浙江科技学院学报，第 34 卷第 6 期，2022 年 12 月
Journal of Zhejiang University of Science and Technology
Vol. 34 No. 6, Dec. 2022

doi：10.3969/j.issn.1671-8798.2022.06.005

基于半监督学习的中文电子病历命名实体识别

张　杰，黄　杰，万　健
（浙江科技学院 信息与电子工程学院，杭州 310023）

摘　要：面向中文电子病历的命名实体识别（named entity recognition，NER）研究已经取得不错的成果，但其中大部分方法依赖于已标注医疗语料而无法充分利用未标注语料，且方法中构建的文本特征相对单一，无法深入获取医疗文本的特征。针对上述问题，设计了一种基于半监督学习的 NER 模型。首先，本模型通过构建多个特征来捕捉病历文本中的语义信息，使用基于转换器的双向编码表征（bidirectional encoder representation from transformers，BERT）训练海量的未标注数据来学习适合中文医疗领域的字向量表示，并使用双向语言模型捕捉每个字的上下文特征向量，以及使用医疗词典结合双向最大匹配算法构建文本的词典特征向量。其次，融合 3 种特征向量后输入由双向门控循环单元、自注意力机制和条件随机场组成的 NER 模型中训练。最后，NER 模型通过预测未标注语料获得候选标注语料，引入自举（bootstrapping）算法筛选置信度高的候选标注语料，将其合并到初始标注语料后迭代训练 NER 模型。试验结果表明，本模型在自建脑血管数据集和中国知识图谱与语义计算大会（China Conference on Knowledge Graph and Semantic Computing，CCKS）发布的 CCKS2017、CCKS2018 数据集上的 F_1 值分别为 90.16%、92.72%和 90.93%，优于其他使用额外特征的 NER 模型和主流神经网络模型。本模型为提高中文电子病历的实体识别精度提供了一种新方法，可应用于实际工程中的 NER 任务。

关键词：中文电子病历；命名实体识别；半监督学习；语言模型；自举算法
中图分类号：TP183　　**文献标志码**：A　　**文章编号**：1671-8798(2022)06-0502-10

On named entity recognition for Chinese electronic medical record based on semi-supervised learning

ZHANG Jie, HUANG Jie, WAN Jian
(School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, Zhejiang, China)

Abstract: Studies on named entity recognition (NER) for Chinese electronic medical record have achieved desirable results, but most of them rely on annotated medical corpus and fail to make full use of unannotated medical corpus, and the text features constructed in the method are relatively single, so it is impossible to obtain the features of medical text in depth. Aiming at the above problems, a NER model based on semi-supervised learning was designed. Firstly, the model constructed multiple features to capture the semantic information in the text of medical

收稿日期：2021-11-09
基金项目：国家自然科学基金项目（61972358）；浙江省重点研发计划项目（2020C03071）
通信作者：万　健（1969—　），男，福建省泉州人，教授，博士，主要从事云计算及大数据研究。E-mail：wanjian@zust.edu.cn。

record, and harnessed the bidirectional encoder representation from transformers (BERT) to train massive unannotated data with a view to learning word vector representation suitable for the Chinese medical field. Meanwhile, the bidirectional language model was applied to capture the context feature vector of each word, combing the medical dictionary with the two-way maximum matching algorithm to construct the dictionary feature vector of text. Secondly, those three eigenvectors were fused and input into the NER model, composed of bidirectional gating loop unit, self-attention mechanism and conditional random field. Finally, the NER model obtained candidate annotated corpora by predicting unannotated corpora, and introduced bootstrapping to screen candidate annotated corpora with high confidence, which were merged into the initially annotated corpora and then iteratively trained the NER model. The experimental results show that the F_1 values of the model in the self-built cerebrovascular data set and the CCKS2017 and CCKS2018 data sets published by China Conference on Knowledge Graph and Semantic Computing (CCKS) are 90.16%, 92.72% and 90.93% respectively, being superior to other NER models and mainstream neural network models using additional features. This model can provide a new method for improving the entity recognition accuracy of Chinese electronic medical record, applicable to NER tasks in practical engineering.

Keywords: Chinese electronic medical record; named entity recognition; semi-supervised learning; language model; bootstrapping method

中文电子病历(chinese electronic medical record,CEMR)具有安全可靠、时效性强、存储便利等优点,已经被广泛应用于医疗体系中[1]。面向中文电子病历的命名实体识别(named entity recognition,NER)指从 CEMR 中识别出医疗相关的实体名称,如疾病名称、病症描述、药物名称等。命名实体识别技术可用于挖掘 CEMR 中包含的医疗信息,基于医疗信息构建的知识库和知识图谱可以辅助医生完成疾病诊断和药物推荐等工作[2]。因此,研究中文电子病历命名实体识别具有现实意义和实际经济价值。

