doi:10.3772/j.issn.2095-915x.2019.06.006

基于预训练 BERT 字嵌入模型的领域实体识别

丁龙 文雯 林强

南华大学计算机学院 衡阳 421001

开放科学（资源服务）标识码（OSID）

摘要：随着医疗信息化的发展，越来越多的医疗信息被数字化的记录下来，这些医疗信息蕴含着丰富的医学知识。如何有效地提高提取和利用海量医疗文本信息成为当下医疗信息化发展的巨大挑战，针对目前医疗文本标注数据的不足以及医疗实体边界模糊的问题，本文提出一种基于大量医疗文献预训练的字嵌入语言表示模型。该模型利用大量的医疗文献对 BERT 模型进行预训练，从而得到 EMR-BERT 模型，再通过 EMR-BERT 对训练文本进行字嵌入向量表示，将结果输到 Bi-LSTM 模型，最后利用 CRF 模型进行输出得到最终的结果。通过多组对比实验证明，EMR-BERT+BiLSTM+CRF 模型最终结果优于目前主流模型。因此，该模型能够有效解决医疗电子病历领域命名实体识别任务下，标注数据不足以及实体边界模糊的问题。

摘要：医疗电子病历；命名实体识别；EMR-BERT；字嵌入；Bi-LSTM；CRF

中图分类号：TP32 G35

Domain Entity Recognition Based on Pre-trained BERT Character Embedding

DING Long WEN Wen LIN Qiang

School of Computer, University Of South China, Hengyang 421001, China

Abstract: With the development of medical informationization, more and more medical information is digitally recorded. These medical information contains a wealth of medical knowledge. How to effectively improve the effective extraction and utilization of massive medical text information has become a huge challenge for the development of medical informationization. In order to

基金项目：湖南省教育厅优秀青年项目（18B279）；湖南省哲学社会科学课题（16YBA323）；湖南省“研究生科研创新”项目（CX20190737）；南华大学研究生教改项目（2016JG029）。

作者简介：丁龙（1995-），硕士，研究方向：自然语言处理、实体识别、知识图谱，E-mail：dragonding.sv@gmail.com；林强（1997-），硕士，研究方向：自然语言处理 、小样本学习、知识图谱，E-mail：linqiang0219@163.com；文雯（1997-），硕士，研究方向：自然语言处理，少样本学习，关系抽取，E-mail：wenwen_rainbow@163.com。

doi:10.3772/j.issn.2095-915x.2019.06.006

solve the problem of insufficient data labeling and blurring of medical entity boundaries, this paper proposes a word embedding language representation model based on a large number of medical literature pre-training, which uses a large number of medical literature to pre-train the BERT model to obtain the EMR-BERT model, and then embeds the text into the training text through EMR-BERT. It means that the result is input to the Bi-LSTM model, and finally the output is obtained by using the CRF model. Through multiple sets of comparison experiments, the results of EMR-BERT+BiLSTM+CRF model is better than the current mainstream model. Therefore, the model can effectively solve the problem of insufficient annotation data and fuzzy boundary of the entity in the medical electronic medical record field.

Keywords: Medical electronic record; named entity recognition; EMR-BERT; character embedding; Bi-LSTM; CRF

引言

医疗电子病历(electronic medical record, EMR)，是一种以数字化记录病历信息的方式，其记录的内容可以是病患的治疗方案、病史等电子文本信息，也可以是医疗图像信息，这些都是病患在就诊中产生的记录数据。医疗电子病历有两大类：门诊病历和住院病历。门诊病历一般只记录患者一次性信息，包括病史，诊断信息和处方，数据分散无序；住院病历一般包含大量患者信息，包括入院诊断、病情变化、辅助检查、治疗措施，医生意见以及出院诊断等信息，其中含有大量的医疗实体，是 EMR 的重点研究内容之一[1]。因此如何准确、高效地提取数据实体，成为医疗电子病历研究中的首要任务。

