基于中文电子病历知识图谱的实体对齐研究

李丽双*
大连理工大学
计算机科学与技术学院
辽宁，大连
lilishaung314@163.com

董姜媛
大连理工大学
计算机科学与技术学院
辽宁，大连
donjyuan@163.com

摘要

医疗知识图谱中知识重叠和互补的现象普遍存在，利用实体对齐进行医疗知识图谱融合成为迫切需要。然而据我们调研，目前医疗领域中的实体对齐尚没有一个完整的处理方案。因此本文提出了一个规范的基于中文电子病历的医疗知识图谱实体对齐流程，为医疗领域的实体对齐提供了一种可行的方案。同时针对基于中文电子病历医疗知识图谱之间结构异构性的特点，设计了一个双视角并行图神经网络(DuPNet)模型用于解决医疗领域实体对齐，并取得较好的效果。

关键词： 医疗知识图谱 ； 中文电子病历 ； 实体对齐 ； 结构异构性 ； 并行图神经网络

Research on Entity Alignment Based on Knowledge Graph of Chinese Electronic Medical Record

Lishuang Li*
School of Computer Science
and Technology
Dalian University of Technology
Dalian, China
lilishaung314@163.com

Jiangyuan Dong
School of Computer Science
and Technology
Dalian University of Technology
Dalian, China
donjyuan@163.com

Abstract

The phenomenon of knowledge overlap and complementarity is common in different medical knowledge graphs. It is urgent to use entity alignment to fuse the medical knowledge graphs. However, according to our research, there is not yet a complete solution for entity alignment in the medical field. Therefore, we propose a standardized entity alignment process based on the Chinese electronic medical record knowledge graph, which provides a feasible scheme for entity alignment in the medical field. Meanwhile, according to the characteristic of the structural heterogeneity of the medical knowledge graph, we design a Dual-view Parallel Graph Neural Network (DuPNet) to solve the problem of entity alignment in the medical field, which achieves good results.

Keywords: Medical knowledge graph , Chinese electronic medical record , Entity alignment , Structure heterogeneity , Parallel graph neural network

基金项目：大连市科技创新基金(2020JJ26GX035)

1 引言

电子病历是信息化医疗健康服务的产物之一，它包含着大量的医学事实。随着国内电子病历的积累，利用自然语言处理技术从电子病历中自动化获取、整合医疗信息具有重要意义。构建电子病历相关的知识图谱是最有效的展示和利用电子病历中医疗信息的方法之一。然而，随着中文医疗知识图谱的广泛构建(奥德玛等, 2019; Xiu, 2020)，不同知识图谱之间存在着知识重叠和互补的现象，这就需要利用知识图谱融合来整合分散在各个知识图谱中的医疗知识，通过知识融合技术建立一个大规模的医疗知识图谱，可以为辅助决策、智能问答等下游应用提供技术支持(刘道文等, 2021; 刘勘和张雅荃, 2020)，从而促进智能医疗的发展。

知识融合中最关键的技术是实体对齐，其目的是判别不同知识图谱中的实体是否指向现实世界中的同一对象。据我们调研，目前中文医疗领域实体对齐的相关研究较少，大多数研究首先通过计算实体名称相似度生成候选实体对，然后再进一步利用结构、属性等信息判断候选实体对之间的相似性，这种方法虽然可以通过候选实体对降低模型复杂度，然而难以保证候选实体对的质量。在通用领域，实体对齐早期也是主要采用基于相似性度量的方法(Bhattacharya和Getoor, 2007; Jiang等, 2014)。随着知识图谱表示学习的兴起，越来越多的研究人员使用知识图谱表示学习解决实体对齐问题，最经典的是基于翻译的模型(Song等, 2021; Lu等, 2021),它们利用TransE(Bordes等, 2013)对三元组编码实现实体对齐。近年来，随着图神经网络的发展，一些研究使用图神经网络建模知识图谱的结构，用实体的邻域信息增强实体嵌入，即利用图卷积递归聚合邻居的嵌入表示来学习中心实体表示，通过计算实体间的嵌入距离实现实体对齐。

基于图神经网络的方法充分地利用了知识图谱的结构信息，提高了实体对齐模型的性能。然而，由于知识来源和构建目的不同，知识图谱之间存在着结构异构性，给此类方法带来了挑战。比如，由不同医院相同科室电子病历构建的两个知识图谱KG1和KG2存在的医疗知识的重叠与互补，造成了它们之间的结构异构性。Li等 (2019)利用知识嵌入和交叉图模型联合的半监督方法缓解结构异构性。Sun等 (2020)用实体邻域信息增强实体嵌入，并且使用图注意力机制为实体的每个邻居学习注意力分数来缓解结构异构性。Chen等 (2021)利用潜在的空间邻域聚合来处理结构异构性。然而这些研究过程仅考虑了结构异构性中的实体邻域异构性，如中心实体“艾滋病”与“AIDS”仅有实体“发烧”这一共同邻居，其余邻居均不同，该方法会为共同邻居“发烧”学习一个较高的权重。此类方法忽略了关系异构性对结构异构性的重要影响。事实上，来源不同的知识图谱往往具有关系独立性，例如，存在于KG1中的某一关系并不一定存在于KG2，导致了知识图谱之间的关系异构性，这是造成知识图谱结构异构性的重要原因。此外，现有的研究(Li等, 2019; Cao等, 2019)认为多层图卷积网络的输出层表示集成了实体的多跳邻域信息，因此他们将网络的输出层表示视为实体的嵌入表示。然而，我们发现随着卷积层数的增加，中心实体聚集的邻域信息呈指数级增长，因此给实体的表示带来了大量的噪声。

