计算机科学
COMPUTER SCIENCE

融合机器阅读理解的中文医学命名实体识别方法

罗媛媛, 杨春明, 李波, 张晖, 赵旭剑

引用本文

罗媛媛, 杨春明, 李波, 张晖, 赵旭剑. 融合机器阅读理解的中文医学命名实体识别方法[J]. 计算机科学, 2023, 50(9): 287-294.

LUO Yuanyuan, YANG Chunming, LI Bo, ZHANG Hui, ZHAO Xujian. Chinese Medical Named Entity Recognition Method Incorporating Machine ReadingComprehension [J]. Computer Science, 2023, 50(9): 287-294.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220900226

融合机器阅读理解的中文医学命名实体识别方法

罗媛媛[1] 杨春明[1,3] 李 波[1] 张 晖[2] 赵旭剑[1,3]
1 西南科技大学计算机科学与技术学院 四川 绵阳 621000
2 西南科技大学数理学院 四川 绵阳 621000
3 四川省大数据与智能系统工程技术研究中心 四川 绵阳 621010
(2306543568@qq.com)

摘 要 医学命名实体识别是自动构建大规模医学知识库的关键,但医学文本中存在实体嵌套现象,采用序列标注的方法不能识别出嵌套中的实体。文中提出了基于阅读理解框架的中文医学命名实体识别方法,该方法将嵌套命名实体识别问题建模为机器阅读理解问题,使用 BERT 建立阅读理解问题和医学文本之间的联系,并引入多头注意力机制强化问题和嵌套实体之间的语义联系,最后用两个分类器对实体开头和结尾位置进行预测。与目前 5 种主流方法相比,该方法取得了最优结果,综合 F1 值达到了 67.65%;与经典的实体识别模型 BiLSTM-CRF 相比,F1 值提升了 7.17%,其中嵌套较多的临床表现实体提升 16.81%。

关键词:命名实体识别;中文医学;嵌套实体;机器阅读理解;多头注意力机制
中图法分类号 TP391.1

Chinese Medical Named Entity Recognition Method Incorporating Machine Reading Comprehension

LUO Yuanyuan[1], YANG Chunming[1,3], LI Bo[1], ZHANG Hui[2] and ZHAO Xujian[1,3]
1 School of Computer Science and Technology, Southwest University of Science and Technology, Mianyang, Sichuan 621000, China
2 School of Mathematics and Physics, Southwest University of Science and Technology, Mianyang, Sichuan 621000, China
3 Sichuan Big Data and Intelligent System Engineering Technology Research Center, Mianyang, Sichuan 621010, China

Abstract Medical named entity recognition is the key to automatically build a large-scale medical knowledge base. However, medical entities are often nested, and it can not be recognized by the sequence labeling method. This paper proposes a Chinese medical named entity recognition method based on reading comprehension framework. It models the nested named entity recognition problem as a machine reading problem, uses BERT to establish the connection between the reading comprehension problem and medical text, and introduces a multi-head attention mechanism to strengthen the semantic connection between the problem and nested named entity, and finally uses two classifiers to predict the beginning and end positions of entities. This method achieves the best results with an F1-score of 67.65% when compared with the current five mainstream methods. Compared with the most classical BiLSTM-CRF, the F1-score improves by 7.17%, and the nested "symptom" entities increase by 16.81%.

Keywords Named entity recognition, Chinese medical, Nested entities, Machine reading comprehension, Multi-head attention mechanism

1 引言

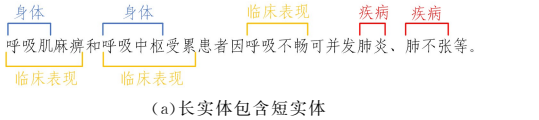
医学实体指蕴含在非结构化医学文本中的疾病、临床表现、身体、医疗程序等概念术语。从海量的医学教材、临床病例、检验报告、医学文献等非结构化医学文本中自动识别医学实体,是构建高质量医学知识库的关键。医学实体通常由专业术语构成,在不同类型的医学文本中常会出现实体嵌套现象,如"呼吸肌麻痹"是一个临床表现实体,同时里面嵌套了身体实体"呼吸肌",如图 1(a)所示。"HLA-DQA1 基因""脑脊"在无明确上下文提示时,既属于身体实体,也是医学检验项目实体,如图 1(b)所示。

由于中文医学文本的复杂性和专业性,要准确识别医学实体,不仅要识别出实体边界,还需明确实体的类别,这使得模型需要具有完备的特征表达能力和极强的特征提取能力。经典的命名实体识别模型采用序列标注的方法,即对医学文本的每个字符打上预设的标签,但当存在嵌套实体时,一个字符存在多个标签,就无法为医学实体打上合适的标签,不能准确识别出嵌套的实体。

因此,本文将嵌套实体识别问题看作是机器阅读理解问题,即通过对医学文本中需要识别的实体类型进行提问,以此

到稿日期:2022-09-23 返修日期:2022-12-02
基金项目:四川省科技厅重点研发项目(2021YFG0031);四川省省级科研院所科技成果转化项目(22YSZH0021)
This work was supported by the Key R&D Project of Science & Technology Department of Sichuan Province(2021YFG0031) and Scientific and Technological Achievements Transformation Project of Sichuan Provincial Scientific Research Institute(22YSZH0021).
通信作者:杨春明(yangchunming@swust.edu.cn)

来明确该类实体的边界。比如要识别图 1(a)中的身体类型实体,提问为:“哪一部分是文本中提到的身体?”即可将身体实体与临床表现实体区分开。同时,由于提问的问句带有先验信息,也能较好地识别出非嵌套的医学实体。该模型首先使用改进的 Chinese-robert-wwm-large[1] 模型构建机器阅读理解的编码部分和交互部分,建立问题和医学文本之间的联系,然后引入多头注意力机制(Multi-Head Attention Mechanism)[2]强化问题和嵌套医学实体之间的语义联系,最后通过全连接层和 softmax 函数计算将最终隐藏状态转化为答案跨度的概率。



