浙江科技学院学报,第 34 卷第 5 期,2022 年 10 月
Journal of Zhejiang University of Science and Technology
Vol. 34 No. 5, Oct. 2022

doi: 10.3969/j.issn.1671-8798.2022.05.007

基于知识图谱的问答系统中属性映射方法研究

叶仕超,雷景生,杨胜英
(浙江科技学院 信息与电子工程学院,杭州 310023)

摘　要: 在基于知识图谱的智能问答系统中,属性映射模块结果的错误传播会导致最终无法得到正确答案,对此提出了一种基于多注意力多维文本的属性映射方法。首先通过拆分问题文本及结合属性信息得到多维文本表示;其次使用长短期记忆网络(long-short-term memory,LSTM)层生成各自的隐层表示;然后输入多注意力机制层后使问句和属性之间的关系及语义信息更加完善,利用属性之间的交互信息及多种角度来加强问句语义信息的理解;最后通过卷积神经网络(convolutional neural networks,CNN)提取局部特征并且采用 softmax 分类器实现属性映射。试验结果表明,在自然语言处理与中文计算会议(NLPCC 2018)中知识库问答(KBQA)任务所提供的开源数据集上,本方法相比主流属性映射模型其性能有显著提升,准确率最高提升 6.62%。本模型可以补足单一文本表示与注意力机制的短板,有效解决属性映射模块中语义歧义的问题,这有助于后续提高智能问答系统的整体性能。

关键词: 智能问答;属性映射;多注意力;多维文本;知识图谱
中图分类号: TP391.1　　**文献标志码**: A　　**文章编号**: 1671-8798(2022)05-0435-09

Research on attribute mapping method in question answering system based on knowledge graph

YE Shichao, LEI Jingsheng, YANG Shengying
(School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou310023, Zhejiang, China)

Abstract: In the intelligent question answering system based on knowledge graph, the error propagation of the result of attribute mapping module can lead to the inability to get the correct answer ultimately, hence an attribute mapping method based on multi-attention and multi-dimensional text came to the rescue. Firstly, the multi-dimensional text representation was obtained by splitting the question text and combining the attribute information; secondly, the long-term and short-term memory (long-short-term memory, LSTM) networks were used to generate their respective hidden layer representations, and then the multi-attention mechanism layer was input to improve the relationship and semantic information between questions and attributes, and the interactive information between attributes and various angles were used to enhance the understanding of semantic information of questions; finally, local features were

收稿日期: 2021-07-05
基金项目: 国家自然科学基金项目(61972357,61672337);浙江省重点研发计划项目(2019C03135)
通信作者: 雷景生(1966—　),男,陕西省韩城人,教授,博士,主要从事电力大数据挖掘研究。E-mail:jshlei@zust.edu.cn。

extracted by virtue of convolution neural network (convolutional neural networks, CNN) and attribute mapping was realized with the aid of softmax classifier. The experimental results show that on the open source data set provided by the knowledge base question and answer (KBQA) task in the natural language processing and Chinese computing conference (NLPCC 2018), the performance of this method is significantly improved compared with the mainstream attribute mapping model, with the accuracy up to 6.62% higher. This model can make up for the deficiency of single text representation and attention mechanism and effectively solve the problem of semantic ambiguity in attribute mapping module, which is conducive to improving the overall performance of intelligent question answering system.

Keywords: intelligent question answering; attribute mapping; multi-attention; multi-dimensional text; knowledge graph

智能问答系统指通过自然语言处理和知识提取等技术手段,对人类语言的思维结构进行理解和分析的一种信息检索系统。随着大数据时代的到来,大量文本、音频、图像等信息的表现形式和载体在互联网上不断产生,这些数据对人们有利用价值,但大量无效信息也会对人们产生干扰。早期传统的问答系统是基于信息检索的方法,在给定的文本中寻找信息,即检索含有相关内容的文本,并从中选取出所需的问题答案[1]。传统的信息检索方式是通过搜索引擎查询和使用特定领域的信息管理系统来实现。这种方式需要用户准确输入关键词,缺乏对自然语言的语义做出解释和分析,并且需要再到相关网页中查询,使得用户体验满意度不高[2]。随着大规模知识库的不断发展,基于知识库问答系统(question answering over knowledge base,KBQA)的相关研究应运而生。知识图谱(knowledge graph)的概念是谷歌在2012年5月所提出的,其目的是使得搜索的质量得到质的飞跃,改善用户体验[3]。作为一种新型的数据表示方式,知识图谱的基本储存形式被称为三元组[4],它能很好地组织和管理互联网信息,是一个高质量的知识库。根据内部实现方法的不同,知识图谱问答系统可分为基于提取方法、语义解析、向量空间建模3种常见类型[5]。

