计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications [doi: 10.15888/j.cnki.csa.008642]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于条件提示与序列标注的电子病历关系三元组识别①

郭宇捷[1,2], 唐珂轲[1,2], 付立军[1,2,3], 于碧辉[1,2], 韩振桥[1,2]

[1](中国科学院大学, 北京 100049)
[2](中国科学院 沈阳计算技术研究所, 沈阳 110168)
[3](山东大学 大数据技术与认知智能实验室, 济南 250100)
通信作者: 付立军, E-mail: fu_lijun@ucas.ac.cn

摘　要: 电子病历是诊疗过程中记录患者健康状况的档案, 文本中分布着大量的医学实体, 其中蕴含着丰富的医学信息. 目前医学领域的关系抽取模型主要是通过关系分类的方法识别两个给定医学实体之间的语义关系. 中文电子病历具有实体高密度分布的特点. 针对这个问题, 本文提出了一种基于条件提示与序列标注的关系三元组识别方法, 将关系三元组识别任务转换为序列标注任务. 关系三元组中的头实体和关系类型作为条件提示信息, 通过序列标注方法识别电子病历文本中与条件提示信息有关联的尾实体. 在中文电子病历数据集上的实验证明本文方法能有效识别中文电子病历中的关系三元组.

关键词: 中文电子病历; 关系抽取; 条件提示; 序列标注; 关系三元组

引用格式: 郭宇捷,唐珂轲,付立军,于碧辉,韩振桥.基于条件提示与序列标注的电子病历关系三元组识别.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/8642.html

Relation Triple Recognition in Electronic Medical Records Based on Condition Hint and Sequence Labeling

GUO Yu-Jie[1,2], TANG Ke-Ke[1,2], FU Li-Jun[1,2,3], YU Bi-Hui[1,2], HAN Zhen-Qiao[1,2]

[1](University of Chinese Academy of Sciences, Beijing 100049, China)
[2](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)
[3](Laboratory of Big Data and Artificial Intelligence Technology, Shandong University, Jinan 250100, China)

Abstract: Electronic medical records are the archives to note patients' health conditions during treatment, where a large number of medical entities are scattered throughout the text and a wealth of medical information is contained. Existing relation extraction models in the medical field mainly utilize the relation classification method to recognize the semantic relation between two medical entities. Chinese electronic medical records have the characteristic of a dense distribution of medical entities in the text. In response, this study proposes a method based on condition hint and sequence labeling to extract relation triples. In this approach, the relation triple recognition task is converted to a sequence labeling task. The head entity and relation type in a relation triple combine to form condition hint information, and the model recognizes tail entities relevant to the condition hint information from the text of electronic medical records by sequence labeling. The experimental results on an electronic medical records dataset show that this method can be applied to recognize relation triples in Chinese electronic medical records.

Key words: Chinese electronic medical records; relation extraction; condition hint; sequence labeling; relation triple

① 基金项目: 国家社科基金 (21BTQ106)
收稿时间: 2021-11-10; 修改时间: 2021-12-13; 采用时间: 2021-12-28; csa 在线出版时间: 2022-06-01

1 引言

电子病历中蕴含着大量的医学实体和概念, 记录了大量的与患者健康状况相关的信息, 是一个丰富的医学知识宝库. 基于电子病例的关系抽取作为医学信息抽取领域的子任务, 旨在从非结构化的电子病历文本中抽取两个医学实体之间的关系, 是构建医学垂直领域知识图谱的关键步骤. 知识图谱的基本组成是形如<头实体, 关系类型, 尾实体>的三元组. 从电子病历中识别医学实体之间的语义关系, 构建医学领域知识图谱对下游医学任务具有重要意义.

目前对电子病历中的关系抽取的研究主要集中在英文的电子病历, 主要是使用关系分类方法分析给定的两个实体的上下文, 从而判断两个实体间的关系所属的类别[1−3]. 然而, 中文电子病历中的医学实体的分布具有高度密集的特点. 例如在图 1(a) 所给出的例子中, “患者伤后出现头部疼痛和鼻腔流液”这个句子记录了患者伤后出现的症状表现, 包含[头部], [疼痛], [鼻腔], [流液]等医学实体, 实体间产生两组关系, 构成了两对医学关系三元组, 分别为<头部, 结构描述, 疼痛>和<鼻腔, 结构描述, 流液>, 其中“结构描述”是预定义的关系类型. 从上述的例子可以看出, [头部]和[疼痛]是一组有关系的实体对, 它们所在的上下文中分布着[鼻腔]和[流液]等医学实体, 这些实体概念的内容对分析[头部]和[疼痛]之间的语义关系并没有起到帮助作用, 当这些无关实体的数量变多时甚至可能会给关系分类模型引入噪声, 阻碍模型做出正确的分类决策.