针对以上问题，本文设计了一个双视角并行图神经网络模型(DuPNet)用于中文电子病历的医疗知识图谱实体对齐。模型分别利用实体交互和关系交互缓解实体邻域异构性和关系异构性，以协同缓解医疗知识图谱的结构异构性。我们利用一个简洁有效的门控机制聚合网络层之间的输出，使得模型在捕获实体多跳邻域信息的同时，缓解由多层卷积引起的噪声问题。

此外，在医疗领域中，目前实体对齐相关研究相对较少，因此医疗领域中的实体对齐尚没有一个完整的处理流程，本文采用上述模型进行实体对齐，同时，针对医疗知识图谱的特点提出了一个规范的基于中文电子病历的医疗知识图谱实体对齐流程。主要贡献如下：

(1)提出了一个规范的基于中文电子病历的医疗知识图谱实体对齐流程，为医疗领域的实体对齐提供了一种可行的方案。

(2)针对基于中文电子病历医疗知识图谱之间结构异构性的特点，设计了一个双视角并行图神经网络(DuPNet)模型用于解决医疗领域实体对齐，并取得较好的效果。

2 相关工作

2.1 通用领域的实体对齐方法

在通用领域，JETEA(Song等, 2021)采用基于翻译的方法并且将实体类型匹配作为约束条件，使用一种迭代的方式将新检测到的对齐实体添加到训练数据中，以促进实体对齐。JTMEA(Lu等, 2021)也采用了基于翻译的方法，引入了一种具有属性增强的知识嵌入模型。然而基于翻译的方法无法充分利用知识图谱的结构信息。随着图神经网络的发展，越来越

多研究者使用基于图神经网络的方法解决实体对齐问题。GCN-Align(Wang等, 2018)首次尝试使用图神经网络进行实体对齐，将跨语言的实体嵌入到一个统一的向量空间中，并且将结构嵌入和属性嵌入相结合，以获得精确的对齐。KECG(Li等, 2019)提出一种基于联合知识嵌入模型和交叉图模型的半监督实体对齐方法，更好地利用种子对齐在整个图上传播。MUGNN(Cao等, 2019)提出了一种多通道的图神经网络框架处理实体对齐问题。AliNet(Sun等, 2020)通过使用门控策略和注意机制聚合多跳邻域，缓解实体邻域异构性。LatsEA(Chen等, 2021)利用潜在的空间邻域聚合来处理实体邻域异构性，并将实体对齐作为最大二部图匹配问题，采用匈牙利算法进行求解。AliNet和LatsEA在聚合邻域信息时认为实体的一跳邻居都同样重要。然而，并不是所有的一跳邻居都对中心实体有积极的贡献。上述基于图神经网络的模型虽然考虑了实体邻域异构性，但它们忽略了知识图谱之间关系异构性对结构异构性的重要影响。

2.2 中文医疗领域的实体对齐方法

目前中文医疗领域实体对齐的相关研究较少。宋文欣 (2018)分别用无监督和有监督的方法对医疗知识库进行实体对齐，首先计算实体指称项相似度生成候选实体对，然后在候选实体对之间得到最终的对齐实体对。蔡娇 (2020)采用基于网络语义标签的实体对齐算法用于遗传病领域的数据库，首先计算疾病名称相似度以生成候选实体对，然后用候选实体对计算多标签综合相似度，根据综合相似度判断实体对齐。这种方法虽然可以通过候选实体对降低模型复杂度，然而难以保证候选实体对的质量。

3 方法

3.1 双视角并行图神经网络实体对齐模型

为解决基于电子病历医疗知识图谱的结构异构性问题，本文设计与搭建了一个双视角并行图神经网络(DuPNet)实体对齐模型。

3.1.1 问题定义

本文将医疗知识图谱定义为$G = (E, R, T)$，其中E代表实体集，R代表关系集，T代表三元组集合，$e \in E$, $r \in R$, $t \in T$ 分别代表任一实体、关系、三元组。假设存在两个异构的医疗知识图谱G和$G' = (E', R', T')$，实体对齐最终目的是找出所有E和E'中指向同一对象的实体对。另外，$\mathbf{E}$和$\mathbf{E}'$分别代表实体特征矩阵，$\mathbf{R}$和$\mathbf{R}'$分别代表关系特征矩阵，均通过随机初始化的方式得到。

3.1.2 DuPNet模型架构

DuPNet从实体交互和关系交互的视角协同缓解医疗知识图谱的结构异构性。模型框架如图1所示。其中(1)和(2)代表由关系相似度矩阵得到的关系匹配度向量。从实体交互的视角来看，使用自注意力机制聚合实体的邻域信息，以缓解实体邻域异构性。从关系交互视角来看，由关系嵌入交互得到关系相似度矩阵，再由关系相似度矩阵得出关系匹配度作为跨图注意力分数聚合邻域信息，以缓解关系异构性。为得到更精确的实体表示，DuPNet利用门控机制聚合隐藏层和输出层的嵌入表示，从而缓解多层卷积引起的噪声问题。