基于提取方法的知识图谱问答系统中一个技术关键点是知识的关联,其难点在于把捕捉到的问题同知识图谱中的信息相联系。Zhou等[6]为了能获得更加深层的文本表示信息,在通过长短期记忆网络(long-short-term memory,LSTM)提取的同时,还结合了卷积神经网络(convolutional neural networks,CNN)和注意力机制,这使得文本表示信息不仅包含上下文信息还对重点文本内容进行特别关注。Hao[7]等根据不同方向的候选答案信息,通过结合交叉注意力机制的神经网络模型来动态地表示问题和相应候选答案的文本信息,将文本内容序列化,同时将知识库信息集成到答案中,以此缓解词汇量不足的问题。Yu[8]等在提出的知识库问答系统中,使用一种将问句和属性基于不同粒度进行匹配的属性映射模型,即残差学习增强的分级双向LSTM(hierarchical residual bidirectional LSTM,HR-BiLSTM);该模型使用残差LSTM通过不同的抽象层比较问题和关系名称,在与实体识别模块的互相配合下使系统整体的准确率显著提高。未知语言现象是自然语言处理任务中的常见问题,其中最典型的则是未登录词,而在输入向量中利用多维表示信息是一个有效的方法[9]。目前学术界主要通过拼接的方式来结合多维表示的特征向量。为了加强文本信息的向量表示能力,聂维民等[10]通过将字符维度信息结合主题维度信息和词维度信息,引入融合门的思想,在降低输出向量维度的同时将以上三种不同维度的特征信息融合,得到的文本向量有效地提高了后续文本分类试验的模型表现。可见,通过结合文本不同维度的特征向量来加强文本信息表示是近些年自然语言处理研究的一个重点。陈功等[11]提出了一种多维度的Web文本表示方法,他将从文本内容中抽取的隐含特征和表层特征与学习用户行为得到的社交特征进行融合,从而提高文本表示能力。江明奇等[12]在解决属性分类问题时,根据问答文本的特点,提出了一种多维文本表示的方法,将切分后的句子通过LSTM模型提取一个隐层特征表示,再将融合后形成的多维文本进行卷积层处理,从而获得最终分类结果。目前,基于知识图谱的智能问答研究,在英文方面已经取得

了很大的成功，但因为中文与英文的差异，如词汇边界、实体歧义等问题，所以单一关系的问答是中文方面目前的主要研究方向[13-14]。综上所述，为了继续研究中文开放领域知识图谱智能问答方面的相关内容，我们在 NLPCC-KBQA 2018 任务所提供的知识图谱和问答语料上做了相关研究，针对智能问答系统中的属性映射模块提出了基于多注意力多维文本的属性映射方法。

1　属性映射模型结构与方法

本研究提出的知识图谱问答系统是基于提取方法的，其流程如图 1 所示。首先需要经过实体识别模块将问句中的实体信息正确地提取出来，然后通过属性映射模块进行问句意图和知识图谱中属性的映射，从而去理解用户问句中的意图。

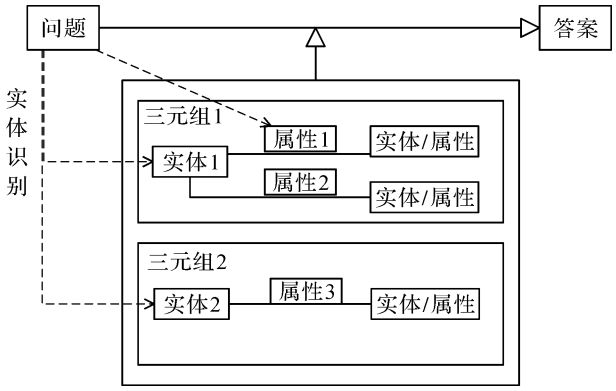

图 1　基于提取方法的知识图谱问答系统流程图

Fig. 1　Flow chart of knowledge graph question answering system based on extraction method