随着深度学习技术和神经网络模型的不断发展,卷积神经网络(convolutional neural network,CNN)、双向长短期记忆网络[3](bidirectional long-short-term memory,BiLSTM)、双向门控循环单元[4](bidirectional gated recurrent unit,BiGRU)、条件随机场(conditional random field,CRF)等模型被广泛应用于 NER 任务中。医疗领域包含众多专业名词,直接将通用领域 NER 模型应用于中文电子病历命名实体识别任务时,模型的识别效果不佳,研究者将医疗领域特征引入 NER 模型,有效地提升了模型的实体识别效果。Wang 等[5]提出一种基于多粒度语义字典和多模态树的 NER 模型,利用多模态树提取词汇特征及词边界特征,试验结果表明该模型能显著提升实体识别的效果。Ji 等[6]提出一种将医疗词汇信息和词语纠错规则引入 BiLSTM-CRF 的集成模型,试验结果表明该模型在准确率和召回率上均优于传统 NER 模型。Yin 等[7]提出一种基于部首级特征和自注意力机制的 NER 模型,试验结果表明该模型能显著提升实体识别的效果。Wu 等[8]提出一种融合多特征的医疗 NER 模型,使用预训练语言模型获得文本的字符特征,引入 BiLSTM 提取文本的部首特征,试验结果表明该模型极大地提高了医疗实体识别的性能。

上述方法引入额外的医疗特征(如词典特征、部首特征等)来提升实体识别效果,但是,这类方法依赖于大规模的标注语料和文本特征。现存的中文电子病历标注语料十分匮乏且规模较小,而未标注的中文电子病历语料却海量且易获得,如何利用未标注语料来提高 NER 效果已成为研究的热点[9]。部分研究者提出利用海量未标注医疗文本进行训练的语言模型,如 Tang 等[10]提出一种融合语言模型和自注意力模型的方法,利用大量的未标注医疗语料训练双向语言模型,深入捕捉医疗文本的语义特征;Wen 等[11]使用双向语言模型和掩码语言模型获取未标注医疗语料中包含的上下文语义特征,分别通过权重迁移和特征融合的方法与字符特征向量进行拼接,有效地提升了 NER 效果;Yu 等[12]使用海量的未标注生物医学文本训练

基于转换器的双向编码表征模型[13](bidirectional encoder representation from transformers,BERT),将自举算法结合 NER 模型进行迭代训练,充分利用有限的标注语料和大量未标注语料,显著提升实体识别效果。以上 3 种方法利用语言模型训练大量未标注语料,通过捕捉文本中的深层语义特征来提高识别的准确率,然而在缺乏大量未标注训练语料或训练语料长度较短的领域,使用该类方法进行命名实体识别则效果欠佳。

针对上述问题,本研究提出一个基于半监督学习的融合双向语言模型(bidirectional language model,BiLM)和自举(bootstrapping)算法的命名实体识别模型(BiLM_BNER 模型),该模型使用大量未标注的医疗文本训练 BERT 和双向语言模型。BERT 用于获取 CEMR 文本的字符向量,BiLM 用于获取医疗文本的上下文语义特征向量。此外,引入医疗词典来构建 CEMR 文本的词典特征向量。为了充分利用未标注医疗语料中包含的语义信息,BiLM_BNER 模型引入自举算法将未标注语料转换为高置信度的标注语料,通过模型的迭代训练来不断扩大标注语料的规模以提升模型的识别效果。

1 BiLM_BNER 模型介绍

1.1 整体模型

图 1 为 BiLM_BNER 模型结构示意图,BiLM_BNER 集成模型包含 BERT 模型、BiLM 模型、词典特征、BiGRU-Attention-CRF 及 bootstrapping 算法。BiLM_BNER 模型训练的具体流程如下:1) 利用未标注医疗文本训练 BERT 和 BiLM,学习医疗文本的字符向量表示和上下文特征向量表示;2) 给定中文医疗文本序列 $W=[w_1,w_2,\cdots,w_m]$,$w_t(t\in[1,\cdots,m])$表示一个中文字符,BERT 模型将文本中的字符转化为向量表示 $\boldsymbol{h}_c=[\boldsymbol{h}_{c1},\boldsymbol{h}_{c2},\cdots,\boldsymbol{h}_{cm}]$,$\boldsymbol{h}_{ct}(t\in[1,\cdots,m])$为字符的向量表示形式;BiLM 获得字符在病历文本的上下文特征向量 $\boldsymbol{h}_f=[\boldsymbol{h}_{f1},\boldsymbol{h}_{f2},\cdots,\boldsymbol{h}_{fm}]$和 $\boldsymbol{h}_b=[\boldsymbol{h}_{b1},\boldsymbol{h}_{b2},\cdots,\boldsymbol{h}_{bm}]$,$\boldsymbol{h}_{ft}$、$\boldsymbol{h}_{bt}(t\in[1,\cdots,m])$分别为表示正向语言模型和反向语言模型的输出结果;构建医疗词典,将词典结合双向最大匹配算法来标注训练文本中的医疗词汇,构建词典特征向量 $\boldsymbol{h}_w=[\boldsymbol{h}_{w1},\boldsymbol{h}_{w2},\cdots,\boldsymbol{h}_{wm}]$,$\boldsymbol{h}_{wt}(t\in[1,\cdots,m])$为字符对应的词典特征向量;3) 融合字符特征向量、上下文特征向量和词典特征向量,将融合结果输入 BiGRU-Attention-CRF 进行训练,训练好的模型通过预测未标注中文电子病历文本得到候选标注语料;4) bootstrapping 算法设定最优阈值来筛选候选标注语料中置信度高的新标注语料,将其加入初始医疗标注数据集后继续训练 NER 模型;5) 重复步骤 4,随着 NER 模型的迭代训练,初始标注数据规模不断扩大,模型的识别效果不断提升,迭代终止条件为设置的最大训练次数或模型的 F_1 值达到最大值。

