

文章编号: 2095-2163(2021)08-0097-06

中图分类号: TP391.1

文献标志码: A

基于 XLNet-BiLSTM 的中文电子病历命名实体识别方法

沈宙锋¹, 苏前敏¹, 郭晶磊²

(1 上海工程技术大学 电子电气工程学院, 上海 201620; 2 上海中医药大学 基础医学院, 上海 201203)

摘要: 中文临床电子病历命名实体识别是实现智慧医疗的基本任务之一。本文针对传统的词向量模型文本语义表示不充分,以及循环神经网络(RNN)模型无法解决长时间依赖等问题,提出一个基于 XLNet 的中文临床电子病历命名实体识别模型 XLNet-BiLSTM-MHA-CRF,将 XLNet 预训练语言模型作为嵌入层,对病历文本进行向量化表示,解决一词多义等问题;利用双向长短时记忆网络(BiLSTM)门控制单元获取句子的前向和后向语义特征信息,将特征序列输入到多头注意力层(multi-head attention, MHA);利用 MHA 获得特征序列不同子空间表示的信息,增强上下文语义的关联性,同时剔除噪声;最后输入条件随机场 CRF 识别全局最优序列。实验结果表明, XLNet-BiLSTM-Attention-CRF 模型在 CCKS-2017 命名实体识别数据集上取得了良好的效果。

关键词: 电子病历; 命名实体识别; XLNet; 多头注意力

Named entity recognition model of Chinese clinical electronic medical record based on XLNet-BiLSTM

SHEN Zhoufeng¹, SU Qianmin¹, GUO Jinglei²

(1 School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China;

2 School of Basic Medical Sciences Shanghai University of Chinese Medicine, Shanghai 201203, China)

[Abstract] The recognition of named entities in Chinese clinical electronic medical records is one of the basic tasks to realize smart medical care. Aiming at the insufficient text semantic representation of the traditional word vector model and the inability of the recurrent neural network (RNN) model to solve the problems of long-term dependence, a Chinese clinical electronic medical record named entity recognition model XLNet-BiLSTM-MHA-CRF based on XLNet is proposed. Use the XLNet pre-training language model as the embedding layer to vectorize the medical record text to solve the problem of ambiguity; use the bidirectional long and short-term memory network (BiLSTM) gate control unit to obtain the forward and backward semantic feature information of the sentence; Then input the feature sequence to the multi-head attention layer (multi-head attention, MHA), use MHA to obtain information represented by different subspaces of the feature sequence, enhance the relevance of context semantics and eliminate noise; finally, input the conditional random field CRF to identify the global maximum sequence. The experimental results show that the XLNet-BiLSTM-Attention-CRF model has achieved good results on the CCKS-2017 named entity recognition data set.

[Key words] electronic medical record; named entity recognition; XLNet; multi-headed attention

0 引言

电子病历(EMR, Electronic Medical Record)是计算机信息系统存储、管理和传输的医疗记录,包含医务人员在为患者诊疗过程中记录的关于患者病史、临床表现、治疗方法等数字化信息^[1]。由于电子病历大多是半结构化和非结构化的状态,对其进行分析和数据挖掘受到严重制约。命名实体识别(named entity recognition, NER)是发现和识别自然文本中的专有名词和有意义的词语,并将其归类到预定义的类别中,是自然语言处理(natural

language processing, NLP)任务中的一项重要分支^[2]。运用命名实体识别技术对电子病历文本进行分析研究,目的是自动地识别并且分类电子病历中的医疗命名实体。

传统的电子病历命名实体识别研究主要分为基于规则和基于机器学习两种方法,基于规则的方法主要依靠领域专家构建的领域词典进行识别,对于词典中没有出现的实体通过手工编辑的规则来识别医疗命名实体^[3]。由于词典构建和规则制定对领域专家的依赖性,基于机器学习电子病历命名实体识别的方法被广泛运用。近年来深度学习在语音识

基金项目: 上海市 2017 年度“科技创新行动计划”基金资助项目(17401970900)。

作者简介: 沈宙锋(1995-),男,硕士研究生,主要研究方向:自然语言处理、数据挖掘;苏前敏(1974-),男,博士,副教授,主要研究方向:医学数据挖掘、智能信息处理;郭晶磊(1975-),女,博士,副教授,主要研究方向:中医方剂文献研究。