3.1.3 实体交互视角

在实体交互视角中，通过自注意力机制迭代地为实体的每个邻居学习精确的自注意力分数，通过在训练的过程中，对重要的邻居赋予较高的权重，来缓解实体邻域的异构性。对于医疗知识图谱G中的任一实体e_i,自注意力分数由实体e_i和它的邻居实体的嵌入表示计算得到。自注意力分数$attn_{ij}^{e}$的计算公式如下:

$$attn_{ij}^{e} = \frac{exp(c_{ij}^{e})}{\sum\limits_{e_k \in \mathcal{N}_1(e_i) \cup \{e_i\}} exp(c_{ik}^{e})}. \tag{1}$$

$$c_{ij}^{e} = \sigma(\mathbf{q}[\mathbf{W}_1\mathbf{e}_i \| \mathbf{W}_2\mathbf{e}_j]). \tag{2}$$

其中c_{ij}^{e}是自注意力系数，代表实体e_j对e_i的重要程度。$e_k \in \mathcal{N}_1(e_i) \cup \{e_i\}$代表实体$e_i$包括自身在内的邻居，$\|$代表向量拼接，$\sigma(\cdot)$ 是激活函数，选择为$LeakyReLU(\cdot)$。$\mathbf{W}_1$,$\mathbf{W}_2$和$\mathbf{q}$是可训练参数。

第二十一届中国计算语言学大会论文集，第211页-第221页，南昌，中国，2022年10月14日至16日。

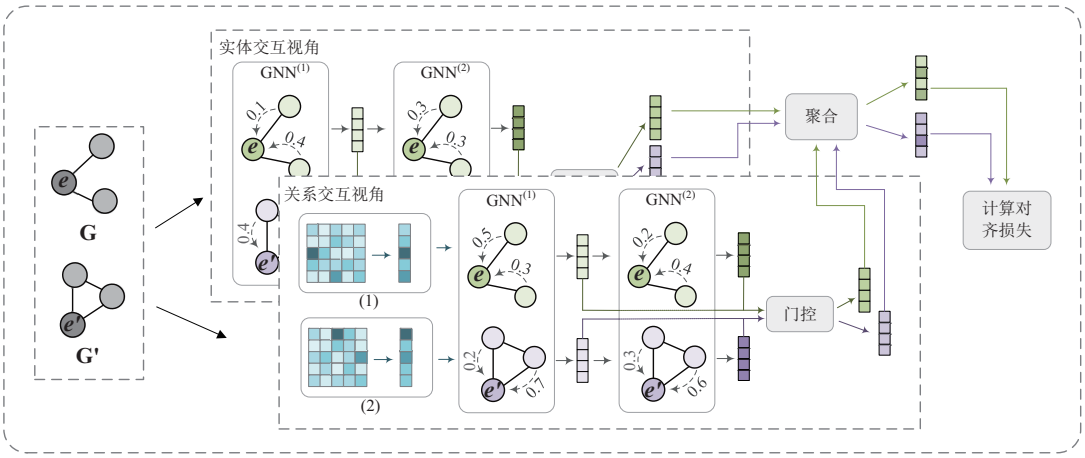

Figure 1: 双视角并行图神经网络实体对齐模型

在图神经网络中，节点的表示是通过递归聚合其邻居的特征向量来学习的。本文利用由公式(1)计算得到的自注意力分数$attn_{ij}^{e}$ 聚合邻居特征向量，得到实体e_i在实体交互视角中第l层的表示$\mathbf{h}_{i,e}^{(l)}$，计算公式如下：

$$\mathbf{h}_{i,e}^{(l)} = \sigma(\sum_{e_j \in \mathcal{N}_1(e_i) \cup \{e_i\}} attn_{ij}^{e} \mathbf{W}_3^{(l)} \mathbf{h}_{j,e}^{(l-1)}). \tag{3}$$

其中，$\mathbf{W}_3^{(l)}$是该视角网络中第l层的权重，$\sigma(\cdot)$为激活函数，选择为$ReLU(\cdot)$。

3.1.4 关系交互视角

在关系交互视角中，使医疗知识图谱G和G'的关系特征矩阵相互作用，得到关系相似度矩阵，然后进行最大池化操作，得到关系匹配向量**Match**。最后利用从关系匹配向量中得到的跨图匹配分数来聚合来自邻居的信息，关系匹配向量的计算公式为：

$$\mathbf{Match} = f_{max}(f_{sim}(\mathbf{R}, \mathbf{R}')). \tag{4}$$

其中，$f_{sim}(\cdot)$代表关系相似度计算函数，定义为$f_{sim}(\mathbf{R}, \mathbf{R}') = \mathbf{R}^T\mathbf{R}'$，$\mathbf{R}$和$\mathbf{R}'$分别代表待对齐的两个医疗知识图谱的关系特征矩阵，为可训练的参数。$f_{max}(\cdot)$代表最大池化操作函数。由关系匹配向量**Match**计算得到跨图匹配分数$attn_{ij}^{r}$，公式如下：

$$attn_{ij}^{r} = \frac{exp(c_{ij}^{r})}{\sum_{e_k \in \mathcal{N}_1(e_i) \cup \{e_i\}} exp(c_{ik}^{r})}. \tag{5}$$

$$c_{ij}^{r} = \underset{(e_i, r_{ij}, e_j) \in T}{\mathbf{Match}}[r_{ij}]. \tag{6}$$