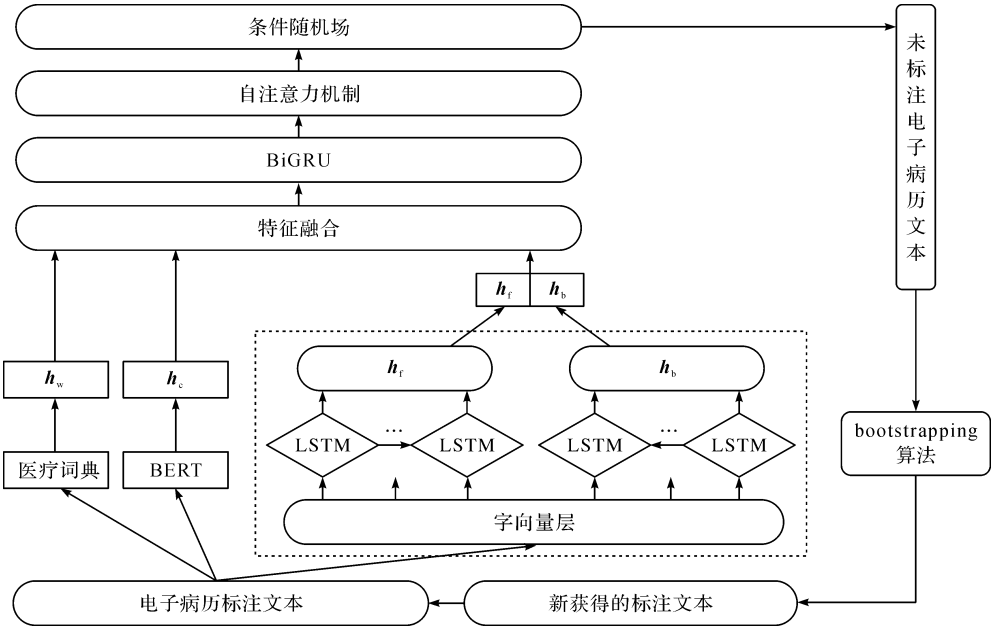

图 1 BiLM_BNER 模型结构示意图

Fig. 1 Schematic diagram of BiLM_BNER model structure

1.2　双向语言模型

语言模型（language model，LM）可以预测句子序列出现的概率，能够从大规模自然语言中学习到丰富的语义知识和内在的规律，目前被广泛应用于自然语言处理任务中[14]。双向语言模型由正向 LM 和反向 LM 组成，LM 包含字向量层、LSTM 层和 softmax 层。图 2 为双向语言模型的结构示意图。

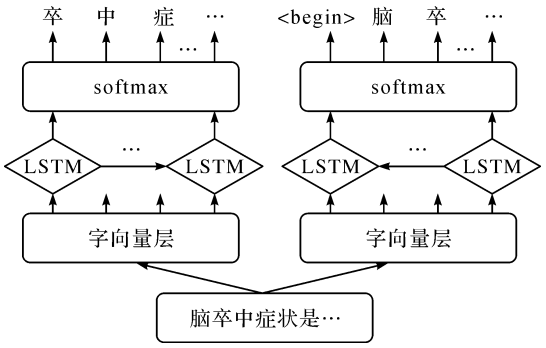

图 2　双向语言模型结构示意图

Fig. 2　Schematic diagram of bidirectional language model

BiLM 的训练流程如下：1）字向量层使用字转换向量模型（word to vector，word2vec）[15]技术将中文字符处理成向量表示形式，将字向量分别输入正向 LSTM 和反向 LSTM 中提取文本特征，正向 LSTM 能够根据历史信息预测当前字符 w_t 的后一个字符为 w_{t+1}，反向 LSTM 能够根据未来的信息预测出其前一个字符为 w_{t-1}；2）将正向 LSTM 和反向 LSTM 输出的特征向量分别输入 softmax 层中计算字符出现的概率。

以正向 LM 为例，输入文本序列表示一个中文字符，通过 word2vec 处理后得到中文字符的向量表示 $\boldsymbol{c}_t$，将 $\boldsymbol{c}_t$ 输入正向 LSTM 单元得到隐藏层的输出向量 $\boldsymbol{h}_{\mathrm{f},t}$。

$$\boldsymbol{h}_{\mathrm{f},t}=\mathrm{LSTM}_{\mathrm{f}}(\boldsymbol{c}_t,\boldsymbol{h}_{\mathrm{f},(t-1)};\boldsymbol{b}_{\mathrm{f}})。\tag{1}$$

式(1)中：$\boldsymbol{h}_{\mathrm{f},t}$为 LSTM 的隐藏层输出；$\boldsymbol{c}_t$ 为字向量；$\boldsymbol{h}_{\mathrm{f},(t-1)}$为上一时刻 LSTM 隐藏层的输出向量；$\boldsymbol{b}_{\mathrm{f}}$ 为正向 LSTM 模型的参数。

LSTM 的输出结果包含了每个标签的预测分值，将 LSTM 的输出向量输入 softmax 层，计算字典中各个字符在当前位置出现的概率

$$\boldsymbol{S}'_{\mathrm{f},t}=\mathrm{softmax}(\boldsymbol{W}_{\mathrm{f}}\boldsymbol{h}_{\mathrm{f},t}+\boldsymbol{b}_{\mathrm{f}})。\tag{2}$$

式(2)中：$\boldsymbol{S}'_{\mathrm{f},t}$为预测下一个字符的概率；$\boldsymbol{W}_{\mathrm{f}}$ 为权重矩阵；$\boldsymbol{b}_{\mathrm{f}}$ 为偏置矩阵；$\boldsymbol{h}_{\mathrm{f},t}$为 LSTM 隐藏层输出向量。

