·医学数据科学·

基于 K-BERT 的中文妇产科电子病历实体识别研究

张 由,李 舫

(上海电力大学计算机科学与技术学院,上海 201306)

摘要:针对利用预训练模型进行中文妇产科电子病历命名实体识别时,BERT 缺乏一定的医疗领域专业知识而导致其识别性能下降的问题,提出了一种基于知识图谱的预训练模型——K-BERT 的命名实体识别模型 K-BERT-BiLSTM-CRF。通过 K-BERT 预训练模型获取包含医学背景知识的语义特征向量,利用双向长短时记忆网络(BiLSTM)与条件随机场(CRF)提取上下文相关特征并且解决标签偏移问题,完成实体识别。利用真实妇产科医疗电子病历数据集进行训练,K-BERT-BiLSTM-CRF 模型的 F_1 值达到了 90.04%。实验表明,相比一般 BERT 的模型,K-BERT-BiLSTM-CRF 命名实体识别模型在中文妇产科电子病历领域上的表现更优异,识别效果更好。

关键词:K-BERT;双向长短时记忆网络;条件随机场;妇产科电子病历;命名实体识别

中图分类号:TP391.1 **文献标识码**:A **DOI**:10.3969/j.issn.1006-1959.2024.01.012

文章编号:1006-1959(2024)01-0065-07

Research on Entity Recognition of Chinese Obstetrics and Gynecology Electronic Medical Records Based on K-BERT

ZHANG You,LI Fang

(College of Computer Science and Technology,Shanghai University of Electric Power,Shanghai 201306,China)

Abstract:When the pre-trained model is used to name entity recognition of Chinese obstetrics and gynecology electronic medical records, BERT lacks certain professional knowledge in the medical field, which leads to the decline of its recognition performance. A pre-trained model based on knowledge graph-K-BERT name entity recognition model K-BERT-BiLSTM-CRF is proposed. The K-BERT pre-training model is used to obtain the semantic feature vector containing the medical background knowledge, and the bidirectional long short-term memory network (BiLSTM) and conditional random field (CRF) are used to extract the context-related features and solve the label offset problem to complete the entity recognition. Using the real obstetrics and gynecology medical electronic medical record data set for training, the F_1 value of the K-BERT-BiLSTM-CRF model reached 90.04%. Experiments show that compared with the general BERT model, the K-BERT-BiLSTM-CRF name entity recognition model performs better in the field of Chinese obstetrics and gynecology electronic medical records, and the recognition effect is better.

Key words:K-BERT;Bidirectional long short-term memory;Conditional random fields;Obstetrics and gynecology electronic medical records;Name entity recognition

电子病历(electronic medical records)是医疗系统信息化的产物,是医务人员在其医疗活动过程中,使用医疗机构信息系统生成的关于病患的文字、符号、图标、图形、数据、影像等数字化信息,并能实现存储、管理、传输和重现的医疗记录[1]。随着医疗信息化的快速发展,电子病历已经成为医疗服务的重要基础设施之一。同时,智慧医疗已经成为医疗行业的新兴发展方向,电子病历作为智慧医疗的重要数据源之一,将对未来医疗的发展产生重要影响。在妇产科领域,电子病历的应用也日益广泛。由于妇产科疾病的特殊性和敏感性,电子病历可以更加全面、准确地记录妇产科患者的病历信息,对于提高医疗质量、优化医疗服务、保障女性健康具有重要意义。因此,研究妇产科电子病历对女性的意义已成为当前的热点问题,有望为妇产科领域的医疗服务提供更为优质的支持。命名实体识别(name entity recognition,NER)是指从非结构化文本中识别出具有特定意义的实体,如人名、地名、机构名等[2]。在电子病历中,由于病历信息的非结构化特点,传统的基于规则、关键词匹配或机器学习的方法无法满足实际需求。因此,采用深度学习等技术对电子病历进行命名实体识别已成为当前的研究热点之一。对于妇产科领域的电子病历数据进行 NER 任务的研究也愈发重要。妇产科领域具有许多特殊的术语和专业知识,对于实现对电子病历中妇产科实体的自动识别,需要在语料库的构建、特征的选择、模型的训练

