DOI: 10.13718/j.cnki.xdzk.2020.10.004

基于 LCN 的医疗知识问答模型

马满福[1]，　刘元喆[1]，　李　勇[1]，　王　霞[2]，
贾　海[2]，　史彦斌[3]，　张小康[4]

1. 西北师范大学 计算机科学与工程学院，兰州 730070；2. 甘肃省人民医院，兰州 730000；
3. 兰州大学 药学院，兰州 730000；4. 兰州七度数聚技术有限公司，兰州 730070

摘要：中文医疗领域分词比较困难，导致现有算法对于医疗问题特征提取不充分，针对中文分词的特点，提出基于 LCN(Lattice CNN，格子卷积神经网络)的医疗知识问答模型. 首先，利用某三甲医院提供的 15 000 份电子住院记录，基于电子住院记录利用 Glove 模型训练医学词向量. 其次，通过各大医疗网站获得大量医学名词及名词间的关系，构建医学知识图谱，并提取知识图谱中的关系词，结合已训练的词向量获取关系向量. 最终，以医学词向量作为模型输入端并利用 LCN 神经网络提取医疗问题特征，计算问题特征与关系向量的相似度，进而训练医疗知识问答模型. 实验表明，LCN 模型准确率可达 89.0%，与同类问答模型比较，提高了 2%.

关　键　词：医疗知识问答；Glove 模型；LCN；知识图谱；电子病历

中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1673-9868(2020)10-0025-12

随着医学领域信息资源的日渐丰富，消费者需要专门解决方案来适应健康相关信息的异质性和特点[1]. 在线医疗保健社区可以为用户提供远程医疗支持，既给用户带来了便利，又有助于积累大量的数据. 同时，与数据量的爆炸性增长相比，医生的数量相当有限[2]. 医疗问答可将患者提出的问题进行整合、分析，利用机器学习算法训练智能问答模型，再利用其自动解答患者的疑问，从而减少医生的工作量.

本文采用知识问答模型构建医疗问答系统解决上述问题，知识问答模型不同于传统的基于文档的问答模型. DBQA(Document-based question answering，文档问答系统)采用自然语言表达方式进行提问，返回包含着答案的文档，用户需要阅读已存在的文档发现相关答案. 而 KBQA(Knowledge Base Question Answering，知识问答系统)通过理解问句的意图，利用人工定制的句法解析树，将自然语言处理为 Select 语句，查询数据库可直接返回答案. 所以，医学知识问答模型需要人工对句法解析树不断添加新词汇和映射机制，成本过大. 近些年基于短文本匹配的知识问答模型逐渐发展起来，但我国的医疗知识问答模型仍存在着一些挑战和限制. 首先，我国还没有现有的中文医疗知识问答库. 其次，针对中文领域的医学词库的构建尚不成熟，没有较好的中文分词工具处理专业领域文本. 由于医学文本的分词效果不佳，导致现有研究诸如疾病预测或医疗问答问题中利用深度学习模型很难提取到医学文本的特征.

收稿日期：2020-09-29
基金项目：国家自然科学基金项目(71764025，61863032，61662070)；甘肃省中医药管理局科研课题(GZK-2019-40)；全国高等院校计算机基础教育教学研究项目(2020-AFCEC-355)；甘肃省教育科学规划课题研究项目(GS[2018]GHBBKZ021，S[2018]GHBBKW007)；西北师范大学青年教师科研能力提升计划项目(NWNU-LKQN-17-9).
作者简介：马满福(1968－)，男，博士，教授，主要从事分布计算和移动计算的研究.

针对以上问题，本文通过各大医疗网站爬取大量医疗问题与疾病常识，将疾病常识存储为“实体—关系—实体”形式的医疗知识图谱[3]. 根据医疗问题和其对应的医疗关系，构建“问题—关系”一对一的医疗知识问答库. 本文还利用LCN模型提取问题特征，LCN中的格子可以提取问句中的所有分词情况并把它们转为特征向量，充分概括问句的特征信息，解决因为分词工具不成熟所导致的特征提取模糊. 基于LCN的问答模型将医疗问答转换为一个选择最佳关系的文本匹配问题，并根据匹配提高问题与标签答案的相似度从而构建问答模型. 当新问题输入时，根据新问题的特征，选取相似度最高的对应答案. 无需人工定制句法分析树，节省人工成本. 最终通过实验，LCN模型准确率可达89.0%，比同类知识问答模型准确率高出2%.

医疗知识问答模型需要先构建医疗知识图谱，近年来多位学者利用不同的方法构建了医疗知识图谱，并结合知识图谱解决医疗问题. Li X等利用膝骨关节炎患者的电子病历文本构建医学知识图谱[4]，以支持诸如知识检索和决策之类的智能医学应用，并促进医学资源的共享. Chai X Q提取生物医学实体之间的关系以构建生物医学知识图谱[5]，并利用知识图谱嵌入方法将知识图谱中的实体和关系转换为低维连续向量. 最后，将已知的病理疾病关系数据用于训练双向长短期记忆网络(Bi-STLM)的疾病诊断模型. Yuan J B等利用弱监督的方法提取医疗文本中的实体与关系词构建医疗知识图谱[6]. 现有研究主要利用自然语言处理的方法抽取电子病历信息构建知识图谱，抽取到的实体或关系存在部分错误. 本文利用爬虫方法对于医疗网站中的实体和关系进行抽取，并存储于Neo4j数据库中，有着较高的准确性.

构建知识图谱成功后，需要将医疗问题和其对应的知识进行匹配，并计算两者的相似度构建医疗知识问答模型. 现有的医学问答系统主要分为文本问答系统和知识问答系统. 近些年，文本问答系统已有多位学者进行了研究，Liu H I等[7]提出了一种基于CNN的自我注意嵌入式模型的中医问答系统，利用LSTM分析问题特征，并通过CNN的卷积核获取特征图，最终通过CNN的池化层提高模型的准确性. Nguyen V等[8]利用文本推断的方法识别具有相似语义的问题，改进了自然语言推理和问题蕴含的方法，进一步完善了医学问答，提出了结合开放领域和生物医学领域以改善语义理解和语义消歧的问答系统MEDIQA. Zhang S等[9]利用答案匹配的方法，提出了一种端到端的字符级多尺度卷积神经框架cMedQA，使用CNN从不同比例的问题或者答案中提取上下文信息，进而通过相似度的计算完成医疗问题与答案的匹配.

