计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(12):55−62 [doi: 10.15888/j.cnki.csa.008229]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于医疗知识图谱的交互式智能导诊系统①

全　威[1,2], 马志柔[1], 刘　杰[1], 叶　丹[1], 钟　华[1]

[1](中国科学院 软件研究所 软件工程技术研究开发中心, 北京 100190)
[2](中国科学院大学, 北京 100049)
通讯作者: 马志柔, E-mail: mazhirou@otcaix.iscas.ac.cn

摘　要: 针对在线问诊中患者主诉医疗信息表述多样化, 医疗知识利用不足的问题, 本文设计实现了基于医疗知识图谱的交互式智能导诊系统. 该系统引入医疗知识图谱提供导诊知识, 通过实体识别和实体链接技术规范化主诉文本中的医疗表述, 利用医疗实体生成知识图谱子图并获取子图语义信息, 融合子图和患者主诉的语义信息得到科室置信度. 当推荐科室置信度低时, 通过多轮交互问询的方式补充患者症状信息, 最终给出推荐科室. 该系统能够为建立快速精准智能医疗体系提供技术支持, 有效提升导诊效率, 缓解医疗资源紧张.

关键词: 知识图谱; 医疗导诊; 实体识别; 实体链接; 多轮交互

引用格式: 全威,马志柔,刘杰,叶丹,钟华.基于医疗知识图谱的交互式智能导诊系统.计算机系统应用,2021,30(12):55–62. http://www.c-s-a.org.cn/1003-3254/8229.html

Interactive Intelligent Diagnosis Guidance System Based on Medical Knowledge Graph

QUAN Wei[1,2], MA Zhi-Rou[1], LIU Jie[1], YE Dan[1], ZHONG Hua[1]

[1](Technology Center of Software Engineering, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: Diversified expression of medical information and inefficient utilization of medical knowledge are encountered in online consultations. This study designs and implements an interactive intelligent diagnosis guidance system based on medical knowledge graphs. The system introduces medical knowledge graphs to provide medical knowledge and relies on entity recognition and entity linking technology to standardize the medical expression in the main condition description text. Moreover, it uses the medical entity to generate knowledge subgraphs and obtain their semantic information and merges the semantic information of the subgraphs and patient condition description to obtain department confidence. When the confidence of the recommended department is low, multi-round interactive inquiry can supplement the patient's symptom information, and finally, the recommended department is determined. The system can provide technical support for building a fast and accurate intelligent medical system to improve the efficiency of diagnosis and alleviate the shortage of medical resources.

Key words: knowledge graph; diagnosis guidance; entity recognition; entity linking; multi-rounds interaction

随着医疗信息技术的发展, 传统医疗健康模式正在向"互联网+医疗健康"转型, 网络预约、远程治疗、在线问诊等医疗服务也逐渐普及. 在线问诊系统依据患者主诉来分配科室, 是医生与患者线上沟通咨询的重要媒介. 大多数患者对自身病情缺乏全面了解, 医学专业知识存在不足, 患者的主诉文本口语化、通俗化、多样化, 识别其中的医疗信息以及精准的科室推荐算法是导诊系统的核心部分. 现有导诊系统通过规则推

① 基金项目: 国家重点研发计划 (2017YFB1002303)
Foundation item: National Key Research and Development Program of China (2017YFB1002303)
收稿时间: 2021-02-25; 修改时间: 2021-03-19; 采用时间: 2021-04-16

© 中国科学院软件研究所 http://www.c-s-a.org.cn

理、文本分类、图谱问答等方法实现. 规则推理方法需要人工设计推理规则, 灵活度不高; 文本分类方法直接预测主诉文本就诊科室, 但导诊是医学的强知识问题, 需要医疗理论知识的支撑保证其可靠性; 传统的图谱问答方法则只在主诉文本较为简单时实用, 无法应用于多医疗实体的复杂主诉文本.

针对上述问题, 本文设计了交互式智能导诊系统, 通过识别患者主诉中的疾病、症状、检查、药品等多种医疗指称, 然后链接到医疗知识图谱对应医疗实体上, 利用医疗实体在医疗图谱中查询子图, 通过图编码器提取子图语义信息, 结合主诉文本语义信息进行导诊. 最后针对推荐科室置信度不高的情况, 使用单轮导诊和多轮交互模块相结合的导诊方式, 通过搜索关联症状让用户选择来补充信息.