作者简介:张由(1993.4-),男,上海人,硕士研究生,主要从事医学自然语言处理及机器学习研究

通讯作者:李舫(1974.4-),女,山西运城人,博士,讲师,主要从事图像分割、图像配准、点集配准及机器学习研究

等方面进行一系列的优化。同时,由于妇产科领域的疾病分类较为复杂,电子病历中的实体种类也更加繁多,对 NER 任务的难度和要求提高了不少。通过 NER 任务可以更加全面、准确地识别和记录妇产科患者的病历信息,为后续的医疗诊疗提供有力支持。

1 研究背景及现状

早期命名实体识别技术主要是基于规则、模板和特征工程等传统机器学习方法,其主要思路是通过手工设计规则、特征和模板等来提取实体特征并识别命名实体。这些方法依赖于领域专家的经验和先验知识,但是往往难以涵盖所有的实体类型和语境,因此在实际应用中往往表现不尽如人意。近年来由于深度学习发展迅速,越来越多的研究人员将相关技术运用到命名实体识别的研究上。这些方法可以自动学习语言特征,无需手工设计特征和规则,可以更好地适应不同的语境和实体类型,大大降低了人工成本。Lample G 等[3]提出双向长短期记忆网络(Bidirectional Long Short-Term Memory,BiLSTM)与条件随机场(Conditional Random Field,CRF)结合的神经网络模型,由于 BiLSTM 的双向结构能够获取上下文的序列信息,并且 CRF 可以处理标签间的依赖关系,因此在命名实体识别任务上得到了相当理想的效果。王若嘉等[4]将 BiLSTM-CRF 引入中文电子病历命名实体识别领域,标注数据集,建立 BiLSTM-CRF 模型对电子病历中症状、检查等 5 类命名实体进行识别,得到了 78.12%的值;李超凡等[5]将词嵌入与 BiLSTM-CRF 进行结合,对病历进行实体识别,进一步提高了模型性能,值达到了 90.8%。Lu NJ 等[6]使用上海一家三甲医院的 11 万条住院和门诊记录结合搜狗词典,将单词边界信息编码为模型输入特征,使用多级嵌入(字符级嵌入、词语级嵌入和字典特征级嵌入)作为 BiLSTM-CRF 输入,达到了 92.68%的值。

2018 年谷歌团队 Devlin J 等[7]所提出来的一种语言预处理模型 BERT(Bidirectional Encoder Representations from Transformer)来表征词向量。BERT 通过在海量的无标注数据上预训练语言模型,然后在命名实体识别任务上进行微调,可以在少量标注数据的情况下取得较好的效果,极大地降低了标注数据的需求。基于 BERT 研究和改进模型被广泛研究以及应用[8],例如谢腾等[9]设计 BERT-BiLSTM-CRF 模型用于中文实体识别,在 MSRA 语料和人民日报语料库上分别达到了 94.65%和 94.67%的值。在中文电子病历命名实体识别领域,Liu ML 等[10]利用网络爬取的数据,构造细粒度 BiLSTM-CRF 分层标签模型,结合包含拼音、字形信息的 BERT 模型,引入额外的标签信息和语义信息,提升了模型的性能,得到了 85.59%的值;张芳丛等[11]将 RoBERTa-WWM 中文预训练模型与 BiLSTM-CRF 结合,设计了 RoBERTa-WWM-BiLSTM-CRF 的中文电子病历命名实体识别模型,由于 RoBERTa 预训练模型对 BERT 进行了改进,使用了中文训练样本,并且使用中文全词遮掩技术,解决了词识别不全及一词多义的问题,提高了识别的准确率,值达到了 89.08%。