其中**Match**$[\cdot]$代表关系匹配度索引操作。T代表知识图谱的三元组集合。利用跨图匹配分数计算实体在关系交互视角中第l层的表示$\mathbf{h}_{i,r}^{(l)}$，计算公式为：

$$\mathbf{h}_{i,r}^{(l)} = \sigma(\sum_{e_j \in \mathcal{N}_1(e_i) \cup \{e_i\}} attn_{ij}^{r} \mathbf{W}_4^{(l)} \mathbf{h}_{j,r}^{(l-1)}). \tag{7}$$

其中$\mathbf{W}_4^{(l)}$是该视角网络中第l层的权重，$\sigma(\cdot)$为激活函数，选择为$ReLU(\cdot)$。

3.1.5 门控聚合

为了获得更准确的实体表示，利用门控机制来聚合网络中隐藏层和输出层的嵌入表示，将其应用于上述两个视角。门控机制在捕获实体的多跳邻域信息增强实体的嵌入的同时去除各层的冗余噪声，从而缓解多层卷积引起的噪声问题。

门控机制的实现细节如下：

$$Gate_l(\mathbf{input}_1, ..., \mathbf{input}_l) = \begin{cases} \mathbf{g}_l \cdot \mathbf{input}_{l-1} + (1 - \mathbf{g}_l) \cdot \mathbf{input}_l, & l = 2 \\ \mathbf{g}_l \cdot Gate_{l-1} + (1 - \mathbf{g}_l) \cdot \mathbf{input}_l, & l > 2 \end{cases} \tag{8}$$

其中，l代表网络层数，$\mathbf{input}_l$代表网络第l层输出，$\mathbf{g}_l$为一组可训练的参数。

以实体交互视角为例，任一实体e_i该视角下嵌入表示为$\mathbf{h}_{i,e}$，公式如下：

$$\mathbf{h}_{i,e} = Gate_l(\mathbf{h}_{i,e}^{(1)}, ..., \mathbf{h}_{i,e}^{(l)}). \tag{9}$$

其中，$\mathbf{h}_{i,e}^{(l)}$为网络第l层的输出表示。同理，任一实体e_i该关系视角下嵌入表示为$\mathbf{h}_{i,r}$，公式如下：

$$\mathbf{h}_{i,r} = Gate_l(\mathbf{h}_{i,r}^{(1)}, ..., \mathbf{h}_{i,r}^{(l)}). \tag{10}$$

实体e_i的最终嵌入表示$\mathbf{h}_i$由门控机制聚合两个视角的输出得到，具体计算公式如下：

$$\mathbf{h}_i = \mathbf{g}_a \cdot \mathbf{h}_{i,e} + (1 - \mathbf{g}_a) \cdot \mathbf{h}_{i,r}. \tag{11}$$

其中$\mathbf{g}_a$为一组可训练的参数，用来控制两个视角的聚合。

3.1.6 对齐损失函数

对齐损失函数由两部分构成，分别是实体对齐损失和三元组对齐损失。其中实体对齐损失函数如下：

$$L_{ent} = \sum_{(e,e') \in A_e^+} \sum_{(e_-,e'_-) \in A_e^-} max\{0, \gamma_1 + dis(\mathbf{e} - \mathbf{e}') - dis(\mathbf{e}_- - \mathbf{e}'_-)\}. \tag{12}$$

其中A_e^+代表实体对齐对正例集合，A_e^-代表实体对齐对负例集合，γ_1为边际超参数，$dis(\cdot)$代表L_2范数，用于计算实体间的距离。

此外，我们引入三元组损失建模实体和关系之间的联系，并将三元组损失函数定义如下：

$$L_{tri} = \sum_{(h,r,t) \in T} \sum_{(h_-,r_-,t_-) \in T_-} max\{0, \gamma_2 + dis(\mathbf{h} + \mathbf{r} - \mathbf{t}) - dis(\mathbf{h}_- + \mathbf{r}_- - \mathbf{t}_-)\}. \tag{13}$$

其中，T代表三元组正例集合，T_-代表三元组负例集合。

综上所述，DuPNet最终的损失函数如下：

$$L = L_{ent} + L_{tri}. \tag{14}$$

3.2 基于中文电子病历医疗知识图谱的实体对齐流程

医疗知识图谱融合的目的是通过整合各个医疗知识图谱中分散的知识来构建一个更加精确和完善的医疗知识库，实体对齐是其中最关键的一步。针对医疗实体对齐中的实际应用，本文提出了一个规范的基于中文电子病历医疗知识图谱的实体对齐流程，如图2所示。首先，由于中文电子病历中知识纷繁复杂，同一医学术语知识图谱中可能存在多个不标准的实体表述。针对这一问题，我们首先构建医学词根库对单个医疗知识图谱进行实体规范化。其次，对医疗知识图谱进行推理(Lan等, 2021)能够补全缺失的知识，基于电子病历的单个医疗知识图谱中的知识往往是不完整的，所以提出利用规则挖掘进行知识推理。经过上述处理，单个医疗知识图谱的知识精度和完整性得到了提升，同时也为后续的实体对齐提供了良好的基础。最后构建医疗实体种子对，用训练集训练上述模型DuPNet的网络参数，以实现医疗知识图谱的实体对齐。下面将以两个由不同医院相同科室的电子病历构建得到的医疗知识图谱为例进行阐述，多个医疗知识图谱对齐即以两个对齐为基础进行迭代处理。