实体的关系或属性就是该意图的具体表现形式。基于多注意力多维文本的属性映射模型结构如图 2 所示。

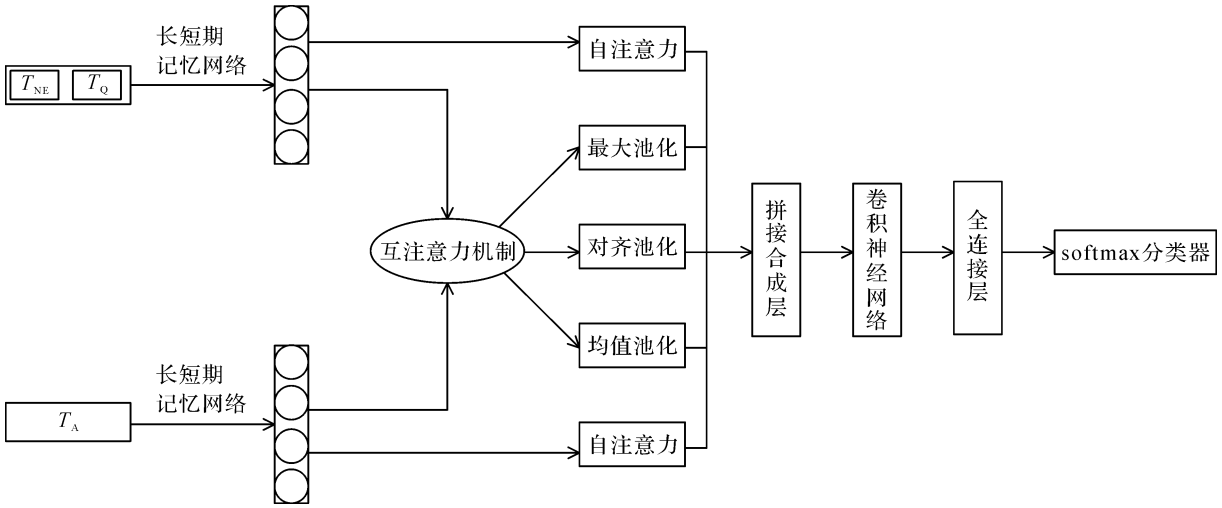

图 2　基于多注意力多维文本的属性映射模型结构

Fig. 2　Structure of attribute mapping model based on multi-attention and multi-dimension text

由图 2 可知，本模型将属性映射任务转化成语句对分类任务。模型将问题文本拆分为问句实体文本 T_{NE} 和描述属性相关文本 T_Q，T_A 是对应的属性文本信息，再将这些文本信息通过 LSTM 层生成各自的隐层表示 h_Q 和 h_A，使用互注意力机制和自注意力机制对它们进行处理，分别得到相似矩阵和自身依赖关系；在相似矩阵上使用最大池化、对齐池化、均值池化来增强问句和属性的交互信息，拼接注意力层的这些隐层表示并通过卷积神经网络和全连接层提取特征，使用 softmax 分类器对用户问句的属性进行分类及预测。

1.1　问题文本的多维表示

与普通文本类型不同，问答系统中的问题文本主要由问句实体和描述属性的相关文本内容组成。在属

性映射任务模块中，描述属性的文本内容是判断属性分类的主要依据，但问句中的实体信息也可以起到一定的辅助作用。不同问题中，实体信息对属性映射的判断起到的作用是不同的，有时也可能会产生干扰。因此，通过将问题切分的方法，充分利用实体信息部分和描述属性文本内容之间的隐层关系，来获得更准确和丰富的文本信息。另外，以往的许多属性映射研究只是单一的注重提取问题的表示，忽略了属性信息与问句之间的联系。因此我们将样本中属性信息提取出来作为输入文本的一维表示，并使用多注意力机制层使问句和属性之间的关系及语义信息更加完善，利用属性之间的交互信息及多种角度来加强问句语义信息的理解。文本表示是本方法中的关键环节，我们通过使用多维文本表示来进行智能问答系统中属性映射模块的试验。

1.2 多注意力机制融合层

一般情况下，对一个句子使用一次注意力机制，然后将训练学习到的表示信息输送给预测层。现在许多模型因仅使用一种注意力机制或变体而不能获得较好的表征信息，但是如果在调用时，通过连接来融合多次注意力机制的表示，会使后面的模型层计算成本大幅增加。因此我们提出的多注意力机制为改善表征学习的过程，通过自注意力机制的方式，在子序列编码层的表示向量上加上标量特征值。本方法可以融合多种注意力类型，多注意力机制融合层由自注意力机制、均值池化、对齐池化和最大池化这四种注意力机制的变体构成。其中，最大池化可以在另一个文本中选择出最重要的表征信息。均值池化可以评估出当前表征信息对于另一文本的重要程度。对齐池化则是通过另一文本信息里较为重要的内容代替当前文本内容。自注意力机制可以计算当前文本内容与句子的相似度，并且获得序列和全局的特征信息，解决了自身文本序列长距离依赖的问题。因此多注意力机制层是结合不同注意力的结果，充分利用各自注意力机制的优点，同时还能给模型提供更好的解释性。