(a)长实体包含短实体

(b)一个实体存在两种类别

图 1 中文医学实体的两种嵌套情况

Fig. 1 Two cases of nested Chinese medical entities

2 相关工作

命名实体识别任务通常被建模为序列标注任务,即对输入序列的每一个字符预测其标签,并计算出联合概率最大的标记组合。早期命名实体识别多采用机器学习的方法,比较经典的模型有隐马尔可夫(Hidden Markov Model,HMM)[3]、支持向量机(Support Vector Machine,SVM)[4]、条件随机场(Conditional Random Fields,CRF)[5]等。但机器学习的方法比较依赖于特征工程,在建立特征时耗时耗力。

随着深度学习在自然语言处理(Natural Language Processing,NLP)任务中的深入研究,基于深度学习的命名实体识别方法受到广泛关注。相比机器学习方法,基于深度学习的方法能容纳更丰富的语义信息,具有较强的特征提取能力。其中最典型的有卷积神经网络(Convolutional Neural Networks,CNN)[6]和循环神经网络(Recurrent Neural Network,RNN)[7]。基于 RNN-CRF[8]的方法在中文命名实体识别任务中取得了很好的效果。Xu 等[9]提出基于双向长短期网络(Bi-directional Long Short-Term Memory,BiLSTM)[10]和 CRF 的医学命名实体识别模型。Tang 等[11]提出了基于注意力的 CNN-LSTM-CRF 模型,用于识别中文临床文本中的实体。

此外,由于中文医学文本具有极强的专业性,且比较依赖语义信息,为了提高中文字词表征的多义性,研究者们在模型中增加预训练方法来对单词进行表征。早期最常见的预训练模型采用了 Word2vec[12]工具训练词向量,之后 BERT[13](Bidirectional Encoder Representations from Transformer)预训练模型被提出,BERT 一度成为了最受欢迎的预训练模型。Dai 等[14]提出基于 BERT-BiLSTM-CRF 的中文电子健康档案命名实体识别模型,结果明显优于非预训练的模型。Li 等[15]提出基于变异 BERT 结构的中国临床命名实体识别模型,利用未标记的特定领域知识,预先训练出未标记的中文医疗文本。

与常规的命名实体识别不同的是,中文医学文本标注语料较少、实体边界模糊、结构嵌套等难点给中文医学文本命名实体识别任务带来了极大的挑战。以往的方法忽略了实体嵌套结构问题,在标注时直接标注长度最大的实体。当嵌套结构较少时,对整体结果影响并不大,但是当嵌套结构较多时,对整体 F1 值的影响较大。

嵌套命名实体识别一般被看作多层次的序列标注问题,即根据嵌套实体的层数,每一层用一个命名实体识别模型进行识别[16]。Xu 等[17]采用双层 BiLSTM-CRF 方法来识别中医药文献中的实体。分层标注的方法存在层次越深、标签分布就越稀疏的问题,训练的难度也随之增加,容易造成层与层之间的错误传播。对此,有学者对此类序列标注模型进行了改进[18]。Sun 等[19]认为序列都是由一个个跨度组成的,给每个跨度打标签能解决嵌套问题,但需要在跨度的选择上加以限制。跨度的选择是一个复杂的问题,且其得到的负样本依然很多。此外,对多个子序列进行分类的计算成本很高,时空复杂度也较高。

另外,嵌套的实体也可被看作是状态之间的转换,并以此构建图来进行识别[20]。Wang 等[21]根据不同形式的词设计不同的动作,通过这些动作来处理不同的实体构建解析树,并根据实体的当前状态来决定是否打标签或是打更高层次的标签。另外,将嵌套实体所在的句子构建为超图[22],能有效捕捉长度不限的重叠的实体,使得实体的边界、类型和头部信息可以在一个框架中共同学习[23]。然而构建超图或者解析树依赖特定的转换系统,需要领域专家,不够一般化,且在构建时容易出现伪结构,在推导时会出现二义性,导致不能确定最终结果。

嵌套命名实体主要是实体的重叠问题(长实体覆盖短实体、同一实体表达出不同的类别),对其进行识别的关键是明确实体在句子中表达的语义,以此来确定实体的边界。机器阅读理解(Machine Reading Comprehension,MRC)[24-25]通过对句子提问来明确句子中实体的语义,能很好地改善实体重叠的问题[26]。如 Cao 等[27]提出基于 BERT 的机器阅读理解框架的中文电子病历嵌套实体识别方法,但未充分利用提问信息,采用不同提问方式得到的效果差异较大。Chiang 等[28]提出的基于 QA-SL 的中文电子健康记录命名实体识别框架就采用了机器阅读理解和序列标注融合的方法,但此方法的嵌套实体较为特殊,并不适合常规的中文医学命名实体识别。

为解决现有方法不能解决中文医学实体嵌套结构的问题,机器阅读理解方法不能完全利用提问信息,本文在 BERT-MRC[29]模型中引入多头注意力机制,进一步捕获医学实体和提问信息之间的依赖关系。实验结果表明,相比其他方法,该方法在中文医学数据集上取得了最好效果。

3 融合机器阅读理解的中文医学命名实体识别模型

3.1 模型描述

中文医学命名实体识别指识别并抽取出与医学临床相关的实体,并将实体归类到预定义好的类别。实体识别任务可被看作是阅读理解过程,即给出不同医学实体的提问或描述(Question),然后在医学文本(Context)中找出对应的答案(Answer)。其基本的流程为:分别对 Question 和 Context 进行嵌入和特征提取,然后在交互层将 Context 和 Question 的语义信息融合,使模型更容易找出 Context 中对应问题的答案部分,最后根据具体的问题任务来预测答案。具体流程如图 2 所示。