现有研究存在以下两方面问题:其一,BERT 模型一般通过大量开放语料库进行预训练,以获得通用的语言表示形式。但是由于开放语料库的专业知识不够充分,导致这些 BERT 模型在垂直领域表现不佳。在处理电子病历命名实体识别任务时,经过维基百科预训练的 Google BERT 表现不佳。如果使用垂直领域的文本直接进行预训练,由于 BERT 通常含有 110 M 以上的参数,训练一个垂直领域的 BERT 需要大量的计算资源以及时间。其二,妇女由于其特殊的生理及病理特点,受情绪以及环境因素影响较男性更为显著,对于隐私保护的要求也更高[12]。近年来移动互联网和智能手机的高速发展,开发妇产科线上问诊平台、智能预问诊、智能分诊导诊等系统对于照顾妇女患者情绪以及保护隐私有着积极的作用。但关于电子病历实体识别的研究采用的数据集大多为全科数据集,暂无专门针对妇产科的研究,这导致了许多妇产科相关应用只能利用全科数据集训练的模型,模型性能往往不够理想,实际应用效果欠佳,对于妇女患者会产生一定的困扰。

基于上述两个问题,结合近期 Liu W 等[13]提出的基于知识图谱的预训练模型 K-BERT,本文提出了一种 K-BERT-BiLSTM-CRF 命名实体识别模型,具体步骤如下:①采用某三甲医院的 300 份真实妇产科电子病历作为数据集,并且进行数据预处理(脱敏、标注等);②将垂直领域知识图谱三元组数据集与已预处理妇产科电子病历已标注数据注入 BERT 进行训练,得到具有领域知识的预训练模型 K-BERT,并获得其特征向量;③将得到的特征向量输入 BiLSTM-CRF 网络,利用 BiLSTM 以及 CRF 获得语

料的上下文序列信息以及纠正错误标签顺序，获得最优标签序列，完成妇产科电子病历文本中的实体识别。

2 研究方法

2.1 模型框架 模型整体结构如图 1 所示。模型整体分为 3 层，分别是 K-BERT 层、BiLSTM 层以及 CRF 层。

2.2 K-BERT K-BERT 是融合知识图谱的语言训练模型，如图 1 所示，模型由知识层、嵌入层、可见层和掩码转换器组成。处理步骤如下：

①将输入的妇产科电子病历文本语句表示为 $S=\{w_0,w_1,\cdots,w_n\}$，其中 w_i 为中文单个字符。

②将 S 输入 K-BERT 模型中，其知识层会自动识别知识图谱中相关的实体，并将 S 扩充成带有实体关系的三元组形式 $w_i,\{r_k,w_j\}$，其中 w_i,w_j,r_k 为医学实体，为实体间关系。这样 S 会变成一个包含实体以及实体间关系的句子树。

以输入“卵巢囊肿会引起下腹疼痛”为例，知识层会以知识图谱为标准，将句子中的“卵巢囊肿”和“下腹疼痛”抽取出来，并且将实体扩充成三元组{卵巢囊肿，belongs_to，妇科}、{“卵巢囊肿”，do_eat，“海参”}、{“下腹疼痛”，has_symptom，“腹泻”}，然后将这 3 个三元组注入原句子中形成句子树，见图 2。

