Computer Science and Application 计算机科学与应用, 2024, 14(10), 58-66
Published Online October 2024 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2024.1410202

Hans汉斯

基于并行神经网络的疾病特征实体识别方法

杨建兴[1,2]，卢照敢[1]，赵柴学正[1]

[1]河南财经政法大学计算机与信息工程学院，河南 郑州
[2]浙江医院信息中心，浙江 杭州

收稿日期：2024年9月2日；录用日期：2024年10月3日；发布日期：2024年10月11日

摘 要

针对静脉血栓栓塞症电子病历文本语义复杂，疾病信息多维性导致的疾病特征学习不彻底、实体识别不准确的问题，本文提出了一种并行神经网络的疾病特征实体识别方法。首先，通过RoBERTa模型，更好地学习到病历实体中的特征信息。然后，通过双向长短期记忆网络，提取病历中的全局特征，再经过并行的迭代膨胀卷积神经网络提取病历中的局部特征。最后，利用CRF推理层修正神经网络输出的疾病特征标签。在医院提供的2000份静脉血栓电子病历上，本方法的平均准确率为85.26%，相对于单纯的卷积神经网络，该方案的识别准确率提高了13.52%。

关键词

神经网络，RoBERTa，实体识别，静脉血栓

Disease Feature Entity Recognition Method Based on Parallel Neural Network

Jianxing Yang[1,2], Zhaogan Lu[1], Chaixuezheng Zhao[1]

[1]Computer and Information Engineering College, Henan University of Economics and Law, Zhengzhou Henan
[2]Information Center, Zhejiang Hospital, Hangzhou Zhejiang

Received: Sep. 2nd, 2024; accepted: Oct. 3rd, 2024; published: Oct. 11th, 2024

Abstract

In response to the intricate semantics of electronic medical records for venous thromboembolism and the multi-dimensionality of disease information that gives rise to incomplete acquisition of disease features and inaccurate entity recognition, this paper presents a parallel neural network-based disease feature and entity recognition approach. Firstly, the language representation RoBERTa model is employed to more effectively acquire the feature information of medical record entities. Subsequently, a

文章引用: 杨建兴, 卢照敢, 赵柴学正. 基于并行神经网络的疾病特征实体识别方法[J]. 计算机科学与应用, 2024, 14(10): 58-66. DOI: 10.12677/csa.2024.1410202

bidirectional long short-term memory network is utilized to extract global features from the medical record, followed by a parallel iterative dilated convolutional neural network for extracting local features from the medical record. Eventually, the CRF inference layer is adopted to rectify the disease feature labels output by the neural network. On the 2000 venous thromboembolism electronic medical records provided by the hospital, the average accuracy of the proposed method is 85.26%, which is 13.52% higher than that of the pure convolutional neural network.

Keywords

Neural Network, RoBERTa, Entity Recognition, Venous Thromboembolism

1. 引言

静脉血栓症(Venous Thromboembolism, VTE)作为全球重大健康问题之一，是住院患者非预期死亡的重要原因，其发病隐匿，极易被误诊、漏诊，被称为“沉默的杀手”。目前，人们采用 VTE 诊断方法主要是基于人工分析电子病历数据，填写线性加权评估表，对不同风险因素赋予权重，最终通过累加评估项分值确定患者的 VTE 风险等级。这种人工评估方式存在表项过多，严重增加医护人员的工作量，医生主动性不高，评估不及时的问题。同一个表项也会因医护人员的经验差异，出现评估结果不一致性的情况，导致评估结果准确度较低，患者错失最佳的治疗时机。

近年来，随着信息化建设在医疗卫生领域的深入推进，医疗大数据时代的到来，以电子病历为核心的临床信息系统生成了大量数据，这些数据蕴含了丰富的医疗信息，是研究患者身体状况的第一手资料，对医疗决策有着重大参考价值[1]。但由于电子病历数据往往以自由文本的形式存储，无法直接使用。因此，通过实体识别(Named Entity Recognition, NER)技术辅助医生智能评估 VTE，从非结构化的文本中准确提取出结构化的、有价值的 VTE 特征数据取代人工的方式，是当前研究的一个重要任务。