命名实体识别(Named Entity Cognition, NER) 是自然语言处理的基本任务之一，其主要任务是识别出文本中的不同类型实体并加以分类，为后续关系抽取等任务进行铺垫。在医疗领域中，其任务是从医疗电子病历中识别并抽取出医疗实体，如病状、手术、病灶大小等，这些医疗实体对临床决策以及个性化医疗服务与建设具有深远影响。

但由于医疗领域数据具有较强的专业性，现如今对于医疗电子病历的实体识别依然存在以下问题：（1）传统基于神经网络的命名实体识别模型需要大量的标记训练数据，然而医疗领域数据专有名词具有较强的专业性，标注成本高，导致准确的标注数据较少。（2）医生书写习惯的不同，同一实体存在多种写法，如“曲安奈德”，“康 A”，“康尼克通”以上三种书写方式均为同一种药物，但是目前的实体识别模型难以联系上下文对上述类型实体进行归类。（3）医疗实体边界模糊，如“肺”与“肺野”中肺即是实体又是组成实体的字符，容易造成实体识别模型识别率下降。

本文首先利用 BERT 中文预训练模型为 EMR-BERT 提供基本参数和字典，用医疗文献数据对 EMR-BERT 模型进行预训练，构建 EMR-BERT 预训练模型。其次对预训练模型进行 Fine-tune 后将训练语料输入到 EMR-BERT 生成字向量表示，接着将字向量输入到 Bi-LSTM Encoder 模型中。最后利用条件随机场(Conditional Random Fields, CRF)模型进行输出，得到最终的 F1 值，结果达到 62.06%，优于当

前最佳模型。

1 相关工作

医疗电子病历实体识别是医疗领域自然语言处理中最常见的任务 [2]。MedLEE、MedKAT 和 CTAKES[3] 等几个具有代表性的传统基于字典匹配的方法都非常依赖于词典的质量，对不断更新的医疗实体的识别效果较差。基于规则的方法具有代表性的有 Shusaku[4] 所提出来的基于规则自动提取的专家系统。该系统对于头疼类疾病医疗文本提取效果较好，但规则的构建往往需要大量该领域的专家以及专业的语言知识，并且还要注意规则之间的冲突与局限性，该方法缺乏鲁棒性和可移植性。

为了解决上述问题，国内外学者提出各类统计机器学习的方法。由于机器学习模型在命名实体识别的出色表现，这些方法与模型逐渐成为了医疗电子病历命名实体识别的主流方案。这些模型包括基于最大熵（Maximum entropy, ME）[5]、支持向量机（Support vector machine, SVM）[6] 的分类模型，基于隐马尔可夫 [7]（Hidden Markov Model, HMM）的生成模型，基于条件随机场（Conditional Random Field Algorithm, CRF）的判别模型等。由 Lafferty 等人 [8] 提出的 CRF 模型结合了 HMM 模型以及 ME 模型的特点，能够较好的解决标签偏移的问题，产生更好的准确性。Ye[9] 等人通过词性，符号，词汇特征以及词的边界等特征组成特征集，将 CRF 用于中文电子病历实体识别中，产生了良好的效果。CRF 模型常和其他的方法混合在一起使用，多为“特征模型”加上 CRF 的结构，CRF 的优点在于擅长使用序列标注的信息并使用 Viterbi 解码获取最佳的序列，从而出色的完成命名实体识别任务。

长短期记忆网络（Long Short Term Memory Network, LSTM）[10] 被广泛的应用于文本的上下文特征提取。因为 LSTM 是用于处理序列数据的神经网络 [11]，可以很好的捕获长距离依赖关系，但是无法获取从后往前的编码信息。双向长短期记忆网络（Bi-directional Long-Short Term Memory, Bi-LSTM）可以更好的捕获双向语义的依赖。到目前为止，大量的工作都致力于使用 Bi-LSTM 来训练字或词的上下文特征再加上 CRF 的方法优化序列参数。相比于只使用 CRF 的方法，Bi-LSTM+CRF 不需要手动定义特征，通过 LSTM 自动提取文本特征并且来获取句子级标记信息，有效地使用过去的输入特征。Bi-LSTM+CRF 模型在医疗电子病历领域取得了优异的表现 [12-15]。