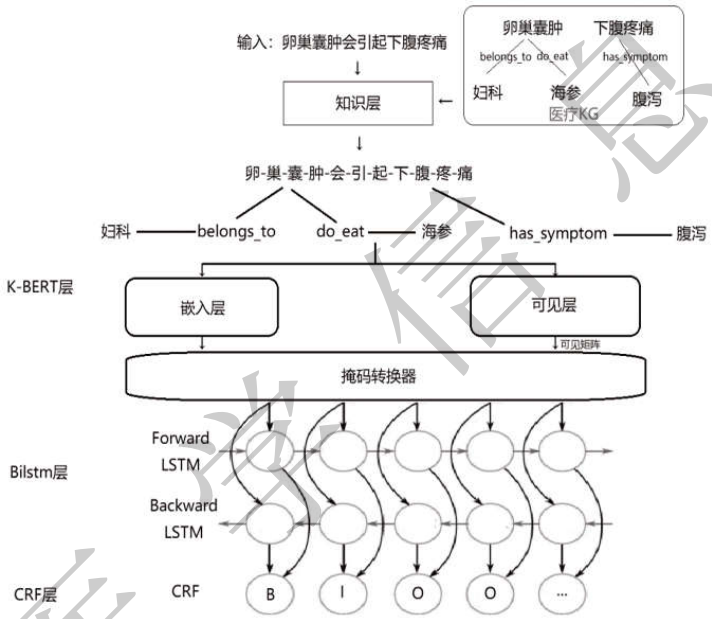

图 1 K-BERT-BiLSTM-CRF

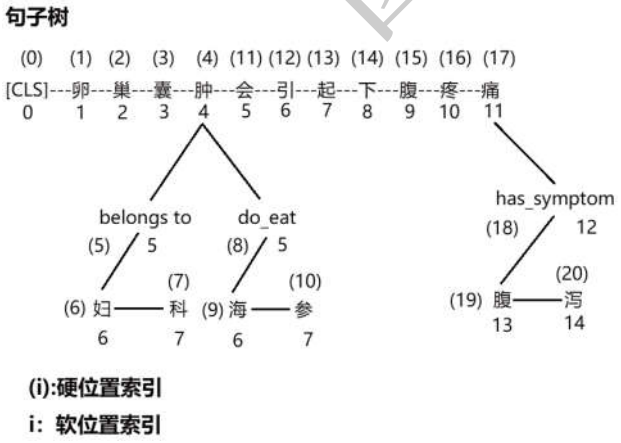

图 2 句子树结构

③将句子树输入嵌入层和可见层。嵌入层主要的作用为给句子树添加软索引位置，见图 2，之后将其铺平。目的是为了将句子树铺平之后仍然不丢失原来的位置信息，以得到正确的序列。每个树干上都标有数字，表达的即是软位置索引，如卵 1 巢 2 囊 3 肿 4 do_eat5 海 6 参 7。如此标记之后，每根树干上表达的均为正确的位置信息。

可见层通过生成一个可见矩阵 M，来限制词与词之间的关系。M 定义如式(1)：

$$M_{ij}=\begin{cases}0, & w_i,w_j \text{ 相互可见}\\ -\infty, & w_i,w_j \text{ 相互不可见}\end{cases}\#(1)\#$$

其中，相互可见的取值为 0，互不可见的取值为，i 与 j 均为硬位置索引。

④将铺平后的句子树以及可见矩阵输入到掩码转换器中。掩码转换器由 12 层掩码自注意力模块堆叠而成，其作用为确保一个词只和同一个树干的上下文有关系。Mask-Self-Attention 的定义如式(2)~式(4)：

$$Q^{i+1},K^{i+1},V^{i+1}=h^iW_q,h^iW_k,h^iW_v\#(2)\#$$

$$S^{i+1}=\mathrm{softmax}\left(\frac{Q^{i+1}K^{i+1}+M}{\sqrt{d_k}}\right)\#(3)$$

$$h^{i+1}=S^{i+1}V^{i+1}\#(4)$$

其中，W_q，W_k，W_v 是模型需要学习的矩阵向量参数；h^i 是隐状态的第 i 个 Mask-Self-Attention 块；d_k 是缩放因子，用于控制训练过程中的梯度稳定性；M 为可见矩阵。如果两个字在同一树干上，则 M_{ij} 的值为 0，之后按照 softmax 进行打分计算；若两个字不在同一树干上，则 S^{i+1} 的得分为 0，M_{ij} 的值为负无穷，也就意味着这两个字相互不可见。

如图 2 所示，如果不加以处理直接输入 BERT，模型会误认为“下腹疼痛”是在“海参”之后，甚至会理解为“海参”会引发“下腹疼痛”，这会对模型性能造成很大的影响。