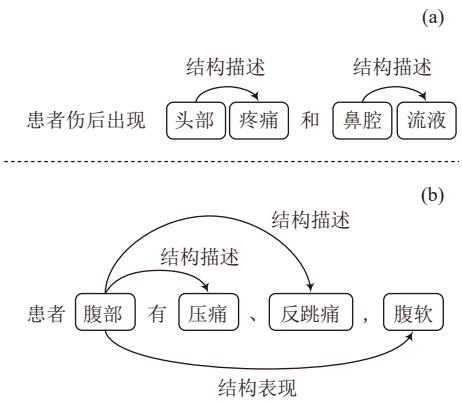

图 1 中文电子病历关系三元组实体分布示意图

中文电子病历文本中医学实体密集分布的特点还会产生多个三元组共享一个实体的情况. 如图 1(b) 所示, 实体[压痛]和实体[反跳痛]都与实体[腹部]具有语义关系, 关系类型为“结构描述”, 实体[腹软]与实体[腹部]也具有语义关系, 关系类型为“结构表现”. 当一个实体参与了多个关系三元组时, 传统的关系分类器容易发生混淆. 综上所述, 中文电子病历文本的实体高密度分布的问题, 给现有的关系抽取模型带来了一定的挑战.

针对使用关系分类方法处理中文电子病历文本中的高密度实体分布问题所面临的挑战, 本文提出了一种基于条件提示与序列标注的关系三元组识别方法. 相较于传统的将实体对以及实体对所在的文本输入模型以求解实体对语义关系的关系分类方法, 本文提出的方法将关系抽取任务转换成一个基于条件提示信息的序列标注任务. 该方法的核心是给定先验条件作为提示, 建模条件提示信息与中文电子病历文本语义的依赖关系, 并从文本中寻找多个能够与条件提示信息相匹配的片段进行标注, 其中条件提示信息定义为关系三元组的头实体和关系类型词, 被抽取的多个片段是关系三元组中的尾实体. 比如在对图 1(b) 中的例子进行关系三元组抽取的过程中, 当条件提示信息为“腹部”和“结构表现”组成的术语时, 基于该术语作为先验条件, 可抽取出片段[腹软]. 当条件提示信息由“腹部”和“结构描述”组成时, 模型抽取内容则变成[压痛]和[反跳痛]两个片段. 即假如一个给定的头实体参与了多个关系三元组, 并且这些关系三元组的关系类型都是相同的, 那么模型的目标是从文本中识别出与给定的头实体和关系类型相关联的全部尾实体片段, 组成若干个关系三元组. 由于条件提示信息的存在, 电子病历文本中与条件提示信息无关联的医学实体将被过滤, 与条件提示信息相关联的医学实体才能被识别.

本文的主要贡献如下:

(1) 组织构建了一批中文电子病历数据集, 在医学领域专家的指导下定义中文电子病历中的实体以及实体对之间的语义关系, 并请医学专业人员对实体和关系数据进行了人工标注.

(2) 针对中文电子病历中实体密集分布的数据特征, 设计了一种基于条件提示与序列标注的关系三元组识别方法, 通过建模条件提示信息与文本序列字符特征的关联, 从文本中识别出与条件提示信息相关的尾实体片段, 从而实现关系三元组的识别. 在中文电子病历数据集上进行的实验证明了该方法的有效性.

2 相关工作

目前在医学领域的关系抽取方法主要有基于规则匹配的方法, 基于特征工程的方法和基于深度学习的方法.