由于机器学习无法直接对原始文本语料进行处理，词嵌入（word embedding）的概念开始出现。词嵌入是文本的学习表示，其中具有相同含义的单词具有相似的表示，这种词和文本的表示方法是深度学习在自然语言处理中的一项重大突破。2013 年由 Mikolov[16] 等提出的 Word2Vec 是浅层神经网络词嵌入学习最流行的技术之一，它通过把词的 one-hot 表示转换为低纬、稠密的向量，每个词通常由数十或数百个维度的实值向量表示。该方法对推动深度学习领域词嵌入模型产生了深远的影响，直到现在，Word2Vec 依然广泛的应用于医疗领域的任务之中，并在很多任务中取得最好的成绩 [17]。

doi:10.3772/j.issn.2095-915x.2019.06.006

在医疗电子病历的记录中，由于医疗实体的特殊性，上下文同一实体的命名可能不同。表 1 节选了病历的部分内容，为了方便观察特地只标注了“肿瘤部位”。通过观察可以看出，右乳与右侧乳房为同一实体。序列标注任务中，Word2Vec 是局部语料库进行训练的，其特征提取是基于滑窗的。在对类似上述文本进行词向量嵌入时，仅仅只能获取句子级别的信息，这样会丢失实体与实体间的关系，造成无法发现实体间上下文关系，有些情况下还可能会造成相同实体标签不一致，即上下文相同实体出现不同标签。为了解决以上问题，Glove[18] 模型被提出。Glove 的滑窗是基于全局语料的，加入全局信息后，能够有效提高上下文中相同实体的正确标注。但是 Word2vec 与 Glove 属于静态的词向量，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。为了解决上述问题，Peter[19] 等人提出了 ELMO 模型，其本质是两个独立训练的单向串联的 LSTM 预训练语言模型，它能够有效解决上下文词义的歧义问题。这种上下文词嵌入的方式与以往传统的词嵌入不同，它可以获取上下文信息并动态地改变词向量编码。实验表明，ELMO 模型在医疗领域数据上的表现已经优于之前的模型 [20]。但是近期由 Google AI 语言研究人员提出的 BERT[21] 模型在多个下游项目中，不仅在开放领域取得的结果优于 ELMO，而且在医疗领域也取得了更好的结果 [22]。

表 1　电子病历样例

结合临床,右乳癌。较前(2015-06-23)片基本变化不著;双肺转移,较前部分缩小,左肺门未见增大淋巴结，部分变化不著,部分略增大;肝转移,较前好转;双侧腋窝、纵隔多发小淋巴结,变化不著;右侧内乳区饱满,变化不著;脑多发转移。2.左肺胸膜局限性增厚,局部钙化,变化不著。右侧乳房上方见不规则软组织密度肿块影,约4.6CM × 3.2CM,边缘浅分叶,局部与周围腺体及乳头区皮肤分界不清,胸肌间脂肪间隙模糊,增强扫描见显著不均匀强化,其内见片状低密度影。

BERT 是一个多层的双向 transformer 编码器。考虑到 BERT 在 transformer 结构进行编码的时候，傅里叶基的位置编码只提供了相对模糊的位置信息，但序列标记问题同时也需要强大的字、词级上下文信息来预测下一个词，相比之下 LSTM 能够更好的捕获字、词级的上下文信息 [23]。最近，Meng 等人 [24] 发现在中文词表示中，数据稀疏问题会导致过拟合的发生，并且大量的 OOV（out-of-vocabulary）限制了模型的学习能力以及分词方法的不统一，导致结果不佳的问题。他们通过大量的实验表明，中文自然语言处理中“字”的表现总是优于“词”的表现。

因此，本文提出一种类似 BERT 结构的预训练 EMR-BERT 模型进行字嵌入表示，将字向量表示信息加入到 Bi-LSTM+CRF 模型中，能够有效地解决多义性，边界模糊以及标记数据不足的问题。