互注意力机制是基于两个向量表示的，所以是成对出现的。它能够同时注意文本序列对，并且获取到基于单词之间的近似交互矩阵

$$\boldsymbol{C}=f(\boldsymbol{q}_i;\boldsymbol{a}_j)。\tag{1}$$

式(1)中：f 为多层感知器函数；$\boldsymbol{C}$ 为近似交互矩阵；$\boldsymbol{q}_i$ 为问句文本；$\boldsymbol{a}_j$ 为属性信息。基于互注意力机制的最常见的变体是结合池化技术，以此通过计算近似交互矩阵得到用户问句和属性中基于字符的注意力系数。结合最大池化技术是分别在纵向和横向的维度上进行最大池化，通过每个字符对其他文本序列上的最大影响来注意每个字，即取最大值得到问句和属性新的表示分别为 $\boldsymbol{q}'$ 和 $\boldsymbol{a}'$：

$$\boldsymbol{q}'=\mathrm{soft}(\max(\boldsymbol{C}))^{\mathrm{T}}\boldsymbol{q};\tag{2}$$

$$\boldsymbol{a}'=\mathrm{soft}(\max(\boldsymbol{C}))^{\mathrm{T}}\boldsymbol{a}。\tag{3}$$

单纯的互注意力机制仅仅只是对相对重要的字符进行了加权和评分，而对齐池化将问句中的第 i 个字符与属性中每个字符分别求权重，然后加权求和，这不仅重新调整了序列对，还学习了每个字符的重要性，最后获得加权向量 $\boldsymbol{q}_i'$ 和 $\boldsymbol{a}_i'$：

$$\boldsymbol{a}_i':=\sum_{j=1}^{s_q}\frac{\exp(c_{ij})}{\sum_{v=1}^{s_q}\exp(c_{iv})}\boldsymbol{q}_j;\tag{4}$$

$$\boldsymbol{q}_i':=\sum_{i=1}^{s_a}\frac{\exp(c_{ij})}{\sum_{v=1}^{s_a}\exp(c_{vj})}\boldsymbol{a}_j。\tag{5}$$

式(4)～(5)中：$\boldsymbol{q}_i'$ 为 $\boldsymbol{a}$ 上选择出与问题最有关联且重要的内容，其通过加权求和的方式代表 $\boldsymbol{q}_i$；$\boldsymbol{a}_i'$ 同理。2017 年 Google 公司提出了自注意力机制的概念，它善于捕捉全局信息和并行计算，可以学习自身序列的长距离依赖表示信息。因此对问句文本 $\boldsymbol{q}$ 和属性信息 $\boldsymbol{a}$ 分别进行自注意力机制处理：

$$W_i':=\sum_{j=1}^{l}\frac{\exp(c_{ij})}{\sum_{v=1}^{l}\exp(c_{iv})}W_j。\tag{6}$$

不同的池化操作可以从不同的视角去获取问句文本和属性信息之间的关系及重要程度。均值池化

主要是基于自身字符在其他文本整体内容的关联性去关注每个字，而最大池化可以根据自身每个字符在另一个文本内容字符上的相关性及重要性去选择要关注的字符。在经过不同的池化操作后，通过压缩函数可以获得相应的标量信息，然后与问句和属性的原始表示相结合。

$$y=f([\bar{\boldsymbol{w}};\boldsymbol{w}]); \tag{7}$$

$$y=f([\bar{\boldsymbol{w}}\odot\boldsymbol{w}]); \tag{8}$$

$$y=f([\bar{\boldsymbol{w}}-\boldsymbol{w}])。 \tag{9}$$

式(7)～(9)中：$\boldsymbol{w}$ 为问句内容或属性信息的字符；$\bar{\boldsymbol{w}}$ 为 $\boldsymbol{w}$ 分别经过注意力机制层之后的表征；f 为压缩函数；$\odot$ 为哈达玛乘积。为了从多个角度获得不同的表示，分别使用了乘积、连接、减法 3 种运算方式，然后通过注意力机制产生的值加强原始表征信息。在对问句和属性单独进行自注意力机制的同时将每个问句属性对进行互注意力机制后再分别使用最大池化、对齐池化和均值池化。然后将产生的 3 个注意力标量和初始的表示进行拼接，从而获得新的表示。

1.3　LSTM 和 CNN 的运用

LSTM 设计的初衷是针对一般递归神经网络存在的长距离依赖问题进行解决和提升，同时相比普通神经网络，在训练中的反向传播时可以避免梯度消失的问题。因此本文方法将问句和属性文本信息特征分别输入 LSTM 用于提取文本的长距离依赖特征，生成的高维向量会学习到更深层次的信息。Hochrelter 等[15]在 1997 年提出了 LSTM 的结构，针对原始循环神经网络(recurrent neural network，RNN)模型的问题进行了改进。Graves[16]等对 LSTM 内部结构和机理进行改动和提升，使其更适合用于自然语言处理的研究。LSTM 的结构如图 3 所示。