文本问答系统返回非结构化文本答案，是概念性片面化的文本，更适用于回答“为什么会患病”之类的问题，需要人们从散乱的答案之中进一步分析获取自己想要知道的信息，不能直接满足人们的需求. 然而现有医疗问答系统则会直接通过分析医疗问题，从知识库中获取和问题相关的实体词，并通过自然语言处理方法构成包含答案的简短语句，人们可从系统返回的答案中直接获得所需信息，更适用于“所患什么疾病”“吃什么药”等问题的答案，显然知识问答系统更适用于医疗领域.

现有的医疗知识问答系统对医疗知识普及以及医生的临床用药决策有着重要的意义和参考价值，近年来人们提出了利用神经网络方法计算问题与答案的相似度. Zou Y等[10]以中医药领域为基础，以中医药网站《本草纲目》的开源数据为数据源，建立了中药知识图谱，根据知识图谱实现自动答疑和辅助处方的功能. Zhu W等[11]提出问答系统Dr-KGQA，利用bilstm-crf提取医学文本中的实体和关系，构建医疗知识图谱，并使用text-CNN将医疗问题和图谱中的关系词进行匹配，构建问答模型. Sadid A H等[12]基于T-Know中医药知识图谱服务系统，抽取电子病历中的三元组，并在知识图谱的基础上，开发了用于单个问题理解和多轮对话的深度学习算法. Zhang Y Y等[13]构建了一种多模态知识感知层次注意网络MKHAN，通过利用多模态知识图谱解决医学问题，通过组合实体结构、语言学和视觉信息来生成路径，并通过利用MKG路径中的顺序依存关系来推断问答互动的基本原理.

基于相似度方法采用了相对统一的RDF表示知识图谱，并且把语义理解的结果映射到知识图谱的本体后生成SPARQL查询解答问题系统，通过本体可将用户问题映射到基于概念拓扑图标识的查询表达式中，相当于知识图谱中的子图，基于相似度算法不断完善对用户问题特征的提取，以便于找出问题到知识图谱子图的最合理映射. 上述方法虽然很全面地抓取了问题中的信息，但并未解决针对专业领域的分词困

难问题，Lai Y X 等[14]提出了 LCN 模型，通过单词格抓取语言问题中的多粒度信息提高匹配的准确性，本文利用 LCN 的多粒度抓取特征方法解决医疗专业分词困难的问题，进而训练医疗知识问答模型.

1　LCN 问答模型构建

构建医疗知识问答模型共分为 4 步：第一步，利用某三甲医院提供的 15 000 份电子住院记录，基于 Glove 算法训练医学词向量. 第二步，通过医疗网站获取大量医学关系三元组，构建医学知识图谱，并提取知识图谱中的关系名词，结合医学词向量构建关系向量. 第三步，利用医疗网站获取大量医学问题，构建“问题—关系”数据库，基于此数据库，利用本文提出的 LCN 模型训练医疗知识问答模型. 第四步，根据已构建的问答模型，向问答模型输入新的医学问题，返回模型中与新问题特征相似度最高的答案，完成医疗问答. 整体流程如图 1.