以句子“脑额叶损伤的表现”为例，正向 LM 输出结果为“额叶损伤的表现＜end＞”。反向 LM 模型的结构和正向 LM 相同，在模型训练时，反向 LM 模型使用反向 LSTM 预测当前输入字符 w_t 的上一个字为 w_{t+1}。对于句子“脑额叶损伤的表现”，反向 LM 的预测结果为“＜begin＞脑额叶损伤的表”。

我们使用大量的未标注医疗文本分别训练正向语言模型和反向语言模型，使得模型学习到医疗文本的语法和语义信息。在训练 BiLM_BNER 模型时，我们将正向语言模型和反向语言模型的 softmax 层去除，固定语言模型的权重，利用 BiLM 提取电子病历文本的上下文语义特征，将正向语言模型和反向语言模型对应位置的特征进行拼接。

1.3　中文医疗语料预训练的 BERT 模型

BERT 模型将多层双向变压器（transformer）模型作为编码器，在训练时能更好地捕捉文本中上下文的语义信息，有效地解决了 word2vec 存在的一词多义问题。图 3 为 BERT 模型处理字向量过程示意图，BERT 模型的输入由字向量、句向量和位置向量叠加生成，BERT 模型通过查询字向量表将中文序列中的字符转换为字向量表示，句子向量用于区分不同的句子，位置向量用于区分句子中不同位置的字符，将三者进行拼接后输入多层 transformer 中提取文本特征，最终的输出向量作为字符特征向量。以“[CLS]头晕[SEP]脑梗死[SEP]”为例，用[CLS]标识句子的开始位置，用[SEP]标识句子的结束位置，$\boldsymbol{E}$ 表示向量表达，Trm 表示 transformer 模型。

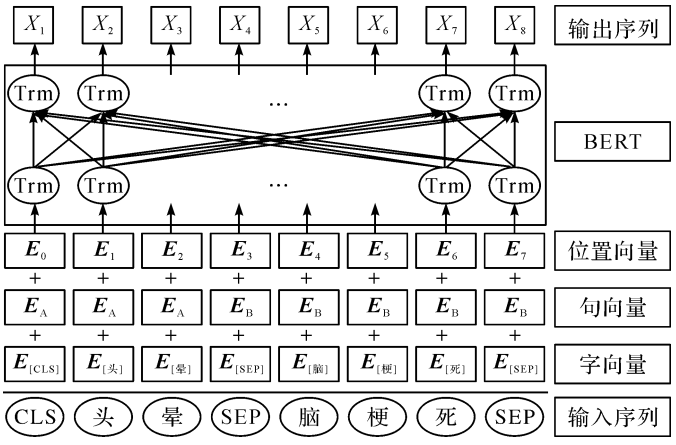

图 3　BERT 模型处理字向量过程示意图

Fig. 3　Schematic diagram of BERT's word vector processing

Lee 等[16]利用海量的英文生物医学语料库再次训练 BERT 模型，获得面向英文医疗领域的 BioBERT 模型。本研究使用中文医疗语料库训练 BERT 模型，使得 BERT 学到适合中文医疗文本的向量表达方式，将面向中文医疗领域的 BERT 模型命名为 CbBERT 模型，如何构建中文医疗训练语料库的过程将在 2.1 节做详细介绍。本研究参照 BERT 模型提出者的思路构建 BERT 模型并设置模型参数，将医疗 txt 文本处理成一个.tfrecord 文件，调用 run_pretraining.py 进行预训练，训练完成后得到一个.ckpt 格式的 TensorFlow 模型，最终利用转换脚本将.ckpt 格式转换为.bin 格式的 PyTorch 模型。

1.4 构建词典特征模块

医疗领域词典中包含丰富的词汇特征，本研究通过引入词汇特征帮助 NER 模型识别实体的边界，从而提升实体识别效果[17]。使用 CCKS2017 和 CCKS2018 电子病历数据集中标注的医疗实体构建一个包含症状描述、检查检验、疾病诊断、身体部位和治疗方案的医疗词典，并获取搜狗医疗词库等互联网中包含的医疗词汇来进一步扩充词汇量。此外，使用分词工具对 CCKS 发布的电子病历和医院真实脑血管电子病历文本进行分词和词性标注，对人工筛选分词结果中符合要求的名词，将其加入医疗词典中。去除医疗词典的重复名词和噪声词汇，得到最终的医疗领域词典统计表，具体见表 1。

表 1 医疗领域词典统计表

Table 1 Statistics table of medical dictionary

词典类别	定义	实体符号	词汇量/个	实例
症状描述	患者对机体生理功能的主观感受，外部观察到患者机体生理功能的客观事实	Symptom	333	胸闷、视物模糊
检查检验	在治疗时涉及到的检查项目	Check	643	核磁共振、B 超
疾病诊断	人群中出现的疾病、综合征	Disease	454	老年痴呆、癫痫
身体部位	行使特定功能的身体器官	Body	1 923	脑干、肾上腺
治疗方案	在治疗时涉及到的手术名称、药物名称	Treat	445	微创术、降压药

