预训练语言模型在中文电子病历命名实体识别上的应用

Application of Pre-Training Language Model in Chinese EMR Named Entity Recognition

吴小雪,张庆辉(河南工业大学,河南 郑州 450001)

Wu Xiao-xue,Zhang Qin-hui(Henan University of Technology,Henan Zhengzhou 450001)

摘 要:中文预训练语言模型能够表达句子丰富的特征信息,并且可解决针对中文出现的"一词多义"问题,是当前自然语言处理任务中普遍使用的模型。研究预训练模型在中文电子病历命名实体识别任务上的应用,为基于深度学习的中文电子病历信息抽取探索一种信的信息优化方法。该文首先介绍了四种语言预训练模型 BERT,ERNIE,ALBERT,NEZHA,并搭建预训练模型、BiLSTM、CRF 的融合结构,在 CCKS2018 中文电子病历数据集上进行医学命名实体识别任务。实验结果表明 NEZHA 取得了当前预训练模型最优的识别结果。

关键词:预训练模型;命名实体识别;电子病历

中图分类号:TP391.1 文献标识码:A 文章编号:1003-0107(2020)09-0061-05

Abstract:Chinese pre-training language model can express rich feature information of sentences and solve the problem of "polysemy" in Chinese.It is widely used in natural language processing tasks.This paper studies the application of pre-training model in Chinese EMR named entity recognition task,and explores an informa- tion optimization method for Chinese EMR information extraction based on deep learning.Firstly,this paper introduces four language pre-training models: BERT,ERNIE,AlBERT,NEZHA,and builds the fusion structure of pre-training model,BILSTM and CRF,and carries out medical named entity recognition task on ccks2018 Chinese electronic medical record data set.The experimental results show that NEZHA achieves the best recognition result of the current pre-training model.

Key words:pre-training model;named entity recognition;electronic medical record

CLC number:TP391.1 Document code: A Article ID:1003-0107(2020)09-0061-05

0 引言

电子病历(Electronic medical recode,EMR)命名实体识别(Name Entity Recognition,NER)[1]是指从给定电子病历文本中识别出能表达医疗概念的实体及实体边界,被作为医疗信息抽取及数据挖掘的主要内容及基础任务之一。通过对电子病历进行命名实体识别,可以获取有效的医疗信息,利用这些医疗信息不仅可以为医生提供决策支持,更可以为患者提供针对性的诊疗方案。

通用领域命名实体识别主要有三类方法,分别是基于规则和词典的方法[2-3],基于统计机器学习的方法以及基于深度学习的方法。其中,由于基于规则和词典的方法受限于领域规则且工作量大,基于机器学习的方法依赖于特征选择及数据量,因而被逐步取代[4-5]。基于深度学习的方法是目前用于命名实体识别的主流方法,该方法可利用深度神经网络模型自行训练数据并提取特征而无须人工干预。Graves[6]等人提出了双向长短期记忆网络(BiLSTM),该神经网络模型可同时记忆当前输入的上下文信息,在命名实体识别任务中得到了广泛的应用。

作者简介:吴小雪(1995-),女,硕士,研究方向为自然语言处理;

张庆辉(1974-),男,河南南阳人,工学博士,副教授,硕士生导师,从事电子信息科学与技术专业的教学与科研工作。

而在中文命名实体识别领域，Wan[7]等人通过在 BiLSTM 之上引入注意力机制，成功获取输入字符之间的依赖关系，从而提升了模型识别精度。由于深度学习方法依赖于训练数据的标注特征，而为非结构化的电子病历数据集进行数据标注仍然是一个较为困难的任务。为此，2018 年 Devlin[8]等人首次提出了预训练语言模型 BERT，该模型采用无监督学习，不需人工干预就可在大规模语料中低成本训练，在多个 NLP 任务中都取得了当时最好的处理效果。李妮等[9]使用 BERT 为词的多义性建模，通过搭建 BERT- IDCNN- CRF 模型，不仅取得了较高的识别结果，更极大地缩短了训练时间。Qiao[10]等人将 BERT 模型用于中文电子病历命名实体识别任务中以提升识别效果。

本文旨在探究预训练语言模型 BERT，ERNIE，ALBERT 以及 NEZHA[11-13]在中文电子病历上进行命名实体识别任务时的性能。首先分别介绍了四种模型的结构及原理，接着搭建预训练模型 +BiLSTM+CRF 框架，下基于 CCKS2018 电子病历数据集进行命名实体识别任务，最后讨论不同的实验结果及其原因。

