文章编号：2095-2163(2020)09-0124-03 中图分类号：TP391.12 文献标志码：A

基于 KV-MemNN 的心血管病自动问答系统设计与实现

黄诗怡，李继云
（东华大学 计算机科学与技术学院，上海 201620）

摘 要：目前心血管病患病率居高不下，死亡率也高居首位，严重影响着居民的生活。心血管病已经成为重大的公共卫生问题，加强心血管病知识的普及势在必行。本文借助于现有的心血管病知识图谱，构建了支持五种需求的问答数据集，并且实现了基于 KV-MemNN 模型的心血管病自动问答系统。通过测试数据的评估，验证了本文设计方案的适用性。

关键词：自动问答；心血管病；知识图谱；KV-MemNN

The design and implementation of CVD question answering system based on KV-MemNN

HUANG Shiyi, LI Jiyun
(School of Computer Science and Technology, Donghua University, Shanghai 201620, China)

【Abstract】 The prevalence of cardiovascular disease (CVD) remains at a high level at present, and the mortality rate has been in the leading position among major diseases for the last few years, leading to not only a severe influence on residents´ life but also a gradually increasing burden on the national economic. This paper draws on the existing CVD knowledge graph and constructs a question answering (QA) dataset of which the questions are designed to fulfill five kinds of needs, and then realizes the CVD question answering process based on the KV-MemNN model. The evaluation result on the test set verifies the applicability of the design scheme used in the paper.

【Key words】 question answering; cardiovascular disease; knowledge graph; KV-MemNN

0 引 言

目前中国心血管病(CVD,cardiovascular disease)病人数量高达 2.9 亿。随着城市化进程的加速，人口老龄化问题突出，及吸烟、身体活动不足、不合理膳食等不良生活习惯的盛行，心血管病患病率持续居高不下，并呈现出上升趋势。心血管病死亡率占据首位，2018 年心血管病死亡分别占农村、城市居民疾病死因的 46.66%、43.80%[1]。心血管病住院总费用快速增长，国家心血管病的负担逐渐加重，心血管病已经成为重大的公共卫生问题。虽然国内已经展开一系列心血管病社区防治工作并取得一定的成效，心血管病医疗质量也在不断提高，但是随着国民心血管病多个危险因素的流行趋势明显，防治工作仍然面临着严峻的挑战[2]。

自动问答系统作为互联网时代信息获取的一种有效途径，为心血管病的防治带来了新的机遇，成为进一步加强心血管病知识普及的一种手段。自动问答系统与书籍、海报、杂志等传统的信息获取方式不同，与传统的搜索引擎也有所区别，目的在于提高信息获取的效率，可以根据用户提出的问句，直接返回精准而简洁的答案。本文以现有的心血管病知识图谱为数据支持，构建一个基于 KV-MemNN(key-value memory network)模型的心血管病自动问答系统，通过挖掘知识图谱中有价值的信息，加速心血管病知识的普及。

1 知识图谱简介及预处理

本文使用的心血管病知识图谱共包含 1 173 个实体，2 381 组实体关系和实体属性。实体包括心血管病、症状以及药物；实体关系包括疾病与症状之间的关系、疾病与药物之间的关系，在此分别定义为相关症状、常用药物关系。一个以心肌梗塞为中心的知识图谱示例，如图 1 所示。可知一种疾病往往对应多个相关症状关系、多个常用药物关系。

为了确保知识图谱的可用性，对其中的数据进行检查与修正。

(1)数据存储于 5 个.csv 文件中，存在个别疾病名称不一致的情况，需要统一修改为正确的名称；

(2)个别疾病名称与数据来源中的名称不同，需要修正为数据来源中的名称；

(3)该知识图谱主要抽取自 39 健康网的疾病百科，又有部分其他来源的补充信息，存在少量表述方式的不一致，个别疾病与症状重名，但数量极少，影响不大，不作处理。此外，疾病的别名内容比较复杂，除

作者简介：黄诗怡(1993-)，女，硕士研究生，主要研究方向：知识图谱、自动问答；李继云(1969-)，女，博士，教授，主要研究方向：数据工程机器学习、人工智能等。

收稿日期：2020-07-15

了包含传统意义上的别称，还可能包含疾病的父类、某种子类、某种特征等等，导致疾病的一些别名可能与其他疾病或症状重名，需要对冗余的别名进行清理。为了完全避免部分别名与其他疾病或症状的重名问题，将一种疾病的所有别名作为一个整体看待。