2　方法

前文给出了医疗电子病历实体识别存在标注数据少、医疗实体边界模糊以及同一实体多种写法的问题，并对目前主流的模型与方法进行分析对比，为解决上述问题，本文提出一种与 BERT 中 token embedding 嵌入表示类似的字

嵌入表示方法，并且使用医疗文献对本文提出的 EMR-BERT 模型进行预训练，得出的字向量再与 Bi-LSTM 结构相结合，最终输出结果。

2.1 BERT

BERT(Bidirectional Encoder Representations from Transformers) 是 Google AI 语言研究人员最近提出的一个模型，该方法的创新是使用一个多层的双向 transformer 编码器。采用双向的 transformer 对字词进行编码，BERT 原文中给出了两种基本模型结构：BERT-Base 和 BERT-Large。BERT-Base 是由 12 层双向 Transformer 编码器模块、涵盖 768 个隐层神经元的 hidden size，以及 12 个自我注意力头数组成。由于 BERT-Large 在后面进行预训练的过程中代价过高，且 google 团队在 2018 年 11 月提出 BERT-Base Chinese 中文预训练模型，就是采用 BERT-Base 的模型框架，所以后文谈到 BERT 所采用的模型结构即为 BERT-Base。在 BERT 中文预训练模型采用字符化的方式处理汉字，使用基于 BPE（Byte-Pair-Encoding）算法双字节编码 WordPiece 来对其它语言进行标注，BERT 神经网络架构如图 1 所示。

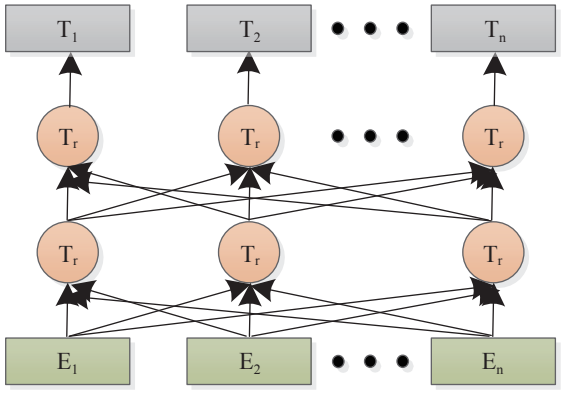

图 1　BERT 模型

图 1 中的箭头表示从一层到下一层的信息流，顶部的 $[T_1,T_2,\cdots,T_N]$ 表示每个输入字的最终的上下文表示。

2.2 预训练EMR-BERT

BERT 是一个深度双向无监督的语言表示模型，使用开放领域语料进行预训练，预训练语料无需标注，这点大大降低了预训练语料的成本。在处理中文下游任务时 BERT 模型采用的是 Google 在开放领域数据集下进行训练的中文预训练模型，在开放领域表现超过了现在主流的模型的结果。

EMR-BERT 采用和 BERT 模型相同的两种方式进行模型预训练过程。第一种预训练方式，我们使用双向 masked 语言模型，采用“随机屏蔽”的方法给输入的 token 标记上 [MASK]，利用上下文来对所标记上 [MASK] 的 token 进行预测任务。但是由于给文本标记上 [MASK]，会影响到后面模型在 fine-tune 时候的表现，影响模型的语言理解能力，所以 google 团队随机选择 15% 的 token。比如生成器选中了“双肺门及纵隔内未见肿大淋巴结”这句话中的“纵”会出现以下三种情况：1. 直接用 [MASK] 来替换所选字，占 80% 的可能性。如：双肺门及 [MASK] 隔内未见肿大淋巴结。2. 随机替换一个字给所选字，占 10% 的可能性。如：双肺门及 [淋] 隔内未见肿大淋巴结。3. 不作出任何改变，占 10% 的可能性。第二种预训练方法主要是为了达到粗粒度（句子级别）关系的水平。第二种方法采用“下一句话预测”方法，首先，它在预训练语料中随机选择两个句子，如句子 A 和句子 B，然后判断 B 是否是 A 的下一句，

doi:10.3772/j.issn.2095-915x.2019.06.006

B 是实际的下一个句子概率为 50%。通过以上两个预训练方法，该模型可以有效的提高模型对单词和句子级别表示。通过上述两种预训练模型，我们可以提高对医疗电子病历中实体的识别效果。