早期的关系抽取任务主要依靠人工分析文本特征构建模板, 用模板在新的文本中匹配符合既定规则的关系三元组, 这要求制定规则的人员拥有丰富的医学领域知识, 因此构建人工规则的代价十分昂贵. 文献 [4] 利用句子上下文的语法结构信息构建了 10 多个模板, 从医学文献中识别蛋白质实体以及实体之间的关系. 文献 [5] 设计了一种简化句法的方法, 将结构复杂的上下文句子分解为简单句, 通过规则从简单句中识别出具有相互作用的药物对. 对于不同的语料而言, 由于背景内容和语言结构的差异, 基于规则模板的方法难以在不同的语料之间迁移, 会导致关系抽取的召回率较低, 因此基于模板规则匹配的方法的泛化能力较差.

医学信息抽取任务的发展使得以特征工程为和核心的机器学习方法广泛应用于医学信息抽取. 2010i2b2/VA 评测任务[6] 的提出吸引了众多研究者关注电子病历中的关系抽取任务. 文献 [7] 手工提取词语特征、短语特征、句法特征等多种特征用于抽取化学-疾病数据集中的关系三元组. 文献 [8] 在支持向量机 (support vector machine, SVM) 的基础上融合句子结构信息, 在提取电子病历实体关系时考虑了句子结构的相似性. 相比于基于模板规则匹配的方法而言, 基于特征工程的方法具有很好的泛化能力, 能移植到不同的语料, 然而特征的选择会极大的影响模型最终的抽取性能, 并且提取语法、句法等特征往往需要用到外部工具, 外部工具自身的误差可能会传递到关系抽取模型中.

近年来, 深度学习方法在医学关系抽取任务中得到了广泛的应用, 其中比较经典的是基于卷积神经网络 (convolutional neural network, CNN) 和循环神经网络 (recurrent neural network, RNN) 的模型[9,10]. 文献 [11] 将词嵌入向量和位置嵌入向量输入 CNN 模型进行药物相互作用的关系抽取. 文献 [12] 利用 CNN 获取句子的局部特征, 并结合最大池化抽取电子健康档案中的实体关系事实. 文献 [13] 利用残余卷积块降低了电子病历实体关系抽取中数据噪声带来的影响. 文献 [14] 提出了一种两阶段的方法提取医学文本中的实体和关系, 其中关系抽取模块利用 CNN 提取单词、实体类型以及位置嵌入的特征. 文献 [15] 利用层次 RNN 模型引入最短依赖识别药物关系. 为了弥补 RNN 处理长距离文本特征能力不足的缺陷, 其变种长短期记忆网络 (long short-term memory, LSTM) 得到了应用[16]. 文献 [17] 通过结合双向长短期记忆网络 (bidirectional long short-term memory, Bi-LSTM) 和多跳自注意力机制获取文本的多重向量表示, 提升捕捉医学实体之间复杂语义信息的效果. 文献 [18] 在深度学习框架中整合最短路径依赖和句子序列表示, 提升了关系抽取的性能. 基于深度学习的方法将文本转换为向量表示, 不需要手工提取复杂的特征.

对于开放领域的实体关系抽取任务, 研究者们提出了许多新范式. 文献 [19] 将实体关系抽取任务转换成阅读理解任务, 依据实体和关系生成不同的问题模板, 通过从上下文中抽取出能够回答该问题的片段的方法识别文本中的实体和关系. 文献 [20] 对重叠关系三元组问题进行了研究, 并提出了一种将文本中的主语实体映射成宾语实体的级联标注框架.

3 关系三元组识别模型

为了应对中文电子病历文本中实体高密度分布带来的问题, 本文提出了一个基于条件提示与序列标注的关系三元组识别方法, 旨在通过捕捉由头实体和关系类型词组合成的提示信息与中文电子病历文本片段的关联, 并从文本中抽取出与条件提示信息相关联的片段, 被抽取出的片段作为尾实体与条件提示信息中的头实体和关系类型构成一个有效的关系三元组. 本文提出的模型如图 2 所示, 主要包括以下几个部分: 关系类型词编码、电子病历文本编码、条件信息交互融合以及解码输出. 模型的输入是电子病历文本、三元组的关系类型词以及三元组的头实体的掩码序列. 模型的输出是基于 BIESO 标注规范[21] 标注的序列, 其中, B 表示尾实体片段的开头, I 表示尾实体片段的中间部分, E 表示尾实体片段的结尾, 若尾实体片段为单字则标记为 S, 其余的无关字符将被标注为 O. 对于 BIES 标签, 标签的后面通过“-”连接尾实体所属的类别, 如“B-描述”“I-描述”“E-描述”, 表示该尾实体为一个“描述”类型的实体. 序列标注的结果经过处理后得到若干条形如<头实体, 关系类型, 尾实体>的关系三元组.