2.3 BiLSTM 与 CRF 长短期记忆网络（LSTM）是 RNN（循环神经网络）的一种变体，解决了 RNN 训练过程中梯度爆炸或梯度消失的问题，使网络能够实现长期记忆，并且捕捉上下文信息，其核心结构为遗忘门、输入门、输出门及记忆单元[14]，其结构用公式表达为：

$$f_t=\sigma(W_f\times[h_{t-1},x_t]+b_t)\#(5)$$

$$i_t=\sigma(W_i\times[h_{t-1},x_t]+b_i)\#(6)$$

$$\widetilde{C}_t=\tanh(W_c\times[h_{t-1},x_t]+b_c)\#(7)$$

$$C_T=f_t\times C_{t-1}+i_t\times C_t\#(8)$$

$$o_t=\sigma(W_o\times[h_{t-1},x_t]+b_o)\#(9)$$

$$h_t=o_t\times\tanh(C_t)\#(10)$$

其中，σ 为激活函数，W 为权重矩阵，h_{t-1} 为上一时刻的输出，x_t 为当前输出，b 为偏置向量，f_t，i_t，o_t 分别为遗忘门、输入门和输出门，h_t 则为 t 时刻的输出。

LSTM 模型不能充分处理妇产科电子病历上下文信息[15]，而 Graves A 等[16]提出的 BiLSTM 模型，对每个输入都分别进行前向 LSTM 与后向 LSTM，然后将同一时刻的两个输出进行合并，这样每个时刻都对应着前向和后向的信息。因此，利用 BiLSTM 模型进行妇产科中文电子病历的特征提取，将会得到完整的上下文信息。BiLSTM 模型由前向 LSTM 和后向 LSTM 组成，其输出如公式（11）所示[17]。

$$h_t=<\overrightarrow{h_t},\overleftarrow{h_t}>\#(11)$$

其中，$\overrightarrow{h_t}$ 为前向特征表示，$\overleftarrow{h_t}$ 为后向特征表示，h_t 为电子病历文本的特征。

BiLSTM 在处理远距离上下文关系方面表现较好，但其却无法处理标签之间的顺序错位问题，例如 I 在 B 前面，B 和 I 中间有 O 等情况。而 CRF 可以通过邻近标签的依赖关系，获得一个最优的标签序列，正好可以弥补 BiLSTM 这一问题。

CRF 可以通过分数函数求出每个序列的得分，对每个分数进行归一化处理，最后采用维特比算法求出最优标签序列[18]。其公式为：

$$S(X,Y)=\sum_{i=0}^{n}A_{yi,yi+1}+\sum_{i=1}^{n}P_{i,yi}\#(12)$$

$$P(Y\mid X)=\frac{e^{s(X,Y)}}{\sum_{Y\in Y_x}S(X,Y)}\#(13)$$

$$Y^*=\mathrm{argmaxs}(X,\widetilde{Y})\#(14)$$

$$\mathrm{s.t.}\widetilde{Y}\in Y_x$$

其中，S 为得分函数，P 为第 i 个元素输出为 y_i 的概率，$A_{yi,yi+1}$ 为 y_i 到 y_{i+1} 的转移概率，$\widetilde{Y}$ 表示真实的标注序列；Y_x 表示所有出现的标注序列，Y^* 为得分最高的序列，即为最优标签序列。

3 实验

3.1 数据预处理 实验数据主要采用某三甲医院的 300 份真实妇产科电子病历数据集，其中包含入院记录、出院小结等文本。数据预处理过程包括：

3.1.1 电子病历脱敏处理 电子病历中记录着患者的姓名、地址、病史等隐私信息，为了保护患者隐私以及降低无关信息对实体识别效果的干扰，在不改变电子病历文本语义表达的前提下，对电子病历的内容进行脱敏处理，得到脱敏的数据集。

3.1.2 人工序列标注 由于电子病历为非结构化数据，因此需要对其进行人工实体标注。以相关实体为对象，对疾病和诊断、检查检验、症状、手术、药物、身体部位 6 类实体进行标注。标注策略选取 BIO，例如实体类别为“疾病”，将该实体的开头标记为“B-DSE”，此实体词的中间字符与结尾字符标记为“I-DSE”，非实体词的其他字符标记为 O，所以该数据集文本对应的标签分别是 B-DSE、I-DSE、B-LAB、I-LAB、B-SYM、I-SYM、B-OPS、I-OPS、B-DRG、I-DRG、B-PAT、I-PAT、O 这 13 类（表 1）。各类预定义类别及其含义信息如下：①疾病和诊断，如卵巢癌、子宫肌瘤等；②检查检验，如 CT、大畸形筛查、生化常规等；③症状，如下腹部疼痛、阴道瘙痒