2.3 EMR-BERT +BiLSTM+CRF

本文所提到的 EMR-BERT 对文本输入做了全新的定义，可以是单句也可以是双句子，并对输入表示的概念进行拓展，它的输入表示由位置嵌入 (position embeddings)，字向量嵌入 (character embeddings)，句子嵌入 (segment embeddings) 三个嵌入叠加组成。位置嵌入记录了字的顺序信息，字向量嵌入记录的是每个字的信息，句子嵌入记录的是每个句子独特的整体信息。

本文模型选用了 EMR-BERT 对已经序列标注的训练语料进行三种形式的嵌入叠加，取代传统的 Word2Vec 词嵌入方式，再将处理好的向量表示加入到 Bi-LSTM 模型，最后利用 CRF 模型输出最终的实体识别结果，模型结构如图 2 所示。

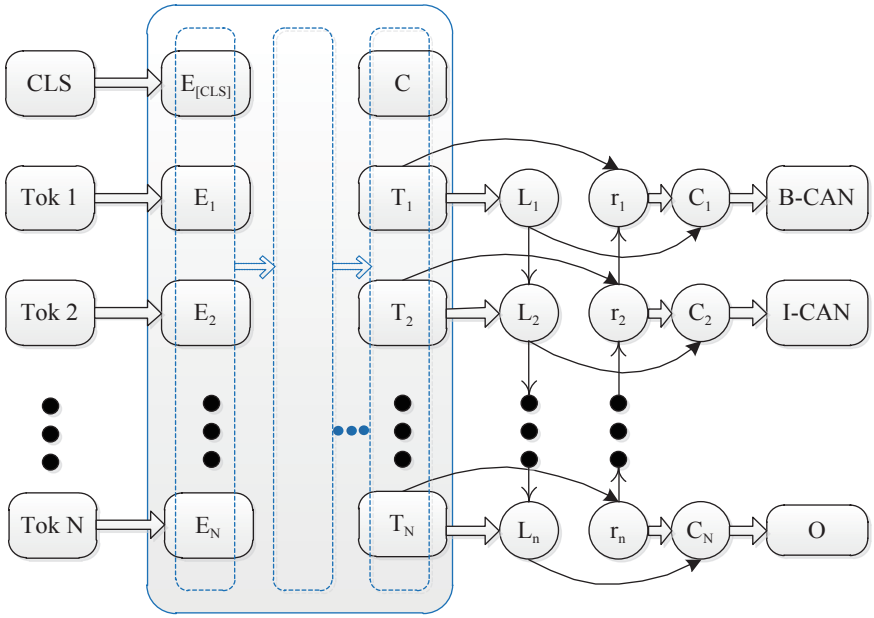

图 2 **EMR-BERT +BiLSTM+CRF** 模型

如图 2 所示，通过 EMR-BERT 模型进行字嵌入表示，产生了输入序列为 {T1,T2,⋯TN}，由 Bi-LSTM 返回输入序列的表示序列 {h1,h2,⋯, hn}。在 t 时刻，EMR-BERT 通过字嵌入给定输入 Tt，通过 Bi-LSTM 隐含层的输出表示 ht, 模型输出可以表示为：

$$i_t = \sigma(W_{Ti}T_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma(W_{Tf}T_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (2)$$

$$C_t = f_t c_{t-1} + i_t \tan h(W_{Tc}T_t + W_{hc}h_{t-1} + b_f) \quad (3)$$

$$O_t = \sigma(W_{To}T_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \quad (4)$$

$$h_t = O_t \tan h(c_t) \quad (5)$$

其中，W 为层与层之间的权重矩阵，b 为偏移向量，c 为神经单元的状态，σ 和 $tanh$ 为两种不同神经元的激活函数，i 为输入门，f 为

遗忘门，*O* 表示输出门。

3 实验

本节首先描述实验数据和评估指标，然后设置实验参数，之后，按顺序依次进行三次实验。第一个实验讨论预训练模型收敛步数，以避免计算资源的浪费；第二个实验对模型进行微调，找到模型的最佳参数；第三个实验，对预训练 BERT 模型进行字嵌入表示，再加上 Bi-LSTM 模型，达到最佳结果，以解决电子病历实体识别中实体多义性，边界模糊以及标记数据不足的问题。

