基于知识图谱的医疗问答系统研究

李　伟，王竣生，秦　鹏
（六盘水师范学院 计算机科学学院，贵州 六盘水 553000）

摘要：随着人工智能技术的不断发展，越来越多的自然语言处理技术应用到医疗行业。如何从海量医疗数据中提炼信息，并根据用户的问题给出针对性较强的回答，是推进医疗智能化的关键问题。文章研究利用 BiLstm+CRF 模型处理医疗领域问答相关数据，基于图数据库 Neo4j 构建一个医疗知识图谱，并在此基础上构建一个问答系统，实现医疗知识的自动问答服务。实验结果表明，该系统可以为用户提出的问题查找匹配准确答案并返回给用户。

关键词：医疗知识；知识图谱；问答系统
中图分类号：TP311　　**文献标识码：**A　　**文章编号：**2096-9759(2023)06-0107-03

Research on Medical Q&A System Based on Knowledge Graph

LI Wei, WANG Junsheng, Qin Peng
(School of Computer Science, Liupanshui Normal University, Liupanshui, Guizhou 553000, China)

Abstract: With the continuous development of artificial intelligence technology, more and more natural language processing techniques are applied to the medical industry. How to extract information from the vast amount of medical data and give more targeted answers based on the user's questions is a key issue in advancing medical intelligence. In this paper, we use the BiLstm+CRF model to process question and answer related data in the medical field, construct a medical knowledge graph based on the graph database Neo4j, and build a question and answer system based on it to realise an automatic question and answer service for medical knowledge. The experimental results show that the system can find the exact answer to the question asked by the user and return it to the user.

Keywords: medical knowledge; knowledge graph; question and answer system

1 引言

随着信息技术的快速发展，网络已经成为人们日常生活中不可缺少的一部分，而网络中的数据也变得越来越丰富，如何从网上数据中获取有用信息，并将其应用于用户问答需求，已成为一个亟待解决的问题。知识图谱作为一种新型的数据库技术，在大数据分析、人工智能等领域有广阔的应用前景，而知识图谱与问答系统相结合，可为用户提供高质量、个性化和准确的问答服务。

知识图谱是由 Google 公司在 2012 年发布的[1]，主要应用在 Google 搜索引擎中，其最初的目的是借助知识图谱来优化 Google 的搜索结果，为用户提供含有完整的搜索结果，方便用户能够尽快获取到所需的信息，为用户提供更好的搜索体验。Siddhant Garg[2]等人提出了一种 TANDA 的有效技术，主要用于自然语言任务的预训练变换模型的精调。Shen[3] 等人提出每个医学答案对应一个有效问题的分布，着重研究了如何整合结构化知识和非结构化知识来生成上述医疗问答对。崔洁[4]等人提出采用自顶向下的方式构建乳腺肿瘤知识图谱，为乳腺肿瘤疾病知识学习与推理奠定了数据基础。

本文研究基于知识图谱，设计一个医疗知识问答系统。首先在网上抓取医疗领域问答相关数据，然后基于图数据库 neo4j 构建一个医疗知识图谱，并在此基础上构建一个问答系统，以来实现医疗知识的自动问答服务。

2 相关技术

2.1 医疗知识图谱构建

知识图谱的建设[5]，一般包括数据获取、数据处理和数据存储等过程。医疗知识图谱要先定义医疗领域相关的实体、

收稿日期：2023-03-10
基金项目：六盘水师范学院校级基金项目（LPSSYZDZK202204）。
作者简介：李伟（1985-），山东聊城人，硕士，副教授，研究方向为数据挖掘、大数据技术。

工作效率和质量，提升用户感知的目的。按稽核人员每天人工处理 230 单计算，使用工具前每单耗时 10 分钟，通过投诉稽核辅助分析工具处理，投诉稽核自动化处理工单总计耗时约 20 分钟，效率提升：230*10/20=115 倍以上，且通过对比分析，准确率达到了 100%。

6 结语

本工具基于 MySQL 数据库技术架构进行开发，操作人员只需要在电脑安装 MySQL 数据库，并在 workbench 内连接 MySQL 服务器创建需求表单导入基础数据进行运算便可得到所需结果。本工具具有较强的可移植性，关联维度的数据可从各省的网络优化管理平台大数据获取，适用于其他省开展应用部署，各省可在符合《中国移动网络数据安全管理办法》的基础下，利用集中投诉分析管理系统中的各种大数据信息，开展投诉工单分析和集中呈现关联等大数据分析应用，可参考本成果投诉稽核辅助分析工具，通过开发相关算法、功能、工具，实现快速识别、智能分析相关问题，提升问题定位准确性及分析效率。

参考文献：

[1] 张伟丽, 江春华, 魏劲超. MySQL 复制技术的研究及应用 [J].计算机科学.2012,(z3).168-170.
[2] 王锐. 搭建 MySQL 数据库主从库平台实现数据备份 [J].电脑编程技巧与维护.2011,(19).30-31.