1 相关工作

1.1 智能导诊系统

导诊系统从原理上大致可分为基于规则模板和基于数据模型两类. 基于规则推理的方法通过人工建立症状、疾病和科室之间的对应规则实现导诊功能. 崔浩等[1] 通过提供图形化的界面让用户输入年龄、性别等个人信息, 选择患病部位及相关症状, 将相关症状作为特征推理匹配得到科室, 推荐给患者. 基于数据模型的方法不需要人工建立规则, 将导诊看作科室分类问题, 从医疗问诊网站爬取大量的问诊数据, 抽取患者病情描述和科室数据, 使用传统机器学习方法或者深度神经网络分类模型作为导诊模型. 郑姝雅[2] 将患者的年龄、性别特征和主诉信息融合后使用 SVM 预测科室. 陆康[3] 利用知识图谱问答的方式进行导诊, 通过识别患者问题中的疾病实体及其意图, 查询知识图谱对应科室进行导诊. 但无法解决问题中出现多个疾病实体时的情况, 在实际问诊中更多的是包含多实体的句子. Liu 等[4] 通过医疗百科网站和电子病历构建了以症状-疾病-科室为核心的医疗知识图谱, 根据用户自身症状计算潜在疾病及其权重, 通过查询医疗知识图谱得到疾病和科室之间的相关度, 将两种权重系数融合后得到科室的相关度.

1.2 医疗实体识别

医疗实体识别是指从医疗文本中识别疾病、症状、检查等实体. 早期的医疗实体识别使用基于词典的方法, 例如 cTAKES[5]、MedKAT[6] 等系统. 虽然词典方法在准确率上有不错的效果, 但由于医疗实体存在多种多样的表述, 词典方法覆盖率不高. 随着标注数据的日益丰富, 机器学习方法成为实体识别技术的主流, 有监督的序列标注模型取得了不错的效果, 如隐马尔可夫 HMM、条件随机场 CRF 等模型. Liu 等[7] 在 CRF 模型中加入 4 种特征对电子病历进行识别. 基于神经网络的方法相较于传统机器学习方法不需要特征工程, 非常灵活. 当数据规模较大时, 明显优于机器学习方法. Almgren 等[8] 提出基于字符的 Bi-LSTM 模型, 通过端到端的方式训练有明显提升. Xu 等[9] 将 Bi-LSTM 和 CRF 结合, 实验证明该方法优于单一模型. 医疗实体构词复杂, 分词时容易出现错误影响实体识别效果, BERT[10] 等预训练语言模型使用字符级的编码能有效规避分词导致的实体识别错误, 同时相较 Glove、Word2Vec 等静态词向量, 能获得动态字向量表示, 得到更丰富的文本语义信息.

1.3 医疗图谱应用

随着知识图谱的发展, 医疗领域知识图谱也日益完善, 在知识推理、智能问答、辅助诊断等智能医疗应用中发挥了重要作用. 侯梦薇等[11] 对医疗知识图谱架构、构建技术及应用现状进行了全面剖析. 奥德玛等[12] 基于大规模医疗文本数据, 利用自然语言处理与文本挖掘技术, 以人机结合的方式研发了一个中文医疗知识图谱. 昝红英等[13] 通过知识图谱中症状的发作部位和所属科室帮助导诊. 汤人杰等[14] 根据电子病历数据建立了症状和疾病之间的权重关系, 用于辅助诊断. 首先清洗电子病历中现病史一栏的否定修饰词, 然后识别其中的症状实体, 最后使用概率图方法计算症状和疾病之间的权重. 基于语义解析的知识图谱智能问答分为实体识别、实体链接和关系预测 3 个阶段, Wang 等[15] 使用编码器-解码器框架, 通过解码器预测关系路径, 增加了对关系路径正确性的验证结构, 并使用 APVA-TURBO 方式训练编码器和解码器, 提升了问答的准确率. Feng 等[16] 在常识问答任务中引入常识知识图谱来增加常识知识, 提出融合了基于路径的推理方法和图神经网络的多跳图编码器 MHGRN, 增强了模型的常识知识多跳推理能力.

然而, 上述研究工作中没有考虑到医疗导诊系统的特殊性. 一方面导诊是医学的强知识问题, 需要医疗理论知识的支撑保证其可靠性; 一方面患者提供的信息不足可能导致单轮导诊成功率低. 因此, 本文考虑通

© 中国科学院软件研究所 http://www.c-s-a.org.cn

过引入医疗知识图谱和多轮交互技术来提升导诊系统的智能化效果.

2 关键技术研究

在进行医疗导诊时, 其关键技术是如何从患者的主诉文本中提取医疗信息以及如何使用医疗知识更好的理解患者文本语义. 本节将介绍医疗信息识别算法和患者主诉推荐算法两个关键技术.

通常患者的主诉文本如表 1 所示.

表 1 主诉文本及对应科室

主诉	科室
最近一个多月以来头皮屑越来越多, 头越来越痒	皮肤科
原先疱疹痊愈后出现, 吃煎炸食物, 红肿更大	口腔科
两年前发现尿潜血, 一年前红细胞位相检查变形红细胞为主, 今年11月份体检开始出现尿蛋白, 间隔十天后复查仍有尿蛋白+−, 在今年7月份也做过尿沉渣,那个时候还没发现蛋白.	肾内科

2.1 医疗信息识别算法

医疗信息识别是指识别出患者主诉中医疗信息,并将其对应到医疗知识图谱中, 得到统一规范的医疗表述.本文首先使用 BERT+Bi-LSTM+CRF 模型识别主诉文本中的症状、疾病、检查、药品等医疗实体指称, 然后通过无监督的字符结合语义匹配的方法将识别的医疗指称链接到医疗知识图谱中的实体, 最后查询图谱中实体的名称得到统一规范的表述.