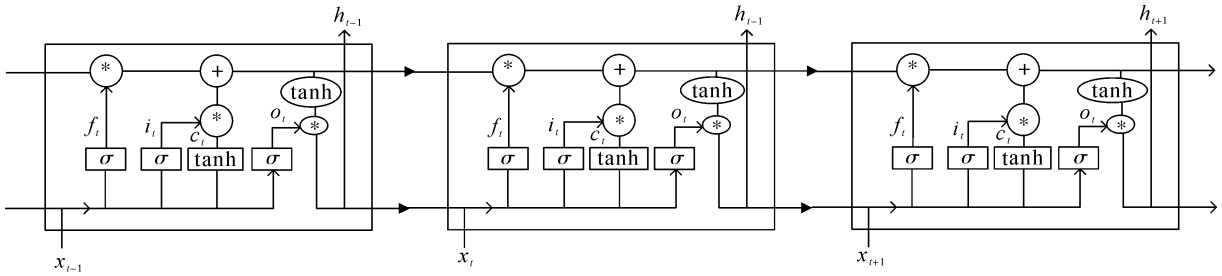

图 3　LSTM 的结构

Fig. 3　Structure of long-short-term memory network

LSTM 的每个神经单元在任意时刻都将进行以下计算：

$$i_t=\sigma(W_i x_t+U_i h_{t-1}+V_i C_{t-1}); \tag{10}$$

$$\tilde{C}_t=\tanh(W_c x_t+U_c h_{t-1}); \tag{11}$$

$$f_t=\sigma(W_f x_t+U_f h_{t-1}+V_f C_{t-1}); \tag{12}$$

$$c_t=f_t\odot c_{t-1}+i_t\odot\tilde{c}_t; \tag{13}$$

$$o_t=\sigma(W_o x_t+U_o h_{t-1}+V_o c_t); \tag{14}$$

$$h_t=o_t\odot\tanh(c_t)。 \tag{15}$$

式(10)～(15)中：c_t 为 t 时刻记忆单元的值；σ 为 sigmoid 函数 f_t；i_t、o_t、$\tilde{C}_t$ 分别为遗忘门、输入门、输出门在 t 时刻的值和候选记忆单元；h_t 为 t 时刻 LSTM 的输出。

CNN 是一种主要通过卷积层、池化层和全连接层组合而成的前馈神经网络[17]。CNN 可以处理高维数据，无须手动选取特征并且能够针对输入的局部信息进行感知和权值的共享。通过卷积神经网络来对经过多注意力层后的高维特征进行建模，卷积后得到新特征。在这个过程中，卷积核的大小为 $h\times d$，h 是滑动窗口的大小，d 是文本向量维度，具体文本特征 y 的计算公式如下：

$$y=\sigma(W\boldsymbol{x}+\boldsymbol{b})。 \tag{16}$$

式(16)中：y 为经过卷积操作后获得的特征；σ 为非线性激活函数；使用的激活函数为 ReLu 函数；$\boldsymbol{W}$ 为卷积核的权重参数矩阵；$\boldsymbol{x}$ 为文本向量表示；$\boldsymbol{b}$ 为偏置项。

1.4 属性映射

通过 softmax 分类器作为输出层来实现对问句文本的属性映射，预测概率值的公式如下：

$$P=\mathrm{softmax}(\boldsymbol{W}\boldsymbol{x}+\boldsymbol{b})。\tag{17}$$

式(17)中：P 为包含多个概率值的预测标签概率集。学习的过程中，损失函数为最小化交叉熵的误差，同时为了防止过拟合加入了正则项，损失函数公式如下式：

$$J(\theta)=\sum_i\sum_j y_{ij}\ln\hat{y}_{ij}+\lambda\|\theta\|^2。\tag{18}$$

式(18)中：$J(\theta)$为属性类别；y_{ij}为问句对应属性的正确类别；$\hat{y}_{ij}$为问句对应属性的预测类别；i 为样本数；$L2$ 正则项系数为 λ。

2 试验结果与分析

2.1 试验数据集及预处理

自然语言处理与中文计算会议(the natural language processing and Chinese computing conference，NLPCC 2018)为知识库问答(KBQA)任务提供了一个开源数据集。该数据集提供了一个包含 24 479 个一一对应的问答数据对和所需的知识图谱，提供的知识图谱包含 6 502 738 个实体、587 875 个相关属性及其构成的 43 063 769 个三元组集合，是目前中文开放领域规模较大的知识图谱。NLPCC 2018 知识库问答任务提供的知识图谱内容具体形式见表 1。

表1 NLPCC 2018 知识库问答任务提供的知识图谱内容

Table 1 Knowledge graph provided by knowledge base question and answer (KBQA) task in NLPCC 2018

实体	属性	属性值/实体
浙江村	别名	浙江村
浙江村	中文名	浙江村
浙江村	主要构成	温州人
浙江村	位置	北京丰台区大红门、木樨园地区
浙江村	辐射范围	周围 5 公里区域
浙江村	中心	木樨园桥环岛