利用医疗词典和双向最大匹配算法[18](bi-direction maximum matching，BMM)标注测试集中的电子病历文本的词语。BMM 算法以匹配词汇量多且词语长度宜长为原则，选取正向最大匹配法和反向最大匹配法的分词最优解。将匹配到的实体根据其类型进行标注，为了标注实体的边界信息，使用“B-”标注实体首字符，使用“I-”标注实体的其他字符，非实体则用 None 标注。以“因头痛伴心悸，检查心电图和颅脑 CT”为例，实体类型标注结果为“None B-Symptom I-Symptom None B-Symptom I-Symptom None None None B-Check I-Check I-Check None B-Check I-Check I-Check I-Check”。将实体类型名称随机初始化成一个 50 维词典特征向量，并将当前字符对应的实体类型名称转化为词典特征向量。

1.5 BiGRU-Attention-CRF 模型

BiGRU 模型用于获取医疗文本的上下文语义特征，但 BiGRU 模型存在长距离依赖问题，其获取语义信息的能力会随着文本序列的增加而减弱，引入自注意力机制可有效解决此问题。自注意力机制能够忽略字词之间的距离，通过计算依赖关系来增大文本中重要字词的权重，从而解决 BiGRU 模型存在的长距离依赖问题。CRF 模型训练过程中能自动地学习标签之间的约束关系，利用维特比算法解码生成最优标注序列。图 4 为 BiGRU-Attention-CRF 模型的序列标注过程示意图。

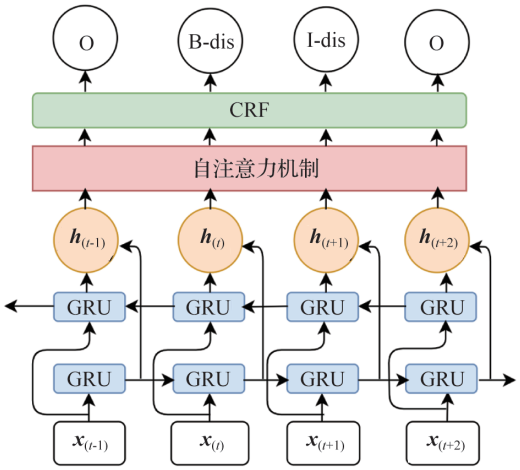

图 4 BiGRU-Attention-CRF 模型的序列标注过程示意图

Fig. 4 Schematic diagram of sequence labeling process of BiGRU-Attention-CRF model

1.6 bootstrapping 算法

bootstrapping 算法可以利用有限的样本资源进行多次重复抽样，从而建立起一个足以代表母体样本分布的新样本[19]。图 5 为 bootstrapping 算法训练流程示意图。bootstrapping 算法的训练过程如下：1）将一个数

据规模较小的标注数据集作为初始数据集，NER 模型在初始数据集上进行训练；2）训练好的 NER 模型用于预测未标注医疗语料，得到候选标注语料，bootstrapping 算法通过设置阈值来筛选候选标注语料中置信度较高的语料，将符合要求的候选标注语料合并到初始标注数据集中，再次训练 NER 模型；3）重复步骤 2，迭代训练 NER 模型，迭代终止的条件是达到设定的最大训练次数或者模型的识别效果不再提升。

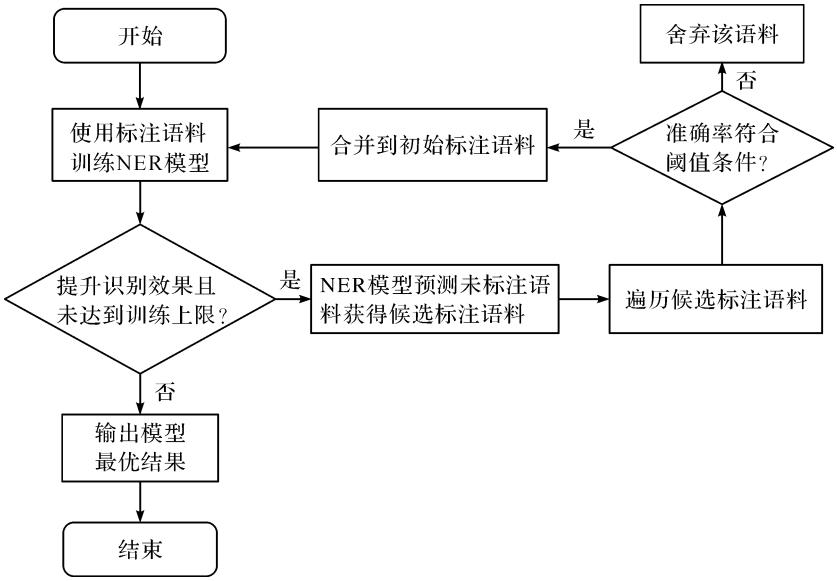

图 5　bootstrapping 算法训练流程示意图

Fig. 5　Schematic diagram of bootstrapping algorithm training process

2　试验及分析

2.1　数据集

2.1.1　医疗领域未标注数据集

利用爬虫技术爬取："疾病百科问药网"的疾病简介、症状描述、医治方案等文本，"寻医问药网"等在线咨询网站中的疾病信息和诊断依据文本，"中国公众健康网"中的疾病症状、诊断、治疗方案等医疗方案文本，以及 CCKS 比赛公开的无标注电子病历及我们从合作医院获得的中文电子病历。对上述医疗文本经过数据清洗，删除包含特殊字符、无关医疗的噪声句子，再经过分句后最终获得一个包含 92 784 个医疗句子、总计大小为 1.5 GB 的结构化医疗未标注数据集。