实体关系和实体属性，具体如下：

（1）知识图谱实体类型：药品、食物、检查、科室、药品大类、疾病、症状。

（2）知识图谱实体关系类型：科室－科室关系、疾病－忌吃食物关系、疾病－宜吃食物关系、疾病－推荐吃食物关系、疾病－通用药品关系、疾病－热门药品关系、疾病－检查关系、厂商－药物关系、疾病症状关系、疾病并发关系、疾病与科室之间的关系。

（3）知识图谱疾病属性：疾病描述、预防措施、病因、患病概率、易患病人群、治疗方式、治疗周期、治疗概率。

构建医疗知识图谱首先使用网络爬虫采集网上医疗领域问答相关数据，包括疾病、症状、药物和饮食等内容。然后处理抓取的数据，采用基于规则的方法提取医疗实体、属性以及实体间的关系，然后将数据存储到 Neo4j 图数据库，医疗知识图谱的构建框架如图 1 所示：

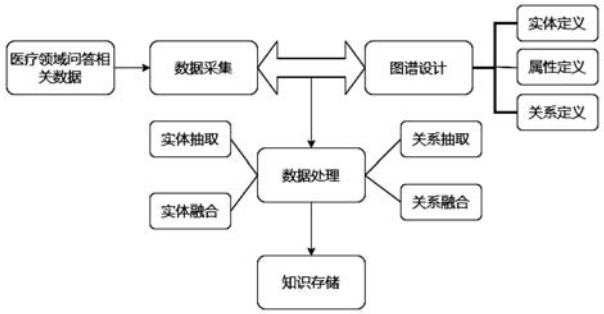

图 1 知识图谱的构建框架

2.2 BiLstm + CRF 模型

采集的医疗相关数据通常涉及广泛的参与人群，潜藏着丰富的医疗价值。然而，它们大多为非结构化数据。为了能够充分利用数据信息，需要抽取和挖掘数据中有用的医疗知识。系统使用已经标注的医学语料，搭建 BiLstm+Crf 进行训练，来提取医疗领域的各种实体，从而为提取实体关系构建知识图谱打好基础。BiLstm + CRF 模型架构图如图 2 所示：

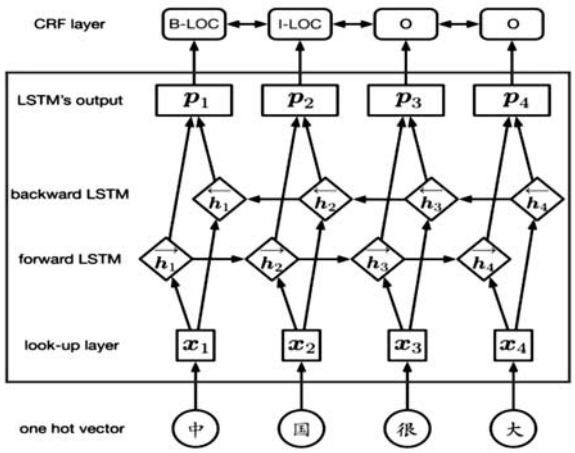

图 2 BiLstm+CRF 模型架构

该模型架构中主要包含两个部分：（1）BiLstm 层：图中红框的部分就是 BiLstm 层，该层的功能是将每个 token 的输入转换为融合了上下文的 token 输出；（2）CRF 层：CRF 层可以预测的标签添加一些约束，以来保证预测的标签是合法有效的。而这些约束可以通过 CRF 层，在训练数据训练过程中自动学习得到。

其中，BiLstm 层包含两个部分：（1）Embedding 层：Embedding 层的主要作用是将每个 token（字）嵌入成为固定长度的词向量，然后送入 BiLstm 层进行计算；（2）BiLstm 层：BiLstm 层作用是利用双向 LSTM 的机制，提取问句文本的上下文语义特征，为每个 token 输出一个状态向量。