收稿日期: 2021-02-06

别、图像识别和视频分析等多个领域取得了重大进展,大量研究人员将深度学习运用到电子病历实体识别领域,通过在大规模的标注数据中训练与学习,可以更好地抽取上下文语义特征进行表示^[4]。

基于深度神经网络的命名实体识别方法,都需要通过词嵌入方法将文本信息转换成序列化向量,目前比较流行的词嵌入方法是2013年由Mikolov等提出的Word2Vec,将传统的词的one-hot表示转换为低维、稠密的向量,每个词都由数十或数百个维度的实值向量表示^[5]。但是Word2vec训练的词向量是静态的,即同样的字在不同的语句中向量表示都是不变的,也就无法获取相同词汇的多种含义,且不能在训练过程随上下文来消除词义的歧义^[6]。电子病历中通常存在一词多义现象,如“疾”这个字在不同的词语中意义不同,既可以是名词疾病,也可以是形容词剧烈;近年来,针对以上问题,学术界提出了许多与上下文有关的词嵌入表示方法,比如ELMO(embeddings from language models)方法和OpenAI-GPT(generative pre-training)方法^[7]。但是,上述两种语言模型的语言表示都是单向的,无法同时获取前后两个方向电子病历文本的语义信息。

为了解决上述问题,本研究拟采用双向自回归预训练语言模型XLNet引入电子病历NER任务中,提出了XLNet-BiLSTM-MHA-CRF命名实体识别模型,并利用该模型对医疗电子病历中预定义的疾病、症状、治疗、检查、身体部位5类实体进行命名实体识别。实验证明,使用预训练语言模型构建词嵌入,并在BiLSTM-CRF中加入多头注意力机制,多角度的提取文本特征,有效提高了命名实体识别的效果。本文所述算法在ccks2017命名实体识别任务中取得F1值为91.74%。

1 XLNet-BiLSTM-Attention-CRF命名实体识别模型

XLNet-BiLSTM-MHA-CRF命名实体识别模型的整体结构如图1所示。模型第一层为XLNet词嵌入层,通过XLNet预训练语言模型,运用低维的字向量对病历中的每一个字进行表示,得到序列化文本输入;第二层是BiLSTM层,利用双向长短时记忆神经网络自动提取句子的前向特征和后向特征进行拼接输入下一层;第三层是MHA层,通过计算多角度的注意力概率获得句子的长距离依赖特征,得到新的特征向量;第四层是CRF层,通过计算对输入的文本特征进行序列标注,输出最优标签。

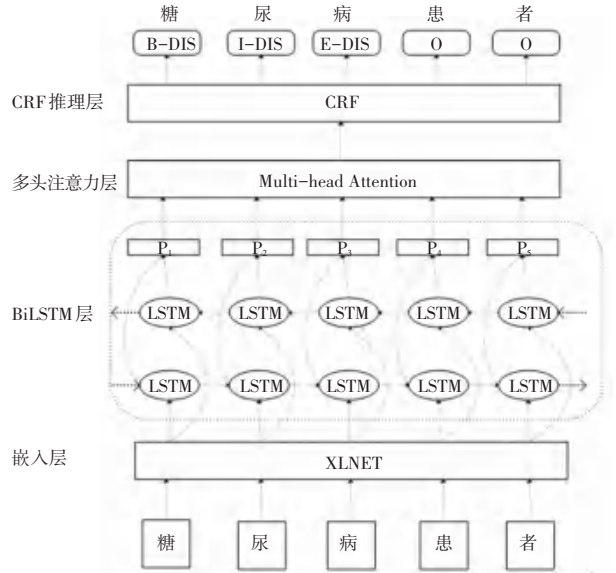


图1 ALBERT-BiGRU-MHA-CRF命名实体识别模型结构图
Fig. 1 The structure diagram of ALBERT-BiGRU-MHA-CRF named entity recognition model