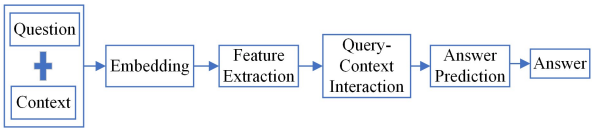


图 2 机器阅读理解流程

Fig. 2 Machine reading comprehension process

因此，利用机器阅读理解进行医学实体识别的任务可被描述为一个有监督的学习问题：给出三元组形式的训练数据(Context，Question，Answer)，其中 Context 表示含有嵌套实体的医学句子，Question 表示对对应实体的问题或描述，Answer 表示应当找出的实体答案。任务目标是学习一个预测器 f，能够将相关句子 Context 与问题 Question 作为输入，返回一个对应的答案 Answer 作为输出。

$$f:(Context,Qestion)\rightarrow Answer \tag{1}$$

其中，$Context$ 是每一个医学文本句子 $X=\{x_1,x_2,\cdots,x_n\}$，n 是句子的长度，实体类别标签集合 $Y=\{y_1,y_2,\cdots,y_k\}$(如疾病、药物和身体)。对于每一个类别标签类型 $y\in Y$，提出一个问句 $Q_y=\{q_1,q_2,\cdots,q_m\}$，其中 m 是问题的长度。每一个标注的实体 $x_{\text{start,end}}=\{x_{\text{start}},x_{\text{start}+1},\cdots,x_{\text{end}-1},x_{\text{end}}\}$ 就是答案，$x_{\text{start,end}}$ 是句子 X 的子串，类型是 y。下标 start，end 表示句子 X 中索引从 start 到 end 的序列，且 start≤end。对每一个句子 X，根据标签类别 y 生成的问题 Q_y，构造出三元组 $(X,Q_y,x_{\text{start,end}})$ 以进行训练。

综上，基于阅读理解的医学实体识别模型 BERT-MHAM-MRC 如图 3 所示，首先采用 BERT 将 Question 和 Context 序列转化为模型可识别的向量表示，并建立起问句和实体语句间的联系，然后进一步使用多头注意力机制聚焦句子中与问题关联的部分，最后融合 BERT 输出和多头注意力输出，用两个分类器分别预测实体答案是开头还是结尾的概率，其中实体的类型就是问句所代表的类型。

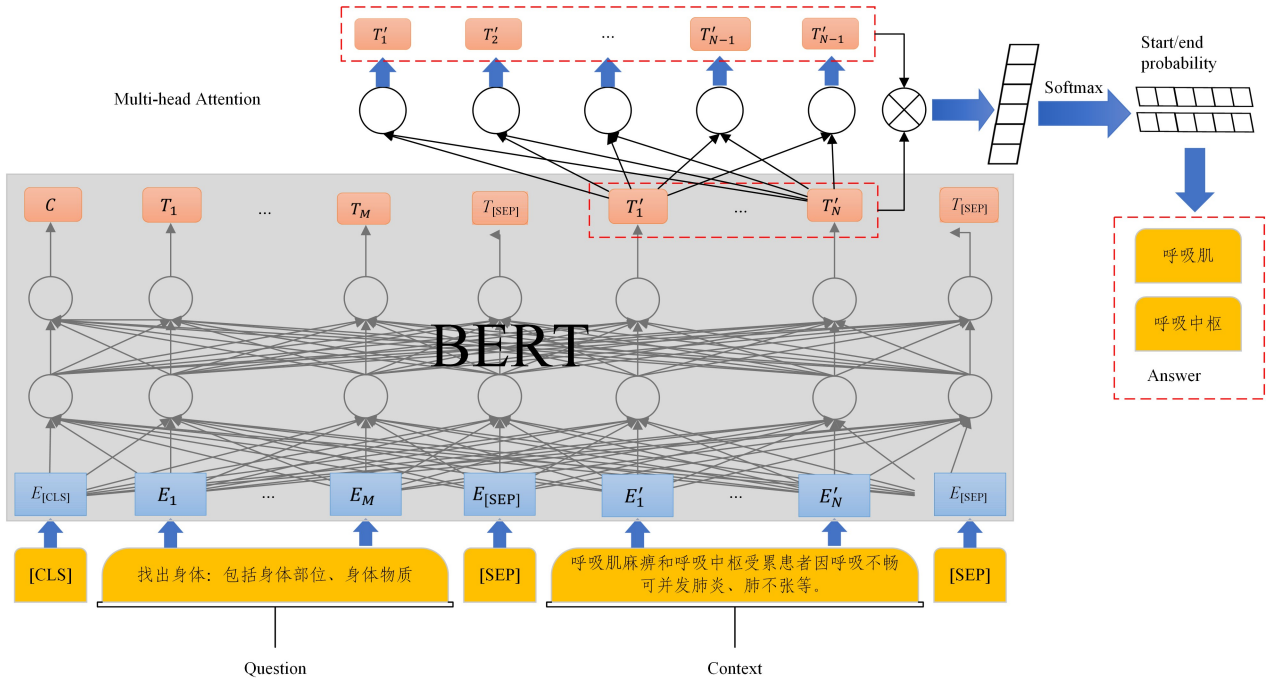


图 3 BERT-MHAM-MRC 模型

Fig. 3 BERT-MHAM-MRC model

3.2 阅读理解问题和医学文本嵌入

为了建立问句与实体所在句子的语义关联，本文采用 BERT[13] 来对输入进行嵌入。与基础 BERT 不同，Chinese-robert-wwm[1] 更加专注于中文数据集。为获取字符级的上下文关系，问句和句子均以字符作为输入，问题 Q_y 和句子 X 分别用[CLS]和[SEP]连接起来，输入如式(2)所示：

$$input=[CLS],q_1,q_2,\cdots,q_m,[SEP],x_1,x_2,\cdots,x_n,[SEP] \tag{2}$$

模型的输入表示由 token embedding(字符嵌入)、segment embedding(分段嵌入)，position embedding(位置嵌入)3 部分组成，如图 4 所示。

最终采用 BERT 最后一层隐藏层的上下文表示矩阵 $\boldsymbol{E}\in R^{n\times d}$ 来进行下一步操作，其中 d 是维度，n 是输入医学文本 Context 的长度。