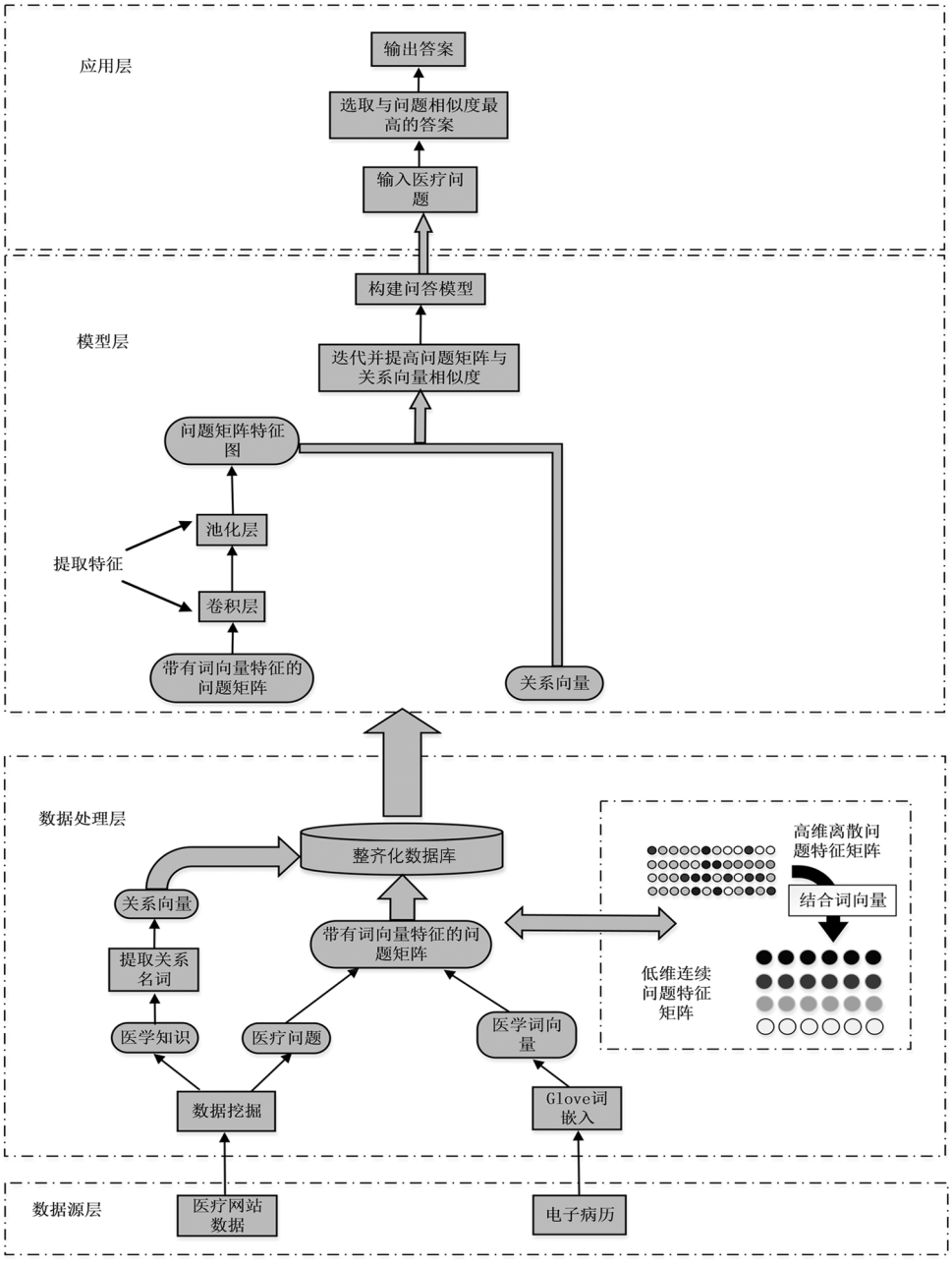

图 1　LCN 问答模型图

1.1 Glove词向量训练

我们采用某三甲医院提供的15 000份电子住院记录作为词向量训练的数据，本文所用的电子病历由多个科室的病患住院记录组成，病例的科室分布如图2.

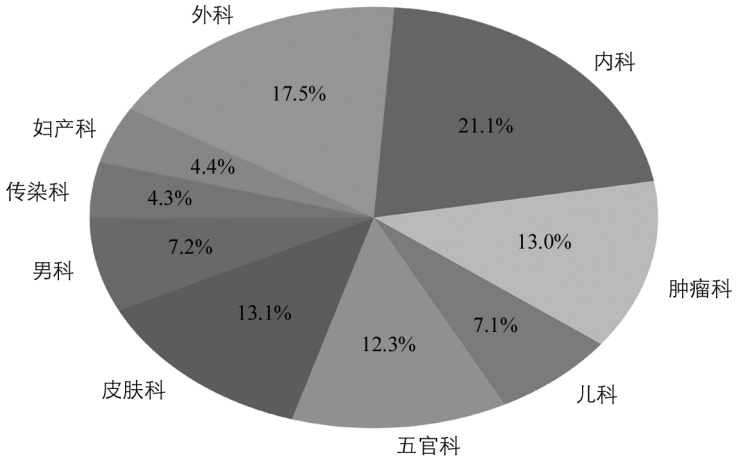

图2 科室分布

图2显示，外科和内科病人所占比重较大，因为现多数常见疾病如“感冒”“咳嗽”被划分到内科之中，“颈椎病”等被划分到外科之中，所以医学词向量词库中多数为外科和内科疾病的有关词语. 肿瘤科、五官科、皮肤科数量仅次于内、外科，妇产科、传染科和儿科中的疾病种类比较少，所占病例比重也较少. 本文选取了外科中的骨伤科作为样式，如图3.

姓名：XXX 性别：男 年龄：50 科室：骨伤科 病区：二病区

床号：22 住院号：123456 民族：汉族 婚姻：已婚 职业：白领

住址：甘肃省 出生地：甘肃省 发病节气：

入院时间：2013年6月9日 记录时间：2014年7月5日

主诉：近期摔倒，下肢、胸背部疼痛.

现病史：患者自述8 h前在家摔倒，当即感觉胸背部疼痛伴活动受限，并伴有双下肢疼痛、麻木，患者在家属的搀扶下回床休息，胸背部疼痛仍不缓解，轻微活动可使胸背部疼痛加重，患者及家属为求进一步诊治遂来我院就诊.

既往史：无既往史

个人史：生于本地，否认长期外地居住史，否认疫区居留史，否认特殊化学品及放射性接触史，否认吸烟，否认饮酒.

婚育史：适龄结婚，配偶及子女均体健.

图3 电子病历样式

电子住院记录涵盖了病人的性别、年龄、所在地等特征信息，病人主诉是病人自述自身的症状或者体征. 现病史为病患发病的最初症状，即从病患发病至本次就诊之间的身体情况，图3中患者的现病史表明近期在家中摔倒，并感觉到下肢疼痛、麻木，并伴随着胸背疼痛. 既往史为患者过去曾经所患的疾病以及身体状况. 个人史和婚育史为患者的个人家庭以及生活习惯. 病人主诉以及病史对医生初步诊断有着辅助作用. 病历中主诉、现病史包含了大量该领域的专属医学名词，本文通过利用jieba分词工具对病历文本进行分词，并进一步使用Glove预训练模型训练词向量.

采用Glove[15]模型训练医学词向量，定义 $\boldsymbol{X}$ 为共现词频矩阵，其中元素 $X_{i,j}$ 为词 j 出现在词 i 环境的次数，“环境”为词 i 周围不超过10个词的范围.