Figure 2: 基于电子病历医疗知识图谱的实体对齐流程

3.2.1 医疗知识图谱实体规范化

电子病历不同于医学书籍和文献，医生记录电子病历的习惯因人而异，导致在知识图谱中对于同一医学术语可能会有多个不同的医学实体表达，例如，对于医学术语“支气管炎”，医学实体“支气管炎”和“支气管炎症”可能同时存在于医疗知识图谱中。本文将这种具有相同词根的不同实体表达同一医学术语的情况称为“多词一义”问题，它使得知识图谱极度冗余。我们首先对每个医疗知识图谱进行实体规范化操作，提高知识图谱中实体的准确度，为后续实体对齐奠定良好的基础。

(1)医学词根库构建

在医疗领域中，医学词根可以代表医学实体中一个有意义的子串，且能够反应该医学实体的重要特征。由于医学术语的多个实体表达中大多包含相同词根，如上述例子中所示，“支气管炎”和“支气管炎症”中都含有相同词根“支气管”。 因此可以通过构造医学词根库推荐得到“多词

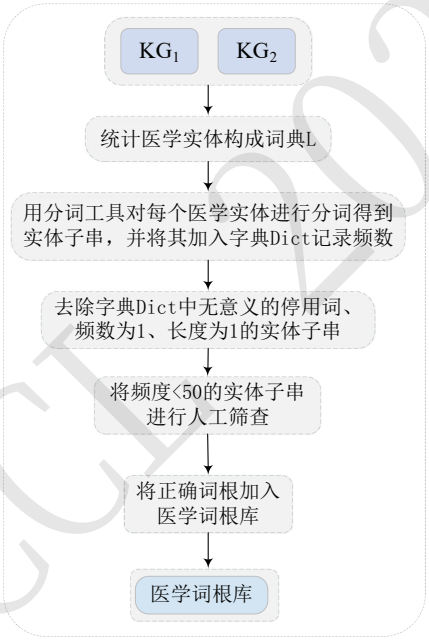

Figure 3: 医学词根库构建流程

一义”候选表，再由医学生对该表中“多词一义”实体进行标注并规范实体名称。医学词根库的构建方法如图3所示。

首先统计两个医疗知识图谱中的医学实体并构成词典L，然后利用北京大学开发的分词工具包pkuseg的医药领域对每个医学实体进行分词得到实体子串，并将其加入字典Dict，记录每个子串出现的频数。之后去除字典Dict中的无意义的停用词、频数为1以及长度为1的子串，因为这些子串对医学实体并没有很好的表征能力。例如子串“其他”、“任何”、“上”等。除此之外，子串频数过高可能会降低对医学实体的表征能力。例如子串“皮肤”出现的频数为743，包含该子串的实体“非黑色素瘤皮肤癌”、“亚急性皮肤型红斑狼疮”和“皮肤幼虫移行症”均不代表同一医学术语。因此我们将字典中频数小于50的子串加入医学词根库，在此之前为保证词根质量，我们先对其进行人工筛查。部分词根如图4所示。

(2)实体规范化

第二十一届中国计算语言学大会论文集，第211页-第221页，南昌，中国，2022年10月14日至16日。

胆固醇	细菌性	动脉硬化	杂音	颈动脉	狼疮性	霉菌	纤维瘤
血斑	粉碎性	并发症	腓肠	流产	结痂	胃肠道	化脓性
刺激	虹膜	心肌酶	湿疹	扭伤	肱骨	坏死性	回盲部
胃窦部	硬化	硫唑	肋骨	遗传性	尿路	硬化	斑丘
紫癜	粥样	体重	胰岛	阻滞	潮红	腋下	恶心
病理学	脑病	剧痛	挛缩	肠梗阻	贲门	继发	染色体

Figure 4: 医学词根示例

规范实体	**实体1**	**实体2**	**实体3**
不稳定型心绞痛	不稳定型心绞痛	不稳定性心绞痛	不稳定心绞痛
支气管炎	支气管炎	支气管炎症	
糖尿病肾病	糖尿病肾病	糖尿病性肾病	
双侧筛窦炎	双侧筛窦炎	双侧筛窦炎症	
原发性恶性肿瘤	原发性恶性肿瘤	原发恶性肿瘤	
急性胆囊炎	急性胆囊炎	胆囊急性炎症	
骨质疏松症	骨质疏松	骨质疏松症	
狼疮性肾炎	狼疮性肾炎	狼疮肾炎	
弓形虫病	弓形虫	弓形虫病	

Figure 5: “多词一义”规范表示例

根据上节得到的医学词根库，利用字符串索引算法推荐得到每个词根的“多词一义”候选表，再由医学生标注出其中正确的“多词一义”实体，并规范每一组“多词一义”实体的名称，由此得到“多词一义”规范表，该表的部分内容如图5所示。在知识图谱中，每一组“多词一义”实体被合并成同一规范实体，与“多词一义”实体相关的三元组中的实体也被其规范实体替代。