3.1 模型输入编码层

如图 2 所示, 本文的模型设计了两个输入网络, 分别用于编码关系类型词和电子病历文本, 编码的 token 为单个中文字符. 为了使模型能够捕获电子病历文本中的每一个字相对于头实体的位置特征, 头实体

的前后位置加入了特殊字符@作为位置标记. 在文本表示的过程中, 使用一个字向量矩阵, 将关系类型词和电子病历文本序列中的每个字映射成高密度的字嵌入向量. E_x是电子病历文本序列的字嵌入向量表示, 其长度为 m. E_r是关系类型词的字嵌入向量表示, 其长度为 n.

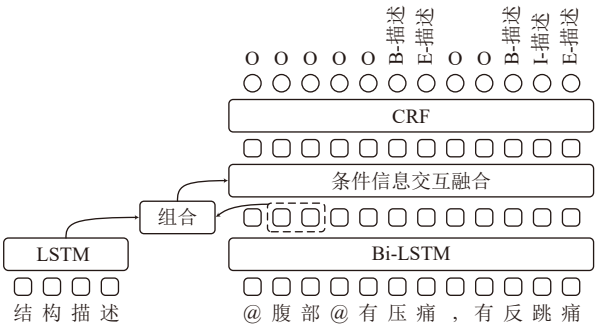

图 2 基于条件提示与序列标注的关系三元组识别模型图

为了获得电子病历文本序列中每个字的特征, 将上一步获得的文本序列字嵌入向量表示E_x利用 Bi-LSTM 进行特征编码. Bi-LSTM 网络能够从前向和后向两个方向对电子病历上下文序列进行编码, 前向编码结果$\overrightarrow{H_x}$和后向编码结果$\overleftarrow{H_x}$经过拼接后, 得到电子病历文本序列的字级别特征表示向量$H_x=[x_1,x_2,\cdots,x_m]$, 其中x_i表示第 i 个字符经过 Bi-LSTM 编码后输出的隐藏层特征向量. 电子病历文本序列的特征表示向量的编码过程的公式如式 (1)–式 (3):

$$\overrightarrow{H_x}=\overrightarrow{LSTM}(E_x)=[\overrightarrow{x_1},\overrightarrow{x_2},\cdots,\overrightarrow{x_m}] \tag{1}$$

$$\overleftarrow{H_x}=\overleftarrow{LSTM}(E_x)=[\overleftarrow{x_1},\overleftarrow{x_2},\cdots,\overleftarrow{x_m}] \tag{2}$$

$$x_i=[\overrightarrow{x_i}\oplus\overleftarrow{x_i}] \tag{3}$$

为了获得关系类型词的特征表示, 将字嵌入映射阶段获得的关系类型词字嵌入向量表示E_r利用 LSTM 网络编码得到特征表示H_r, 编码关系类型词特征表示的公式如式 (4):

$$H_r=LSTM(E_r)=[h_1^r,h_2^r,\cdots,h_n^r] \tag{4}$$

其中, h_j^r 表示第 j 个字经过 LSTM 编码后输出的隐藏层特征向量. 取H_r的最后一个隐藏层状态h_n^r作为关系类型词的特征表示h_{rel}.

3.2 条件信息交互融合层

为了应对电子病历关系三元组识别中存在的实体密集分布以及一个实体参与了多个关系三元组的情况, 在本文提出的模型中设计了一个条件信息交互融合层, 用于建模由头实体+关系类型组成的条件提示信息与尾实体之间的依赖, 使得模型的输出层对电子病历文本序列进行标注的时候能够考虑先验条件, 依据不同的先验条件标注不同的尾实体片段. 条件信息交互融合层主要包含两个步骤: 1) 创建条件提示信息; 2) 条件提示信息与电子病历文本交互.