考虑到 NLPCC 2018 中知识库问答任务提供的开源数据集并非属性映射任务领域数据集，且该数据集提供了 24 479 个问答数据对，问答数据对并未明确标注实体所在位置，为获得数据集中问句中的实体信息，本文通过远程监督的方法，基于 NLPCC2018-KBQA 所给三元组数据构成的知识图谱，远程监督标注 NLPCC2018-KBQA 问答对数据中问句的实体序列。基于远程监督的问答对问句实体标注算法流程如图 4 所示。

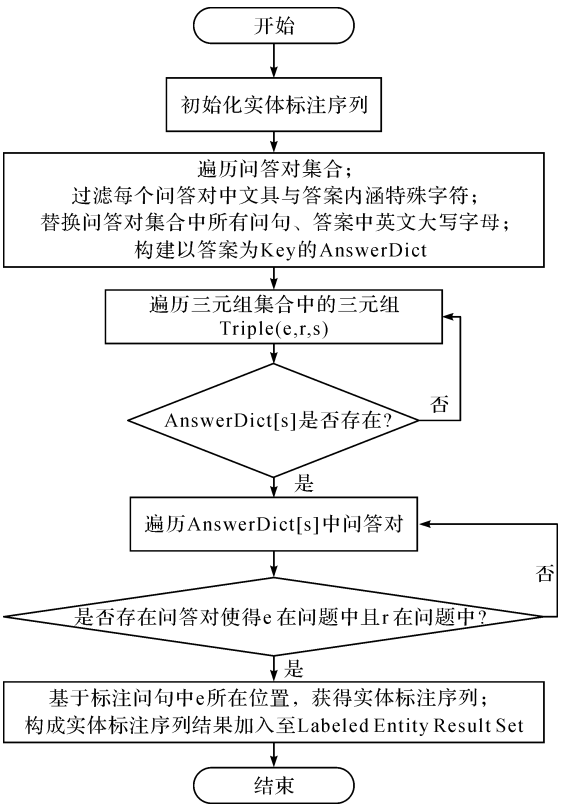

图 4 基于远程监督的问答对问句实体标注算法流程图

Fig. 4 Algorithm flow chart of question-answer pair question entity labeling based on remote supervision

将数据集里包含的 24 479 个问答数据对、mention2id 文件及 43 063 796 个知识图谱三元组的集合通过基于远程监督的问答对问句实体标注算法进行标注。该算法首先形成长度为 21 269 的映射字典，即将问答对建立表格时词典的大小，也是数据中不重复的答案个数。然后从问句中查找是否包含对应三元组中的实体，如果不包含该实体，通过 mention2id 文件，把实体名替换成它对应的同义词继续查找。即使这样还是有实体在对应的问句中查找不到，面对这种情况就舍弃掉这部分数据。最后，通过基于远程监督的问答对问句实体标注算法实际成功标注有效问答对 15 640 项。

将找到实体信息的 15 640 问答对，通过基于知识图谱的关系抽取负采样算法，构造属性映射模块所需的样本数据。试验中分别对每一问答对问句实体样例进行 1 次正采样与 4 次负采样，使每一个样例正样本与负样本比值不要与知识图谱平均实体相关个数差距太大。采样后最终获得 75 325 个属性映射样本。负采样属性映射样本示例见表 2。

表 2　负采样属性映射样本示例

Table 2　Samples of negative sample attribute mapping

序号	问句内容	属性	正负标记
0	比较好的福尔摩斯探案全集译者是谁呀	译者	1
1	比较好的福尔摩斯探案全集译者是谁呀	追赠	0
2	比较好的福尔摩斯探案全集译者是谁呀	幅面	0
3	比较好的福尔摩斯探案全集译者是谁呀	费用	0
4	比较好的福尔摩斯探案全集译者是谁呀	后期	0

根据本试验设计，将数据集拆分，80％作为训练集，10％作为验证集，10％作为测试集，将保存在验证集上效果比较好的模型用于测试集上进行评估。

2.2　评估指标的选择

评价指标体系包括以下几部分：精确率 P、召回率 R、F_1 值、宏平均精确率 $\overline{P}$、宏平均召回率 $\overline{R}$、宏平均 $\overline{F}_1$ 值、准确率 A，用以评估属性映射模块的准确程度。

准确率 A 是模型最终分类正确的样本数量与样本总数的比值。

$$A=\frac{T_{\mathrm{P}}+T_{\mathrm{N}}}{T_{\mathrm{P}}+F_{\mathrm{P}}+T_{\mathrm{N}}+F_{\mathrm{N}}}。\tag{19}$$

式(19)中：T_{P} 为实际为正且预测为正的样本数；F_{P} 为实际为负但预测为正的样本数；F_{N} 为实际为正但预测为负的样本数；T_{N} 为实际为负预测为负的样本数。