3.2.2 规则挖掘补全医疗知识图谱

现有医疗知识图谱通常由人工或半自动的方式构建，普遍存在不完备的问题。本文通过挖掘医疗知识图谱中潜在的规则来填补实体间缺失的关系从而达到补全的目的。首先，专家从现有的两个基于中文电子病历的异构医疗知识图谱中归纳出潜在的规则。之后，在每个医疗知识图谱中进行规则匹配，得到推理出的三元组。最后，为保证规则推理得到的三元组的质量，需要人工对推理出的三元组进行筛选。

(1)规则归纳

由于医学知识图谱存在精度要求高且复杂度高等特点，为保证补全三元组的正确性，我们请专家为每个医疗知识图谱归纳出规则集B。具体规则由前提三元组和结论三元组组成，其中，结论三元组可以由一系列的前提三元组推理得出。例如，[治疗改善疾病(x,y)]∧[疾病显示症状(y,z)]⇒[对症治疗(x,z)]。

(2) 规则落地

将由规则归纳得到的规则集合B应用于医疗知识图谱，给定一条规则$\beta \in B$，查找该知识图谱中满足该条规则的所有前提三元组，并依据规则推理出结论三元组，若结论三元组不存在于原来的知识图谱中则添加至原有知识图谱，即完整了一次三元组的补全操作。例如根据上述规则得到：[治疗改善疾病(泼尼松,肾病综合症)]∧[疾病显示症状('肾病综合症,蛋白尿)]⇒[对症治疗(泼尼松,蛋白尿)]。

(3)人工筛选

对于由规则落地推理出的结论三元组，虽然能够确保逻辑上的正确性，然而，有些医疗知识非常复杂，结论三元组仍然可能存在错误的情况，为进一步保证补全的结论三元组的准确性，专业的医学生对补全的结论三元组所表达的医疗知识进行确认，筛选出正确的结论三元组，将其补充到原医疗知识图谱中。

3.2.3 DuPNet用于电子病历医疗知识图谱实体对齐

(1)构建医疗实体种子对流程

基于图神经网络的实体对齐模型需要已知的实体种子对作为训练集和测试集，使得模型为待对齐实体学习到相近的嵌入表示。在医疗领域中，实体种子对主要由医学实体与其别名、简称等组成。例如，疾病实体“AIDS”与“艾滋病”为一组种子对。

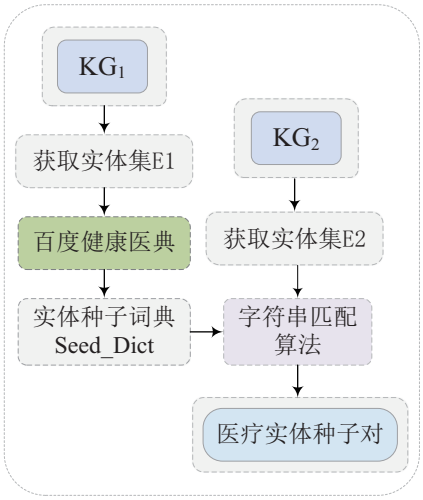

Figure 6: 医疗实体种子对构建流程

然而电子病历的内容中仅含有病人诊断治疗全过程的原始记录，并不直接保存有与疾病、治疗等实体相关的别名、简称等信息，因此给基于电子病历知识图谱的医疗实体种子对构建带来了困难。针对这一问题，本文提出的实体种子对标注流程如图6所示。

对于两个医疗知识图谱KG1和KG2，首先，获取KG1的实体集E1。然后从权威的健康知识科普平台“百度健康医典”利用网络爬虫技术为E1中的实体获得别名、简称等信息，构成实体种子词典Seed_Dict。最后将实体种子词典Seed_Dict与KG2实体集E2进行字符串匹配得到实体种子对。

(2)电子病历医疗知识图谱实体对齐

利用上述流程得到的医疗实体种子对作为训练集和测试集，在训练过程中通过最小化损失函数使得训练集中实体种子对的嵌入距离逐渐相近。训练完成后的模型具备了识别对齐实体对的能力，可以为待对齐实体学习到相近的嵌入表示，实现实体对齐。

4 实验

4.1 基于电子病历的医疗知识图谱数据详情

基于电子病历的医疗知识图谱KG_1和KG_2是对不同医院相同科室的电子病历进行三元组抽取得到的，其详细数据如表1中原始数据所示。经过构建“多词一义”规范表进行实体规范化后，KG_1和KG_2的实体数量分别减少494个和510个，修正后的医疗知识图谱为SKG_1和SKG_2。然后，在SKG_1和SKG_2基础上进行规则挖掘补全知识图谱，经由规则推理、人工筛选后得出的新三元组数量分别为11,639个和11,803个，补全后的医疗知识图谱为$CSKG_1$和$CSKG_2$。

	知识图谱	实体数量	关系数量	三元组数量
原始数据	KG_1	19,540	13	112,902
	KG_2	19,727	13	111,425
实体规范化	SKG_1	19,046	13	112,902
	SKG_2	19,217	13	111,425
规则挖掘补全	$CSKG_1$	19,046	13	124,541
	$CSKG_2$	19,217	13	123,228