我们定义 $X_i=\sum_j X_{i,j}$，X_i 为所有词出现在词 i 环境的次数，求得词 k 出现在词 i 环境的条件概率 $P_{i,k}$，

$$P_{i,k}=P(k \mid i)=\frac{X_{i,k}}{X_i} \tag{1}$$

给定矢量 $\boldsymbol{w}_i$，$\boldsymbol{w}_j$，$\boldsymbol{w}_k$ 可计算损失函数 J，

$$J=\sum_{i,j,k}^{N}\left(\frac{P_{i,k}}{P_{j,k}}-g(\boldsymbol{w}_i,\boldsymbol{w}_j,\boldsymbol{w}_k)\right)^2 \tag{2}$$

由于复杂度过高，利用 $\boldsymbol{w}_i$，$\boldsymbol{w}_j$ 和 $\boldsymbol{w}_k$ 的矢量特性进行相减和内积，得到

$$\frac{P_{i,k}}{P_{j,k}}=\exp((\boldsymbol{w}_i-\boldsymbol{w}_j)^T\boldsymbol{w}_k) \tag{3}$$

并进行指数运算

$$\frac{P_{i,k}}{P_{j,k}}=\frac{\exp(\boldsymbol{w}_i^T\boldsymbol{w}_k)}{\exp(\boldsymbol{w}_j^T\boldsymbol{w}_k)} \tag{4}$$

将公式 1 带入

$$\log(X_{i,j})-\log(X_i)=\boldsymbol{w}_i^T\boldsymbol{w}_j \tag{5}$$

将 X_i 拆分为两个标量 b_i，b_j

$$\log(X_{i,j})=\boldsymbol{w}_i^T\boldsymbol{w}_j+b_i+b_j \tag{6}$$

得到最终损失函数 J

$$J=\sum_{i,j}^{N}(\boldsymbol{w}_i^T\boldsymbol{w}_j+b_i+b_j-\log(X_{i,j}))^2 \tag{7}$$

将调试所得的词向量 $\boldsymbol{w}_i$ 集合构建医学词向量库 T，且 $\boldsymbol{w}_i\in T$.

1.2　医疗知识图谱构建

本文利用爬虫手段，对医学网站中的疾病实体和实体间关系进行抽取. 设知识图谱表示为 G，图中节点与关系定义为

$$\{G=\langle M,B\rangle\} \tag{8}$$

M 表示节点集合，B 表示边的集合. 每 2 个节点和 1 个关系组成 1 个三元组 z，定义为

$$\{z=(l_h,l_r,l_t)\mid l_h,l_t\in M,l_r\in B\} \tag{9}$$

定义 l_h 为头实体，l_t 为尾实体，l_r 为实体关系，从语法角度解释为

$$\text{三元组}=(\text{主语},\text{谓语},\text{宾语}) \tag{10}$$

以(病毒性感冒，并发疾病，鼻炎)为例，设“病毒性感冒”为头实体 l_h，“鼻炎”为尾实体 l_t，“并发疾病”为实体间关系 l_r，知识图谱样式如图 4.

图 4 的左半部分，病毒性感冒的用药为川贝枇杷糖浆和小柴胡颗粒，该病的检查方法是血常规，传染方式为呼吸道传染，症状为鼻塞和流鼻涕. 还可看出，病毒性感冒和鼻炎为并发疾病，图的右半部分显示了鼻炎的用药、传染方式等属性. 通过遍历词向量库 T，获取图谱中的关系词向量 f_{an}，

$$f_{an}=T_{l_r},\ T_{l_r}\in T \tag{11}$$

1.3　构建 LCN 问答模型

针对 CNN 无法提取不同分词下的语义特征问题，本文采用 LCN 模型，LCN 模型引入了词格，词格包含了一个单词针对于不同分词的所有可能的上下文，如图 5.

LCN 有效地避免了专业领域分词不准确带来的弊端. 由于每个词格包含了一个词的不同上下文，所以会产生多个特征向量，采用合并法处理多个特征向量，如图 6.

针对于任意一个晶格 $B=\langle V,E\rangle$，V 是问题中所有可能出现的字词的集合，E 是相邻两个字词组成的边的集合，设问题中字 w 处的晶格核尺寸为 n，通过卷积的操作，该晶格输出的特征向量可以表示为

$$\begin{aligned}&\beta_1,\beta_2,\cdots,\beta_n=\text{softmax}(\boldsymbol{W}_c(\boldsymbol{v}_{w_1}:\cdots:\boldsymbol{v}_{w_n})+\boldsymbol{b}_c^T)\\&\mid\forall i,\omega_i\in(\omega_i,\omega_{i+1})\in E,\omega_{\left[\frac{n+1}{2}\right]}=\omega\end{aligned} \tag{12}$$