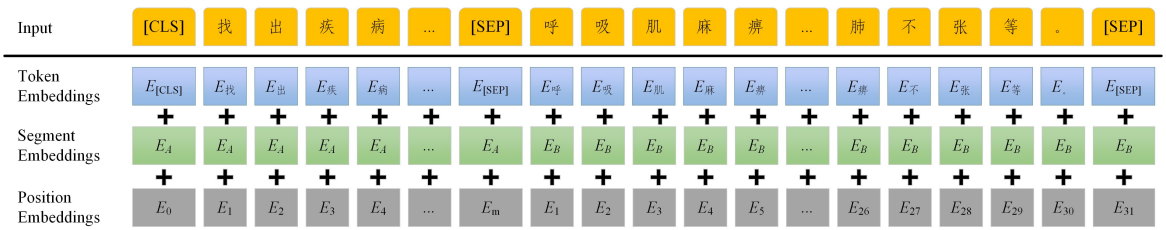


图 4 BERT-MHAM-MRC 模型的输入表示

Fig. 4 Input representation of BERT-MHAM-MRC model

3.3 多头注意力机制强化语义联系

尽管 BERT 充分利用了句子中的字符信息和位置信息，但针对某些实体类型的问句信息的利用依然较少[30]。注意力机制(Attention Mechanism)[31]本质为查询语句($\boldsymbol{Q}$)到目标语句($\boldsymbol{K}$-$\boldsymbol{V}$)的映射，通过将有限的注意力权重分配给不同的特征向量，能快速筛选出对实体贡献较大的关键信息。$\boldsymbol{Q}$,$\boldsymbol{K}$,$\boldsymbol{V}$ 都由输入的特征向量得到，能获取输入向量中局部关注的信息。多头注意力是进行多次注意力计算后的结果，使模型可以从不同空间学习语义特征。本文获取到 BERT 输出 $\boldsymbol{E}\in R^{n\times d}$ 后，采用多头自注意力(Multi-Head Self-Attention)[2]为其重新分配权重，这样能缓解 BERT 对实体问题信息利用不充分的问题。

注意力 *Attention* 的计算公式如式(3)所示，$head_i$ 是单头的注意力得分，最后将所有单头注意力进行拼接，得到多头注意力的输出 $MultiHead$。

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{3}$$

$$head_i=Attention(\boldsymbol{Q}\boldsymbol{W}_i^{\mathrm{Q}},\boldsymbol{K}\boldsymbol{W}_i^{\mathrm{K}},\boldsymbol{V}\boldsymbol{W}_i^{\mathrm{V}}) \tag{4}$$

$$MultiHead=Contact(head_1,head_2,\cdots,head_e) \tag{5}$$

其中，$\boldsymbol{Q}=\boldsymbol{K}=\boldsymbol{V}=\boldsymbol{E}$；$d_k$ 为缩放因子，用于缓解内积过大产生的梯度弥散问题；$\boldsymbol{W}_i^{\mathrm{Q}}$,$\boldsymbol{W}_i^{\mathrm{K}}$,$\boldsymbol{W}_i^{\mathrm{V}}$ 为神经网络权重参数。最终将经过多头自注意力的输出和原本的 BERT 输出按照 1∶1 的比例联合起来作为整体输出，并将这个输出用于分类。

3.4 答案预测

为了能准确识别出嵌套实体，将上一步融合的多头注意力和 BERT 的输出向量连接起来作为整体的输出表示，由于输出已经包含了答案实体的信息，具备了生成答案的条件，此时用两个 softmax 分类器分别预测每一个 token 是开头还是结尾的概率，并将该概率映射到坐标，然后将开始和结尾坐标进行 sigmoid 匹配约束，得到最终答案。通过 softmax 分类器得到的开始或结束的概率计算如下：

$$P_{\mathrm{start}}=\mathrm{softmax}(\boldsymbol{E}\cdot\boldsymbol{T}_{\mathrm{start}})\in R^{n\times 2} \tag{6}$$

$$P_{\mathrm{end}}=\mathrm{softmax}(\boldsymbol{E}\cdot\boldsymbol{T}_{\mathrm{end}})\in \boldsymbol{R}^{n\times 2} \tag{7}$$

其中，$\boldsymbol{T}_{\mathrm{start}}$,$\boldsymbol{T}_{\mathrm{end}}\in R^{d\times 2}$ 是在训练过程中学习到的参数矩阵，$\boldsymbol{E}$ 是上层多头自注意力和 BERT 融合的输出，P 代表该位置是实体开始或结束的概率。随后需要将概率映射到实际坐标，对每一行的概率做 argmax，能得到两个长度为 n 的 0-1 序列。如第 k 个位置是 1，说明第 k 个位置就是实体开始或结束的位置。

$$\hat{I}_{\mathrm{start}}=\{i\,|\,\mathrm{argmax}(P_{\mathrm{start}}^{(i)})=1,i=1,\cdots,n\} \tag{8}$$

$$\hat{I}_{\mathrm{end}}=\{j\,|\,\mathrm{argmax}(P_{\mathrm{end}}^{(j)})=1,j=1,\cdots,n\} \tag{9}$$

得到开始坐标和结尾坐标的 0-1 序列后，需要将开始和结尾的实际坐标匹配才能得到最终的实体答案，对 $\hat{I}_{\mathrm{start}}$ 中每个为 1 的位置 i，和 $\hat{I}_{\mathrm{end}}$ 中每个为 1 的位置 j(满足 $i\leqslant j$ 的连续字符序列 $x_{i,j}$)，计算 $x_{i,j}$ 是实体且类型是 y 的概率 $P_{i,j}$ 的公式如式(10)所示，采用 sigmoid 函数进行匹配约束，其中 $\boldsymbol{m}\in R^{1\times 2d}$ 是需要学习的向量。

$$P_{i_{\mathrm{start}},j_{\mathrm{end}}}=\mathrm{sigmoid}(\boldsymbol{m}\cdot concat(\boldsymbol{E}_{i_{\mathrm{start}}},\boldsymbol{E}_{j_{\mathrm{end}}})) \tag{10}$$