2.1.2　CCKS CNER 医疗数据集

CCKS2017 数据集包含已标注的 400 份病历文档，将其中 300 份作为训练集用于模型训练，100 份作为测试集，数据集中标注了 5 类实体（症状体征、检查检验、疾病诊断、治疗方案、身体部位）。CCKS2018 数据集包含一个训练集和一个测试集，分别包括 600 个医疗记录和 400 个医疗记录，标注了 5 类实体（解剖部位、症状描述、独立症状、药物、手术）。2 个数据集的实体类型和数量统计分别见表 2 和表 3。

表 2　CCKS2017 数据集的实体类型及数量统计

Table 2　Statistics table of entity type and quantity in CCKS2017 data set　　个

数据集	症状体征	检查检验	疾病诊断	治疗方案	身体部位	实体总数
训练集	7 831	9 546	722	1 048	10 719	29 866
测试集	2 311	3 143	553	465	3 021	9 493

表 3　CCKS2018 数据集的实体类型及数量统计

Table 3　Statistics table of entity type and quantity in CCKS2018 data set　　个

数据集	解剖部位	症状描述	独立症状	药物	手术	实体总数
训练集	9 472	2 484	3 712	1 221	1 329	18 218
测试集	6 339	918	1 327	813	735	10 132

2.1.3 自建脑血管疾病数据集

我们从合作医院提供的 400 份临床电子病历中提取现病状、既往史、入院诊断、住院经过等医疗文本，总计获得 1 062 个文本，按句号划分，总计获得 4 920 个句子。使用 pkuseg 分词工具对句子进行分词，并组织 3 名医学专业人员结合专家构建的医疗词表对分词结果进行标注，获得一个脑血管疾病标注数据集。将标注完的 80%的句子作为训练集，剩下的 20%用于测试集。考虑医疗文本特征和后续构建医疗知识图谱所需的实体类型，我们定义了 4 种实体类型：1）身体部位，包含身体器官，如“冠状动脉”“小脑半球”“右侧上颌窦”；2）症状体征，由患病导致的不适，如“意识障碍”“头晕”“行走不稳”；3）疾病诊断，包含疾病名称、综合征，如“高血压”“脑梗死”“脑血栓”；4）治疗方案，包括治疗过程中涉及到的手术名称和药物名称，如“冠脉造影术”“阿司匹林片”。

脑血管疾病数据集的实体类型及数量统计见表 4。

表 4 脑血管疾病数据集的实体类型及数量统计

Table 4 Statistics table of entity type and quantity in cerebrovascular data set 个

数据集	症状体征	疾病诊断	治疗方案	身体部位	实体总数
训练集	3 756	786	1 526	4 394	10 462
测试集	1 211	493	568	1 361	3 633

2.2 试验基本设置

试验所用服务器配置如下：CPU 为 i7-8700，运行内存为 16 GB，显卡为 GeForce RTX 2080Ti。模型在 Ubuntu 操作系统下使用 PyTorch 1.12 进行训练。

BiLM_BNER 模型的参数主要包含 CbBERT 模型、BiLM 模型及 BiGRU-Attention-CRF 的参数。CbBERT 使用 BERT-base-Chinese-uncased 模型参数进行初始化，所以两者参数相同，共计 12 层双向 transformer，transformer 的隐藏层的维度设置为 768、注意力机制头数为 12、模型参数数量为个。BiLM 模型设置字向量维度为 128，LSTM 隐藏层的维度设置为 128，句子的批处理大小为 32 句，最大训练次数为 50，使用自适应时刻估计(adaptive moment estimation，Adam)算法优化模型参数，引入随机失活方法防止过拟合问题，BiLM 的参数配置见表 5。

BiGRU-Attention-CRF 设置最大训练次数为 100，批处理大小为 32 句，GRU 隐藏层的维度设置为 128，BiGRU-Attention-CRF 的参数配置见表 6。

表 5 BiLM 的参数配置

Table 5 Parameter table of BiLM

名称	值
字向量维度	128
LSTM 隐藏层维度	128
随机失活	0.5
优化器	Adam
学习率	0.001
批处理大小/句	32
训练次数	50

表 6 BiGRU-Attention-CRF 的参数配置

Table 6 Parameter table of BiGRU-Attention-CRF

名称	值
学习率	0.001
批处理大小/句	32
随机失活	0.5
优化器	Adam
训练次数	100
GRU 隐藏节点数	128
BiGRU 层数	1

2.3 评价指标

NER 任务中，通常采用准确率(precision，P)、召回率(recall，R)和 F_1 值作为评价模型性能的指标，3 个评价指标的计算公式如下：

$$\begin{cases} P=\dfrac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{P}}}\times 100\%; \\ R=\dfrac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{N}}}\times 100\%; \\ F_1=\dfrac{2PR}{P+R}。 \end{cases} \tag{3}$$

式(3)中：T_P 为预测结果和实际结果均为正例的样例个数；F_P 为实际结果为反例且预测结果为正例的样例个数；F_N 为实际结果为正例而预测结果为负例的样例个数。

2.4　试验结果与分析

2.4.1　bootstrapping 算法最优阈值试验

设计相邻阈值间隔为 0.1，阈值区间为 0.1 和 1.0 的试验，探究 bootstrapping 算法在不同医疗数据集上的最优阈值。BiLM_BNER 模型的 F_1 值随阈值变化示意图如图 6 所示。