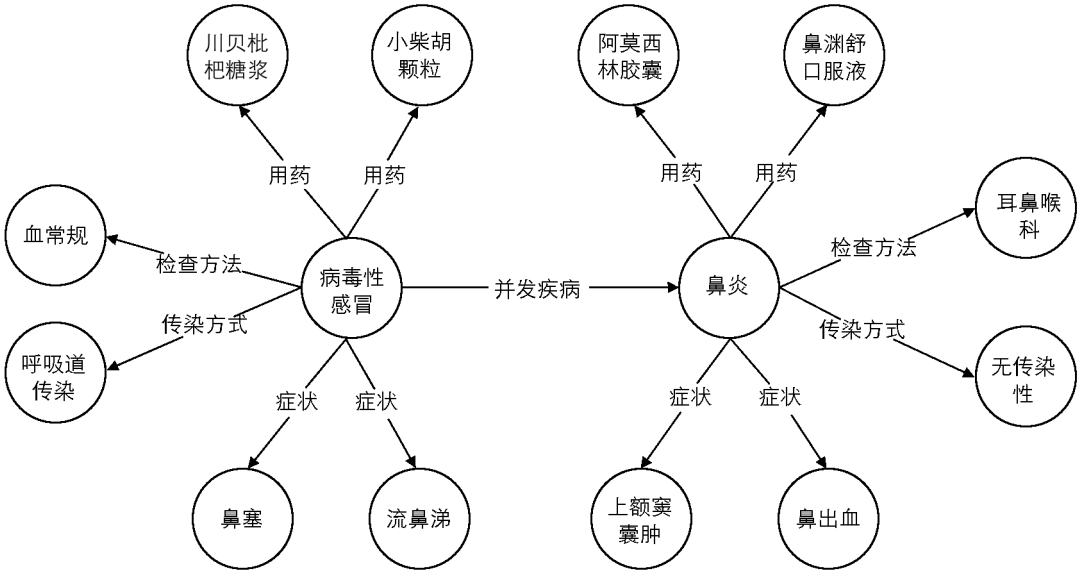

图4 知识图谱式样

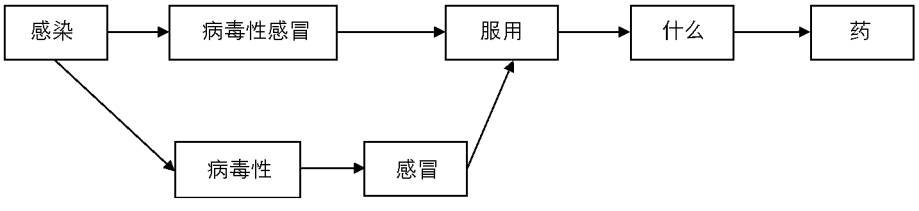

图5 分词样例

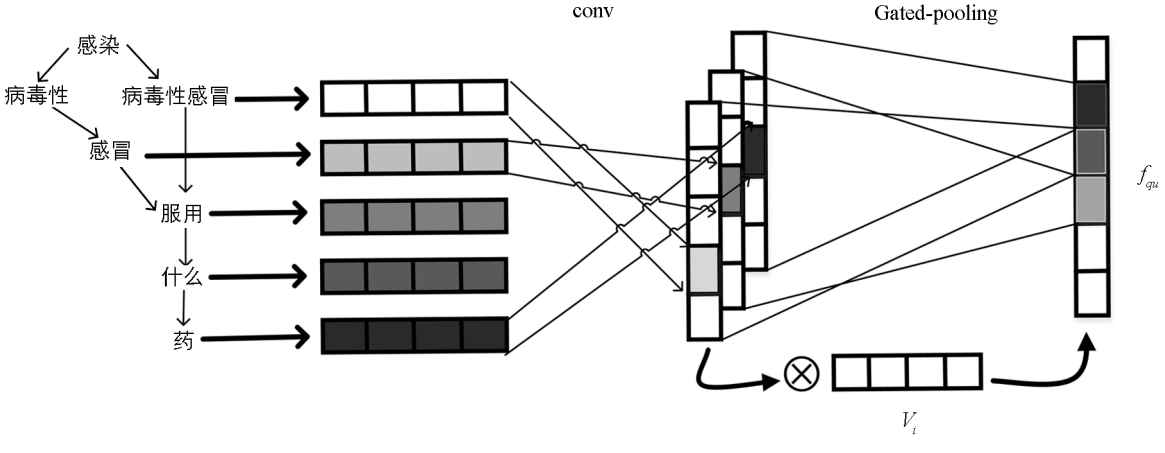

图6 LCN模型图

其中softmax是激活函数，$\boldsymbol{v}_{w_i}$ 是与该层中单词 $\boldsymbol{w}_i$ 相对应的输入向量，同时 $v_{w_i} \in T$，$(\boldsymbol{v}_{w_1}:\cdots:\boldsymbol{v}_{w_n})$表示合并不同词向量，$\boldsymbol{W}_c$ 为卷积核参数矩阵，$\boldsymbol{b}_c$ 为偏置向量. 对于多个特征向量，通过池化求得表示向量

$$f_{qu}=\text{gated-pooling}\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_n\}=\sum_{i=1}^{n}\beta_i\times t_i \tag{13}$$

其中，t_i 为固定参数，β_i 为池化操作的门控权重，n 为词格所承载上下文范围，通过调节 β_i 和 n 可控制表示向量，f_{qu} 为所求得的问题表示特征向量. 将 f_{qu} 带入到残差链接神经网络计算评分 s，

$$s=\text{sigmoid}(\boldsymbol{W}_2\text{ReLu}(\boldsymbol{W}_1(f_{qu}\cdot f_{an})+\boldsymbol{b}_1^T)+\boldsymbol{b}_2^T) \tag{14}$$

$$L = -\sum_{i=1}^{N}[y_i \log(s_i) + (1-y_i)\log(1-s_i)] \tag{15}$$

其中 y_i 为第 i 个词所对应的标签，L 为交叉熵损失.

综上所述，本文提出了基于 LCN 的问答模型，如图 7.