4 实验与结果分析

4.1 数据集与评价指标

本文使用 CHIP2020 Task1 所发布的中文医学文本命名实体识别数据集，将医学文本命名实体分为 9 类，分别是疾病、临床表现、药物、医疗设备、医疗程序、身体、医学检验项目、微生物类和科室。

表 1 列出了 9 种命名实体的类型、描述和示例。标注之前对文章进行自动分词处理，所有的医学实体均已正确切分。原标注文件中“临床表现”实体存在嵌套，该实体内部允许存在其他 8 类实体。另外限制每一条数据的最大长度为 160 个字符，最后总共有 15000 条数据和 61791 个实体，按照 8∶1∶1 的比例随机划分为训练集、验证集和测试集，实体分布如表 2 所列。

表 1 CHIP2020 数据集详细说明

Table 1 Detailed description of CHIP2020 dataset

实体类型	描述	子类	样例
疾病(disease)	疾病指导致病人处于非健康状态的原因或者医生对病人做出的诊断，并且是能够被治疗的	疾病或综合征；中毒或受伤；器官或细胞受损	尿潴留、泌尿系感染
临床表现(symptom)	临床表现是疾病的表现，泛指患者的不适感觉以及通过检查得知的异常表现	症状；体征	呼吸困难、阵发性喘憋
医疗程序(procedure)	医疗程序泛指诊断或治疗所采取的措施、方法及过程	检查程序；治疗或预防程序	免疫学方法检测、抗体检测
医疗设备(equipment)	医疗设备泛指诊断或治疗所使用的工具、器具、仪器等	检查设备；治疗设备	显微镜
药物(drug)	药物指用来预防、治疗及诊断疾病的物质	药物	EBV 疫苗
医学检验项目(item)	医学检验项目指检查涉及的体液检查项目、重要生理指标以及其他检查项目，本文规定“医疗检验项目”主要针对人体而言，是能够通过设备或实验检测出的项目，并且能够被量化，有其对应的测量值或指标值	医学检验项目	渗透压、肾溶质负荷、热能密度、黏稠度
身体(body)	身体泛指细胞、组织以及位于人体特定区域的由细小物质成分组合而成的结构、器官、系统、肢体，另外包括身体产生或解剖身体产生的物质等	身体物质；身体部位	脾、脾脏
科室(department)	科室主要指医院或医疗机构所设有的科室	科室	眼科
微生物类(microbes)	微生物类包括细菌、病毒、真菌以及一些小型的原生生物、显微藻类等在内的一大类生物群体，另外包括微生物类产生的毒素、激 素、酶等	微生物类	寄生虫成虫、寄生虫

表 2 CHIP2020 数据集实体分布情况

Table 2 Entities distribution of CHIP2020 dataset

数据集	疾病	临床表现	医疗程序	医疗设备	药物	医学检验项目	身体	科室	微生物类
训练集	12824	9894	5075	709	3112	2006	14026	259	1506
验证集	1525	1165	619	103	409	237	1848	30	205
测试集	1490	1209	638	76	409	338	1823	59	197

按照机器阅读理解的输入格式，需要对输入的医学文本分别构建问题，将训练集、验证集和测试集构造为机器阅读理解能识别的(Context，Question，Answer)三元组数据。以图 1(a)中的医学文本为例，“Context：呼吸肌麻痹和呼吸中枢受累患者因呼吸不畅可并发肺炎、肺不张等”，需要构建如表 3 所列的 Question 和 Answer，其中 Question 是基于关键字的方式进行构造的。评价指标精确率 Precision、召回率 Recall 和 F1 值的计算公式如下所示：

$$Precision=\frac{TP}{TP+FP} \tag{11}$$

$$Recall=\frac{TP}{TP+FN} \tag{12}$$

$$F1=\frac{2*Precision*Recall}{Precision+Recall} \tag{13}$$

其中，TP 表示正确识别当前实体类别的样本数量；FP 表示错误识别当前类别的样本数量；FN 表示本属于当前类别但没有被识别到的样本数量；$TP+FP$ 则表示被识别为当前类别的所有样本数量；$TP+FN$ 则表示被标注为当前类别的所有样本数量。

表 3 机器阅读理解三元组数据格式例子

Table 3 Example of machine reading comprehension triple data format

Question	Answer
找出疾病：包括疾病或综合症、中毒或受伤、器官或细胞受损	[肺炎，肺不张]
找出临床表现：包括症状、体征	[呼吸肌麻痹，呼吸中枢受累，呼吸不畅]
找出医疗程序：包括检查程序、治疗或预防程序	—
找出医疗设备：包括检查设备和治疗设备	—
找出药物：包括药物	—
找出医学检验项目：包括医学检验项目	—
找出身体：包括身体部位和身体物质	[呼吸肌，呼吸中枢]
找出科室：包括科室	—
找出微生物类：包括微生物类	—

4.2 实验及参数设置

BERT-MHAM-MRC 模型在 CHIP2020 上的参数设置如下：BERT 采用哈工大改进的 Chinese-robert-wwm-large 模型；优化器采用 Adamw；学习率是 8×10^{-6}；设置 BERT 隐藏单元为 1536；添加注意力头数的个数为 8；最小批处理尺寸为 6；Dropout 为 0.5。

由于机器阅读理解中问题带有实体的先验知识，对最终结果的影响较为明显[30]，因此构造合适的问题是非常重要的。在实际构造问题的过程中，可根据对实体类型的不同维度(如实体解释说明、实体位置等)构造出不同问题，每种方法会产生不同的效果。一般来说，问题构造需要尽可能区别出实体类别，对问题的形式没有特别要求。在 BERT-MHAM-MRC 模型上比较了两种不同的问题构造方法对中文医学命名实体的影响，一种是基于关键词的构造方法，即表 3 所用的提问方法；另一种是按照注释指南(Annotation Guideline Notes)方式，即数据构建者提供的实体类型说明，采用表 1 中所列数据集的描述。两种方法的结果如表 4 所列。