Table 1: 基于电子病历的医疗知识图谱数据详情

	ZH_EN			JA_EN			FR_EN		
	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR	Hits@1	Hits@10	MRR
JETEA(2021)	42.7	75.0	-	36.4	72.4	-	36.5	71.8	-
JTMEA(2021)	42.2	75.9	53.0	38.7	72.4	50.0	37.1	75.6	46.0
GCN-Align(2018)	41.3	74.4	54.9	39.9	74.5	54.6	37.3	74.5	53.2
KECG(2019)	47.8	83.5	59.8	49.0	84.4	61.0	48.6	85.1	61.0
MuGNN(2019)	49.4	**84.4**	61.1	50.1	85.7	62.1	49.5	87.0	62.1
LatsEA(2021)	52.2	76.3	61.3	53.9	77.2	62.5	53.8	78.7	63.2
Alinet(2020)	**53.9**	82.6	62.8	**54.9**	83.1	64.5	55.2	85.5	65.7
DuPNet	52.8	83.9	**63.2**	53.7	**85.8**	**64.8**	**55.7**	**87.7**	**66.8**

Table 2: DuPNet在标准数据集上的实验结果

4.2 实验结果与分析

4.2.1 评价指标

遵循前人工作(Chen等, 2017; Cao等, 2019; Sun等, 2020)，本文采用Hits@k和MRR作为模型的评价指标。其中，Hits@k代表前k个候选实体中正确对齐实体的百分比，MRR代表正确对齐实体排名倒数的平均值。

4.2.2 DuPNet处理结构异构性的能力测试

DuPNet旨在解决结构异构性这一重要问题，为验证DuPNet缓解结构异构性的能力，本文将DuPNet在标准数据集DBP15K(Sun等, 2017)上与解决结构异构性的模型进行对比，该类模型均未使用预训练模型。其中GCN-Align(Wang等, 2018)是利用图神经网络解决实体对齐的首次尝试，KECG(Li等, 2019)、MUGNN(Cao等, 2019)、LatsEA(Chen等, 2021)和Alinet(Sun等, 2020)均是解决结构异构性的经典模型。另外将DuPNet与最新的基于翻译的模型JETEA(Song等, 2021)和JTMEA(Lu等, 2021)进行对比。实验结果如表2所示，其中黑体代表最优结果，下划线代表次优结果。

从表2中可知，基于图神经网络的方法普遍优于基于翻译的方法，这是因为基于图神经网络的方法能够充分利用知识图谱的结构信息。DuPNet全面优于JETEA(Song等, 2021)和JTMEA(Lu等, 2021)。与解决结构异构性的模型相比，DuPNet除了在ZH_EN的Hits@1和Hits@10上是次优结果，在JA_EN的Hits@1上与次优结果持平，在其他所有数据集的所有指标上都是最优结果，证明了DuPNet在处理结构异构性方面的优越性。这是因为DuPNet提出的双视角交互和门控机制起了非常重要的作用。一方面，双视角交互综合考虑了实体邻域异构性和关系异构性，可以使得对齐实体学习到更加相似的表示。另一方面，DuPNet通过门控机制对网络的隐层和输出层进行聚合，可以学习到实体更精确的表示，在保留多跳邻域信息的同时有效地去除各层的噪声。

4.2.3 DuPNet在电子病历医疗知识图谱上的实验结果

	Hits@1	Hits@5	Hits@10	Hits@50	MRR
DuPNet(w/o ent)	73.9	82.7	86.5	92.6	78.1
DuPNet(w/o rel)	75.5	82.7	86.3	94.5	78.5
DuPNet(w/o gate)	72.5	82.9	86.1	94.0	77.1
DuPNet	76.1	84.1	86.8	94.5	79.4
DuPNet-Bert	84.3	92.3	94.5	99.2	87.6

Table 3: DuPNet在电子病历医疗知识图谱上的实验结果

(1)DuPNet总体结果

医疗知识图谱KG_1和KG_2经由实体规范化和规则挖掘补全后得到的知识图谱$CSKG_1$和$CSKG_2$。我们对$CSKG_1$和$CSKG_2$按照3.2.3节的方法构建医疗实体种子对，得到的实体种子对数量为910，为保证模型训练充分，将实体种子对的60%作为训练集，40%作为测试集。DuPNet在医疗知识图谱$CSKG_1$和$CSKG_2$上的实验结果如表3所示，Hits@1值达到76.1%，Hits@10达到86.5%。

(2)利用Bert提高DuPNet在实际应用中的性能

为提高模型在实际应用中的性能，我们在DuPNet的基础上引入Bert预训练模型。由于DuPNet模型中实体表示是通过随机初始化得到的，因此实体的嵌入表示中仅包含结构信息。在医疗领域，医学实体的名称蕴含着丰富的语义信息，能够反应医学实体的重要特征，因此我们用Bert对医学实体名称编码，用带有丰富语义信息的词嵌入初始化实体表示来增强实体对齐。我们将引入Bert后的模型命名为DuPNet-Bert。

DuPNet-Bert在医疗知识图谱$CSKG_1$和$CSKG_2$上的实验结果如表3所示，在Hits@1上达到84.3%，在Hits@10上达到94.5%。与DuPNet相比，DuPNet-Bert在Hits@1、Hits@10和MRR上分别提高了8.2%，8.0%和8.2%，充分验证了Bert能够给实体嵌入表示带来有意义的语义信息。