3.1 实验数据与评价标准及参数设置

本文模型的预训练语料使用了 2000 份中文医学文献，训练语料来自中国健康信息处理学术会议 (CHIP2018) 评测一：中文电子病历中临床医疗实体及属性抽取任务中发布的 600 份电子病历训练数据。首先对文本进行必要的预处理（去除特殊字符、英文大小写以及部分标点符号等），并在处理好的文本中加入了噪音。按照 5:1 的比例对原始语料中肿瘤原发部位以及转移部位进行人工标注，采用 BIO 形式的序列标注法，B 表示实体的开始，I 表示实体的延续，O 表示非实体部分。数据样例如表 2 所示。

表 **2** 标注样例

Token	Label
右	B-BODY
乳	I –BODY
术	O
后	O
缺	O
如	O

本文采用了自然语言处理任务中常见的 3 种评价指标，即准确率 (Precision)，召回率 (Recall)，综合指标 F1 值，其中：

1) 肿瘤部位的识别准确率：P = N3 / N2

2) 肿瘤部位的识别召回率：R = N3 / N1

3) 肿瘤部位的综合指标：F1 = (2 PR/(P+R))

N1 表示语料中实际实体的数量；N2 表示模型识别出的实体数量；N3 表示模型正确识别出的实体数量，P、Q 分别是准确率与召回率的缩写。

3.2 实验方法的参数设置

我们选取了与癌症相关的 2000 篇中文医疗文献，作为 EMR-BERT 的预训练语料。我们的第一组实验旨在发现预训练模型收敛时的步数。本文提出的对 EMR-BERT 模型的预训练过程中，会消耗大量计算资源，且模型会在一定步数下收敛，继续进行预训练会降低模型的准确率。因此，当模型收敛时，应当停止预训练过程，以减少资源的浪费。根据上述问题，本文将在实验一讨论模型收敛的步数，基本参数为：batch size：32；maximum sequence length：128，学习率为 3e-5；其他参数保持默认值（masked language model probability = 0.15 以及 max predictions per sequence = 20）。

在第二组实验中，本文通过对比实验找到模型最佳的参数设置，其他参数保持默认值。

3.3 实验比较与结论

为了能够对两种预训练方法在预训练模型中收敛速度进行探讨，我们将预训练步数的参数设定在 12 万步。在图 3 中横坐标表示步数，纵坐

doi:10.3772/j.issn.2095-915x.2019.06.006

标表示精确率，黄色的实线表示下一句预测的精确率随步数的变化情况，蓝色虚线表示 masked 语言模型的精确率随步数的变化情况。在图中我们可以清楚的看到，在 4 万步之前预训练模型在双向 masked 语言模型和下一句预测两种方法上的表现都随着步数的增加在提高，结果发现模型在 4 万步的时候已经收敛，继续提高步数，预训练模型并不会使模型的精确率继续上升。