2.1.1 医疗实体识别

深度学习模型可以从医疗实体识别训练数据的标注中自动学习到医疗指称的上下文规律, 并且具备优秀的泛化能力. 本系统使用 BERT+Bi-LSTM+CRF 模型识别患者主诉中的医疗实体指称, 可以将模型分为 BERT, Bi-LSTM 和 CRF 三层网络结构, 该模型结构如图 1 所示.

第 1 层是利用预训练的 BERT 语言模型获取患者主诉文本的字向量, 记作序列$X=(x_1,x_2,\cdots,x_n)$, 通过预训练语言模型得到的字向量能够有效提取患者主诉文本中的特征信息.

第 2 层是 Bi-LSTM 层, 即双向长短时记忆神经网络. 第 1 层得到的序列 X 作为 Bi-LSTM 各个时间步的输入, 得到 Bi-LSTM 层的前向隐状态序列$\overrightarrow{h_t}$和后向隐状态序列$\overleftarrow{h_t}$, 将前向和后向隐状态序列按照时间步拼接得到完整的隐状态序列, 记作$H=(h_1,h_2,\cdots,h_n)$. 之后通过线性层将隐状态序列映射到 s 维 (即标注集的标签类别数目), 将映射后的序列记作$L=(l_1,l_2,\cdots,l_n)$, 其中, l_{ij}代表字x_i对应的每个类别标签的y_j的得分.

第 3 层是条件随机场层 CRF, 在命名实体识别任务中, 某一位置的标签不仅和自身有关, 还要考虑到前后位置的信息. CRF 网络不仅考虑到当前位置的标注概率分布, 还会考虑到之前位置的标签概率分布. 通过 Viterbi 算法解码得到概率最大的标签路径.

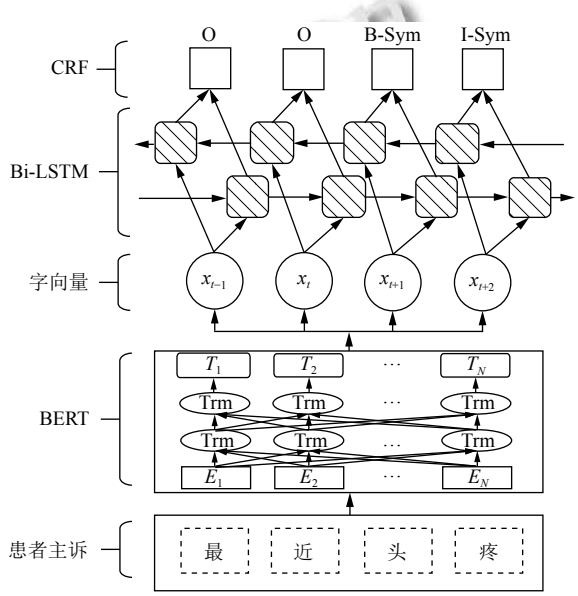

图 1 患者主诉医疗实体识别模型

2.1.2 医疗实体链接

医疗实体链接是将主诉文本中的医疗指称映射到医疗知识图谱中对应的实体, 分为候选实体生成和候选实体排序两个阶段.

第 1 阶段. 候选实体生成: 为医疗指称在医疗知识图谱中找到相关的候选实体列表. 抽取医疗知识图谱中的所有实体, 对每个实体构建 n-Gram (n=1,2,⋯,10)到实体的倒排索引表. 实体识别模型识别患者主诉文本得到医疗指称, 查询倒排索引表得到对应的候选实体列表, 计算医疗指称和列表中的每个元素之间的 Jaccard 相似度, 选择前 100 作为最终的候选实体列表.

第 2 阶段. 候选实体排序: 对候选实体列表中的候选实体排序, 选择排名最高的候选实体作为医疗指称的映射实体. 使用字符结合语义的方式对候选实体打分, 进行排序. 使用 Jaccard 相似度和相对编辑距离作为字符匹配的指标, 通过 fastText 中的 Skip-Gram 语言模型训练主诉文本得到词向量, 从语义的角度对候选

© 中国科学院软件研究所 http://www.c-s-a.org.cn

实体打分, 加权字符匹配度和语义匹配度得到候选实体的最终匹配分数, 最后选择分数最高的候选实体作为医疗指称的映射实体.