1 预训练模型

2018 年 BERT 模型的横空问世，为自然语言处理任务迎来了新的高潮。预训练模型凭借其强大的特征提取功能，对于不同的任务而具有较强的通用性，成为了研究人员的研究热点，就目前而言性能较为突出的有 ERNIE 模型，ALBERT 模型，NEZHA 模型，下文将针对以上模型给出具体的介绍。

1.1 BERT

BERT 是谷歌发布的以 Transformers 的 Encoder 结构作为核心架构的大型语言预训练模型。BERT 的使用包括两个方面，一方面使用大量未标注的数据针对两个不同的任务 MLM (Masked Language Model)、NSP (Next Sen- tence Prediction)对模型进行联合训练，建立能学习到不同的语义信息的模型；另一方面，针对不同的下游自然语言处理任务，用已经标注的数据微调 BERT 模型参数，以实现最优效果。结构上，BERT 采用双向 Transformers 编码结构，自下而上分别引入位置向量、自注意力机制、残差连接及正则化。位置向量被用于储存输入序列中字在句子中的位置信息，对于奇、偶字向量维度分别使用正弦、余弦函数线性映射，如式(1)和式(2)所示。

$$PE(pos,2n)=\sin(pos/10000^{2n/d_{model}}) \quad (1)$$

$$PE(pos,2n+1)=\cos(pos/10000^{2n/d_{model}}) \quad (2)$$

PE 表示位置向量，pos 表示字在句子中的位置，n 表示字向量的维度，PE 随着字向量维度变大而周期性的改变，从而产生一种包含位置信息的纹理，以便模型学习到自然语言时序特性。自注意力机制创造性的在原有注意力的基础上引入多头(Multi Head)操作，使句子中的每一个字都包含句子中其他字的信息，同时学习句子的可能的多重含义的表达，用 Q，K，V 表示三个包含输入句向量信息的矩阵，则注意力机制如式(3)所示。

$$Attention(Q,K,V)=softmax(\frac{QK^T}{\sqrt{dK}})\cdot V \quad (3)$$

残差连接即将自注意力矩阵和模型最初的输入直接相加，以便使得模型在训练时梯度直接反向传播到最原始层；最后模型使用正则操作以加快训练速度，加速收敛。

值得一提的是，BERT 模型的训练时以字作为最小粒度，模型的输入为三个向量叠加输入，包括表示句子的字向量(Token Embedding)、表示字在句子中的位置信息的位置向量(Position Embedding)、表示句子对信息的向量(Segment mbedding)，且这三个向量的维度相同，在做融合操作时直接采用向量相加即得模型输入，用 $X_{Embedding}$ 表示 Transformer 模型输入向量，则：

$$X_{Embedding}=X_{TokenEmbedding}+X_{PositionEmbedding}+X_{SegmentEmbedding} \quad (4)$$

为了适应不同场景、不同语种的任务需求，谷歌在最初发布 BERT 时推出了结构大小不同的模型。其中适用于中文的 BERT- Base- Chinese，采用了使用中文维基百科语料训练，模型共搭建了 12 个 Transfomer 编码块，768 个隐藏层，且在每个注意力层分割了 12 个注意力头，共计 110M 参数量，正因如此而具有不错的泛化能力，可适用于基于中文电子病历的命名实体识别任务。

1.2 ERNIE

ERNIE(Enhanced Representation through kNowledge IntEgration)是由百度推出的基于知识增强的中文预训练模型，该模型使用大量中文语料进行无监督和弱监督训练，并在多个中文自然语言处理任务上取得了较好的成绩。在结构上，ERNIE 沿袭传统预训练模型的结构风格，使用多个 Transfomer[14]编码器组合作为核心架构，利用 Transfomer 的自注意力机制捕获包含上下文语义信息的句子表示同时在数量上仍采用 12 个 Transfomer 解码单元，756 个隐藏层，8 个注意力头。

ERNIE 自提出来经历了两次改进，分别在 BERT 模型的基础上改进了遮掩词策略、预训练任务种类。BERT 在预训练时采用的对输入序列进行随机遮掩(Mask)并预测的策略，在一定程度上能学习到基于单个字符的语义表示，但对于学习较大语义单元的完整表示却表现不佳。基于此，ERNIE 改进了 BERT 的遮掩策略，预训练时对