1.1 XLNet预训练语言模型

XLNet模型是CMU与谷歌团队在2019年提出的一个基于Bert优缺点的广义自回归预训练方法,在传统的自回归语言模型上实现了双向预测^[8]。通过在transformer模块内部使用attention mask方法得到输入文本不同的排列组合,让模型充分提取上下文信息进行训练,克服了Bert模型在Mask机制下的有效信息缺失。XLNet的掩码机制示例如图2所示,当模型输入句子为[糖,尿,病,患,者],随机生成的一组序列为[病,糖,患,者,尿],那么在计算重排列后的“糖”字来说就可以利用到“病”字的信息,所以在第一行只保留了第三个位置的信息(用实心表示),其他的位置的信息被遮掉(用空心表示)。再比如重排后的“尿”字位于最后一个位置,其余四字的信息都可以利用,即第二行除第二个位置外全部以实心表示。

现有的预训练语言模型大多数采用transformer架构,但是在捕捉长距离依赖关系上还存在不足^[9]。为了解决这样的问题,XLNet采用引入循环机制(RNN)和相对位置编码的transformer-xl架构。通过RNN提取上一片段隐状态长距离依赖信息,存储在片段之间的memory单元,供下一片段的预测使用,充分捕捉长距离文本特征。片段之间的信息传递方式如图3所示,虚线框中表示前一片段提取的记忆信息,通过memory单元传递给下一片段,实现了信息的传递。

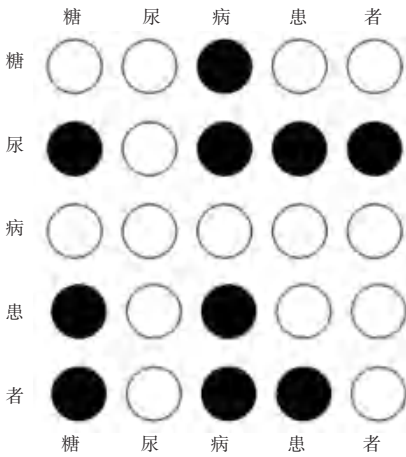


图 2 XLNet 模型掩码机制示例图

Fig. 2 XLNet model mask mechanism example diagram

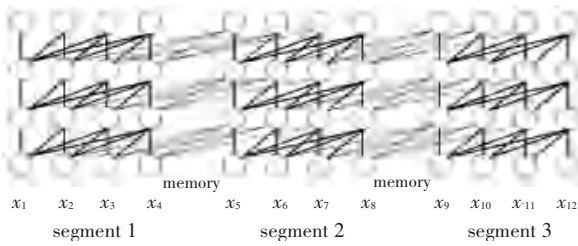


图 3 XLNet 循环机制片段信息传递图

Fig. 3 XLNet cycle mechanism fragment information transfer diagram

在位置编码方面,采用相对位置编码替代绝对位置编码,解决词的多义性问题,增强文本特征提取的完整性,加入相对位置编码后的 self-attention 公式(1)如下:

$$A_{i,j}^{rel} = E_{x_i}^T W_q^T W_{k,E} E_{x_j} + E_{x_i}^T W_q^T W_{k,R} R_{i-j} + u^T W_{k,E} E_{x_j} + v^T W_{k,R} R_{i-j} \quad (1)$$

其中, E_{x_i}, E_{x_j} 分别表示 i, j 的文本向量; W 代表权重矩阵; R_{i-j} 代表 i, j 的相对位置; u^T, v^T 是需要学习的参数; $W_{k,E}, W_{k,R}$ 分别为学习到基于内容的 **key** 向量和基于位置的 **key** 向量。

基于 transformer-xl 的 XLNet 预训练语言模型,通过 attention mask、循环机制和相对位置的编码,克服了自回归语言模型单项传递信息的不足,充分利用上下文的语义信息提取潜在的内部关系,训练出特征更加完整的词向量表示。

1.2 双向长短时记忆网络 (BiLSTM) 模型

针对传统的循环神经网络 (RNN) 在处理序列标注问题时出现梯度消失和梯度爆炸的现象, Hochreiter 和 Schmidhuber 在 1997 年提出长短时记忆网络 (long short term memory, LSTM)^[10], 该网络是在 RNN 基础上的改进,其单元结构如图 4 所示。通过设置遗忘门、输入门和输出门 3 种门限机