2. 问题分析

早期实体识别依赖于词典和规则，需专家手动制定，效率低下且应用受限。随后，机器学习的到来，如隐马尔可夫[2]、最大熵[3]、支持向量机[4]及条件随机场[5]模型逐步替代传统方法，减少人工干预但需大量标注数据，且效果受特征选择的影响。深度学习的兴起，神经网络可自动提取语义，减少特征工程的依赖，如循环神经网络(Recurrent Neural Network, RNN) [6]、卷积神经网络(Convolutional Neural Network, CNN) [7]和 Transformer [8]等模型得到了广泛应用。

深度神经网络技术在实体识别领域成效显著，但在实际医疗场景，VTE 疾病特征实体识别准确率较低，主要原因是中文电子病历文本存在语义复杂、疾病信息多维性，即存在同一疾病特征词在不同的语义环境有不同的含义，识别疾病特征实体需结合上下文多维度语义提取文本特征，导致 RNN 出现长距离依赖问题。本文围绕这个问题，提出了并行神经网络实体识别模型(RoBERTa-BiLSTM-Parallel-IDCNN-CRF, RBPIC)，提高 VTE 疾病特征实体识别准确率。

3. 基于并行神经网络的命名实体识别模型

VTE 实体识别 RBPIC 模型主要包括两个部分：预训练 RoBERTa 模型、BiLSTM-Parallel-IDCNN-CRF

模型。整体结构，如图 1 所示。

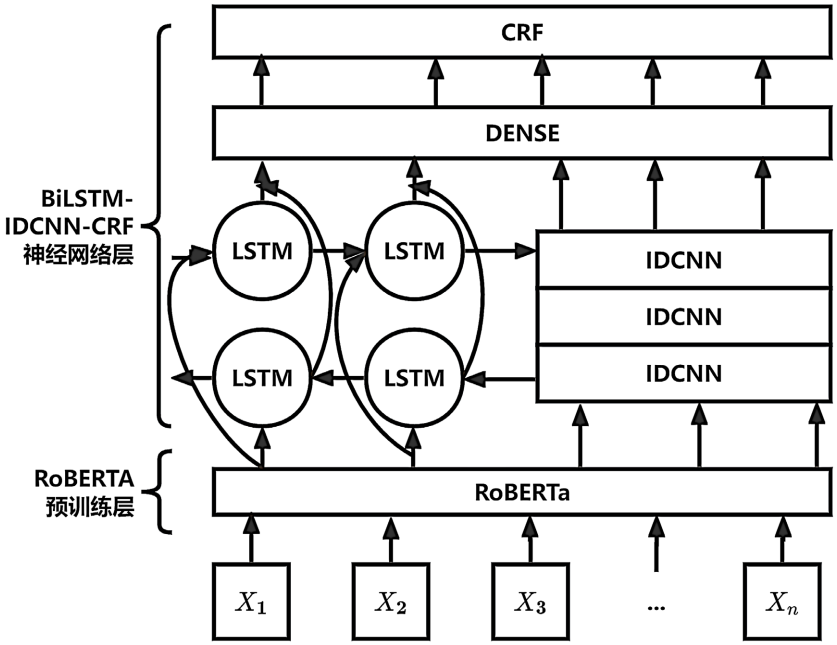

Figure 1. RBPIC model structure
图 1. RBPIC 模型结构图

3.1. RoBERTa 预训练模型

RoBERTa 的嵌入层接收字符级别语料 $X=\{X_1,X_2,X_3,\cdots,X_n\}$。通过求和不同嵌入层处理的 X 序列，得到嵌入向量 $E=\{E_1,E_2,E_3,\cdots,E_n\}$ 输出到 RoBERTa 的 Trm 层。RoBERTa 嵌入层，如图 2 所示。在 RoBERTa 模型的嵌入层中为字符引入位置信息，即采用位置编码策略，通过正弦和余弦函数对字符进行编码，从而将字符的位置信息转化为特征矩阵。这种方式使相同字符在不同位置，能够形成各自不同的特征矩阵，更好地学习到病历文本中的疾病特征。