精确率 P 是模型在所有预测为正例的样本中真正样本所占的比例。

$$P=\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{N}}}。\tag{20}$$

召回率 R 是在所有真正样本中模型分类正确的样本所占的比例。

$$R=\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{N}}}。\tag{21}$$

召回率 R 和精确率 P 是一对负相关的指标，通常 P 越高，R 越低；反之亦然。单一的召回率和精确率并不能客观全面地评价模型效果，而以 P 和 R 为基础的调和平均 F_1 值可以较好地评价模型的综合效果。

$$F_1=\frac{2PR}{P+R}。\tag{22}$$

如果在多个混淆矩阵上综合考察 P 和 R，需要计算各混淆矩阵上的 P 和 R 并求出平均值，从而得到宏平均精确率 $\overline{P}$、宏平均召回率 $\overline{R}$ 及宏平均 $\overline{F}_1$ 值。

$$\overline{P}=\frac{1}{n}\sum_{i=1}^{n}P;\tag{23}$$

$$\overline{R}=\frac{1}{n}\sum_{i=1}^{n}R;\tag{24}$$

$$\overline{F}_1=\frac{2\overline{P}\overline{R}}{\overline{P}+\overline{R}}。\tag{25}$$

式(23)～(25)中：i 为测试集中的样本数。

2.3　对比试验的模型选择

使用 2.1 节预处理后的试验数据针对属性映射任务进行对比试验的设计。通过基于变压器的双向编码器表示模型(bidirectional encoder representation from transformers，BERT)根据具体任务进行微调

的方式进行词嵌入，词嵌入后文本内容的向量维度为 100 维，LSTM 隐藏层维度设置为 100，随机失活参数设置为 0.4，卷积核大小分别为 3、4、5，共 128 个卷积核，学习率为 0.001。对比试验选取的模型分别为 TextCNN、TextRNN、DPCNN(deep pyramid convolutional neural networks)、TextRCNN、PKU 团队[18]和 Lai[19]提出的模型、InsunKBQA[20]及我们提出的基于多注意力多维文本的属性映射模型。

2.4 结果分析

表 3 是对比试验模型与我们提出的基于多注意力多维文本的属性映射模型在 2.1 节预处理后的数据集上试验后所得出的各项指标评估值。由表 3 可知，与基于卷积神经网络的模型相比，本文方法比 TextCNN 和 DPCNN 分别提高了 6.43%和 5.54%的准确率；与基于 LSTM 的 TextRNN 模型相比，本文方法提高了 6.62%的准确率。TextRCNN 是综合了 RNN 和 CNN 二者优点的分类模型，与它相比本文方法提高了 2.94%的准确率。PKU 团队[18]和 Lai 等[19]提出的模型分别在 NLPCC-KBQA 2016 年和 2017 年的比赛中取得了最好成绩。InsunKBQA[20]是周博通等基于 NLPCC-KBQA 数据集，利用提取的方法设计的问答系统，其中包含实体识别模块与属性映射模块。在同源数据集上，本文方法与上述 3 个模型相比，准确率分别提高 5.95%、4.17%、2.36%。由此可见，我们提出的基于多注意力多维文本的模型在属性映射任务中有较好的表现。这表明利用多注意力机制层和多维文本表示有助于理解用户问句的意图，能有效提升属性映射任务的精度。

表 3 属性映射对比试验得出的各项指标评估值

Table 3 Evaluation value of each index obtained by comparative test of attribute mapping %

模型	P	R	F_1	A
TextCNN	88.13	89.80	88.95	90.36
TextRNN	90.59	85.70	88.08	90.17
DPCNN	86.19	92.40	89.19	91.25
TextRCNN	94.57	88.80	91.59	93.85
PKU 团队	89.47	90.10	89.79	90.84
Lai	92.35	91.70	92.02	92.62
InsunKBQA	93.78	92.00	92.88	94.43
本文方法	97.06	95.60	96.32	96.79

为了探究多注意力机制融合层内部不同模块之间的相互影响及整体性能，我们设计了相应的消融试验。通过分别移除模型中的均值池化、对齐池化、最大池化和自注意力机制模块，与原始模型进行对比，来观察各模块对模型的影响。属性映射消融试验结果见表 4。

表 4 属性映射消融试验结果

Table 4 Experimental results of attribute mapping ablation %

模型	F_1	A
本文方法	96.32	96.79
删除均值池化模块后的模型	91.89	95.47
删除最大池化模块后的模型	92.15	96.04
删除对齐池化模块后的模型	78.26	79.52
删除自注意力机制模块后的模型	91.17	93.63