2.2 患者主诉科室推荐算法

在进行科室推荐时, 利用医疗知识图谱中症状、疾病等实体和科室实体之间的关系来帮助科室推荐. 但完整的医疗知识图谱一般包含几十万甚至上百万三元组, 考虑到效率和噪声问题, 采用从完整的医疗知识图谱中查询主诉文本相关的子图帮助导诊. 通过医疗信息识别算法得到患者主诉文本中在医疗知识图谱中对应的医疗实体, 称作主诉实体. 通过查询主诉实体和科室实体在医疗知识图谱中的相关路径得到主诉文本相关子图. 通过图编码器得到子图语义表示, 图编码器结构参考 MHGRN[16] 模型. 但仅使用图编码网络无法处理患者主诉中识别不到医疗实体的情况. 所以通过预训练语言模型 BERT 提取患者主诉的语义信息, 将主诉文本的语义信息和子图信息融合后进行科室推荐.

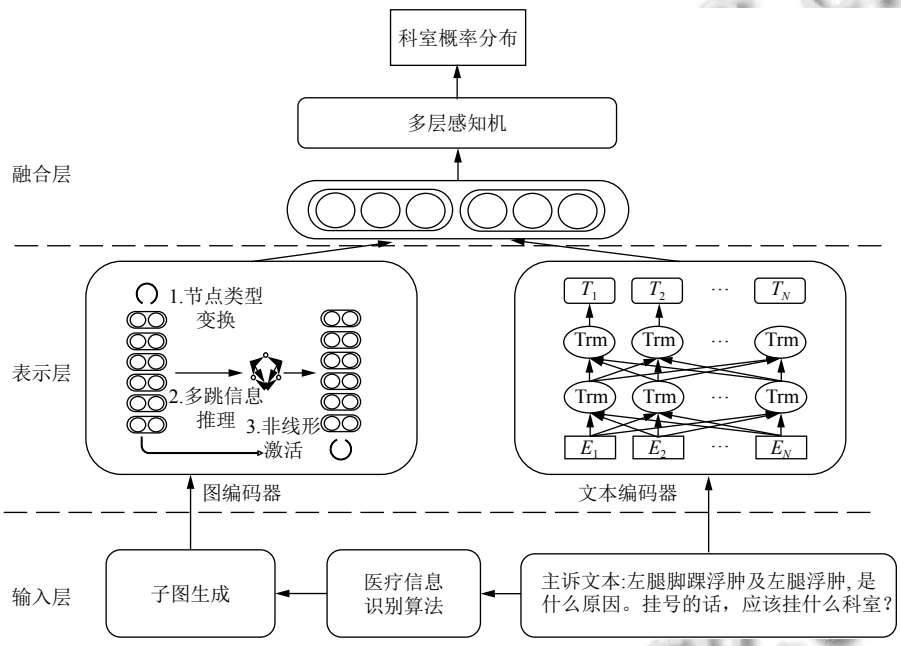

图 2 患者主诉科室推荐模型

图 2 是患者主诉科室推荐模型的结构, 分为输入层、表示层和融合层 3 层.

(1) 输入层: 主要是对患者主诉的预处理以及生成患者主诉相关子图.

文本输入处理: 将患者主诉按字 (token) 切分, 加上 BERT 模型要求的 [CLS] 和 [SEP] 特殊符号, 例如: “[CLS] 左腿脚踝浮肿及左腿浮肿, 是什么原因. 挂号的话, 应该挂什么科室?[SEP] ”, 将输入文本序列记作 $x \in R^{d_L}$, d_L 为患者主诉长度.

子图生成: 通过医疗信息识别算法得到主诉中的实体, 将主诉中的医疗实体称作主诉实体, 利用查询主诉实体-主诉实体, 主诉实体-科室实体之间的路径构建子图. 为了限制子图的规模, 过滤长度大于 4 的路径, 最后用邻接矩阵的形式表示, 记作 $M \in R^{r\times n\times n}$, r 为医疗知识图谱关系数, n 为子图中实体节点数目. 子图生成算法的伪代码如算法 1 所示, 输入是知识图谱 KG, 主诉中的实体集合 E_q, 科室实体集合 E_{dep}, 最大路径长度 $hop_{\max}$, 输出是子图对应的邻接矩阵形式 M.

算法 1. 子图生成算法(KG, E_q, E_{dep}, $hop_{\max}$)

```
1) V:={},E:={},M:=[]
2) def update_paths(KG,V,E,e_i,e_j,path_selector):
3)     paths:=find_paths(KG,e_i,e_j)
4)     for path in paths do:
5)         if path_selector(path) do:
6)             V.add(path.nodes)
7)             E.add(path.edges)
8)     end for
9) for e_{q_i} in E_q do:
10)     for e_dep in E_dep do:
11)         update_paths(KG,V,E,e_{q_i},e_dep, len(path)≤hop_max)
```


© 中国科学院软件研究所 http://www.c-s-a.org.cn

12) end for
13) for e_{q_j} in E_q do:
14) *update_paths*(KG,V,E,e_{q_j},e_{dep}, *len*(*path*)==2)
15) end for
16) *update_neighbors*(KG,e_{q_i})
17) end for
18) M:=*get_adj_matrix*(V,E)

(2) 表示层: 主要由文本编码器和图编码器组成. 文本编码器负责提取患者主诉中的文本语义特征, 模型中使用 BERT 预训练语言模型作为文本编码器, 得到文本语义向量$t \in R^{d_n}$. 图编码器提取患者主诉相关医疗子图中的语义特征, 使用 MHGRN[16] 的图编码网络, 得到子图语义向量$q \in R^{d_m}$.