制看,选择性的处理信息的遗忘和传递,以此来捕获文本序列长距离依赖信息,有效解决了梯度消失的问题^[11]。

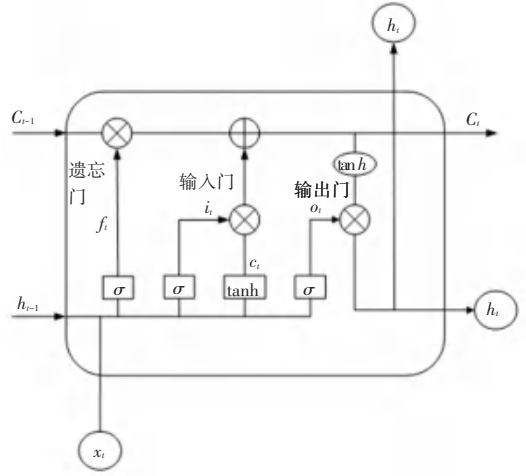


图 4 LSTM 单元结构图

Fig. 4 LSTM unit structure diagram

LSTM 网络一个单元的隐藏层运算过程如式(2)~式(7)所示:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (2)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (3)$$

$$\tilde{C}_t = \tan h(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (4)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (5)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (6)$$

$$h_t = o_t * \tan h(C_t) \quad (7)$$

其中, f_t, i_t, o_t 和 C_t 分别表示 t 时刻的遗忘门、输入门、输出门和记忆细胞; σ 是 sigmoid 激活函数, $\tan h$ 是双曲正切激活函数; W, b 分别表示连接两层的权重矩阵和偏置向量; x_t 为输入向量, h_{t-1} 为 $t-1$ 时刻的输出; h_t 则为 t 时刻的输出; \tilde{C}_t 表示中间状态。

由于 LSTM 只能处理当前单元之前的信息而无法获得之后的信息,于是提出双向长短时记忆网络,就是采用两层 LSTM,分别获取文本序列的前向信息和后向信息进行拼接得到最终的隐藏层特征表示,充分捕捉上下文语义信息,有效提高命名实体识别的效果。

1.3 多头注意力 (MHA) 模型

2017 年谷歌机器翻译团队将多个自注意力进行结合创造性提出多头注意力模型 (multi-head attention model)^[12], 具体结构模型如图 5 所示。将文本序列 $X = (X_1, X_2, \dots, X_n)$ 输入 BiLSTM 层,输出的矩阵 $Y = (Y_1, Y_2, \dots, Y_n)$ 作为 Q、K、V 的输入,放

缩点积注意力 (scaled dot-product attention) 单元共有 h 层, 每一层的注意力计算如式(8)所示; 将 h 个单头注意力输出进行拼接, 同时做一个线性变换如式(9), 得到的 MHA 为第 t 个字的 h 头注意力权重输出。多头注意力模型在 NER 任务中可以充分捕捉句子的长时序依赖关系, 获取全局特征。

$$head_i = \text{softmax} \frac{\text{e}^{\frac{Q \cdot K^T}{\sqrt{d_k}}}}{\text{e}} \cdot V = \text{softmax} \frac{\text{e}^{Y_t W_i^Q (Y_t W_i^K)^T}}{\text{e}} \cdot (Y_t W_i^V), \quad (8)$$

$$MHA_t = \text{concat}(head_1, head_2 \cdots head_h) W^O. \quad (9)$$

其中, W_i^Q 、 W_i^K 、 W_i^V 为需要训练的参数权重; $\sqrt{d_k}$ 为 k 维度的平滑项; $\text{softmax}()$ 为归一化函数; $\text{concat}()$ 为拼接函数; W^O 为权重参数。

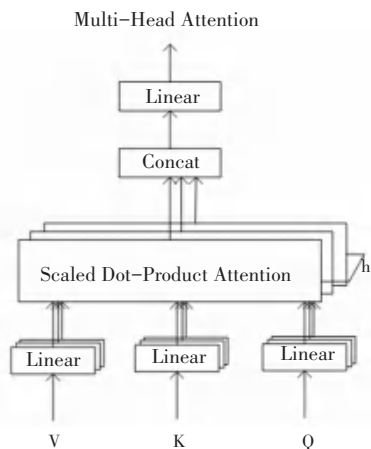


图5 多头注意力模型

Fig. 5 The model of Multi-head attention

1.4 条件随机场 (CRF) 模型

Lafferty 在 2001 年提出线性条件随机场 (CRF) 模型, 计算给定随机变量序列 $X = (X_1, X_2, \cdots, X_n)$ 的条件下, 随机变量序列 $Y = (Y_1, Y_2, \cdots, Y_n)$ 的条件概率分布 $P(Y | X)$ [13]。模型假设随机变量序列满足