等;④手术,如肿瘤减积术、剖宫产等;⑤药物,如顺铂、头孢等;⑥身体部位,如宫颈、腹腔等。指定专业医学团队按照上述规则分别对 300 份数据进行标注,并且每份数据都进行两次以上的标注,最大程度确保标注的准确率。

3.1.3 数据分组 将人工序列标注好的数据按照 4:1 的比例将数据集随机划分为训练集和测试集。训练集和测试集中每个预定义类别的实体个数见表 2。

3.2 实验环境 使用 kaggle 平台(www.kaggle.com)进行模型训练以及测试,框架基于 Pytorch[19],具体实验环境设置见表 3。

3.3 实验参数设置 经过多次实验之后,选取各个模型效果最优的超参数配置,见表 4。

表 1 命名实体分类

医疗实体类别	序列标注	标注示例	医疗实体类别	序列标注	标注示例
疾病与诊断	DSE	卵 B-DSE	症状	SYM	下 B-SYM
		巢 I-DSE			腹 I-SYM
		癌 I-DSE			部 I-SYM
检查检验	LAB	大 B-INF			疼 I-SYM
		排 I-INF			痛 I-SYM
		畸 I-INF	药物	DRG	顺 B-DRG
手术	OPS	剖 B-OPS			铂 I-DRG
		宫 I-OPS	身体部位	PAT	腹 B-PAT
		产 I-OPS			腔 I-PAT
非医疗实体	O	无 O			

表 2 数据集中各类实体数量统计

数据集	疾病与诊断	检查检验	症状	手术	药物	身体部位	总计
训练集	1717	375	1263	829	2891	147	7222
测试集	430	91	316	204	718	36	1795
总计	2147	466	1579	1033	3609	183	9017

表 3 实验环境设置

项目	环境
操作系统	Ubuntu 20.04.4 LTS
CPU	Intel(R) Xeon(R) CPU @ 2.20GHz
GPU	Telsa P100 16GB
内存	16GB
Python 版本	3.7.12
Pytorch 版本	1.12.0+cu116

表 4 模型参数设置

超参数	值
Dropout	0.5
Epoch	50
Batch_size	64
LSTM 隐藏层维度	768
序列最大长度	128
学习率	0.0001

3.4 模型评估 使用命名实体识别领域最常用的评价指标:精确率 Precision、召回率 Recall 和值。其中,精确率 P 指在所有被预测为正的样本中,实际为正的样本概率;召回率 R 指在实际为正的样本中被预测为正的样本概率;F_1 值为召回率和精确率的调和平均值,如公式(15)-公式(17)所示。

$$P=\frac{TP}{TP+FP}\#(15)$$

$$R=\frac{TP}{TP+FN}\#(16)$$

$$F_1=\frac{2PR}{P+R}\#(17)$$

其中,TP 是正确识别的实体的数量,FP 是将其他文本错误识别为实体的数量,FN 是文本中未能识别为实体的实体词数量。

3.5 实验结果 为了评估 K-BERT-BiLSTM-CRF 的

有效性，采用 BiLSTM-CRF、BERT-BiLSTM-CRF、BERT-IDCNN-CRF、RoBERTa-wwm-BiLSTM-CRF 作为对比模型，其在实体识别方面都有广泛的应用。①BiLSTM-CRF 模型：这是 NER 领域的经典模型；在训练时采用静态词向量；②BERT-BiLSTM-CRF 模型：在 BERT 模型基础上，将 BERT 与 BiLSTM-CRF 模型结合，在 NER 任务上实现了更好的效果；③BERT-IDCNN-CRF 模型：在 BERT 模型基础上，将 IDCNN 代替常规的 BiLSTM；④RoBERTa-wwm-BiLSTM-CRF 模型：RoBERTa-wwm 模型将 BERT 字符级掩码替换为词级掩码，可进一步提升实体识别能力。在相同数据集、相同超参数的情况下对这 5个模型分别进行训练，得到了各模型的各个实体以及总体实验结果，见表 5。