文本编码器:

$$t = BERT(x) \tag{1}$$

图编码器:

首先通过图嵌入初始化子图中节点的表示h_i, 再根据子图中节点的类别进行转化.

$$x_i = U_{\phi(i)}h_i + b_{\phi(i)} \tag{2}$$

其中, $\phi(i)$是节点对应类型, 节点类型分为主诉文本医疗实体对应节点 (主诉节点)、科室节点和中间节点 3 种, U 和 b 是对应类别的参数.

将子图中所有长度为 k 的关系路径定义为ϕ_k, 关系路径最大长度为 K, $1 \le k \le K$.

$$\phi_k = \{(j,r_1,\cdots,r_k,i)|(j,r_1,j_1),\cdots,(j_{k-1},r_k,i) \in \mathcal{E}\} \tag{3}$$

使用多层的关系图卷积网络 (Relational Graph Convolution Networks, RGCNs) 得到节点在不同跳数下的表示如式 (4):

$$z_i^k = \sum_{(j,r_1,\cdots,r_k,i)\in\phi_k} \left[\alpha(j,r_1,\cdots,r_k,i)/d_i^k\right] RGCN(x_j) \tag{4}$$

其中, $\alpha(j,r_1,\cdots,r_k,i)/d_i^k$ 是相关系数, z_i^k 是节点 i 在第 k 跳的表示.

通过注意力机制融合不同跳数的节点表示:

$$z_i = \sum_{k=1}^{K} softmax(bilinear(t,z_i^k)) \cdot z_i^k \tag{5}$$

为了避免遗忘节点的原始信息, 通过短跳连接 (shortcut connection) 将节点原始嵌入和最新的节点表示融合.

$$h_i' = \sigma(Vh_i + V'z_i) \tag{6}$$

其中, V和V'是可学习的参数, σ是非线性激活函数.

最后对图中科室节点$\{h_i'|i \in \mathcal{D}\}$做注意力池化 (attentive pooling) 得到子图表示 g, $\mathcal{D}$是子图中科室节点集合.

(3) 融合层: 将患者主诉的文本语义特征向量和子图语义特征向量拼接后得到$q \in R^{d_m+d_n}$通过多层感知机将维度映射到 m 维 (m 维是科室数目), 记作$o \in R^m$, o_i表示推荐第i个科室的概率. 选择概率最大的科室y^*进行推荐.

$$o = MLP(t \oplus g) \tag{7}$$

$$y^* = \text{argmax}(o) \tag{8}$$

3 系统设计与实现

为了解决医疗导诊中的问诊文本表述口语化和症状和科室关系不明显的问题, 本文设计了一套基于医疗知识图谱的交互式智能导诊系统. 系统实现采用的程序开发语言为 Python 语言、深度学习框架为 PyTorch、Web 服务框架为 Flask. 整个系统的组织架构如图 3 所示.

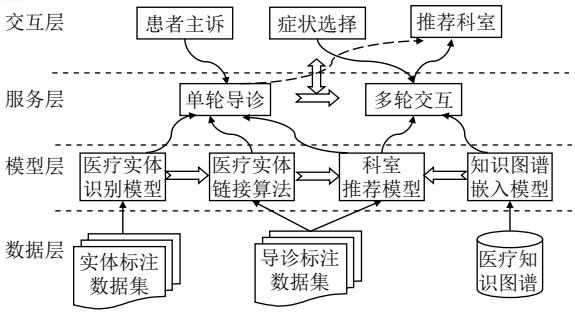

图 3 系统结构图

3.1 数据层

数据层包括医疗知识库和训练系统中模型所使用的语料. 下面简要介绍数据的构建过程及组成.

3.1.1 医疗知识图谱

本文将中科院软件所刘焕勇根据医疗百科网站整理的知识图谱和华东理工大学在 OpenKG 上发布的中文症状三元组关系库进行合并, 将合并后的医疗知识图谱作为导诊系统的知识库, 为导诊提供帮助.

下面简单介绍一下两个知识图谱的情况. 医疗百科网站知识图谱包括 4.4 万实体, 类别包括疾病、药品、食物、科室、检查、症状等 11 种关系, 37 万三元组, 本体结构以疾病实体为核心. 中文症状关系库包含 13.3 万实体, 类别有疾病、症状、检查、科室、部位、药品 6 类, 疾病相关科室、症状相关检查、检查相关部位等 22 类关系, 97 万三元组. 本体结构上, 疾病、科室、检查、症状、部位之间的关系稠密, 更适合导诊场景.