马尔可夫性, 式(10):

$$P(Y_i | X, Y_i, \cdots, Y_N) = P(Y_i | X, Y_{i-1}, Y_{i+1}), \quad (10)$$

式中, X 表示输入观测序列, Y 表示对应的状态序列。

在电子病历命名实体识别任务中, 每个字的标签与其相邻的标签都存在制约关系, 例如 O 标签后面不会是 I 标签, I-DIS 不会跟 B-BOD 后面。CRF 能够根据前一层网络的输出结果, 结合上下文语义标签信息得到每个字对应的标签序列出现的最优概率。

设 MHA 模型的输出序列为 X , 其中一个预测序列为 Y , 则可以得出评估分数 $S(X, Y)$, 式(11):

$$S(X, Y) = \sum_{i=0}^n M_{y_i, y_{i+1}} + \sum_{i=1}^n P_{i, y_i}, \quad (11)$$

式中: $M_{y_i, y_{i+1}}$ 表示从 y_i 标签到 y_{i+1} 标签的转移概率; P_{i, y_i} 表示第 i 个字被标记为 y_i 的概率; n 为序列长度。

最后采用极大似然法求解最大后验概率 $P(y | x)$, 获得模型的损失函数值, 式(12):

$$\log P(y | x) = S(x, y) - \sum_{i=0}^n S(x, y_i). \quad (12)$$

2 实验及结果分析

2.1 试验数据与标注策略

本次实验选取的是 ccks-2017 任务二中 400 份医疗标注数据作为数据集, 并按照 7 : 2 : 1 的方式分为训练集、测试集、预测集。该数据集共包括 39 539 个实体, 分为症状、疾病、治疗、检查、身体部位 5 类, 共 7 183 句话。本文采用 BIOE 的标注方式, 即 B 代表实体首字; I 代表实体中间部分; E 代表实体尾字; O 代表该字不属于规定的实体类别。各类别实体符号及数量见表 1。

表 1 医疗实体标注符号

Tab. 1 Medical entity notation

序号	实体类别	开始标签	中间标签	结尾标签	训练集	测试集	预测集
1	疾病 Disease	B-Dis	I-Dis	E-Dis	893	255	127
2	症状 Symptom	B-Sym	I-Sym	E-Sym	7 100	2 028	1 014
3	检查 Check	B-Che	I-Che	E-Che	8 884	2537	1 268
4	治疗 Cure	B-Cur	I-Cur	E-Cur	1 059	303	151
5	身体部位 Body	B-Bod	I-Bod	E-Bod	9 618	2 748	1 374

2.2 评价指标

实体识别和关系抽取实验通常采用准确率、召

回率和 $F1$ 值指标评价模型的优劣:

$$\text{精确率: } P = \frac{T_p}{T_p + F_p} \times 100\%, \quad (13)$$

$$\text{召回率: } R = \frac{T_p}{T_p + F_N} \times 100\%, \quad (14)$$

$$\text{F1 值: } F_1 = \frac{2 \times P \times R}{P + R} \times 100\%. \quad (15)$$

其中: T_p 表示测试集中的正例被正确预测为正例的个数; F_p 表示测试集中的负例被误分类为正例的个数; F_N 表示测试集中的正例被误分类为负例的个数。

2.3 实验环境与参数设置

本文实验的命名实体识别模型基于 TensorFlow 框架,具体实验环境设置见表 2。

表 2 实验环境

Tab. 2 Experimental environment

项目	环境
操作系统	Windows10
CPU	I7-10750H@ 2.60GHz
GPU(显存大小)	RTX2060(8G)
python 版本	3.6.5
tensorflow 版本	1.14.0