完整的短语及实体遮掩,从而获得基于知识增强的句子语义。ERNIE2.0[15]在 ERNIE 的基础上引入持续学习及多任务训练的方式,获取文本的词法、语法、句法等语义信息。值得注意的是,ERNIE 对每个预训练任务都进行编号,使得模型的输入在原有 BERT 的基础上增加了一个表示训练任务的编号向量,模型根据不同的编号向量对应相应的目标函数训练模型。由于采用了多任务训练,ERNIE 利用持续学习理念,把上一个任务训练的参数作为下一个任务的起始参数,从而使模型汇总多个任务的训练参数。

1.3 ALBERT

ALBERT (A LITE BERT) 是谷歌发布的针对 BERT 模型存在的模型参数量大、对训练设备要求过高等问题的新型轻量预训练模型。模型的改进主要针对两方面,其一面向模型参数量众多问题,另一方面面向模型性能问题。对于第一个问题,模型采用了对参数矩阵进行因式分解及跨层参数共享的方式,切断输入字向量维度模型输出隐藏层维度的联系,使模型仅训练经低维映射的输入向量,同时对模型采用全部参数共享的方式,使得模型参数收敛更为稳定。对于第二个问题 ALBERT 在预训练时改变预训练任务,将应用于 BERT 是 NSP 更换为 SOP (Sentence Order Prediction),SOP 在训练时正向句子对保持不变,具有正确的上下文顺序 ;对于负向句子对,模型将原本具有连贯性的句子顺序交换,再输入模型训练,从而使模型学习到句子之间的连贯性,提升模型性能。

1.4 NEZHA

NEZHA(NEural contextualiZed representation for CHinese lAnguage understanding)是华为诺亚方舟实验室发布的面向中文自然语言理解任务的预训练模型。NEZHA 相较于 BERT 的改进主要体现在两个方面,其一为模型本身的改进,其二为训练过程的优化。与 BERT 使用绝对位置编码获取文本位置信息不同,NEZHA 使用相对函数位置编码,获取句子中不同字词之间的相对位置信息,并以词嵌入的形式融合模型输入,这种模型改进方式使得模型在微调下游任务时具有更强的扩展性,即使遇到比预训练中序列长度更长的序列时,依然可以发挥作用。除此之外 NEZHA 分别使用了混合精度训练和 LAMB 训练器以提升训练速度,提升模型性能。NEZHA 模型在每次迭代训练中混合使用单精度浮点格式 FP32,和半精度浮点格式 FP16 更新主权重,使得模型训练速度提升 2- 3 倍,而 LAMB 训练器则可以适应高速训练下的大 batch size 且不用手动调整学习率,极大地提升了模型训练速度。

2 基于预训练模型的中文电子病历命名实体识别

基于预训练的中文电子病历命名实体识别模型如图 1 所示。本文将命名实体识别任务转化为序列标注任务,由于循环神经网络自身的序列特性,即每个神经元不仅可以执行层间信息传递,更能将信息在上下文间流通,基于此,本文在使用预训练模型的基础上使用双向循环神经网络做特征提取,为了提升标注序列准确性,使用条件随机场生成最优标注序列。

