词频信息，而是重点关注语句的语义信息，使文本的相似度计算方式更合理，且检索到的文本语义相关性更高，避免在匹配文本时引入知识噪声，影响问答系统性能。同时，本模型将文本、问句、知识图谱信息融合，弥补了查询子图实体信息的不完整性，问答效果明显增强。



(a) Hit@1指标结果 (b) *F*1指标结果

图 4 模型与对比模型在 WikiMovies-10K 数据集下实验结果

Fig. 4 Experimental results of the proposed model and comparison models under the WikiMovies-10K dataset



(a) Hit@1指标结果 (b) *F*1指标结果

图 5 模型与对比模型在 WebQuestionsSP 数据集下实验结果

Fig. 5 Experimental results of the proposed model and comparison models under the WebQuestionsSP dataset

在 WikiMovies-10K 数据集中知识图谱完整性为 30% 的条件下，且仅采用知识图谱作为数据源时，本模型的表现没有达到最佳，但与取得最佳表现的模型在指标数值上差距极小，在 Hit@1 和 *F*1 指标上分别降低 0.3% 和 0.2%。原因是当知识图谱的完整性过低时，缺乏文本知识提供推理依据，采用图神经网络模型的推理能力受到限制。在相同条件下，本模型同时采用文本和知识图谱作为数据源时，本模型的表现比其他模型更出色，说明在知识图谱完整性较低时，文本知识为图注意力网络的推理提供了依据和背景知识，对推理的准确性有较大贡献。

在 2 种评价指标下，模型使用文本与知识图谱相融合的表现比仅使用知识图谱作为数据源的表现更佳，

且有较大幅度提升,验证了本文所提模型中的检索器起到重要作用,扩充了模型的知识源,为图神经网络提供推理依据和实体背景知识。

当只采用知识图谱作为数据源时,本模型依然保持优异性能。KV-EF模型未采用图神经网络进行推理,而是将知识图谱中的三元组以固定格式转换为记忆单元,忽略了图神经网络对于知识推理的优势,问答效果不佳。本模型与其他模型均采用图神经网络进行知识推理,且本模型采用图注意力网络作为知识表示,在实体信息中充分融入不同邻接实体与边的信息,为不同的邻接实体赋予不同权重值,解决其他模型所采用的图卷积网络对所有邻接实体都具有相同权重的问题。因此,本模型在只采用知识图谱作为数据源时,不仅能与其他同样采用图神经网络的模型有接近性能,绝大多数情况下甚至能取得领先优势。实验不仅证明图注意力网络具有较强推理能力,同时验证了本模型为不同邻接实体赋予不同注意力得分的合理性。

此外,本文提出的方法在上海汽车集团股份有限公司和上海保隆汽车科技股份有限公司的汽车零部件维修数据集上进行了测试,有效缓解企业在工业数据领域应用知识图谱过程中知识来源受限、问答系统准确性不高等问题,实现了本方法在特定工业应用场景的有效验证。

4　结束语

针对知识图谱的不完整性制约问答系统性能的问题,重点研究通过检索文本对不完整知识图谱问答的作用,提出一种新模型。该模型的检索器部分充分利用问句的语义信息检索相关文本,弥补知识图谱的不完整性,为图注意力网络的推理提供依据,增强模型整体推理能力。该模型的知识融合器部分利用图注意力网络对知识图谱中的实体进行表征,分别对问句、文本进行再表征,使其包含知识图谱的实体信息,得到最终融合知识图谱、问句、文本的融合知识表征。因其完整、准确包含知识信息,对提升问答任务的效果具有显著作用。在2种公共数据集的实验中证明,该模型与前人提出的方法相比,存在一定优势。在未来工作中,团队将关注更多汽车制造企业的知识决策案例,进一步提升本模型在相关领域数据分析与处理过程中的鲁棒性和泛化能力。

参考文献

[1] Wu P Y, Zhang X W, Feng Z Y. A survey of question answering over knowledge base[C]//China Conference on Knowledge Graph and Semantic Computing. Singapore: Springer, 2019: 86-97.

[2] Savenkov D, Agichtein E. When a knowledge base is not enough: question answering over knowledge bases with external text data[C]//Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. New York: ACM, 2016: 235-244.

[3] Chen D Q, Fisch A, Weston J, et al. Reading wikipedia to answer open-domain questions[C]//55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017: 1870-1879.

[4] Sun H T, Dhingra B, Zaheer M, et al. Open domain question answering using early fusion of knowledge bases and text[C]//2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018: 4231-4242.

[5] Xiong W H, Yu M, Chang S Y, et al. Improving question answering over incomplete KBs with knowledge-aware reader[C]//57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 4258-4264.

[6] Han J L, Cheng B, Wang X. Open domain question answering based on text enhanced knowledge graph with hyperedge infusion[C]//EMNLP 2020. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 1475-1481.

[7] Yu D H, Yang Y M, Zhang R H, et al. Knowledge embedding based graph convolutional network[C]//Proceedings of the Web Conference 2021. New York: ACM, 2021: 1619-1628.

[8] Cai L, Yan B, Mai G C, et al. TransGCN: coupling transformation assumptions with graph convolutional networks for link prediction[C]//10th International Conference on Knowledge Capture. New York: ACM, 2019: 131-138.

[9] Haveliwala T H. Topic-sensitive PageRank[C]//11th International Conference on World Wide Web. New York: ACM, 2002: 517-526.

[10] Velikovi P, Cucurull G, Casanova A, et al. Graph attention networks[C]//6th International Conference on Learning Representations. Vancouver, BC, Canada: ICLR, 2018: 2920-2931.

[11] 李德栋. 基于图注意网络的文本增强知识图谱表示学习[D]. 上海: 华东师范大学, 2020.
Li D D. Text-enhanced knowledge graph representation learning based on graph attention network[D]. Shanghai: East China Normal University, 2020. (in Chinese)

[12] Mozafari J, Fatemi A, Moradi P. A method for answer selection using DistilBERT and important words[C]//2020 6th International Conference on Web Research (ICWR). Tehran, Iran: IEEE, 2020: 72-76.

[13] Liu W, Zhou P, Zhao Z, et al. K-BERT: enabling language representation with knowledge graph[C]//34th AAAI Conference on Artificial Intelligence/32nd Innovative Applications of Artificial Intelligence Conference/10th AAAI Symposium on Educational Advances in Artificial Intelligence. New York: AAAI, 2020, 34: 2901-2908.

[14] Fu X Y, Zhang J N, Meng Z Q, et al. MAGNN: metapath aggregated graph neural network for heterogeneous graph embedding [C]//Proceedings of The Web Conference 2020. New York: ACM, 2020: 2331-2341.

[15] Yih W T, Richardson M, Meek C, et al. The value of semantic parse labeling for knowledge base question answering[C]//54th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016: 2:201-206.

[16] Miller A, Fisch A, Dodge J, et al. Key-value memory networks for directly reading documents[C]//2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016: 1400-1409.

[17] Bollacker K, Evans C, Paritosh P, et al. Freebase: a collaboratively created graph database for structuring human knowledge [C]//ACM SIGMOD International Conference on Management of Data. New York: ACM, 2008: 1247-1250.

[18] Sun H T, Bedrax-Weiss T, Cohen W. PullNet: open domain question answering with iterative retrieval on knowledge bases and text[C]//2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 2380-2390.

（编辑 侯 湘）