(3)消融实验

为了验证DuPNet中各个模块的有效性，我们进行了详细的消融实验，结果如表3所示。

首先，去除实体交互模块，并将该实验表示为DuPNet(w/o ent)。实体交互模块的去除导致DuPNet性能整体降低，在Hits@1、Hits@10和MRR上分别下降了2.2%、0.3%和1.3%，这是因为实体交互视角通过给邻居赋予精确的权重缓解实体邻域异构性。实验结果证明了实体交互视角的有效性。

然后，去除关系交互模块，并记为DuPNet(w/o rel)。与DuPNet相比，DuPNet(w/o rel)在Hits@1、Hits@10和MRR上下降了0.6%、0.5%和0.9%，原因在于关系交互视角通过使用关系匹配度能够充分缓解关系异构性，实验结果证明了关系交互视角的有效性。

最后，去除门控机制，用平均池化代替，并表示为DuPNet(w/o gate)。门控机制被去除后，DuPNet在Hits@1、Hits@10和MRR上分别下降了3.6%、0.7%和2.3%。这是因为门控机制在捕获实体多跳邻域信息的同时能够有效去除多层卷积带来的噪声。实验结果证实了门控机制的有效性。

5 结论

医疗知识图谱中知识重叠和互补的现象普遍存在，利用实体对齐进行医疗知识图谱融合成为迫切需要。然而据我们调研，目前在医疗领域的知识图谱实体对齐尚没有完整的处理方案。针对医疗知识图谱的特点提出了一种规范的电子病历医疗知识图谱实体对齐流程，为中文医疗领域的实体对齐提供了一种可行的方案。针对基于电子病历知识图谱结构异构性的特点，设计了一个双视角并行图神经网络模型并用于医疗知识图谱实体对齐，实验结果证明了该模型处理结构异构性的优越性，并且按照上述流程进行了实际的基于中文电子病历知识图谱实体对齐，取得了较好的效果。

参考文献

奥德玛, 杨云飞, 穗志方, 等. 2019. 中文医学知识图谱CMeKG构建初探. 中文信息学报, 33(10):1-9.

Bhattacharya Indrajit, Getoor Lise. 2007. Collective entity resolution in relational data. *Information Sciences*, 1(1):1-36.

Bordes Antoine, Usunier Nicolas, Garcia-Duran Alberto, et al. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems*, 2787–2795.

蔡娇. 2020. 基于遗传病领域的实体对齐研究. 硕士学位论文. 苏州大学.

Cao Yixin, Liu Zhiyuan, Li Chengjiang, et al. 2019. Multi-Channel Graph Neural Network for Entity Alignment. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1452-1461.

Chen Muhao, Tian Yingtao, Yang Mohan, et al. 2017. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 1511-1517.

Chen Wei, Chen Xiaoying, Xiong Shengwu. 2021. Global Entity Alignment with Gated Latent Space Neighborhood Aggregation. *China National Conference on Chinese Computational Linguistics*, 371-384.

Jiang Yong, Wang Xinmin, Zheng Haitao. 2014. A semantic similarity measure based on information distance for ontology alignment. *Information Sciences*, 278:76-87.

Lan Yinyu, He Shizhu, Liu Kang, et al. 2021. Path-based knowledge reasoning with textual semantic information for medical knowledge graph completion. *BMC Medical Informatics Decis Mak*, 21-S(9): 335.

Li Chengjiang, Cao Yixin, Hou Lei, et al. 2019. Semi-supervised entity alignment via joint knowledge embedding model and cross-graph model. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2723–2732.

刘勘, 张雅荃. 2020. 基于医疗知识图谱的并发症辅助诊断. 中文信息学报, 34(10):85-93,104.

刘道文, 阮彤, 张晨童, 等. 2021. 基于多源知识图谱融合的智能导诊算法. 中文信息学报, 35(01):125-134.

Lu Guoming, Zhang Lizong, Jin Minjie, et al. 2021. Entity alignment via knowledge embedding and type matching constraints for knowledge graph inference. *Journal of Ambient Intelligence and Humanized Computing*, 1-11.

宋文欣. 2018. 面向医疗领域的实体对齐研究. 硕士学位论文. 哈尔滨工业大学.

Song Xiuting, Zhang Han, Bai Luyi. 2021. Entity Alignment Between Knowledge Graphs Using Entity Type Matching. *International Conference on Knowledge Science, Engineering and Management*, 578-589.

Sun Zequn, Hu Wei, Li Chengkai. 2017. Cross-lingual entity alignment via joint attribute-preserving embedding. *International Semantic Web Conference*, 628-644.

Sun Zequn, Wang Chengming, Hu Wei, et al. 2020. Knowledge graph alignment network with gated multi-hop neighborhood aggregation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):222-229.

Wang Zhichun, Lv Qingsong, Lan Xiaohan, et al. 2018. Cross-lingual knowledge graph alignment via graph convolutional networks. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 349-357.

Xiu Xiaolei. 2020. Construction of a Digestive System Tumor Knowledge Graph Based on Chinese Electronic Medical Records: Development and Usability Study. *JMIR Med Inform*, 8(10):e18287.

Zhu Hao, Xie Ruobing, Liu Zhiyuan, et al. 2017. Iterative Entity Alignment Via Joint Knowledge Embeddings. *International Joint Conference on Artificial Intelligence*, 4258–4264.