表 4 两种问题生成方式的 F1 值

Table 4 F1 values of two question generation methods

(单位：%)

提问方式	疾病	临床表现	医疗程序	医疗设备	药物	医学检验项目	身体	科室	微生物类	综合
注释指南	77.17	55.06	63.09	62.60	79.10	48.79	66.74	69.31	71.35	66.84
关键字	77.42	55.84	63.60	65.08	79.76	49.82	67.93	71.43	71.73	67.65

另外，由于要对比普通序列标注方法和 BERT-MHAM-MRC 方法的优劣，因此普通序列标注方法采用最大标记法，即存在长实体中嵌套短实体的情况，只标注最长实体，预测时也只需预测出最长实体。选取 4 种序列标注模型和 1 种融合机器阅读理解的模型与 BERT-MHAM-MRC 进行对比。

CRF[5]：CRF 是给定一组输入序列的条件下，另一组输出序列的条件概率分布模型。使用基于 CRFsuite 库的轻量级 sklearn-crfsuite 工具包，使用的特征为“前一个词，当前词，后一个词，前一个词＋当前词，当前词＋后一个词”。CRF 也是序列标注模型的基线模型。

BiLSTM-CRF[10]：使用双向 LSTM 网络和 CRF 的经典模型。

BERT-CRF：采用 BERT 预训练和 CRF 相结合的模型进行命名实体识别。

BERT-BiLSTM-CRF[14]：在经典的 BiLSTM-CRF 模型上添加 BERT 预训练模型。

BERT-MRC[29]：采用 BERT 做机器阅读理解的方法进行命名实体识别，也是机器阅读理解模型的基线模型。

图 5 给出了 6 种模型在 CHIP2020 数据集上的整体 F1 值对比，具体准确率、召回率和 F1 值如表 5 所列。

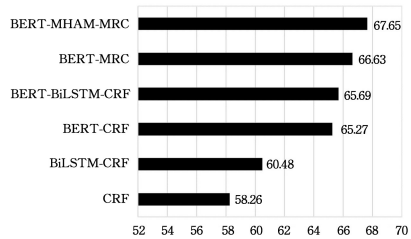


图 5 6 种模型整体 F1 值对比

Fig. 5 Comparison of overall F1 values of six models

表 5 不同模型在各个实体类别上的效果

Table 5 Effects of different models on various entity classes

(单位:%)

模型	指标	疾病	临床表现	医疗程序	医疗设备	药物	医学检验项目	身体	科室	微生物类	综合
CRF	*P*	70.92	42.09	61.77	64.00	73.51	50.67	59.56	68.57	66.67	60.92
	R	70.77	35.33	50.54	45.07	64.76	32.76	59.30	66.67	64.05	55.82
	*F*1	70.84	38.41	55.59	52.89	68.86	39.79	59.43	67.61	65.33	58.26
BiLSTM-CRF	*P*	70.27	41.63	66.98	45.54	63.55	37.35	61.98	64.86	50.94	59.23
	R	71.90	36.73	65.95	64.79	79.28	52.16	62.51	66.67	88.24	61.79
	*F*1	71.08	39.03	**66.46**	53.49	70.55	43.53	62.24	65.75	64.59	60.48
BERT-CRF	*P*	73.33	51.33	61.78	50.00	80.26	58.50	60.74	78.26	76.97	63.70
	R	80.05	49.96	62.86	55.56	78.01	37.23	71.60	72.00	73.41	66.92
	*F*1	76.55	50.64	62.32	52.63	79.12	45.50	65.73	**75.01**	75.15	65.27
BERT-BilSTM-CRF	*P*	72.51	52.02	64.28	56.76	80.71	40.57	63.71	74.28	76.74	64.33
	R	81.60	49.41	64.18	63.64	81.33	42.86	68.21	70.27	76.30	67.10
	*F*1	76.79	50.68	64.23	60.00	81.02	41.68	65.88	72.22	**76.52**	65.69
BERT-MRC	*P*	77.47	63.37	66.04	75.44	77.12	57.47	70.13	86.84	81.51	71.02
	R	76.02	48.22	59.97	56.58	79.95	41.79	62.87	61.11	58.91	62.75
	*F*1	76.74	54.77	62.85	64.66	78.51	48.39	66.30	71.74	68.39	66.63
BERT-MHAM-MRC	*P*	75.34	56.59	64.38	74.55	76.48	58.75	66.38	83.33	73.26	67.46
	R	79.62	55.10	62.84	57.75	83.33	43.25	69.56	62.50	70.26	67.83
	*F*1	**77.42**	**55.84**	63.60	**65.08**	**79.76**	**49.82**	**67.93**	71.43	71.73	**67.65**

4.3 实验结果与分析

4.3.1 不同问题生成方式的对比分析

从表 4 中可以发现,相比基于注释指南构造问题的方法,基于关键字的构造方法的 F1 值提升了 0.81%。这表明对实体的描述并没有让 BERT 获取到最相关的信息,而关键词直接指向关键信息,可在一定程度上提升模型的效果。因此在后续的实验中采用基于关键字的构造问题方式。

4.3.2 模型对比与分析

从图 5 可以发现,BERT-MHAM-MRC 相比其他 5 种模型都取得了最佳效果。相较于 CRF 模型,BERT-MHAM-MRC 模型的 F1 值提升了 9.39%,比经典 BiLSTM-CRF 模型提升了 7.17%,相比 BERT-CRF 和 BERT-BiLSTM-CRF 分别提升了 2.38%和 1.96%,这两个模型的结果差距不大。BERT-BiLSTM-CRF 相比 BiLSTM-CRF 提升较大,F1 值提升了 5.21%,说明 BERT 预训练模型能显著影响实验结果。BERT-MRC 也采用机器阅读理解的方法识别命名实体,相比之前的序列标注模型提升也较大,说明机器阅读理解模型能有效解决命名实体识别任务。BERT-MHAM-MRC 在 BERT-MRC 的基础上增加了多头注意力机制,F1 值提升了 1.02%,说明增加多头注意力使模型更加专注于医学文本中实体的关键信息,能有效提升在中文医学文本上的识别效果。