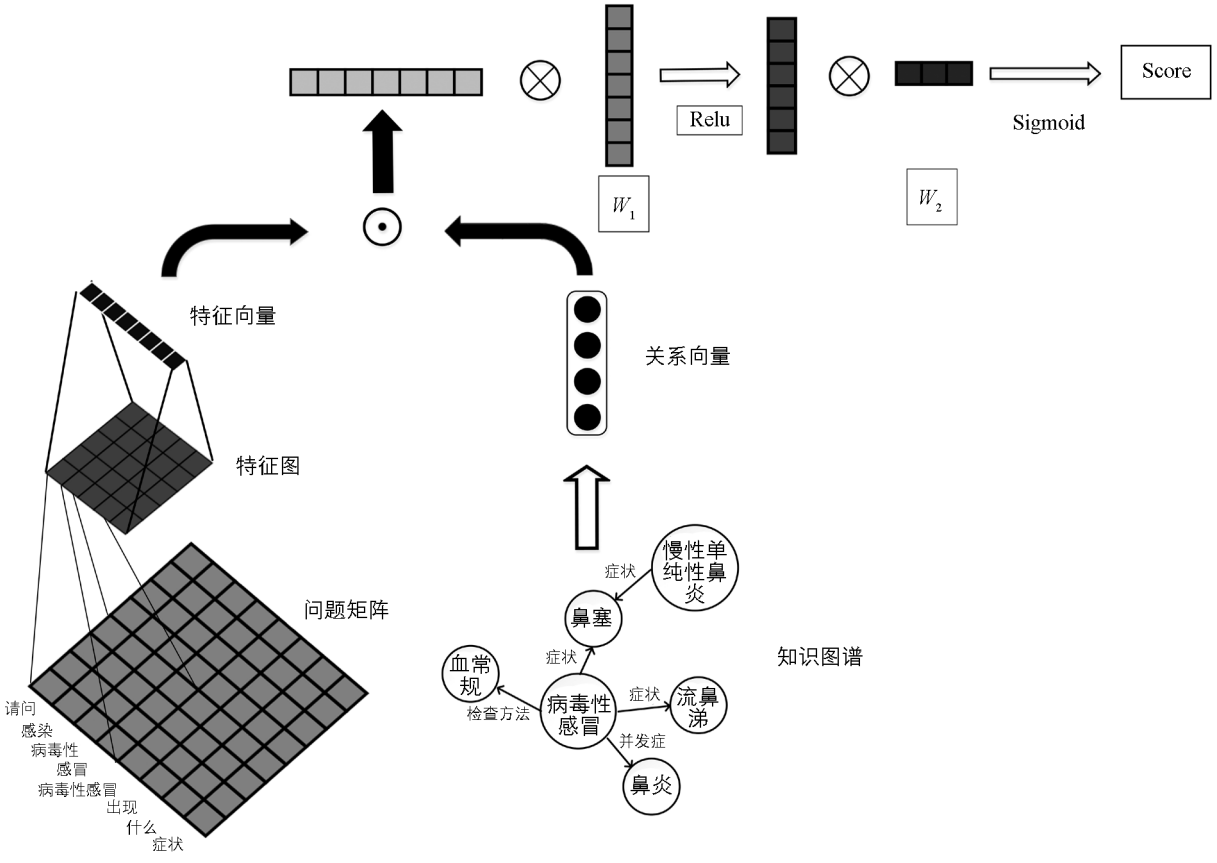

图 7　医疗问答模型总图

2　实验与结果分析

2.1　知识问答数据集构建

本文利用爬虫手段，抽取寻医问药网、39 健康网等中国医疗网站中的结构化数据与医学问题，网站如图 8.

图 8 上半部分中，“鼻旁窦支气管综合征”为疾病主体，“并发疾病——支气管扩张”“传染方式——无传染性”是每个疾病都拥有的相关属性和其对应的属性值，利用数据爬虫的方法，抽取网站中的疾病主体及其相关属性和属性值. 利用 1.2 节中的知识图谱构建方法，构建“(主语，谓语，宾语)”样式的知识图谱，例如“(鼻旁窦支气管综合征，并发疾病，支气管扩张)”.

图 8 下半部分中，通过爬虫法爬取医疗问题，例如“鼻旁窦支气管综合征有哪些并发症”，由人工制定规则的方法，将问题与正确关系的三元组一一对应，并组合成我们的数据集，如表 1.

表 1　模型训练数据样式

项　目	数　据　样　式
问题	鼻旁窦支气管综合征有哪些并发症
正确关系三元组	鼻旁窦支气管综合征，并发症，支气管扩张
候选关系三元组	
	1. 鼻旁窦支气管综合征，发病人群，无特定人群
	2. 鼻旁窦支气管综合征，症状表现，呼吸困难
	3. 鼻旁窦支气管综合征，治愈率，90%

图8 医疗网站

我们利用机器学习中较为传统的二八定律法则[16]划分数据集，将训练集和测试集按照4∶1的比率划分，如表2，第2列和第3列为不同数据部分所对应的问题和关系的数量.

表2 ‘问题—关系’数据统计

项 目	训练集/个	测试集/个
问题对	24 000	6 000
正确关系三元组	24 000	6 000
候选关系三元组	132 000	36 000

训练集中有24 000个问题和132 000个三元组，测试集有6 000个问题和36 000个三元组，每个问题仅包含1个正确三元组，若干个候选三元组，其中候选三元组包括1个正确答案，其余为错误答案，用于负采样. 本文共爬取有3 000种疾病，5 000种症状，每个疾病或症状占有大约10～15个三元组.

2.2 评价指标

对于KGQA，每个问题仅1个正确关系三元组，因此仅使用准确率(ACC)和MRR计算，

$$ACC=\frac{TP+TN}{TP+TN+FP+FN} \tag{16}$$

TP,TN,FP,FN分别表示正类预测为正类数、负类预测为负类数、负类预测为正类数、正类预测为负类数.

MRR是国际通用的搜索算法的评价机制，

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\mathrm{rank}_i} \tag{17}$$

其中$|Q|$为搜索总个数，rank_i 是第 i 个搜索结果，即把标准答案在被评价系统给出结果中的排序取倒数作为它的准确度，再对所有的问题取平均.

2.3　词向量构建

本文使用 Glove 算法对医学词向量进行训练，设定 Glove 模型的中心词环境大小为 10，词向量长度为 50，训练过程经过 100 词的调整和微调，最终获得训练后的医学词向量. 本文所构建的词向量库，包括 15 000 个词向量，5 000 个字向量.

2.4　模型比对

LCN 模型中，定义训练集批处理量 train_batch_size 为 64，测试集批处理量 test_batch_size 为 32，全连接层采用 dropout 为 0.3，激活函数选用 Relu，learning_rate 为 0.8，衰减因子为 0.9，优化器选用 Adam.

2.4.1　分词工具实验对比

LCN 模型主要解决了因分词导致的模型提取特征模糊问题，所以为了验证模型的有效性，首先与不同分词工具做对比实验. 本文分别使用 CNN 结合 HanLP，jieba，FudanNlp 和 CTB(斯坦福)分词器与 LCN 做对比实验，如表 3.

表 3　分词工具实验对比

模型	准确率	*MRR*
CNN＋jieba	0.85	0.90
CNN＋HanLP	0.83	0.89
CNN＋FudanNLP	0.86	0.88
CNN＋PKU	0.84	0.87
CNN＋char	0.87	0.89
LCN	0.89	0.91

利用 CNN 结合不同分类器可知，更适用于中文的 jieba 分词器明显好于其他分词器，但是，由于医学分词的专业性，限制了传统中文分词器的优势，所以字符级别的 CNN＋char 的效果好于 jieba 分词器. 尽管 CNN＋char 整体性能很好，但是由于医学专业领域的分词限制，无法提取到医疗文本的主要特征. LCN 的多粒度抓取可以使医学文本中的字信息和词汇信息互补，使得在分词极不准确的情况下仍然保持优势.