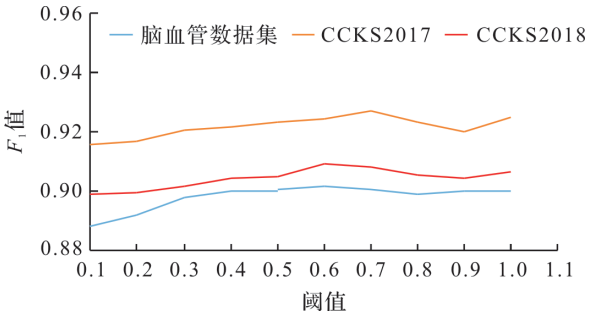

图 6　BiLM_BNER 模型的 F_1 值随阈值变化示意图

Fig. 6　Diagram of F_1 value variation with threshold value in BiLM_BNER model

当阈值为 0.6 时，BiLM_BNER 模型在自建脑血管数据集和 CCKS2018 数据集的 F_1 最高值分别为 90.16%和 90.93%。当阈值为 0.7 时，BiLM_BNER 模型在 CCKS2017 数据集的 F_1 值最高为 92.72%。试验结果表明，最优阈值与标注数据集的数据规模和实体类型数量、未标注数据集的数据规模有关，当标注数据较少时，需要降低阈值才能获得更多的候选标注数据。

2.4.2　不同命名实体识别模型对比试验

设计对比试验来验证 BiLM_BNER 模型在中文电子病历实体识别任务的有效性，试验对照组包括通用领域 NER 模型、引入额外特征(包括部首特征和词典特征)的 NER 模型和基于预训练语言模型的 NER 模型。具体如下：1) 通用领域 NER 模型，如 BiLSTM-CRF、BiGRU-CRF、BERT-CRF 及 Qiu 等[20]提出的残差扩张卷积神经网络(RDCNN-CRF)模型。2) 引入额外特征的 NER 模型，如 Li 等[21]提出的融合医疗词典信息和汉字偏旁部首特征的中文医疗 NER 模型，Wang 等[17]提出的结合领域字典和 BiLSTM-CRF 的 NER 模型。3) 基于预训练语言模型的 NER 模型，如 Luo 等[22]提出的基于笔画 ELMO 和多任务学习的电子病历 NER 模型，Tang 等[10]提出的一种融合预训练双向语言模型和自注意力机制的 NER 模型。不同命名实体识别模型对比试验结果见表 7。

表 7　不同命名实体识别模型对比试验结果

Table 7　Contrast test results of different named entity recognition models　　%

模型	自建脑血管数据集			CCKS2017 数据集			CCKS2018 数据集		
	P	R	F_1	P	R	F_1	P	R	F_1
BiLSTM-CRF	87.76	85.65	86.70	88.55	88.25	88.40	88.02	87.35	87.68
BiGRU-CRF	88.06	85.80	86.92	88.65	88.73	88.69	88.12	87.60	87.86
BERT-CRF	88.25	86.35	87.29	88.96	89.12	89.04	88.75	88.10	88.42
Qiu 等[20]				90.63	92.02	91.32			
Li 等[21]	89.65	88.23	88.93	92.06	91.15	91.60	89.59	89.54	89.56
Wang 等[17]				90.83	91.64	91.24			
Luo 等[22]						91.75			90.05
Tang 等[10]				90.83	91.64	91.24			
BiLM_BNER(本文模型)	90.52	89.81	**90.16**	92.86	92.59	**92.72**	91.30	90.56	**90.93**

试验结果表明，与通用领域 NER 模型相比，引入额外医疗特征(包括部首特征和词典特征)的 NER 模型和基于预训练语言模型的 NER 模型可有效地提升其识别精度。本研究提出的 BiLM_BNER 将字符特征、词典特征和上下文语义特征进行融合，提取电子病历文本的深层语义特征，将 bootstrapping 算法结合 NER 模型，利用小规模标注语料和大规模未标注语料进行迭代训练，筛选新的置信度高的标注语料添加到初始标注语料，不断扩大标注语料的规模，因此，BiLM_BNER 在脑血管数据集及 CCKS2017 和 CCKS2018 医疗数据集的准确率、召回率和 F_1 值上均优于对比试验中的其他 NER 模型，从而证明了本

模型在中文电子病历NER任务中的优良表现。

2.4.3 消融试验

设计消融试验来分析和验证CbBERT、BiLM、词典特征和bootstrapping算法对提升NER效果的作用，试验结果见表8。

表8 消融试验结果

Table 8 Ablation test results %

模型	自建脑血管数据集			CCKS2017数据集			CCKS2018数据集		
	P	R	F_1	P	R	F_1	P	R	F_1
BiGRU-Attention-CRF(Baseline)	87.62	85.30	86.44	88.30	88.12	88.21	88.05	87.45	87.75
dictionary+Baseline	88.86	86.50	87.66	90.13	89.66	89.89	89.58	88.72	89.15
CbBERT+Baseline	87.99	85.73	86.85	88.65	88.37	88.51	88.32	87.76	88.04
BiLM+Baseline	88.13	86.23	87.17	89.12	89.20	89.16	88.96	88.38	88.67
dictionary+BiLM+CbBERT+Baseline	89.76	88.58	89.17	90.58	90.32	90.45	90.05	89.48	89.76
bootstrapping+Baseline	88.65	86.65	87.64	89.96	89.80	89.88	89.75	89.15	89.45
BiLM_BNER(本文模型)	90.52	89.81	**90.16**	92.86	92.59	**92.72**	91.30	90.56	**90.93**