其中，CRF 层主要目的是使用 crf 马尔科夫线性链来让模型学习到约束的规则。使用这些约束，会大大降低标签序列预测中非法序列出现的概率。而约束的实现则是使用 CRF 损失函数来实现的。如果标签一共有 tag_size 个，那么 BiLSTM 的输出维度就是 tag_size，表示的是每个词 wi 映射到 tag 的发射概率值(feats)，设 BiLSTM 的输出矩阵为 P，其中 $P_{i,j}$ 代表词 wi 映射到 tag j 的非归一化概率。

利用 softmax 函数，为正确的 tag 序列 y 定义一个概率值，其公式为：

$$p(y|X) = \frac{e^{S(X,y)}}{\sum_{\tilde{y}\in Y_X} e^{S(X,\tilde{y})}} \quad (1)$$

训练的目标就是最大化概率 p（y|X），最大化用对数似然（因为 p（y|X）中存在指数和除法，对数似然可以化简这些运算），对数似然形式如下：

$$log(p(y \mid X) = log\frac{e^{s(X,y)}}{\sum_{y' \in Y_X} e^{s(X,y')}} = S(X,y) - log(\sum_{y' \in Y_X} e^{s(X,y')}) \quad (2)$$

将损失函数 Loss 定义为-log（p(y|X)），就可以利用梯度下降法来进行网络的学习了。其公式为：

$$\text{Loss} = \log\left(\sum_{\tilde{y}\in Y_X} e^{S(X,\tilde{y})}\right) - S(X,y) \quad (3)$$

实现了 BiLstm 和 CRF 层后，还需要将二者组装成最终的模型。模型训练完成后，使用训练好的模型来实现命名实体识别预测。

2.3 知识存储

根据识别出的医疗实体，解析复杂属性，复杂属性如下：

（1）症状（symptom）提取“疾病—症状”关系

（2）并发症（acompany）提取“疾病—并发症”关系

（3）治疗科室（cure_department）提取“疾病—科室”关系

（4）通用药品（common_drug）提取“疾病—通用药品”关系

（5）推荐用药（recommand_drug）提取“疾病—热门药品”关系

（6）忌口（not_eat）提取“疾病--忌吃食物”、“疾病—宜吃食物”、“疾病—推荐食物”关系

（7）检查方法（check）提取“疾病—检查”关系

（8）药品详情（drug_detail）提取“药品—厂商”关系

通过判断疾病信息中是否存在相关的复杂属性，并从其中提取了实体和实体关系。在返回数据时，需要对实体进行去重，以避免在 Neo4j 中生成重复的实体节点。实现解析并存储所有类型节点的方法，对数据进行解析，并将返回的节点数据存入 Neo4j 数据库。然后调用函数得到存储实体和实体间关系的变量。

知识图谱中主要包含两类节点，一类为中心疾病节点，包含各种疾病属性；另一类为普通实体节点，即药品、食物等，不包含属性。然后，分别调用创建知识图谱中心疾病的节点和创建普通实体节点模块函数。

最后，实现存储关系的方法负责将一种类型的关系存储

到 Neo4j 数据库，参数如下：

（1）start_node：字符串，起始节点的类型标签

（2）end_node：字符串，终止节点的类型标签

（3）edges：列表，每个元素是一个元组，分别是起始节点 name 和终止节点 name

（4）rel_type：字符串 关系类型标签

（5）rel_name：字符串 关系的 name 属性(中文名称)

最后，使用 Cypher 语言直接执行 Neo4j 的语句，为每一对实体关系创建对应的边。实现解析数据并存储所有关系边的方法，得到存储实体和实体间关系的变量。

3 问答系统设计

问答系统主要对用户输入的问题进行分析，查询已构建的医疗知识图谱得到答案并返回给用户。问答系统处理流程如图 3 所示，系统首先对用户输入的问题进行实体提取和问题分类，将提取的结果转换为 cypher 语句，然后使用转换后的 cypher 语句从 Neo4j 图数据库中查询结果，并将结果生成答案返回给用户。

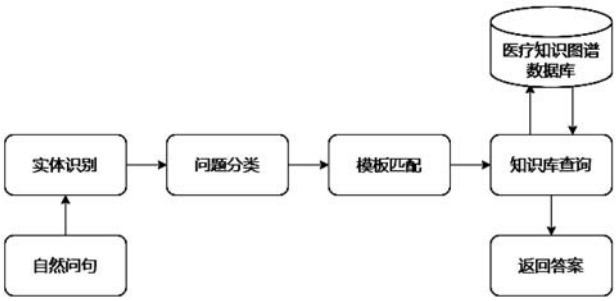

图 3 问答系统处理流程

对于用户输入的自然问句，通过 BiLstm+CRF 模型进行数据预处理和实体识别，然后通过分类的方法进行问句意图识别，在分类类别中进行规则模板匹配，这样可以大大提高问答系统的搜索效率。问题分类采用基于统计的分类方法中的 SVM 模型，通过计算问句中每个词的 TF-IDF 值，得出自然问句相应的特征向量，然后使用 SVM 分类器判断问句的所属类别。语料匹配需要计算问句实体和知识图谱中实体的相似度，本文选择使用余弦距离来计算向量的相似度，公式如下：

$$d = \frac{A * B}{|A| * |B|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} * \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (4)$$