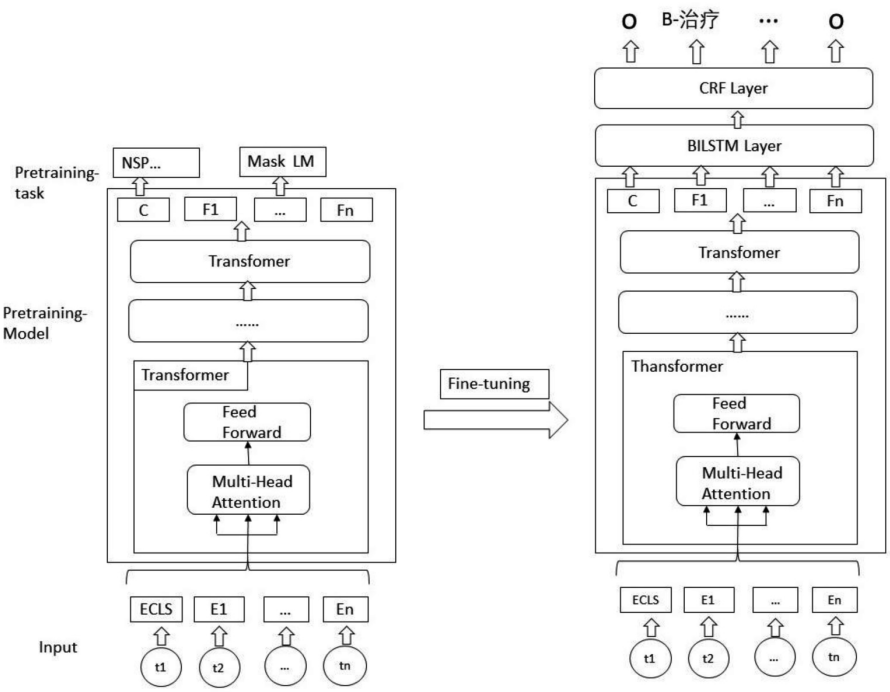

图 1 基于预训练的中文电子病历命名实体识别模型

3 实验

3.1 实验数据及训练设置

本文实验数据主要基于 CCKS2018 的来自医院的真实电子病历数据集合，涉及疾病类型种类约为 20 种，涉及五类实体类型分别命名为 " 疾病 "、" 检查 "、" 症状"、" 治疗 "、" 症状 "。为了提升实验的有效性，本文分别按照一定比例截取数据集，并将其分为训练集、验证集及测试集。同时为了防止模型过拟合，及解决数据集体积较小的问题，本文在原有训练集的基础之上采用 10 折交叉算法，即每次训练时选取 9 个子集数据训练模型，并使用 1 个子集对模型验证，直到所有的子集都完整的经历了训练及测试。

为了保证各层网络之间的输入输出的方差一致性，本文实验采用 Vavier 方法初始化参数，从而将参数的变化尺度限制在一定范围之内，避免变化尺度在最后一层网络中爆炸或者弥散，表示参数，则具体参数设置区间如下：

$$\theta-U\left[-\frac{\sqrt{6}}{\sqrt{n_i+n_{i+1}}},\frac{\sqrt{6}}{\sqrt{n_i+n_{i+1}}}\right] \tag{5}$$

模型优化阶段，本文采用 Adam 算法优化网络参数，同时为了提升模型训练 F1 值，在训练阶段采用早停止策略，设定迭代阈值 20，训练时迭代次数大于 20 且 F1 分数没有进一步增加，即认为训练已经收敛并终止训练，一旦获得某个训练集上的最高 F1 值，则保留当前模型参数。

3.2 实体标签及评价标准

考虑电子病历实体类型的特殊性、结构复杂性，本文采用 BIO 标注策略，将 O 标记于不属于任何概念的实体类型，B 标注于实体开头、I 标注于实体中间；在使用 CRF 获取最优序列标注时，规定一个完整准确的标注应符合以上标注规则，即实体开头及实体结尾部分一致才认为标注正确。

为了评价模型的准确及有效性，本文采用实体识别同用评价标准：查准率(precision)、查全率(recall)及 F1(F-measure)值。通常查准率用来表示实体识别结果中识别正确的实体占识别总实体的数量比例，查全率即识别结果中总实体数量与预测文件中实体数量比例。查准率和查全率定义决定它们作为一组互相矛盾的存在，F1 值可以很好地平衡查准率及查全率，因此被广泛用于命名实体识别评价标准。

3.3 实验结果

由于本文需搭建不同模型，因此本文实验环境在不同的软硬件系统中执行。实验以模型为基础设计了多组对比实验，从而验证不同模型及方法在中文电子病历命名实体识别中的效果。不同模型实验结果对比如表 1 所示。实验分为三个部分，首先在未引入任何外部资源的前提下，以 word2vec 方式获取词特征，整合 CRF 获取基准模型精度，并因此而获得了 66.6%的基线 F 值；其次分别使用 BERT，ERNIE，ALBERT，NEZHA 预训练模型搭建预训练模型、条件随机场框架验证预训练语言模型的可用性及高效性。最后引入双向循环神经网络至模型中添加特征提取环节，实验结果如表 2 所示。

表 1 不同模型实验结果对比

模型	P/%	R/%	F/%
Word2vec+CRF	73.69	77.66	69.01
BERT+CRF	90.68	92.20	90.54
ERNIE+CRF	92.90	94.06	93.37
ALBERT+CRF	90.12	91.88	87.68
NEZHA+CRF	93.10	94.21	93.58

表 2 加入双向循环神经网络的实验结果

模型	P/%	R/%	F/%
Word2vec+BILSTM+CRF	82.31	84.21	75.60
BERT+BILSTM+CRF	91.79	92.78	91.43
ERNIE+BILSTM+CRF	93.22	94.30	93.11
ALBERT+BILSTM+CRF	91.78	89.76	90.12
NEZHA+BILSTM+CRF	93.89	94.31	95.08