处理后的知识图谱以 SPO 三元组（subject，predicate，object）的形式统一存储，并用实体表保存疾病、药物、症状三种实体。

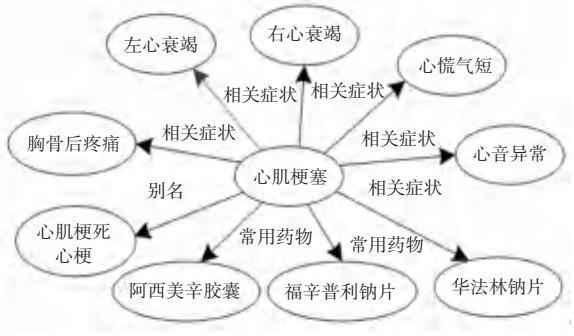

图 1　以心肌梗塞为中心的知识图谱示例

Fig. 1　A knowledge example centered on myocardial infarction

2　基于 KV-MemNN 模型的自动问答

本文运用 Facebook AI 研究院的 Miller 等人提出的 KV-MemNN 模型实现基于心血管病知识图谱的自动问答，这个过程主要分为两部分：问答数据集构建和自动问答实现。问答数据集构建为自动问答提供了数据支撑，自动问答实现则包含了最为关键的数据处理和模型计算过程。

2.1　问答数据集构建

问答数据集依赖于知识图谱构建。针对知识图谱中的关系和属性，设计不同类型的问句模板。已知知识图谱中包含疾病与药物之间的常用药物关系、疾病与症状之间相关症状关系、疾病的别名属性，以此设计根据疾病查询常用药物、相关症状、别名三种类型的问句模板。为了进一步丰富问句类型，对常用药物、相关症状两种关系进行逆向拓展，得到药物与疾病之间的依存关系、症状与疾病之间的依存关系，以此设计根据药物查询可治疗疾病、根据多个症状排查疾病两种类型的问句模板。问句模板通过 39 健康网、百度知道以及使用搜索引擎检索到的相关网页信息搜集整理所得，与知识图谱中对应的三元组进行整合。

整合过程分为两种情况：其一是问句模板与单个实体整合，以相关症状关系为例，由于一种疾病对应多种症状，问句模板与同一种疾病的若干三元组进行整合，得到的问句以多个症状为答案；其二是问句模板与多个实体整合，只应用于相关症状的逆向关系，找到多个症状所对应的共同疾病，将问句模板与这些症状进行整合，得到的问句以共同疾病为答案。

问答数据集由五种问答数据整合，最终得到 13 062组问答数据，每种问答数据以大约 8∶1∶1 的比例分别作为训练集、验证集、测试集。训练集中包含所有问答模板。

2.2　自动问答实现

自动问答基于 KV-MemNN 模型来实现。KV-MemNN 模型建立于 Weston 等人的记忆网络和 Sukhbaatar 等人的端对端记忆网络的基础之上，其体系结构如图 2 所示。模型将数据源中的内容存储于结构为（键，值）的记忆中，从而为知识源的编码提供了更大的灵活性，使得模型能通过键来寻找与问题相关的记忆，并得到这些记忆相对应的值。因此，键的设计应该包含与问题匹配的特征，值的设计应该包含与答案匹配的特征。

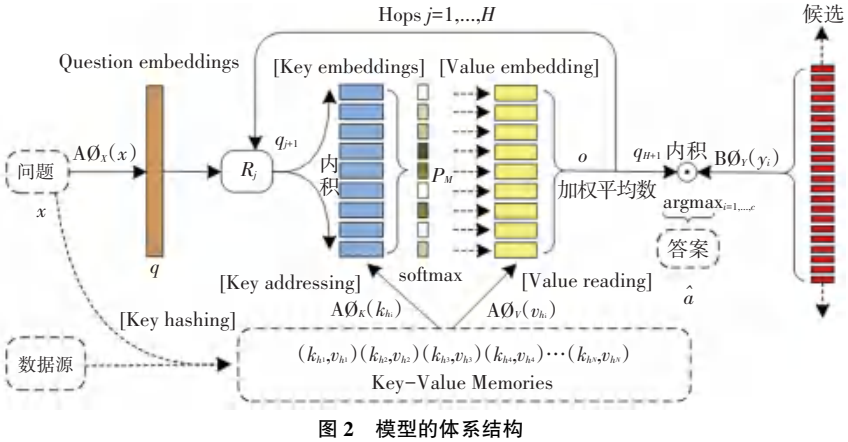

图 2　模型的体系结构

Fig. 2　Architecture of the model

本文以心血管病知识图谱作为模型的数据源，自动问答实现的步骤如下：

步骤 1　对问句进行预处理，包括问句分词和去停用词。问句分词以实体表为辅助，在预先完成

实体匹配的基础上进行。因为疾病、药物、症状 3 种实体名称中专有名词比较多,并且长短不一,直接分词可能造成混乱,从而导致问句原意完全改变,而这种方法可以避免这些名词对分词的干扰,保证准确识别出问句中的所有实体。本文选用哈工大 LTP 进行分词,所得结果可以满足分词的需求。问句的停用词包括标点符号、语气词、以及“请问”、“目前”等对问答过程不造成影响的短语。