2.4.2　深度学习模型对比

LCN 以 CNN 为基础，以字和词混合多粒度医疗文本为输入，提取文本特征. 所以本文选择了 LSTM[17]，Bi-GRU[18]，AMPCNN[19] 3 种深度学习模型进行比较，分别计算模型的准确率(Accuracy)和 *MRR*，如表 4.

表 4　模型准确率结果

模型	准确率	*MRR*
LSTM	0.79	0.83
Bi-GRU	0.81	0.80
AMPCNN	0.85	0.87
LCN	0.89	0.91

AMPCNN 包含了由字符级别和单词级别的 CNN 构成，字符级别用来抓取字符串特征并计算字符串的相似性，单词级别利用最大缓冲带(AMP)检测语句中的谓词含义. 但是 AMPCNN 并不能细节地抓取上下文信息，而本文使用的 LCN 可以抓取谓词上下文的所有可能性，包括每个词的字符与单词状态，

所以准确率较高. LSTM 提取医疗问句中的语序信息，通过分析问句时间序列特征计算与答案之间的相似度，但是模型中的错误信息导致结果准确率较低. Bi-GRU 性能高于传统的 LSTM，因为 Bi-GRU 中的门控单元不但可传递节点之间的隐藏状态，而且可重置节点中的错误信息，可有效解决训练中反向传播的梯度消失问题，但上面两种模型无法解决关系序列在不同分词情况下的特征提取，本文所采用的 LCN 不但将关系名称分解为单词，关系上下文内容也被分解为字与词的混合模式，同时利用关系级别和单词级别的序列分别提取本地信息和全局信息，对于问题的特征抽取更加全面，使得特征图内容更加丰富，准确率更高.

2.5 词向量对比

在基于 LCN 的条件下，为了验证 Glove 的有效性，分别利用 Cbow[20]，Skip-gram[21]，Bert[22] 和 Glove 编译的医学词向量，比较预测的准确性. 通过利用不同的词向量所得结果如图 9.

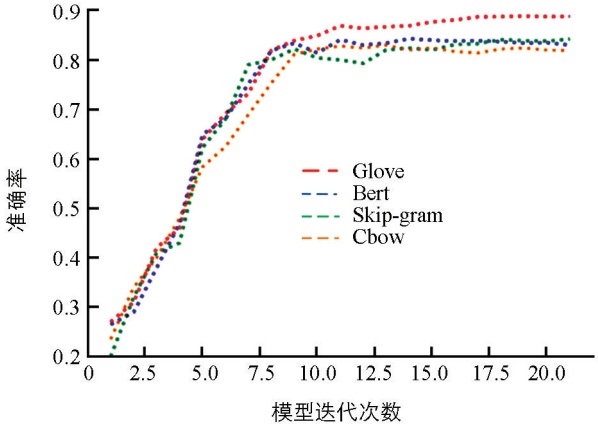

图 9 词向量模型对比

图 9 中 Skip-gram 和 Cbow 属于早期提出的词向量模型，Skip-gram 要预测每个词作为中心词时周围词的情况，模型更加精细，所以准确率高于 Cbow. 然而前两者只考虑词的局部信息，忽略了中心词与预测范围外的其他词，Bert 利用 Transformer 进行编码，充分考虑了上下文特征情况下预测其他词，表现更佳. 本文采用的 Glove 词向量统计了词的共现率，获得的医疗词特征信息更加充分，效果优于其它算法.

从图 9 中还可看出，本文的 Glove+LCN 模型在模型迭代次数为 15 时准确率逐渐趋于饱和，后维持在 89%左右.

2.6 问答模型应用

利用本文的医疗知识问答模型，关于医疗问答的简单应用如表 5.

表 5 问答模型应用

问 题	答 案
鼻旁窦支气管综合征有哪些并发症	鼻旁窦支气管综合征有着支气管扩张、咯血等综合症
鼻旁窦支气管综合征能治吗	鼻旁窦支气管综合征治愈的概率为 90%
患有鼻旁窦支气管综合征出现什么症状	鼻旁窦支气管综合征会出现喘息、呼吸困难等症状
鼻旁窦支气管综合征需要做什么检查	鼻旁窦支气管综合征需要做胸部透视

3 总 结

本文通过爬取医疗网站信息构建医学知识图谱，并进一步构建了医学知识问答库，利用 Glove 模型将电子病历训练为医学词向量，将医学词向量作为输入端，利用 LCN 模型提取问题的特征，并且计算问题特征和答案之间的相似度，进而训练模型完成医疗问答，不仅省去了传统问答模型人工定义规则这一过程，而且在实验中通过与其它问答模型对比，效果优于其它深度学习模型. 但是 LCN 还存在以下不足：

1) LCN 的核心是利用 CNN 中的卷积核提取问题特征，但是忽略了问题的时序特征.

2) 模型对于“问题—关系”的 1 对 1 模式准确率较高，但是对于一个问题多个答案的 1 对 n 模式，其问答表现不佳.

后续将 RNN 入到模型中，提升对问题时序性特征的关注，并且针对 1 对 n 的问答情况更新模型.

参考文献：

[1] BEN ABACHA A, DEMNER-FUSHMAN D. A Question-entailment Approach to Question Answering [J]. BMC Bioinformatics, 2019, 20(1): 511-520.

[2] ZHANG S, ZHANG X, WANG H, et al. Multi-Scale Attentive Interaction Networks for Chinese Medical Question Answer Selection [J]. IEEE Access, 2018(6): 74061-74071.

[3] DENG W, GUO P P, YANG J D. Medical Entity Extraction and Knowledge Graph Construction [C] //2019 16th International Computer Conference on Wavelet Active Media Technology and Information Processing, IEEE, 2019: 41-44.