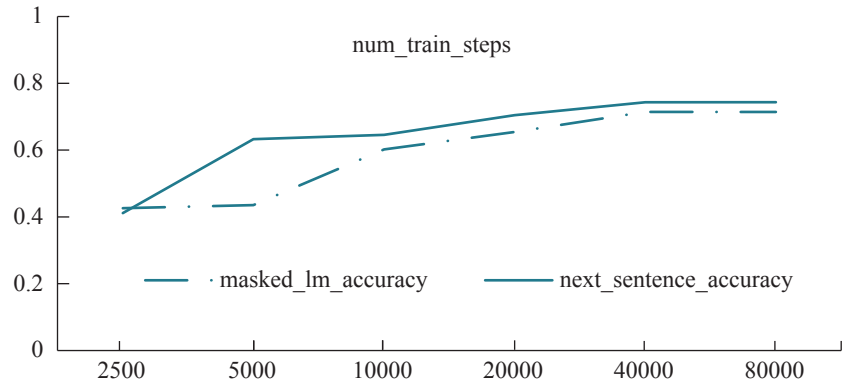

图 **3** 预训练模型在不同步数的表现

第二组实验通过微调本文提出的 EMR-BERT+Bi-LSTM+CRF 模型与 BERT-base+Bi-LSTM+CRF 模型来观察本文提出的方法是否能够达到更好的实体识别效果。可调参数 batch size 设定了为 32、64，学习率设定为 2e-5、3e-5、5e-5。

表 **3** 模型的微调

MODEL		Learning rate								
		2e-5			3e-5			5e-5		
		P	R	F1	P	R	F1	P	R	F1
Batch Size=32	BERT-base+ Bi-LSTM+CRF	59.65	60.32	59.98	59.05	58.49	58.77	44.87	59.76	51.26
	EMR-BERT+ Bi-LSTM+CRF	60.54	58.81	59.66	60.90	59.84	60.37	49.57	59.05	53.89
Batch Size=64	BERT-base+ Bi-LSTM+CRF	52.45	64.71	57.94	52.61	67.89	59.28	49.36	57.94	53.30
	EMR-BERT+ Bi-LSTM+CRF	59.83	57.22	58.50	55.39	66.83	62.06	50.34	59.13	54.38

结合表 3，从上表可以发现无论是 batch size 等于 32 还是 64，EMR-BERT+ Bi-LSTM+CRF 都在学习率为 3e-5 时候达到最优效果，BERTBASE+ Bi-LSTM+CRF 在 Batch size = 32，学习率 =2e-5 时，达到最高 F1 值为 59.98，EMR-BERT+ Bi-LSTM+CRF 在 Batch size = 64，学习率 =3e-5 时，达到最高 F1 值为 62.06，通过对上表的分析，我们发现，EMR-BERT+ Bi-LSTM+CRF 能够取得更优结果。

第三组实验我们将我们提出的 EMR-BERT+ Bi-LSTM+CRF 与当前主流的模型进行比较，实验结果如表 4 所示。

表 4 模型对比

Model	Precision	Recall	F1
$BERT_{BASE}$	57.92	55.83	56.86
EMR-BERT	56.12	60.31	58.14
Word2Vec+Bi-LSTM+CRF	55.73	59.62	57.61
$BERT_{BASE}$+Bi-LSTM+CRF	59.65	60.32	59.98
EMR-BERT+Bi-LSTM+CRF	59.42	64.94	62.06

通过比较表 4 中的几组实验结果表明，在对肿瘤部位的实体识别中，本文提出的 EMR-BERT+ Bi-LSTM+CRF 取得了最佳的 F1 值，说明了我们模型的有效性。通过比较 $BERT_{BASE}$ 和 EMR-BERT，证明基于医学语料预训练可以提高模型对医学实体的识别效果。通过比较 EMR-BERT 和 EMR-BERT-Bi-LSTM-CRF， 我们可以证明 LSTM 比 EMR-BERT 对字符级上下文识别的效果更好。但是可以发现我们模型的精度下降了 0.23 个百分点，经过对上表的比较分析，我们发现训练数据属于医疗电子病历，但我们的预训练模型是基于大量的非特定领域医学文献预训练的，虽然可以识别出比以前开放领域的 $BERT_{base}$ 模型更多的医疗实体，但它还会导致识别出与肿瘤部位无关的医疗实体，导致准确率降低。因此，在今后的工作中，我们会考虑加入更多的关于肿瘤疾病研究的文献语料，来提高模型对肿瘤实体的识别效率，以达到更高的识别精确率。

4 结语

本文针对医疗电子病历的标注数据少，医疗实体边界模糊以及同一实体多种写法的问题。首先，我们预先训练了一个基于大量医学文本的无监督语言表示模型，解决了标注数据缺少的问题。其次利用 BERT 结合 Bi-LSTM + CRF，得到词级、句子级和上下文级的特征表示，有效地解决了同一实体的不同命名和实体边界的模糊性问题。最后，通过比较现有的主流先进模型，我们取得了最好的结果，表明能够有效地完成医疗电子病历实体识别任务。

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] Wang Y, Wang L, Rastegar-Mojarad M, et al. Clinical information extraction applications: A literature review[J]. Journal of Biomedical Informatics, 2018, 77:34-49.