表 5 实验结果(%)

模型	实体类型	P	R	F_1	模型	实体类型	P	R	F_1
BiLSTM-CRF	疾病与诊断	84.77	82.14	83.43	BERT-IDCNN-CRF	药物	87.19	87.64	87.41
	检查检验	85.71	83.05	84.36		身体部位	74.61	75.70	75.15
	症状	81.91	75.56	78.61		综合	86.48	84.41	85.43
	手术	77.57	74.02	75.75	RoBERTa-wwm-	疾病与诊断	92.94	87.11	89.93
	药物	80.86	78.91	79.87	BiLSTM-CRF	检查检验	90.20	90.07	90.13
	身体部位	71.11	74.02	72.54		症状	87.43	82.31	84.79
	综合	82.58	80.24	81.39		手术	84.53	80.48	82.46
BERT-BiLSTM-CRF	疾病与诊断	88.50	80.86	84.51		药物	88.61	85.91	87.24
	检查检验	90.48	85.00	87.65		身体部位	81.38	82.02	81.70
	症状	89.58	81.90	85.57		综合	89.79	88.46	89.12
	手术	75.55	75.1	75.32	K-BERT-BiLSTM-CRF	疾病与诊断	94.58	94.63	94.6
	药物	83.94	80.05	81.95		检查检验	91.43	91.21	91.32
	身体部位	78.91	77.53	78.21		症状	89.88	85.47	87.62
	综合	83.41	83.12	83.26		手术	87.38	87.41	87.39
BERT-IDCNN-CRF	疾病与诊断	77.16	75.33	76.23		药物	91.32	87.22	89.22
	检查检验	91.43	87.88	89.62		身体部位	86.07	85.2	85.63
	症状	85.99	81.38	83.62		综合	90.34	89.75	90.04
	手术	87.22	80.48	83.71					

在所有模型中，“手术”以及“身体部位“识别率相比另外 4 类医疗实体值较低。其主要原因为：“手术”实体普遍长度较长，嵌套内容较多，例如“人工智能辅助技术腹腔镜下全子宫+双侧附件切除术+盆底重建术（阴道髂耻韧带固定术/阴道骶前固定术/腹壁悬吊术）”，该实体中又包含着身体部位，因此在预测过程中会产生实体边界预测错误的现象，从而导致实体识别错误；“身体部位”F_1 较低的原因主要在于该实体数量在整个数据集中占比较低，采集到的训练数据不够均衡影响了部分类别识别的准确性，这一点可以从表 2 中看出。

从表 5 中可以看出，本文提出的 K-BERT-BiLSTM-CRF 识别模型取得了较好的效果，相比其他 4种 BiLSTM-CRF 模型，准确率、召回率以及 F_1 值均有不同程度的提高。该模型在 6 类医疗实体上的 F_1 值均是最高的。这是因为 K-BERT 将医疗知识图谱内容融入 BERT 进行预训练，充分利用了知识图谱在垂直领域的优势，进一步优化了文本的语义特征表示，增强了语义理解，因此取得了较好的电子病历命名实体识别效果。

综上所述，本研究提出的 K-BERT-BiLSTM-CRF 实体识别模型对真实妇产科电子病历的识别正确率最高，可以被运用于妇产科相关领域中。

4 总结

对于中文妇产科电子病历文本数据，本文提出了一种基于 K-BERT，结合 BiLSTM-CRF 的实体识别模型 K-BERT-BiLSTM-CRF，其中 K-BERT 预训练模型结合了医学垂直领域知识图谱的内容，可以更准确地表示医疗相关内容的上下文语义，BiLSTM 和 CRF 可以进一步提取上下文相关特征并且解决

标签偏移问题。实验结果表明,K-BERT-BiLSTM-CRF 训练结果的精确率,召回率以及 F_1 值均高于现有命名实体识别模型,取得了更好的识别效果,在妇产科电子病历文本命名实体识别任务上具有一定的优势。

参考文献:

[1]卫生部.电子病历基本规范(试行)[J].中国卫生质量管理,2010,17(4):22-23.