[4] LI X, LIU H Y, ZHAO X, et al. Automatic Approach for Constructing a Knowledge Graph of Knee Osteoarthritis in Chinese [J]. Health Information Science and Systems, 2020, 8(1): 1-8.

[5] CHAI X Q. Diagnosis Method of Thyroid Disease Combining Knowledge Graph and Deep Learning [J]. IEEE Access, 2020(8): 149787-149795.

[6] YUAN J B, JIN Z W, GUO H, et al. Constructing Biomedical Domain-specific Knowledge Graph with Minimum Supervision [J]. Knowledge and Information Systems, 2020, 62(1): 317-336.

[7] LIU H I, NI C C, HSU C H, et al. Attention Based R&CNN Medical Question Answering System in Chinese [C] // 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), IEEE, 2020: 341-345.

[8] NGUYEN V, KARIMI S, XING Z C. ANU-CSIRO at MEDIQA 2019: Question Answering Using Deep Contextual Knowledge [C] //Proceedings of the 18th BioNLP Workshop and Shared Task, Florence, Italy. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 478-487.

[9] ZHANG S, ZHANG X, WANG H, et al. Chinese Medical Question Answer Matching Using End-to-End Character-Level Multi-Scale CNNs [J]. Applied Sciences, 2017, 7(8): 767-775.

[10] ZOU Y, HE Y, LIU Y. Research and Implementation of Intelligent Question Answering System Based on Knowledge Graph of Traditional Chinese Medicine [C] //2020 39th Chinese Control Conference (CCC), IEEE, 2020: 4266-4272.

[11] ZHU W, NI Y, XIE G T, et al. The Dr-KGQA System for Automatically Answering Medication Related Questions in Chinese [C] //2019 IEEE International Conference on Healthcare Informatics (ICHI), IEEE, 2019: 1-6.

[12] SADID A H, ZHAO S Y, VIVEK D, et al. Clinical Question Answering using Key-Value Memory Networks and Knowledge Graph [C] //TREC, 2016.

[13] ZHANG Y Y, QIAN S S, FANG Q, et al. Multi-modal Knowledge-aware Hierarchical Attention Network for Explainable Medical Question Answering [C] //Proceedings of the 27th ACM International Conference on Multimedia, Nice France. New York, NY, USA: ACM, 2019: 1089-1097.

[14] LAI Y X, FENG Y S, YU X H, et al. Lattice CNNs for Matching Based Chinese Question Answering [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33: 6634-6641.

[15] YUSUF N, YUNUS M A M, WAHID N, et al. Enhancing Query Expansion Method Using Word Embedding [C] // 2019 IEEE 9th International Conference on System Engineering and Technology (ICSET), IEEE, 2019: 232-235.

[16] CRAFT R C, LEAKE C. The Pareto Principle in Organizational Decision Making [J]. Management Decision, 2002, 40(8): 729-733.

[17] EL-GANAINY N O, BALASINGHAM I, HALVORSEN P S, et al. On the Performance of Hierarchical Temporal Memory Predictions of Medical Streams in Real Time [C] //2019 13th International Symposium on Medical Information and Communication Technology (ISMICT), IEEE, 2019: 1-6.

[18] WANG D Y, SU J L, YU H B. Feature Extraction and Analysis of Natural Language Processing for Deep Learning English Language [J]. IEEE Access, 2020(8): 46335-46345.

[19] YIN W P, YU M, XIANG B, et al. Simple Question Answering by Attentive Convolutional Neural Network [EB/OL].

[2020-08-26]. https://arxiv.org/abs/1606.03391.

[20] WANG X, DU Y T, LI X L, et al. Embedded Representation of Relation Words with Visual Supervision [C] //2019 Third IEEE International Conference on Robotic Computing (IRC), IEEE, 2019: 409-412.

[21] IHM S Y, LEE J H, PARK Y H. Skip-Gram-KR: Korean Word Embedding for Semantic Clustering [J]. IEEE Access, 2019(7): 39948-39961.

[22] WANG Z, HUANG Z, GAO J. Chinese Text Classification Method Based on BERT Word Embedding [C] //Proceedings of the 2020 5th International Conference on Mathematics and Artificial Intelligence, 2020: 66-71.

An LCN-Based Medical Knowledge Base Question Answering Model

MA Man-fu[1], LIU Yuan-zhe[1], LI Yong[1], WANG Xia[2], JIA Hai[2], SHI Yan-bin[3], ZHANG Xiao-kang[4]

1. *College of Computer Science and Engineering, Northwest Normal University, Lanzhou 730070, China;*
2. *The People's Hospital of Gansu Province, Lanzhou 730000, China;*
3. *College of Pharmacy, Lanzhou University, Lanzhou 730000, China;*
4. *Lanzhou Qidu Data Technology CO., Ltd., Lanzhou 730070, China*

Abstract: Word segmentation in the Chinese medical field is difficult, resulting in inadequate extraction of medical problem features by the existing algorithms. This paper proposes a medical knowledge base question answering model based on LCN (Lattice Convolutional Neural Network) for the characteristics of Chinese word segmentation. First, 15 000 electronic medical records provided by a first-class hospital at Grade 3 are used to train medical word vectors with the Glove model. Then, a large number of medical nouns and their intra-relations are obtained through major medical websites to construct a medical knowledge map. The relationship words in the knowledge graph are extracted and, combined with the trained word vectors, the relationship vectors are obtained. Finally, the medical word vector is used as the model input and the LCN neural network is used to extract the medical problem features. The model is trained by calculating the similarity of the question features and relationship vectors. Experiments show that the accuracy rate of the LCN model is as high as 89.0%, which is an improvement of 2% compared with similar question answering models.

Key words: medical knowledge base question answering; Glove; LCN (Lattice Convolutional Neural Network); medical knowledge graph; electronic medical record

责任编辑 周仁惠