[3] Savova G K, Masanz J J, Ogren P V, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications[J]. Journal of the American Medical Informatics Association, 2010, 17(5): 507-513.

[4] Tsumoto S. Automated extraction of medical expert system rules from clinical databases based on rough set theory[J]. Information sciences, 1998, 112(1-4): 67-84.

[5] Zhao J. Research on Conditional Probabilistic Model and Its Application in Chinese Named Entity Recognition[D]. Harbin: Harbin Institute of Technology, China, 2006.

[6] Lin X D, Peng H, Liu B. Chinese named entity recognition using support vector machines[C]. In: Proceedings of the 2006 International Conference on Machine Learning and Cybernetics. Guangzhou, China: IEEE. 2006: 4216-4220.

[7] 姜维 , 王晓龙 , 关毅 , 等 . 基于多知识源的中文词法分析系统 [J]. 计算机学报 , 2007(1):137-145.

[8] Habibi M, Weber L, Neves M, et al. Deep learning with word embeddings improves biomedical

doi:10.3772/j.issn.2095-915x.2019.06.006

named entity recognition[J]. Bioinformatics, 2017, 33(14):37-48.

[9] Ye F, Chen Y Y, Zhou G G, et al. Intelligent recognition of named entity in electronic medical records[J]. Chinese Journal of Biomedical Engineering, 2011, 30(2):256-262.

[10] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8):1735-1780.

[11] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv:1508.01991, 2015.

[12] Chalapathy R, Borzeshi E Z, Piccardi M. Bidirectional LSTM-CRF for Clinical Concept Extraction[J]. arXiv preprint arXiv:1611.08373, 2016.

[13] Boag W, Sergeeva E, Kulshreshtha S, et al. CliNER 2.0: Accessible and Accurate Clinical Concept Extraction[J]. arXiv preprint arXiv:1803.02245, 2018.

[14] Guohai Xu, Chengyu Wang, and Xiaofeng He. Improving clinical named entity recognition with global neural attention. APWeb-WAIM 2018, LNCS 10988, Macau, China. 2018: 264-279.

[15] Boag W, Wacome K, Naumann T, et al. CliNER: A lightweight tool for clinical named entity recognition[C]. AMIA Joint Summits on Clinical Research Informatics, 2015.

[16] 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.

[17] Yanshan, Wang, Sijia, et al. A Comparison of Word Embeddings for the Biomedical Natural Language Processing.[J]. Journal of biomedical informatics, 2018(87):12–20.

[18] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation[C]. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.

[19] Peters, Matthew E, Neumann, Mark, Iyyer, Mohit, et al. Deep contextualized word representations[J]. arXiv preprint arXiv:1802.05365, 2018.

[20] Zhu H, Paschalidis I C, Tahmasebi A . Clinical Concept Extraction with Contextual Word Embedding[J]. arXiv preprint arXiv:1810.10566, 2018.

[21] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. http://arxiv.org/abs/1810.04805, 2018.

[22] Si Y Q, Wang J Q, Xu H, et al. 2019. Enhancing Clinical Concept Extraction with Contextual Embedding[J]. arXiv:1902.08691 [cs]. ArXiv:1902.08691.

[23] Wang C, Li M, Smola A J. Language Models with Transformers[J]. arXiv preprint arXiv:1904.09408, 2019.

[24] Meng Y, Li X, Sun X, et al. Is Word Segmentation Necessary for Deep Learning of Chinese Representations?[J]. arXiv preprint arXiv:1905.05526, 2019.