从表 5 可以发现,不同类型的实体识别结果差异较大,BERT-MHAM-MRC 在不同实体上的提升差异也较大。综合表 2 中的内容,这种差异与实体的数量、长度和组成结构呈现一定的相关性,但同时也与实体嵌套数量呈现较大的相关性。对 CHIP2020 中每一类型实体包含嵌套的数量进行统计(即此类实体中含有其他实体),结果如表 6 所列。从表 6 可以发现,临床表现实体存在嵌套的情况最多,占总嵌套数量的 98.78%。结合表 6 和表 5 进行分析,BERT-MHAM-MRC 相比 CRF 在临床表现实体上 F1 值提升了17.43%,比 BiLSTM-CRF 提升了16.81%。由于 CRF,BiLSTM-CRF,BERT-CRF 和 BERT-BiLSTM-CRF 均采用最大标注法进行实体识别,存在嵌套的情况被忽略。例如,"呼吸肌麻痹"是一个"临床表现"实体,但其中嵌套了"呼吸肌"这样一个身体实体。序列标注模型极易只标注"呼吸肌"这个身体实体,这也是所有的序列标注模型中临床表现实体效果都不太好的原因。而 BERT-MHAM-MRC 采用对实体类型提问的方式,能分别识别临床表现实体和其他实体,解决了嵌套的问题。

表 6 CHIP2020 中每类实体存在嵌套的数量

Table 6 Number of nested entities of each type in CHIP2020

疾病	临床表现	医疗程序	医疗设备	药物	医学检验项目	身体	科室	微生物类
1	2843	15	0	1	4	13	1	0

BERT-MHAM-MRC 不仅在嵌套实体上的识别效果较好,在非嵌套和嵌套数量极少的实体类别上表现也较好,分别都有不同程度的提升。这说明 BERT-MHAM-MRC 不仅能较好地解决嵌套命名实体识别问题,在非嵌套的实体识别上表现也较好。

BERT-BiLSTM-CRF 和 BERT-CRF 结果相差并不大,实际上,BERT-BiLSTM-CRF 花费的时间是 BERT-CRF 的数倍,但在某些实体上的效果却不如 BERT-CRF,说明简单地叠加模型提升效果不一定明显。但 BERT-MHAM-MRC 针对阅读理解对中文医学问题利用不充分的缺点增加了多头注意力机制,在 BERT-MRC 模型上的提升还是比较明显。

从表 5 中还可以发现疾病、药物实体的效果较好,原因之一是这两种实体结构单一、分布密集。除此之外,科室和微生物实体虽然分布稀疏,但最终 F1 值都达到了 70%以上,这是因为这两种实体实际类别较少且结构更加简单,如"内科""外科""儿科""XX 真菌""XX 病毒"等。但 BERT-MHAM-MRC 的效果却不如序列标注模型的效果,或许是因为序列标注模型在结构更加单一的数据上表现更好,而机器阅读理解虽然增加了问题交互模块,但其中复杂的运算反而对此类数据不够友好。而身体实体分布最密集,但是效果却不如疾病实体,因为

身体实体包含了维生素、细胞、蛋白质等组成元素，构成结果相对复杂。另外，医学检验项目实体的效果相对最差，经探究发现，医学检验项目实体中存在大量英文字符和特殊字符，如“胸部X线检查”“尿VMA”“PRA”等专业检查指标，并且还有“α”“β”等特殊的字符，这些原因极容易导致实体识别错误。不管是传统命名实体识别方法还是融合机器阅读理解的方法，都不能有效识别中英文夹杂的实体。

4.4 消融实验

对于中文医学命名实体识别来说，采用的BERT-MHAM-MRC模型结合了BERT和多头注意力机制进行机器阅读理解建模，能同时识别出嵌套和非嵌套的实体。但常规序列标注模型不能识别出嵌套的实体，为了探究BERT和多头注意力对实验的影响，对4.2节的两个基线模型添加消融实验，实验结果如表7所列。

表7 消融实验

Table 7 Ablation experiment

（单位：%）

Settings	$F1$	提升
CRF	58.26	
＋BiLSTM	60.48	＋2.22
＋BERT	65.27	＋7.01
＋BiLSTM＋BERT	**65.69**	**＋7.43**
BERT-MRC	66.63	
＋Multi-head Attention(Ours)	**67.65**	**＋1.02**

常规序列标注CRF模型分别加入BiLSTM网络和BERT预训练模型后F1值分别提升了2.22%和7.01%，而融合了BiLSTM和BERT的模型F1值提升了7.43%，说明序列标注模型中，BERT预训练模型对整个实验效果的提升较大。BiLSTM网络也能提升实验效果，两者结合对整个模型的提升会更大，说明BERT和BiLSTM网络都能提升模型效果，且两者结合提升效果会更大。

另外，用BERT构建机器阅读理解模型能解决嵌套命名实体识别的问题，因此其F1值比所有序列标注模型都更好。加入多头注意力机制F1值提升1.02%，说明在机器阅读理解过程中引入多头注意力机制能强化阅读理解的效果，提升整个实验的效果。

结束语 本文提出的BERT-MHAM-MRC模型通过将命名实体识别任务建模为机器阅读理解问题，充分利用BERT语言模型捕获句子字符的上下文信息，使句子和问题字符的上下文信息接近于实体类别的问题信息，然后利用多头注意力网络强化问题和嵌套实体的联系，增加了阅读理解的效果。实验结果表明，该模型在中文医学实体识别任务中取得了更好的效果，尤其是在存在较多嵌套的实体中提升更加明显，相比经典的BiLSTM-CRF模型，整体F1值提升了7.17%，其中在嵌套较多的临床表现实体上F1值提升16.81%，充分证明了BERT-MHAM-MRC能有效解决嵌套问题，但模型针对复杂的中英文夹杂实体、包含特殊字符的实体识别效果依然较差。在未来的工作中，希望能加入额外的规则来提升中英文夹杂实体的效果，另外希望能够对整个模型进行优化，提高模型的效率。

参 考 文 献

[1] CUI Y，CHE W，LIU T，et al. Revisiting Pre-Trained Models for Chinese Natural Language Processing[C]// Empirical Methods in Natural Language Processing. 2020. 657-668.