创建条件提示信息需要将头实体的特征表示和关系类型词的特征表示进行融合. 本文利用一个头实体掩码序列, 从电子病历文本序列的特征表示中获取头实体的特征表示. 假设在电子病历文本序列中, 头实体文本片段的位置跨度定义为P_{sub}, 则头实体掩码序列M_{sub}定义如下:

$$M_{\text{sub}}[i]=\begin{cases}1, i\in P_{\text{sub}}\\ 0, i\notin P_{\text{sub}}\end{cases} \tag{5}$$

头实体的特征表示h_{sub}由式 (6), 式 (7) 计算得到:

$$h'_{\text{sub}}=\frac{1}{end-start+1}\sum_{i=start}^{end}H_x\circ M_{\text{sub}} \tag{6}$$

$$h_{\text{sub}}=\tanh(W_{\text{sub}}h'_{\text{sub}}+b_{\text{sub}}) \tag{7}$$

其中, $start$ 和 end 是头实体在文本中的起始位置和结束位置, tanh 是激活函数, W_{sub}和b_{sub}是可训练的权重和偏置. 融合头实体特征表示与关系类型词的特征表示, 创建条件提示信息的方法如下:

$$h_{\text{ref}}=h_{\text{sub}}+h_{\text{rel}} \tag{8}$$

为了从电子病历文本中识别出与条件提示信息相关联的尾实体片段, 让条件提示信息与电子病历文本产生交互是非常必要的, 这个过程能够让模型区分电子病历文本序列中每一个中文字符 token 相对于条件提示信息的关联程度, 从而使得模型能够依据不同的条件提示信息识别不同的尾实体片段. 本文将条件提示信息的特征表示与电子病历文本序列的每一个字符的特征表示进行拼接, 并利用 Bi-LSTM 网络进行编码, 获得更高级的融合条件提示信息的文本特征表示$H_u=[u_1,u_2,\cdots,u_m]$, u_t是融合条件提示信息的字符特征表示, 其计算公式如式 (9)–式 (12):

$$v_t=[x_t;h_{\text{ref}}] \tag{9}$$

$$\overrightarrow{H_u}=\overrightarrow{LSTM}([v_1,v_2,\cdots,v_m])=[\overrightarrow{u_1},\overrightarrow{u_2},\cdots,\overrightarrow{u_m}] \tag{10}$$

$$\overleftarrow{H_u}=\overleftarrow{LSTM}([v_1,v_2,\cdots,v_m])=[\overleftarrow{u_1},\overleftarrow{u_2},\cdots,\overleftarrow{u_m}] \tag{11}$$

$$u_t=[\overrightarrow{u_t}\oplus\overleftarrow{u_t}] \tag{12}$$

3.3 解码输出层

在解码过程中充分考虑标签之间的关联性有利于

准确的识别实体的边界, 得到最优的标签序列. 本文利用条件随机场 (conditional random fields, CRF) 作为解码输出层[22], CRF 解码过程涉及一个转移矩阵 V 和一个状态序列 Z. 转移矩阵 V 用于学习标签之间的依赖关系, V_{ij} 表示第 i 个标签转移到第 j 个标签的得分. 状态序列$Z=[z_1,z_2,\cdots,z_m]$是 CRF 层的输入序列, 由上一步获得的融合条件提示信息的字符特征表示u_t计算得到:

$$z_t = W_u u_t \tag{13}$$

其中, z_t是第 t 个字对应于每一个标签的得分, W_u是可训练的模型权重参数.对于一个预测序列$Y=[y_1,y_2,\cdots,y_m]$, 它的解码得分公式定义如式 (14):

$$S(Z,Y)=\sum_{t=1}^{m}V_{y_{t-1},y_t}+\sum_{t=1}^{m}Z_{t,y_t} \tag{14}$$

从输入序列 Z 解码得到每一个可能的预测序列 Y 的条件概率计算如式 (15):

$$P(Y|Z)=\frac{\exp(S(Z,Y))}{\sum_{Y'\in Y_Z}\exp(S(Z,Y'))} \tag{15}$$

其中, Y_Z是全部可能从输入序列 Z 解码得到的预测序列的集合. 模型训练的目标是最大化正确标签序列的对数似然概率. 在解码过程中通过维特比算法得到分数最高的标签序列.