本试验将BiGRU-Attention-CRF作为基准模型，引入CbBERT、BiLM、词典特征、BiLM-CbBERT-词典特征进行试验。结果表明，引入CbBERT、BiLM、词典特征、BiLM-CbBERT-词典特征后，模型在脑血管数据集上的F_1值分别提升0.41%、0.73%、1.22%和2.73%，在CCKS2017数据集上F_1值分别提升0.3%、0.95%、1.68%和2.24%，在CCKS2018数据集上F_1值分别提升0.29%、0.92%、1.4%和2.01%。试验结果表明引入额外医疗特征可以提升NER任务的识别效果，融合多个医疗特征可以实现多角度提取文本语义信息，提升效果更加显著。

将bootstrapping算法引入基准模型后，模型在脑血管数据集、CCKS2017数据集和CCKS2018数据集的F_1值分别提升1.2%、1.67%和1.7%。将bootstrapping算法结合NER模型，利用小规模标注语料和大量易获得的未标注文本迭代训练模型，有效地提高了模型的识别精度。

3 结 语

本研究提出一种基于半监督学习的BiLM_BNER模型，通过构建中文电子病历文本的字符特征向量、上下文特征向量和词典特征向量，融合3个特征向量后输入BiGRU-Attention-CRF模型中进行序列标注，BiGRU-Attention-CRF可通过预测未标注语料得到候选标注数据。bootstrapping算法能筛选置信度高的候选标注语料，将其合并到初始标注语料进行迭代训练。BiLM_BNER模型在CCKS2017数据集、CCKS2018数据集和脑血管疾病数据集中均获得令人满意的实体识别效果。下一步我们将不断完善医疗词典，并尝试向模型引入词性特征、分词特征等，研究更高效的多特征融合方法，以提高模型的识别精度。

参考文献：

[1] 杨飞洪，张宇，覃露，等. 中文电子病历的命名实体识别研究进展[J]. 中国数字医学，2020，15(2)：4.

[2] 张俊遥. 基于深度学习的中文命名实体识别研究[D]. 北京：北京邮电大学，2019.

[3] HOCHREITER S，CHMIDHUBER J. Long short-term memory[J]. Neural Computation，1997，9(8)：1735.

[4] CHO K，VAN M B，GULCEHRE C，et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha：EMNLP，2014：1724.

[5] WANG C，WANG H，ZHUANG H，et al. Chinese medical named entity recognition based on multi-granularity semantic dictionary and multimodal tree[J]. Journal of Biomedical Informatics，2020，111：103583.

[6] JI B，LIU R，LI S，et al. A hybrid approach for named entity recognition in Chinese electronic medical record[J].

BMC Medical Informatics and Decision Making,2019,19(2):149.

[7] YIN M, MOU C, XIONG K, et al. Chinese clinical named entity recognition with radical-level feature and self-attention mechanism[J]. Journal of Biomedical Informatics,2019,98:103289.

[8] WU Y, HUANG J, XU C, et al. Research on named entity recognition of electronic medical records based on RoBERTa and radical Level Feature[J]. Wireless Communications and Mobile Computing,2021,2021(10):1.

[9] LIU Z, XU Y, YU T, et al. CrossNER: evaluating cross-domain named entity recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York City: AAAI,2021:13452.

[10] 唐国强,高大启,阮彤,等.融入语言模型和注意力机制的临床电子病历命名实体识别[J].计算机科学,2020,47(3):211.

[11] WEN G, CHEN H, LI H, et al. Cross domains adversarial learning for Chinese named entity recognition for online medical consultation[J]. Journal of Biomedical Informatics,2020,112:103608.

[12] YU H, MAO X L, CHI Z, et al. A robust and domain-adaptive approach for low-resource named entity recognition [C]//2020 IEEE International Conference on Knowledge Graph. Nanjing: IEEE,2020:297.

[13] DEVLIN J. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2018-10-11)[2022-04-25]. https://arxiv.org/abs/1810.04805.

[14] QIU X, SUN T, XU Y, et al. Pre-trained models for natural language processing: a survey[J]. Science China Technological Sciences,2020,63(10):26.

[15] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in Neural Information Processing Systems. Lake Tahoe: NIPS,2013:3111.

[16] LEE J, YOON W, KIM S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics,2020,36(4):1234.

[17] WANG Q, ZHOU Y, RUAN T, et al. Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition[J]. Journal of Biomedical Informatics,2019,92:103133.

[18] GAI R L, GAO F, DUAN L M, et al. Bidirectional maximal matching word segmentation algorithm with rules[J]. Advanced Materials Research,2014,926:3368.

[19] ABNEY S. Bootstrapping[C]//Proceedings of the 40th annual meeting of the Association for Computational Linguistics. Groningen: ACL,2002:360.

[20] QIU J, ZHOU Y, WANG Q, et al. Chinese clinical named entity recognition using residual dilated convolutional neural network with conditional random field[J]. IEEE Transactions on Nano Bioscience,2019:306.

[21] LI X, ZHANG H, ZHOU X H. Chinese clinical named entity recognition with variant neural structures based on BERT methods[J]. Journal of Biomedical Informatics,2020,107:103422.

[22] 罗凌,杨志豪,宋雅文,等.基于笔画 ELMo 和多任务学习的中文电子病历命名实体识别研究[J].计算机学报,2020,43(10):1943.