针对数据库的基本情况, 在合并时以中文症状关系库为基础, 医疗百科网站知识图谱作为补充, 去掉冗

© 中国科学院软件研究所 http://www.c-s-a.org.cn

余食物实体, 得到导诊系统中使用的医疗知识图谱, 具体情况如表 2 所示.

表 2 知识图谱信息表

知识库	实体	关系	三元组
医疗百科知识图谱	34 648	11	377 058
中文症状关系库	133 653	22	974 521
医疗知识图谱	149 848	22	1062 533

最后整理得到的医疗知识图谱本体关系结构如图 4 所示.

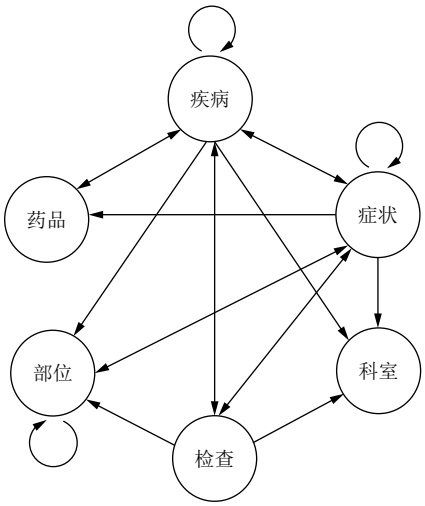

图 4 医疗知识图谱本体结构图

3.1.2 实体标注数据集

实体标注数据集使用来自“瑞金医院 MMC 人工智能辅助构建知识图谱大赛”的瑞金糖尿病数据集, 共有 45 929 条数据, 将其划分为训练集 4 万条, 验证集 2929 条, 测试集 3000 条, 涉及疾病名称、病因、临床表现、部位等 15 类实体类型, 标注采用 BIO 实体标注体系, 加上特殊标注符号“<sos>”、“<pad>”、“<eos>”共 34 个标注, 本文使用该数据训练医疗实体识别模型.

3.1.3 导诊标注数据集

数据语料从好大夫网站上爬取患者和医生的对话数据, 从中抽取患者的主诉及其就诊科室信息构造数据集, 数据集共有 36 400 条数据. 按照 8:1:1 的比例划分为训练集、验证集和测试集. 其中每份数据由患者主诉文本和其对应的科室组成, 共有骨科、神经内科、肝胆外科等 26 个科室, 本文使用该数据训练和评估科室推荐模型.

3.2 模型层

模型层是导诊系统所使用的核心模型. 下面介绍模型训练及效果验证.

3.2.1 知识图谱嵌入模型

知识图谱嵌入 (knowledge graph embedding)[17] 将知识图谱中的实体 (entity) 和关系 (relation) 嵌入到连续向量空间, 在方便计算的同时保留了知识图谱中的结构信息.

本文使用两种知识图谱嵌入方法得到实体和关系的语义向量表示, 一种是经典的 Trans-E[18] 方法, 一种是基于图注意力机制的图嵌入方法[19]. 并在医疗知识图谱上进行了链接预测 (link prediction) 实验, 链接预测是衡量知识图谱嵌入效果的常用方法, 将知识图谱中实体和关系的内容映射到连续向量空间中, 对知识图谱中的实体或关系进行预测, 即给出头实体和关系, 预测尾实体 (h,r,?) 以及给出尾实体和关系预测头实体 (?,r,t) 两种知识图谱的补全任务. 评价指标使用 hit@k, 即正确实体在预测时排名$\leq k$ 的频率. 实验中使用 hit@1, hit@3, 以及 hit@10 三个指标. 实验结果如表 3 所示.

表 3 链接预测实验对比结果

模型	hit@1	hit@3	hit@10
Trans-E	0.1176	0.2059	0.3244
图注意力模型	0.2042	0.2963	0.3932

可以发现图注意力模型得到的实体向量和关系向量明显优于 Trans-E 方法, 所以在系统的后续模块中使用图注意力模型方法得到的图嵌入表示.

3.2.2 医疗实体识别模型

数据集及模型: 使用第 3.1.2 节中介绍的实体标注数据集训练模型, 模型结构使用第 2.1.1 节中介绍的 BERT+Bi-LSTM+CRF.

模型验证与分析: 采用准确率、召回率和 $F1$ 值来评价模型. 在测试集中准确率 76.60%, 召回率 79.07%, $F1$ 值 77.82%. 识别错误的原因主要有两种, 一种是英文缩写实体类别错误 (例如 FFA, Free Fat Acid 是 Test 类别, 识别为 Drug 类别), 这是因为英文缩写本身不具备信息, 而 Test 和 Drug 实体周围的上下文比较相似, 所以容易误判. 另外一种是标注信息不全, 预测了正确的实体, 但在标注中未标注出来. 例如: 从“应注意影像学检查并不是诊断和手术指征依据, 多用于术前协助术式选择”中预测 “影像学检查”为 Test 类别, 但是在标注中未出现, 所以还是算作了错误识别.