4 实验过程与结果评估

4.1 数据集

本文从某三甲医院获取了一批门诊病历数据, 数据的形式为中文电子病历, 主要内容包括主诉、现病史、既往史、体格检查、辅助检查、初步诊断等. 选择了其中的 2 000 篇进行关系三元组识别任务的研究. 在医学领域专家的指导下, 定义了实体和关系的标注规范, 并组织一批医学专业人员对中文电子病历中的实体和关系进行人工标注. 对于中文电子病历中的实体, 本文确定了 11 种实体类型, 并设计了一套实体类型优先级规则, 实体对中具有较高优先级的实体将作为头实体, 较低优先级的实体将作为尾实体. 依据优先级由高到低, 这 11 种实体类型分别是疾病、结构、观察、表现、检查、描述、方位、限定、治疗、药物、用法. 对于中文电子病历中的实体对的关系, 本文将头实体类型和尾实体类型进行拼接, 定义为该实体对的关系. 在本文的研究中, 共确定了 20 种粗粒度的关系类型, 部分关系三元组 schemas 如表 1 所示.

表 1 部分关系三元组 schemas

头实体类型	关系类型	尾实体类型
结构	结构表现	表现
	结构描述	描述
	结构观察	观察
观察	观察描述	描述
	观察表现	表现
	观察方位	方位
表现	表现方位	方位
	表现限定	限定
描述	描述方位	方位
药物	药物用法	用法
疾病	疾病描述	描述

本文设计了一个数据预处理算法, 依据中文电子病历文本以及对应的实体关系标注文件生成形如[关系类型, 文本片段, 头实体掩码序列, 标注序列]的数据样本, 具体方法如算法 1.

算法 1. 中文电子病历数据预处理算法

1) 利用滑动窗口方法, 将电子病历文本切分成若干个文本片段;
2) 对于每一个文本片段, 查找该片段中的实体对; 对于每一个实体对, 利用优先级规则区分头尾实体, 生成头实体掩码序列;
3) 检查该实体对能否形成有效的关系, 若有关系, 则创建标注序列, 其标注的内容是尾实体片段, 并生成数据样本;
4) 若该实体对无关系, 则将头实体的实体类型词和尾实体的实体类型词进行拼接, 检查拼接后的术语是否存在于预定义的关系集合中, 若存在, 则创建一个全为 O 的标注序列, 并生成数据样本, 若不存在, 则不生成数据样本;
5) 标注序列合并, 若生成的数据样本中, 某两条数据头实体、关系类型、文本片段均相同, 则将其标注序列中非 O 的部分合并.

4.2 实验设置

本文实验使用的训练数据、验证数据和测试数据按照 8:1:1 的比例进行划分. 字向量利用随机初始化方法生成 50 维的向量, Bi-LSTM 编码器的隐藏层的维度设置为 128, 批处理的大小设置为 64, 训练过程中的参数优化算法为 Adam, 学习率设置为 0.001, dropout 设置为 0.5 以防止过拟合.

在实验中使用精确率 (*Precision*), 召回率 (*Recall*), *F*1 值作为模型的评价指标.

$$Precision=\frac{TP}{TP+FP} \tag{16}$$

$$Recall=\frac{TP}{TP+FN} \tag{17}$$

$$F1=\frac{2\cdot Precision\cdot Recall}{Precision+Recall} \tag{18}$$

在实验过程中发现实体对没有关系的情况比较多,

为了实现样本均衡, 随机选择一部分无关系样本用于实验.

4.3 实验结果分析

为了验证本文提出的模型的性能, 本文进行了对比实验. 在之前的研究中关系抽取任务通常被视为关系分类任务, 因此通常将 CNN 和 RNN 作为基准模型. 将本文模型分别与 RNN 模型、结合最大池化的 CNN 模型 CNN-Max 进行比较. 在构建对比模型的实验数据时, 对每一条文本片段插入位置标记<e1>、</e1>、<e2>、</e2>指示两个实体在文本片段中的起始位置和结束位置. 对比模型的输出层使用一个全连接层将编码特征映射成具体的关系类别. 实验结果如表 2 所示.