用户提出的问题大多属于事实类型，因此可以采用基于规则模板匹配的方法来查询。系统根据用户问题类型和实体生成对应的 cypher 查询语句，来检索 Neo4j 图数据库，并将返回结果通过模板生成文本返回给用户。

4 实验分析

实验以网上医疗问答数据作为数据来源，来构建了一个医疗知识图谱。采用 Neo4j 图数据库存储医疗知识图谱，问答系统中采用了相似度计算和规则模板匹配的方式来完成。

为了验证所采用算法模型的有效性，实验采用三个评价指标，分别为准确率（P）、召回率（R）、F 值（F1），公式分别如下：

$$P = \frac{TP}{TP + FP} \quad (5)$$

$$R = \frac{TP}{TP + FN} \quad (6)$$

$$F1 = \frac{2 \times P \times R}{P + R} \quad (7)$$

其中，TP 表示正样本被正确识别的数量，FP 表示误报的负样本数量，FN 表示漏报的正样本数量。F1 值是准确率和召回率的合成指标，取值介于 0 到 1 之间，F1 值越大，代表模型的综合性能越好。

实验分别采用了基于规则的模板匹配和基于相似度匹配算法，来查找用户问题答案，两种方式的实验结果如图 4 所示。通过对实验结果进行分析和评估，相对于基于相似度匹配，基于规则的模板匹配在查找结果答案的准确率和召回率都较低。因此，系统可以采用结合两种方式进行问答匹配，优先考虑系统的搜索效率，先使用基于规则模板匹配进行查找，当匹配都失败后，再采用基于相似度计算的方式进行查找。

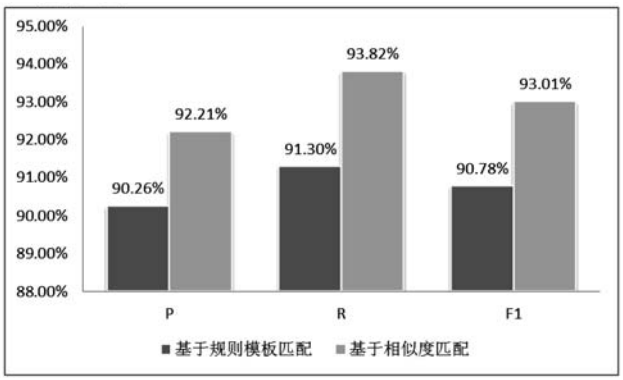

图 4 两种方式的实验结果

5 结语

本文首先介绍了基于知识图谱的问答系统研究现状，然后提出使用 BiLstm+ CRF 模型进行实体识别来构建知识图谱，使用 Neo4j 图数据库存储医疗知识图谱，接着设计了医疗问答系统的处理流程，最后进行了实验分析，得出系统采用基于规则的模板匹配和基于相似度匹配算法相结合的方式进行问答匹配。目前基于知识图谱的智能问答系统还处在发展阶段，很多研究技术还不是很成熟，因此还有很多地方需要深入研究和完善。下一步的研究准备增加语料，扩展医疗知识图谱，以来提高问题回答的准确度。

参考文献：

[1] Fensel D, Simsek U, Angele K, et al. Introduction: what is a knowledge graph[M]//Knowledge Graphs. Springer,Cham, 2020:1-10.

[2] Siddhant Garg, et al. TANDA: Transfer and Adapt Pre-Trained Transformer Models for Answer Sentence Selection[C]. AAAI 2020, Nov 2019.

[3] Sheng Shen, et al.On the Generation of Medical Question-Answer Pairs[C].AAAI 2020, Nov 2019.

[4] 崔洁, 陈德华, 乐嘉锦. 基于 EMR 的乳腺肿瘤知识图谱构建研究[J].计算机应用与软件, 2017, 34(12):122-126.

[5] 李涛,王次臣,李华康.知识图谱的发展与构建[J].南京理工大学学报:自然科学版, 2017, 41(1):22-34.