实验参数具体设置:BiLSTM 模型的隐藏层大小为 128,网络层数为 1,选取 Relu 作为模型的激活函数。在训练阶段将 Dropout 的比例设置为 0.1,批次大小设置为 16,最大序列长度为 128,学习率设置为

$1e-5$,丢失率为 0.1,使用 Adam 优化器进行训练。

2.4 实验结果分析

为了验证本文提出 XLNet-BiLSTM-MHA-CRF 模型的性能,将其和以下 3 组模型进行对比:

- (1) BiLSTM-CRF 模型;
- (2) Bert-BiLSTM-CRF 模型;
- (3) XLNet-BiLSTM-CRF 模型。

不同模型的实验对比结果见表 3,可以看出 XLNet-BiLSTM-MHA-CRF 模型的精确率、召回率和 F1 值在症状、疾病、治疗、检查、身体部位 5 类医疗实体上都是最高的,相比于 BiLSTM 基线模型分别提高了 3.46%、1.14%、2.31%。

在所有模型中,症状、检查和身体部位 3 个医疗实体的 F1 值普遍较高,而疾病和治疗的实体识别 F1 值恰恰相反。通过分析发现这两个类别的训练数据量明显过少,导致在模型训练过程中出现严重的过拟合现象;另外,疾病实体和治疗实体多为长词结构,例如“左眼眶软组织裂伤”、“左桡骨远端骨折切开复位内固定术”等,而症状、检查和身体部位的实体结构简单且训练数据量大,模型可以充分学习该类实体的文本特征。因此,在后期对电子病历的实体识别中可以通过增加语料库来提高模型的精度,同时可以进一步研究长词实体的结构,挖掘更深层次的语义信息,例如引入词典信息来增加语义特征,提高模型的泛化能力。

表 3 各模型实验对比结果

Tab. 3 The results of each model experiment were compared

模型	评价指标	实体类型					整体
		疾病	症状	检查	治疗	身体部位	
BiLSTM-CRF	P	76.52	94.15	93.27	72.56	83.95	88.61
	R	75.84	95.76	91.80	75.77	85.35	90.27
	F1	76.18	94.95	92.53	74.13	84.64	89.43
Bert-BiLSTM-CRF	P	78.16	94.59	94.25	73.51	84.76	89.73
	R	76.43	95.89	93.81	76.43	86.71	90.14
	F1	77.29	95.24	94.03	74.94	85.72	89.93
XLNet-BiLSTM-CRF	P	80.73	95.02	94.96	74.98	85.61	91.43
	R	77.84	96.10	94.67	75.32	88.59	90.82
	F1	79.26	95.56	94.81	75.15	87.07	91.12
XLNet-BiLSTM-MHA-CRF	P	81.61	95.47	95.28	75.65	87.48	92.07
	R	78.85	96.41	95.72	76.23	89.79	91.41
	F1	80.21	95.94	95.50	75.94	88.62	91.74

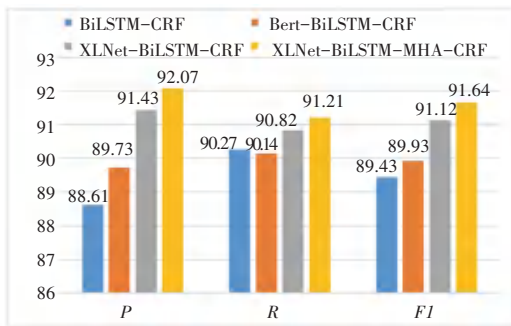


图6 各模型实验结果对比图

Fig. 6 Comparison of experimental results of each model

各模型实验结果对比如图6所示。明显可以看出,基于预训练语言模型XLNet和Bert的模型的表现均比BiLSTM-CRF模型的效果要好,主要是因为后者使用的是传统的Word2vec获取的词向量,无法解决一词多义和同一词的问题,同时也证明了预训练语言模型构造的动态词向量可以提高文本内在语义信息的表达能力;通过对比发现XLNet-BiLSTM-CRF模型比基于Bert的模型表现高0.5%~2%,主要是因为XLNet通过attention mask和transformer-xl模块弥补了Bert的不足,导致识别效果的提升。本文提出的模型相比XLNet-BiLSTM-CRF模型在精确率、召回率和F1上面均有小幅提升,说明加入多头注意力机制可以使文本信息表示更加完整。