步骤 2 数据的向量表示。首先,构建一个词汇表,本文的词汇表由知识图谱以及训练集中的词汇组成,以词汇的出现频数倒序排列,预留空字符处于排列首位;其次,以各个词汇在词汇表中的位置作为索引 id,将知识图谱和训练集、验证集、测试集中的文本信息转换为数值形式,使得知识图谱的三元组、问答数据集的问句与答案,均能够使用向量形式来表达。知识图谱中的三元组转换为键-值对形式,即向量 $(k_1,v_1),\cdots,(k_M,v_M)$,作为模型的记忆。键由 SPO 三元组中的 subject 和 predicate 组成,值用 object 来表示,使得键与值分别包含与问句、答案匹配的特征。

步骤 3 Key hashing,也就是找出所有与问句有关的记忆。首先,借助知识图谱,构建实体与其相关记忆之间的哈希表;其次,以问句中的实体为媒介,找出与各个实体相关的记忆 $(k_{h_1},v_{h_1}),\cdots,(k_{h_N},v_{h_N})$。

步骤 4 训练部分。首先是 Key addressing 和 Value reading 的迭代过程。在寻址阶段计算问句与相关记忆中每个键的相关性评分,公式(1):

$$p_{h_i}=Softmax(A\emptyset_X(x)\cdot A\emptyset_K(k_{h_i})). \quad (1)$$

其中,$\emptyset_{\cdot}$ 表示 D 维特征映射,A 是一个 $d\times D$ 的特征矩阵。在读取阶段,以相关性评分为权重,对值加权求和,得到一个输出向量 o,公式(2):

$$o=\sum_i p_{h_i}A\emptyset_V(v_{h_i}). \quad (2)$$

假定 $q=A\emptyset_X(x)$,使用输出向量 o 对其更新,本文采用的是 $q_{j+1}=R_j(q_j+o)+b_j$,其中 R、b 分别为 $d\times d$、$d\times 1$ 的矩阵,则寻址阶段的公式修改为公式(3):

$$p_{h_i}=Softmax(q_{j+1}^T A\emptyset_K(k_{h_i})). \quad (3)$$

然后是结果预测。经过 H 轮迭代后得到的预测结果为式(4):

$$\hat{a}=argmax_{i=1,\cdots,C}Softmax(q_{H+1}^T B\emptyset_Y(y_i)). \quad (4)$$

其中,y_i 代表所有候选输出,本文中 B 的取值与 A 保持一致。

模型通过 Adam 优化算法来最小化交叉熵损失,从而实现模型中矩阵 A、B、$R_1,\cdots,R_H$ 以及 $b_1,\cdots,b_H$ 的更新。训练过程使用若干个 epoch,每个 epoch 对训练集中的问答数据进行一轮训练,如果一个问答数据中包含多个答案,则问答数据分为多个训练数据依次进行训练。验证集中的问答数据用于评估最佳模型,即准确率最高的模型,准确率的评估标准是所有候选答案中排名首位的预测答案是否正确。

2.3 结果评估

最终模型在 $d=230$,$Hops=5$ 时取得,对测试集的评估结果见表 1。从评估结果可知,对于大多数问答数据排名首位的预测结果是准确的,并且无论是单个或是多个答案的问答数据,各个答案的相关排名都是比较靠前的。

表 1 评估结果

Tab. 1 Evaluation results

准确率/%	MRR	MAP
96.5	0.981	0.975

对测试集中的五种问答数据进行分类评估,其结果见表 2。已知前四种为多个答案的问答数据,最后一种为单个答案的问答数据,两种数据都能得到较好的结果。此外,经过对训练集、验证集、测试集多次数据分配可知,不同的数据分配会影响五种问答数据各自的结果,但是测试集的整体结果差异不大。

表 2 五种问答数据评估结果

Tab. 2 Evaluation results of the five-kind QA data

类型	准确率/%	MRR	MAP
根据疾病查询常用药物	80.6	0.903	0.826
根据药物查询可治疗疾病	100.0	1.000	0.968
根据疾病查询相关症状	100.0	1.000	0.838
根据多个症状排查疾病	96.6	0.981	0.982
根据疾病查询别名	100.0	1.000	1.000

3 结束语

本文以面向心血管领域的知识图谱为基础,构建了心血管病问答数据集,实现了基于 KV-MemNN 模型的自动问答系统,使得心血管病知识图谱能够以更容易接受的方式服务于非医学专业人员,对心血管病知识的传播具有一定的意义。同时也存在一些不足之处需要优化,其一是需要医学专家的指导对知识图谱进行更加全面的检查与错误纠正;其二是知识图谱的内容不够丰富,有待进一步扩充;其三是问答数据集的模板库还不够广泛,需要进一步搜集;此处可以将预训练的词向量应用于模型,进一步提高准确率。

参考文献

[1] 国家卫生健康委员会. 2019 中国卫生健康统计年鉴[M]. 北京:中国协和医科大学出版社, 2019.

[2] 国家心血管病中心. 中国心血管病报告 2018[R]. 北京:中国大百科全书出版社, 2019.