表 2 模型实验结果对比 (%)

关系类型	RNN			CNN-Max			本文模型		
	精确率	召回率	*F*1值	精确率	召回率	*F*1值	精确率	召回率	*F*1值
观察描述	0.8792	0.8732	0.8762	0.9791	0.995	0.987	0.988	0.9957	0.9919
结构表现	0.8054	0.751	0.7773	0.9202	0.9932	0.9553	0.9844	0.9736	0.979
结构描述	0.5499	0.7187	0.6231	0.9525	0.9895	0.9706	0.9828	0.9915	0.9872
结构方位	0.3539	0.4598	0.4	0.9954	0.9972	0.9963	0.9882	1	0.9941
表现限定	0.5535	0.1102	0.1838	0.8844	0.95	0.916	0.8988	0.9277	0.913
结构观察	0.5286	0.3113	0.3918	0.9396	0.9979	0.9679	0.9205	0.9959	0.9567
结构结构	0.3125	0.034	0.0613	0.6217	0.9728	0.7586	0.5472	0.9062	0.6824
观察表现	0.4235	0.1188	0.1856	0.9379	0.8977	0.9174	0.9667	0.8878	0.9255
观察方位	0.45	0.135	0.2077	0.9749	0.97	0.9724	0.9817	0.9817	0.9817
检查检查	0.7321	0.6508	0.6891	0.9688	0.9841	0.9764	0.9	1	0.9474
药物用法	0.5	0.1429	0.2222	0.972	0.9456	0.9586	0.912	0.9569	0.9339
表现方位	0	0	0	0.92	0.8519	0.8846	0.5455	0.9231	0.6857
结构检查	1	0.0256	0.05	0.7907	0.8718	0.8293	0.8846	1	0.9388
检查描述	0	0	0	0.5	0.7273	0.5926	0.5789	1	0.7333
描述限定	0	0	0	0.8	0.6667	0.7273	0.7692	0.8333	0.8
疾病描述	0	0	0	1	0.7778	0.875	0.9048	0.7037	0.7917
表现表现	0	0	0	0.25	0.25	0.25	0.6667	0.25	0.3636
观察限定	0	0	0	1	0.7333	0.8462	0.9167	0.7333	0.8148
描述方位	0	0	0	0.9048	0.8261	0.8636	1	1	1
描述描述	0	0	0	0	0	0	1	1	1
平均	0.7943	0.8101	0.7907	0.9684	0.9652	0.9660	**0.9796**	**0.9765**	**0.9777**

从实验结果中可以发现, 本文提出的模型在精确率、召回率和 *F*1 值上分别达到 0.9796、0.9765 和 0.9777, 表现优于基准模型. 对比基础的 RNN 模型, 本文模型在精确率、召回率和 *F*1 值上分别提升了 18.53%、16.64% 和 18.7%. 对比 CNN-Max 模型, 本文模型在精确率、召回率和 *F*1 值上分别提升了 1.12%、1.13% 和 1.17%. 实验结果验证本文模型能有效的应用于识别中文电子病历中的医学关系三元组.

5 结论与展望

本文设计了一种基于条件提示与序列标注的中文电子病历关系三元组识别方法, 将关系抽取任务建模成从电子病历文本中识别与条件提示信息相关的三元组尾实体片段的序列标注任务, 其中条件提示信息为头实体和关系类型组成的先验知识. 本文的模型聚焦于构建条件提示信息与文本序列的关联, 过滤掉文本序列中与条件提示信息无关的实体概念. 在中文电子病历上的实验结果表明, 本文模型的精确率达到 97.96%, 召回率达到 97.65%, *F*1 值达到 97.77%, 表现优于基准模型, 实现了对中文电子病历中的医学关系三元组的识别.

在未来的研究工作中, 计划将当前工作延伸至更具挑战性的场景, 如医学文献中的实体关系抽取. 为获得更丰富的文本序列表示, 可考虑加入预训练语言模型. 对于模型识别无关实体的情况, 可以考虑引入句法依赖, 限制识别实体片段的结果空间.

参考文献

1 He B, Guan Y, Dai R. Convolutional gated recurrent units for medical relation classification. Proceedings of 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Madrid: IEEE, 2018. 646–650. [doi: 10.1109/BIBM.2018.8621228]