3 结束语

本文提出XLNet-BiLSTM-MHA-CRF的医疗电子病历命名实体识别模型,使用预训练语言模型向量,在大规模语料中训练得到的动态词替代传统的静态词向量,对电子病历进行序列化表示,有效解决一词多义等问题,让上下文的语义表示更加准确;使用广义自回归预测模型XLNet可以有效弥补Bert模型的不足;加入MHA机制可以捕获电子病历文本中的长距离依赖特征。在ccks2017任务二数据

集中实验结果表明,基于XLNet-BiLSTM-MHA-CRF模型F1值为91.64%,相较于其他模型达到较好的识别效果,能够较好地完成医疗电子病历的命名实体识别任务,对医学领域的实体识别研究具有一定参考价值。由于本实验数据仅有400份电子病历数据,实体种类较少且实体数量不平衡,因此后期需要获取更多的电子病历数据来丰富模型的识别种类,为挖掘中文电子病历中隐藏的医疗信息做准备。

参考文献

- [1] 杨锦锋,于秋滨,关毅,等. 电子病历命名实体识别和实体关系抽取研究综述[J]. 自动化学报, 2014, 40(8): 1537-1562.
- [2] 徐菁. 面向中文知识图谱的开放式文本信息抽取关键技术研究[D]. 长沙:国防科技大学,2018.
- [3] 周昆. 基于规则的命名实体识别研究[D]. 合肥:合肥工业大学,2010.
- [4] 付文博,孙涛,梁藉,等. 深度学习原理及应用综述[J]. 计算机科学,2018,45(6A):11-15,40.
- [5] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed Representations of Words and Phrases and their Compositionality [J]. Advances in Neural Information Processing Systems, 2013, 26:3111-3119.
- [6] 李颖. 基于BERT-DPCNN的垃圾弹幕识别改进及应用[D]. 上海:上海师范大学,2020.
- [7] 余同瑞,金丹,韩晓臻,等. 自然语言处理预训练模型的研究综述[J]. 计算机工程与应用, 2020, 56(23): 18-28.
- [8] 李舟军,范宇,吴贤杰. 面向自然语言处理的预训练技术研究综述[J]. 计算机科学, 2020, 47(3): 162-173.
- [9] 郭晓然,罗平,王维兰. 基于Transformer编码器的中文命名实体识别[J]. 吉林大学学报(工学版), 2021, 51(3): 989-995.
- [10] 杨丽,吴雨茜,王俊丽,等. 循环神经网络研究综述[J]. 计算机应用, 2018, 38(S2): 1-6,26.
- [11] 储德平,万波,李红,等. 基于ELMO-CNN-BiLSTM-CRF模型的地质实体识别[J/OL]. 地球科学: 1-22 [2021-02-06]. <http://kns.cnki.net/kcms/detail/42.1874.P.20201109.1600.008.html>.
- [12] 鄂海红,张文静,肖思琪,等. 深度学习实体关系抽取研究综述[J]. 软件学报, 2019, 30(6): 1793-1818.
- [13] 应玉龙,李森,乌达巴拉,等. 基于条件随机场的蒙古语词性标注方法[J]. 计算机应用, 2010, 30(8): 2038-2040.

(上接第96页)

- [8] Castañeda Aviña Perla Rubi, Tlelo Cuautle Esteban, de la Fraga Luis Gerardo. Single-Objective Optimization of a CMOS VCO Considering PVT and Monte Carlo Simulations[J]. Mathematical and Computational Applications,2020,25(4).
- [9] 孟海舟. 一种通信用低噪声电荷泵锁相环的研究与设计[D]. 上海:上海交通大学,2013.

- [10] 张勇生. CMOS低相位噪声环形压控振荡器的设计[D]. 武汉:武汉理工大学,2012.
- [11] 成玉. 集成电路版图设计技巧[J]. 电子技术与软件工程,2018(22):78-79.
- [12] 尹海丰. 宽频率范围低抖动锁相环的研究与设计[D]. 哈尔滨:哈尔滨工业大学,2009.