[2]Gandhi H,Attar V.Extracting Aspect Terms using CRF and Bi-LSTM Models [J].Procedia Computer Science,2020,167(1):2486-2495.

[3]Lample G,Ballesteros M,Subramanian S,et al.Neural architectures for named entity recognition [EB/OL].(2016-03)[2023-03-10].https://www.researchgate.net/publication/305334469_Neural_Architectures_for_Named_Entity_Recognition.

[4]王若嘉,魏思毅,王纪民.BiLSTM-CRF 模型在中文电子病历命名实体识别中的应用研究[J].文学与数据,2019,1(2):53-66.

[5]李超凡,马凯.基于词嵌入和 BiLSTM-CRF 模型的医疗记录实体识别方法[J].中国数字医学,2022,17(4):32-36.

[6]Lu NJ,Zheng J,Wu W,et al.Chinese clinical named entity recognition with word-level information incorporating dictionaries [C]//2019 International Joint Conferenceon Neural Networks.Budapest:IEEE,2019:1-8.

[7]Devlin J,Chang MW,Lee K,et al.BERT:Pre-training of deep bidirectional transformers for language understanding [EB/OL].(2019-05-24)[2023-03-10].https://arxiv.org/pdf/1810.04805.pdf.

[8]Yang B,Li D,Yang N,et al.Intelligent judicial research based on BERT sentence embedding and multi-level attention CNNs [EB/OL]. (2019 -09 -21) [2023 -03 -10].https://kns.cnki.net/KCMS/detail/detail.aspx?dbcode=IPFD&filename=JKDZ201909002038.

[9]谢腾,杨军安,刘辉.基于 BERT-BiLSTM-CRF 模型的中文实体识别[J].计算机系统应用,2020,29(7):48-55.

[10]Liu ML,Zhou XS,Cao Z,et al.Team MSIIP atCCKS 2019 Task 1 [C]//2019 China Conference on Knowledge Graph and Semantic Computing. Hangzhou: ChineseInformation Processing Society of China,2019:1-11.

[11]张芳丛,秦秋莉,姜勇,等.基于 RoBERTa-WWM-BiLSTM-CRF 的中文电子病历命名实体识别研究[J].数据分析与知识发现,2022,6(2):251-262.

[12]Bekaert S,Van Hecke A,Remmen R,et al.Women´s privacy and confidentiality concerns when consulting with health care providers about a sexually transmitted infection[J].J Obstet Gynecol Neonatal Nurs,2018,47(4):512-520.

[13]Liu W,Zhou P,Zhao Z,et al.K-BERT: Enabling Language Representation with Knowledge Graph [J].Proceedings of the AAAI Conference on Artificial Intelligence,2020,34 (3):2901-2908.

[14]Ma X,Tao Z,Wang Y,et al.Long short-term memory neural network for traffic speed prediction using remote microwave sensor data [J].Transp Res Part C Emerg Technol,2015,54:187-197.

[15]Fukada T,Schuster M,Sagisaka Y.Phoneme boundary estimation using bidirectional recurrent neural networks and its applications[J].Syst Comput Jpn,1999,30(4):20-30.

[16]Graves A,Schmidhuber J.Framewise phoneme classification withbidirectional LSTM and otherneural network architectures [J].Neural Network,2005,18(5-6):602-610.

[17]Cornegruta S,Bakewell R,Withey S,et al.Modelling radiological language with bidirectional long short-term memory networks [EB/OL].(2016-09-27)[2023-03-10].https://arxiv.org/pdf/1609.08409.pdf.

[18]Zweig G,Nguyen P,van Compernolle D,et al.Speech Recognition with Segmental Conditional Random Fields: A Summary of the JHU CLSP 2010 Summer Workshop[C]//Proceedings of 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).2011:5044-5047.

[19]Paszke A,Gross S,Chintala S,et al.Automatic differentiation in PyTorch[EB/OL].(2017-10-28)[2023-03-10].https://openreview.net/pdf?id=BJJsrmfCZ.

收稿日期:2023-03-13;修回日期:2023-03-28
编辑/肖婷婷