3.2.3 科室推荐模型

数据集及模型: 使用第 3.1.3 节中的导诊标注数据集训练模型, 模型结构使用第 2.2 节介绍的结构.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

模型验证及分析: 采用准确率评估模型效果. 实验对比了基于 BERT 的分类方法和基于知识图谱问答的 APVA-TURBO 方法. 准确率为 75.14%, 较前两个方法准确率分别提升了 2.09% 和 2.47%. 科室推荐错误的原因主要是因为部分症状可去多个科室就诊, 但数据标注只标签单个科室. 例如泌尿外科和男科均可治疗男士性功能异常相关疾病.

3.3 服务层

服务层分为单轮导诊和多轮交互两个部分, 具体流程如图 5 所示, 患者主诉先通过单轮导诊部分得到推荐科室的置信度 (概率), 如果置信度小于设定的经验阈值 k ($0<k<1$), 则进入多轮交互部分, 通过和患者交互完成导诊, 交互能补充患者症状信息, 从而解决主诉中医疗信息不足的问题.

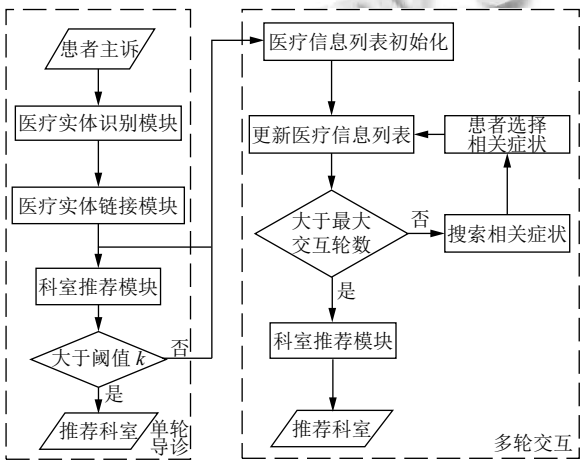

图 5 线上科室导诊流程图

3.3.1 单轮导诊

单轮导诊部分根据患者主诉进行导诊, 输入患者主诉, 输出推荐患者就诊的科室及其概率. 首先通过医疗实体识别模型识别患者主诉中的医疗指称, 然后利用基于字符和语义匹配的实体链接方法得到主诉中的医疗实体, 最后通过科室推荐模型推荐科室, 如果推荐科室的置信度 (概率) 大于设定阈值 k, 则将该科室返回给患者.

3.3.2 多轮交互

当患者主诉文本提供的信息太少时, 无法准确推荐合适的科室给患者, 需要和患者进行交互, 弄清楚患者的主要症状. 信息不足时, 单轮导诊推荐科室的置信度较低, 当低于设定阈值 k 时, 将进入多轮交互模块.

(1) 医疗信息列表初始化: 用医疗实体链接模型得到的实体初始化医疗信息列表, 如果未链接到医疗实体则初始化为空列表.

(2) 更新医疗信息列表: 将患者选择的相关症状加入医疗信息列表中, 将用户未选择的相关症状加入黑名单中, 避免下次搜索相关症状时搜索到用户拒绝过的相关症状.

(3) 搜索相关症状: 如果当前交互轮数小于最大交互轮数, 则基于医疗信息列表和黑名单搜索相关症状. 搜索方法: 遍历医疗信息列表中的实体, 利用图嵌入模型得到的实体向量计算得到实体之间相关度分数, 通过类型约束和黑名单过滤, 最后选择相关度分数最高的 3 个症状作为相关症状.

(4) 患者选择相关症状: 将搜索得到的 3 个相关症状让用户选择, 询问用户是否有这 3 个症状, 患者通过回复相关症状的名称进行选择, 如无则回复“无”.

(5) 最终科室推荐: 在当前交互轮数超出最大交互轮数时, 利用医疗历史信息列表中的医疗实体改写患者主诉后, 再次输入科室推荐模型, 得到最终的科室推荐结果.

3.4 交互层

交互层负责处理用户输入的患者主诉文本、多轮交互中用户选择的症状信息, 以及将推荐的科室展示给患者.

3.5 系统展示

系统导诊界面效果图如图 6 所示, 界面内容包括患者主诉的输入, 患者主诉实体识别和链指的结果, 从医疗知识图谱搜索得到的患者主诉相关子图的展现, 以及概率最大的前 5 个科室的结果展现.