根据表 4 的结果我们可以看出，移除均值池化模块和移除最大池化模块后模型性能都只是略微下降，移除自注意力机制模块后模型性能下降程度相比较而言略微明显。可见，在中文领域自注意力机制模块对文本内容的长期依赖作用，可以提升属性映射任务中问句和属性内容的特征表示能力，从而增强属性映射任务整体性能。而影响模型性能最显著的是对齐池化模块，移除该模块后多注意力机制融合层的作用大幅下降，因此该模型的多注意力机制融合层中对齐池化模块的重要程度最高。由此可以得出，本研究结合这几种注意力机制所提出的多注意力多维文本模型在属性映射任务上是有效的。

3 结 语

属性映射是可以提升知识图谱问答系统整体性能的重要组成部分，本文针对智能问答中属性映射部分进行相关探究。难点在于中文的表述形式不同，容易出现语义歧义现象。针对该问题，本文提出了基于多注意力多维文本的属性映射模型以解决单一文本表示和注意力机制的局限性，同时利用基于远程监

督的问答对问句实体标注算法和基于知识图谱的关系抽取负采样算法，将 NLPCC 2018 中知识库问答(KBQA)任务提供的开源数据集转变为可用于属性映射任务的数据格式。在经过预处理后的数据集上，本文设计的模型取得了较为理想的成绩，设计的消融试验表明相对单一注意力机制，本研究提出的多注意力机制融合层对模型性能实现了显著提升。可见，本研究提出的基于多注意力多维文本的属性映射模型能够有效提高问答系统中属性映射任务的精度。然而在试验设计中依然存在一些不足之处，例如没有进行实体消歧、实体统一，这可能会导致出现属性值与正确答案不匹配的问题。此外，模型未涉及其他垂直领域的问答知识语料，今后我们将围绕以上不足点做进一步探究。

参考文献：

[1] 李跃艳，王昊，邓三鸿，等. 近十年信息检索领域的研究热点与演化趋势研究：基于 SIGIR 会议论文的分析[J]. 数据分析与知识发现，2021，5(4)：13.

[2] 郭肇强，周慧聪，刘释然，等. 基于信息检索的缺陷定位：问题，进展与挑战[J]. 软件学报，2020，31(9)：2826.

[3] 马忠贵，倪润宇，余开航. 知识图谱的最新进展、关键技术和挑战[J]. 工程科学学报，2020，42(10)：1254.

[4] 王媛，时恺泽，牛振东. 一种用于实体关系三元组抽取的位置辅助分步标记方法[J]. 数据分析与知识发现，2021，5(10)：71

[5] 武书钊，李功权，卜明伟. 基于知识图谱的自杀倾向检测问答系统构建[J]. 计算机工程与应用，2021，57(22)：304.

[6] 陶志勇，李小兵，刘影，等. 基于双向长短时记忆网络的改进注意力短文本分类方法[J]. 数据分析与知识发现，2019，3(12)：21.

[7] 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. Vancouver：Association for Computational Linguistics，2017.

[8] YU M，YIN W，HASAN K S，et al. Improved neural relation detection for knowledge base question answering[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver：Association for Computational Linguistics，2017.

[9] DOS S C，GATTI M. Deep convolutional neural networks for sentiment analysis of short texts[C]//Proceedings of the 25th International Conference on Computational Linguistics. Dublin：Dublin City University and Association for Computational Linguistics，2014：69.

[10] 聂维民，陈永洲，马静. 融合多粒度信息的文本向量表示模型[J]. 数据分析与知识发现，2019，3(9)：45.

[11] 陈功，黄瑞章，钟文良. 基于社交特征的多维度文本表示方法[J]. 计算机工程与科学，2016，38(11)：2348.

[12] 江明奇，沈忱林，李寿山. 面向问答文本的属性分类方法[J]. 中文信息学报，2019，33(4)：125.

[13] 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.

[14] 童国烽. 基于知识库的开放域知识问答系统研究[D]. 南京：南京大学，2018.

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

[16] GRAVES A. Supervised sequence labelling with recurrent neural networks[M]//KACPRZYK J. Studies in Computational Intelligence. Heidelberg：Springer Berlin，2012.

[17] 袁祯泷. 关于卷积神经网络的研究进展报告[J]. 自动化应用，2020(7)：87.

[18] LAI Y，LIN Y，CHEN J，et al. Open domain question answering system based on knowledge base[C]//The Fifth Conference on Natural Language Processing and Chinese Computing & The Twenty Fourth International Conference on Computer Processing of Oriental Languages (NLPCC-ICCPOL 2016). Cham：Springer，2016：722.

[19] LAI Y，JIA Y，LIN Y，et al. A chinese question answering system for single-relation factoid questions[C]//National CCF Conference on Natural Language Processing and Chinese Computing. Cham：Springer 2018：124.

[20] 周博通，孙承杰，林磊，等. InsunKBQA：一个基于知识库的问答系统[J]. 智能计算机与应用，2017，7(5)：150.