通过对比传统识别方法与加入预训练模型方法，可以看到模型的召回率由 77.66%上升为 92.20%，大幅度的提升证明了预训练模型在中文电子病历命名实体识别任务上的有效性。通过比较不同预训练模型之间的训练结果，可以看出随着模型更迭，任务识别效率也在逐渐上升，其中 NEZHA 在本文领域表现出了优异的效果；对比加入双向循环神经网络后的模型发现，相较于传统模型，双向循环神经网络在具有预训练模型的框架中优势并不明显，这也说明了预训练模型可以充分获取电子病历语义信息。总的来说，可以得出以下结论：(1)预训练语言模型能大幅提升中文命名实体的效率；(2)就 BERT，ERNIE，ALBERT，NEZHA 而言，NEZHA 取得了较高的识别效果；(3)对于已有预训练模型的结构来说，双向循环神经网络所发挥的作用相对较低。

4 结束语

本文使用了四种预训练语言模型在中文电子病历上进行了命名实体识别任务。论文首先对四种模型展开了详细介绍，其次使用电子病历数据集微调模型参数，并采用了一定策略以获取最优模型，实验结果表明预训练模型在中文电子病历命名实体识别上具有较好的表

现,其中 NEZHA 更是取得了 94.31%的召回率,较以往而言具有明显提升。在未来的工作中,我们将会考虑进一步针对电子病历复杂语义方面做出深入研究,与此同时将继续挖掘使用于中文电子病历的大型数据集,以期取得更好的效果。

参考文献:

[1]Wu G,Tang G,Wang Z,et al.An Attention-Based BiLSTM -CRF Model for Chinese Clinic Named Entity Recognition[J].IEEE Access,2019,(7):113942-113949.

[2]Song M,Yu H,Han W S.Developing a hybrid dictionary-based bio-entity recognition technique[J].BMC Medical Informatics and Decision Making,2015,15(1):1-8.

[3]Zhao Y.Research on Entity Recognition in Traditional Chinese Medicine Diet [C].2017 9th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC),2017: 284-287.

[4]Huang Z,Xu W,Yu K.Bidirectional LSTM-CRF Models for Sequence Tagging[J].arXiv:1508.01991,2015.

[5]Xia Y,Wang Q.Clinical Named Entity Recognition: ECUST in the CCKS-2017 Shared Task 2 [C]. Proceedings of the Evaluation Task at the China Conference on Knowledge Graph and Semantic Computing (CCKS 2017),China,2017:49-54.

[6]Graves A,Schmidhuber J.Framewise phoneme classifica- tion with bidirectional LSTM and other neural network architectures [J].Neural Networks,2005, 18(5-6):602-610.

[7]Wu G,Tang G,Wang Z,et al.An Attention-Based BiLSTM -CRF Model for Chinese Clinic Named Entity Recognition[J].IEEE Access,2019,(7):113942-113949.

[8]Devlin J,Chang M W,Lee K,et al.BERT: Pre-training of Deep Bidirectional Transformers for Language Understan- ding[J].arXiv:1810.04805,2018.

[9]李妮,关焕梅,杨飘,等.基于 BERT-IDCNN-CRF 的中文命名实体识别方法[J].山东大学学报(理学版),2020, 55(1):102-109.

[10]Qiao R,Yang X,Huang W.Medical named entityrecogni- tion based on Bert and model fusion[C]. CCKS,Hangzhou,China,2019.

[11]Devlin J,Chang M,Lee K,et al.Pre-training of Deep Bidi- rectional Transformers for Language Understanding[J].aXxiv:1810.04805,2018.

[12]Sun Y,Wang S,Li Y,et al.ERNIE: Enhanced Representation through Knowledge Integration [J].arXiv:1904. 09223,2019.

[13]Liu Y,Ott M,Goyal N,et al.RoBERTa: A Robustly Opti- mized BERT Pretraining Approach[J].arXiv:1907. 11692,2019.

[14]Vaswani A,Shazeer N,Parmar N,et al.Attention is all you need [C].2017 31st Conference on Neural Information Processing Systems (NIPS 2017),Long Beach,CA,2017:5998-6008.

[15]Sun Y,Wang S,Li Y,et al.ERNIE 2.0: A continual Pre-training Framework for Language Understanding [J]. arXiv:1907.12412,2019.