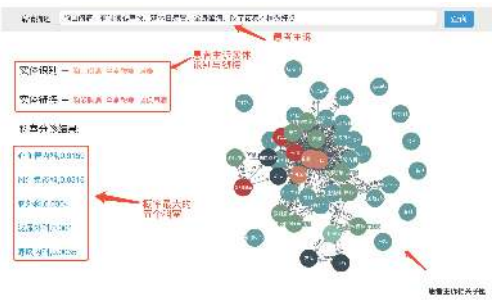

图 6 系统导诊界面效果图

4 结束语

本文介绍了基于医疗知识图谱的交互式智能导诊系统的设计与实现. 该系统利用现有医疗知识图谱资源, 整理得到一个适合导诊的医疗知识图谱, 并在此基础上, 通过实体识别、实体链接技术识别患者主诉中

© 中国科学院软件研究所 http://www.c-s-a.org.cn

的医疗信息, 利用预训练语言模型、图嵌入学习、图神经网络等技术查询得到问诊文本相关的子图, 实现了结合子图和文本语义信息的科室推荐模型. 针对推荐科室置信度不高的情况, 引入知识图谱, 采用单轮和多轮交互结合的方式进行科室推荐. 解决了在复杂的多医疗实体主诉文本上进行导诊的问题, 提升了导诊系统的智能度.

参考文献

1 崔浩, 刘丰源. 智能导诊服务机器人的设计与实现. 计算机应用与软件, 2020, 37(7): 329–333. [doi: 10.3969/j.issn.1000-386x.2020.07.055]

2 郑姝雅. 面向在线问诊平台的精准导医模型构建研究 [硕士学位论文]. 南京: 南京大学, 2020. [doi: 10.27235/d.cnki.gnjiu.2020.000205]

3 陆康. 基于知识图谱的医疗导诊问答系统的设计与实现 [硕士学位论文]. 武汉: 华中科技大学, 2020.

4 Liu DW, Ma ZY, Zhou YM, *et al*. Intelligent hospital guidance system based on multi-round conversation. 2019 IEEE International Conference on Bioinformatics and Biomedicine. San Diego: IEEE, 2019. 1540–1543. [doi: 10.1109/BIBM47256.2019.8983286]

5 Savova GK, Masanz JJ, Ogren PV, *et al*. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications. Journal of the American Medical Informatics Association, 2010, 17(5): 507–513. [doi: 10.1136/jamia.2009.001560]

6 Coden A, Savova G, Sominsky I, *et al*. Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model. Journal of Biomedical Informatics, 2009, 42(5): 937–949. [doi: 10.1016/j.jbi.2008.12.005]

7 Liu KX, Hu QC, Liu JW, *et al*. Named entity recognition in Chinese electronic medical records based on CRF. 2017 14th Web Information Systems and Applications Conference (WISA). Liuzhou: IEEE, 2017. 105–110. [doi: 10.1109/WISA.2017.8]

8 Almgren S, Pavlov S, Mogren O. Named entity recognition in swedish health records with character-based deep bidirectional lstms. Proceedings of the 5th Workshop on Building and Evaluating Resources for Biomedical Text Mining. Osaka: ACL, 2016. 30–39.

9 Xu K, Zhou ZF, Hao TY, *et al*. A bidirectional LSTM and conditional random fields approach to medical named entity recognition. International Conference on Advanced Intelligent Systems and Informatics. Cham: Springer, 2017. 355–365. [doi: 10.1007/978-3-319-64861-3_33]

10 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019. 4171–4186. [doi: 10.18653/v1/n19-1423]

11 侯梦薇, 卫荣, 陆亮, 等. 知识图谱研究综述及其在医疗领域的应用. 计算机研究与发展, 2018, 55(12): 2585–2599. [doi: 10.7544/issn1000-1239.2018.20180623]

12 奥德玛, 杨云飞, 穗志方, 等. 中文医学知识图谱 CMeKG 构建初探. 中文信息学报, 2019, 33(10): 1–9. [doi: 10.3969/j.issn.1003-0077.2019.10.001]

13 咎红英, 韩杨超, 范亚鑫, 等. 中文症状知识库的建立与分析. 中文信息学报, 2020, 34(4): 30–37. [doi: 10.3969/j.issn.1003-0077.2020.04.004]

14 汤人杰, 杨巧节. 基于医疗知识图谱的智能辅助问诊模型研究. 中国数字医学, 2020, 15(10): 5–8. [doi: 10.3969/j.issn.1673-7571.2020.10.002]

15 Wang Y, Zhang RC, Xu C, *et al*. The APVA-TURBO approach to question answering in knowledge base. Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe: ACL, 2018. 1998–2009.

16 Feng YL, Chen XY, Lin BY, *et al*. Scalable multi-hop relational reasoning for knowledge-aware question answering. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2020. 1295–1309. [doi: 10.18653/v1/2020.emnlp-main.99]

17 Bordes A, Usunier N, Garcia-Duran A, *et al*. Translating embeddings for modeling multi-relational data. Proceedings of the 26th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2013. 2787–2795.

18 Socher R, Chen DQ, Manning CD, *et al*. Reasoning with neural tensor networks for knowledge base completion. Proceedings of the 26th International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2013. 926–934.

19 Nathani D, Chauhan J, Sharma C, *et al*. Learning attention-based embeddings for relation prediction in knowledge graphs. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: ACL, 2019. 4710–4723. [doi: 10.18653/v1/P19-1466]

© 中国科学院软件研究所 http://www.c-s-a.org.cn