[2] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017：6000-6010.

[3] MORWAL S，JAHAN N，CHOPRA D . Named entity recognition using hidden Markov model [J]. International Journal on Natural Language Computing，2012，4(1)：15-23.

[4] JU Z，WANG J，ZHU F. Named entity recognition from biomedical text using SVM[C]// 2011 5th International Conference on Bioinformatics and Biomedical Engineering. IEEE，2011：1-4.

[5] SONG S，ZHANG N，HUANG H. Named entity recognition based on conditional random fields[J]. Cluster Computing，2019，22(3)：5195-5206.

[6] GUI T，MA R，ZHANG Q，et al. CNN-Based Chinese NER with Lexicon Rethinking[C]// International Joint Conference on Artificial Intelligence. Macao，China，2019：4982-4988.

[7] CHOWDHURY S，DONG X，QIAN L，et al. A multitask bi-directional RNN model for named entity recognition on Chinese electronic medical records [J]. BMC Bioinformatics，2018，19(17)：75-84.

[8] OUYANG E，LI Y，JIN L，et al. Exploring n-gram character presentation in bidirectional RNN-CRF for Chinese clinical named entity recognition[C]// CEUR Workshop Proceedings. 2017：37-42.

[9] XU K，ZHOU Z，HAO T，et al. A bidirectional LSTM and conditional random fields approach to medical named entity recognition [C]// International Conference on Advanced Intelligent Systems and Informatics. Cham：Springer，2017：355-365.

[10] HUANG Z，XU W，YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv：1508.01991，2015.

[11] TANG B，WANG X，YAN J，et al. Entity recognition in Chinese clinical text using attention-based CNN-LSTM-CRF[J]. BMC Medical Informatics and Decision Making，2019，19(3)：89-97.

[12] MIKOLOV T，CHEN K，CORRADO G，et al. Efficient estimation of word representations in vector space[J]. arXiv：1301.3781，2013.

[13] DEVLIN J，CHANG M W，LEE K，et al. Bert：Pre-training of deep bidirectional transformers for language understanding[J]. arXiv：1810.04805，2018.

[14] DAI Z，WANG X，NI P，et al. Named entity recognition using BERT BiLSTM CRF for Chinese electronic health records[C]// 2019 12th International Congresson Image and Signal Processing，Biomedical Engineering and Informatics(CISP-BMEI). 2019：1-5.

[15] 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.

[16] JU M, MIWA M, ANANIADOU S. A neural layered model for nested named entity recognition[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, Louisiana, 2018: 1446-1459.

[17] XU H, LIU H, JIA Q, et al. A nested named entity recognition method for traditional Chinese medicine records[C]// International Conference on Artificial Intelligence and Security. Cham: Springer, 2021: 488-497.

[18] ZHENG C, CAI Y, XU J, et al. A Boundary-aware Neural Model for Nested Named Entity Recognition[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 357-366.

[19] SUN L, SUN Y, JI F, et al. Joint Learning of Token Context and Span Feature for Span-Based Nested NER [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 2720-2730.

[20] MARINHO Z, MENDES A, MIRANDA S, et al. Hierarchical nested named entity recognition[C]// Proceedings of the 2nd Clinical Natural Language Processing Workshop. 2019: 28-34.

[21] WANG B, LU W, WANG Y, et al. A Neural Transition-based Model for Nested Mention Recognition[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 1011-1017.

[22] LU W, ROTH D. Joint mention extraction and classification with mention hypergraphs[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon, Portugal, 2015: 857-867.

[23] KATIYAR A, CARDIE C. Nested Named Entity Recognition Revisited[C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018: 861-871.

[24] LEVY O, SEO M, CHOI E, et al. Zero-shot relation extraction via reading comprehension[J]. arXiv: 1706.04115, 2017.

[25] LIU J, CHEN Y, LIU K, et al. Event extraction as machine reading comprehension[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020: 1641-1651.

[26] LIU S, ZHANG X, ZHANG S, et al. Neural machine reading comprehension: Methods and trends[J]. Applied Sciences, 2019, 9(18): 3698.

[27] CAO J, ZHOU X, XIONG W, et al. Electronic Medical Record Entity Recognition via Machine Reading Comprehension and Biaffine[J]. Discrete Dynamics in Nature and Society, 2021, 2021(9): 16408371-1-16408371-8.

[28] CHIANG Y L, LIN C H, SUNG C L, et al. Nested Named Entity Recognition for Chinese Electronic Health Records with QA-based Sequence Labeling[C]// Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing. 2021: 18-25.

[29] LI X, FENG J, MENG Y, et al. A unified MRC framework for named entity recognition[J]. arXiv: 1910.11476, 2019.

[30] YANG P, CONG X, SUN Z, et al. Enhanced Language Representation with Label Knowledge for Span Extraction[J]. arXiv: 2111.00884, 2021.

[31] MNIH V, HEESS N, GRAVES A. Recurrent models of visual attention[J]. Advances in Neural Information Processing Systems, 2014, 27: 2204-2212.

LUO Yuanyuan, born in 1998, postgraduate, is a member of China Computer Federation. Her main research interests include knowledge graphs and natural language processing.

YANG Chunming, born in 1980, associate professor, is a member of China Computer Federation. His main research interests include nature language processing and machine learning.

(责任编辑:杨雪敏)