2 He B, Guan Y, Dai R. Classifying medical relations in clinical text via convolutional neural networks. Artificial Intelligence in Medicine, 2019, 93: 43–49. [doi: 10.1016/j.artmed.2018.05.001]
3 Raj D, Sahu S, Anand A. Learning local and global contexts using a convolutional recurrent network model for relation classification in biomedical text. Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). Vancouver: Association for Computational Linguistics, 2017. 311–321. [doi: 10.18653/v1/K17-1032]
4 Blaschke C, Valencia A. The frame-based module of the SUISEKI information extraction system. IEEE Intelligent Systems, 2002, 17(2): 14–20. [doi: 10.1109/MIS.2002.999215]
5 Segura-Bedmar I, Martínez P, de Pablo-Sánchez C. A linguistic rule-based approach to extract drug-drug interactions from pharmacological documents. BMC Bioinformatics, 2011, 12(S2): S1. [doi: 10.1186/1471-2105-12-S2-S1]
6 Uzuner Ö, South BR, Shen SY, *et al*. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association, 2011, 18(5): 552–556. [doi: 10.1136/amiajnl-2011-000203]
7 Alam F, Corazza A, Lavelli A, *et al*. A knowledge-poor approach to chemical-disease relation extraction. Database, 2016, 2016: baw071. [doi: 10.1093/database/baw071]
8 Zhai PJ, Huang X, Zhang BB, *et al*. Relation extraction based on fusion dependency parsing from Chinese EMRs. Scientific Programming, 2020, 2020: 8658040. [doi: 10.1155/2020/8658040]
9 Zeng DJ, Liu K, Lai SW, *et al*. Relation classification via convolutional deep neural network. Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Dublin: ACL, 2014. 2335–2344.
10 Zhang DX, Wang D. Relation classification via recurrent neural network. arXiv: 1508.01006, 2015.
11 Liu SY, Tang BZ, Chen QC, *et al*. Drug-drug interaction extraction via convolutional neural networks. Computational and Mathematical Methods in Medicine, 2016, 2016: 6918381. [doi: 10.1155/2016/6918381]
12 Tian B, Zhang Y, Liu KX, *et al*. Deep learning based information extraction framework on chinese electronic health records. Proceedings of the 30th International Conference on Software Engineering and Knowledge Engineering. Redwood City: Research Inc. and Knowledge Systems Institute Graduate School, 2018. 86–91. [doi: 10.18293/SEKE2018-040]
13 Zhang ZC, Zhou T, Zhang Y, *et al*. Attention-based deep residual learning network for entity relation extraction in Chinese EMRs. BMC Medical Informatics and Decision Making, 2019, 19(2): 55. [doi: 10.1186/s12911-019-0769-0]
14 Suárez-Paniagua V, Zavala RMR, Segura-Bedmar I, *et al*. A two-stage deep learning approach for extracting entities and relationships from medical texts. Journal of Biomedical Informatics, 2019, 99: 103285. [doi: 10.1016/j.jbi.2019.103285]
15 Zhang YJ, Zheng W, Lin HF, *et al*. Drug-drug interaction extraction via hierarchical RNNs on sequence and shortest dependency paths. Bioinformatics, 2018, 34(5): 828–835. [doi: 10.1093/bioinformatics/btx659]
16 Zhang S, Zheng DQ, Hu XC, *et al*. Bidirectional long short-term memory networks for relation classification. Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation. Shanghai: ACL, 2015. 73–78.
17 Zhang TX, Lin HF, Tadesse MM, *et al*. Chinese medical relation extraction based on multi-hop self-attention mechanism. International Journal of Machine Learning and Cybernetics, 2021, 12(2): 355–363. [doi: 10.1007/s13042-020-01176-6]
18 Li ZH, Yang ZH, Shen C, *et al*. Integrating shortest dependency path and sentence sequence into a deep learning framework for relation extraction in clinical text. BMC Medical Informatics and Decision Making, 2019, 19(1): 22. [doi: 10.1186/s12911-019-0736-9]
19 Li XY, Yin F, Sun ZJ, *et al*. Entity-relation extraction as multi-turn question answering. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 1340–1350. [doi: 10.18653/v1/p19-1129]
20 Wei ZP, Su JL, Wang Y, *et al*. A novel cascade binary tagging framework for relational triple extraction. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, 2020. 1476–1488 [doi: 10.18653/v1/2020.acl-main.136.]
21 Yang J, Liang SL, Zhang Y. Design challenges and misconceptions in neural sequence labeling. Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe: Association for Computational Linguistics, 2018. 3879–3889.
22 Lafferty JD, McCallum A, Pereira FCN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of 18th International Conference on Machine Learning. Williamstown: ACM, 2001. 282–289.

(校对责编: 孙君艳)