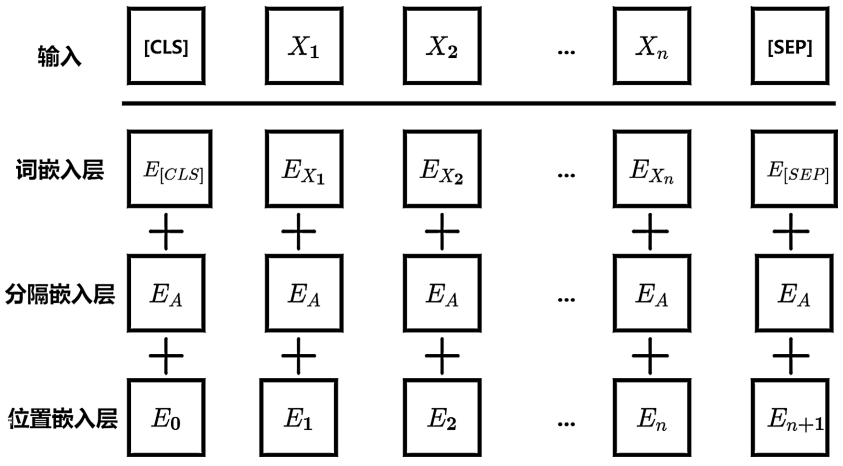

Figure 2. RoBERTa embedding layer
图 2. RoBERTa 嵌入层

嵌入层生成的向量 $E=\{E_1,E_2,E_3,\cdots,E_n\}$ 输入至 RoBERTa 模型，最终输出结果为 $T=\{T_1,T_2,T_3,\cdots,T_n\}$。

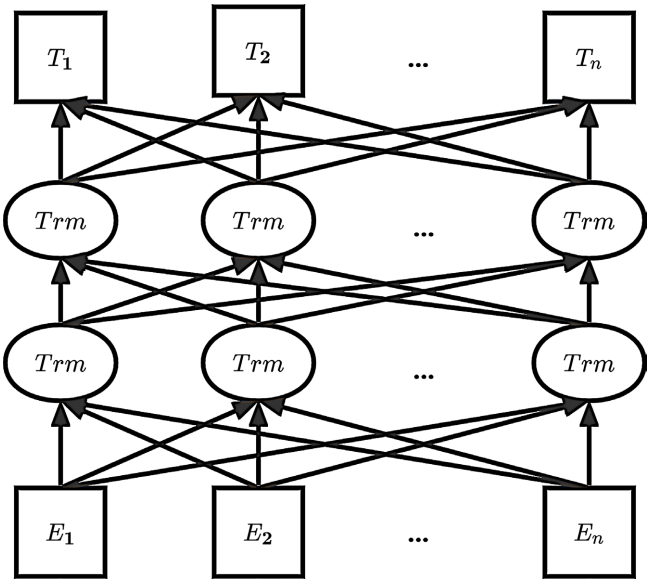

Figure 3. RoBERTa structure
图 3. RoBERTa 网络结构

RoBERTa 的架构如图 3 所示，包含 12 个 Transformer (Trm)层。每个 Trm 层由全连接前馈神经网络、多头自注意力机制，以及残差连接与归一化层共同组成。在经过 Trm 层的编码后，Softmax 层对输出向量 T 进行归一化处理，随后将处理后的向量传递给下一个神经网络层。

3.2. BiLSTM-IDCNN-CRF 神经网络层

RNN 可以较好地处理时序序列，但在中文电子病历中，由于语义复杂，疾病信息多维性，识别疾病特征实体需结合上下文语义多维度提取文本特征。比如“TPR”，在病历文书中通常指“体温”、“脉搏”、“呼吸”三个指标，但是在心功能检查报告中指“外周血管总助力”。“TPR”文本在不同位置上所呈现的含义截然不同。在病历文本语义复杂的场景中，RNN 模型需要提取句子中较远的语义特征，导致模型出现长距离依赖的问题。

长短期记忆网络(Long Short-Term Memory Neural Networks, LSTM)通过引入“门”机制和“细胞状态”来高效处理信息。门结构控制信息的出入流动，“细胞状态”可以长期保持存储状态，提升 LSTM 在长序列数据中捕捉依赖关系的性能[9]。

LSTM 主要由三个核心组成部分构成：输入门、遗忘门和输出门。输入门的主要功能是控制新信息的引入，遗忘门负责决定哪些信息应该被保留以及哪些信息应当被丢弃，而输出门则用于确定最终需要输出的信息。具体如公式(1)输入门、公式(2)遗忘门、公式(3)输出门所示。

$$input_i=\sigma\left(W_{input}\cdot\left[T_i,L_{i-1},C_{i-1}\right]+b_{input}\right) \tag{1}$$

$$forget_i=\sigma\left(W_{forget}\cdot\left[T_i,L_{i-1},C_{i-1}\right]+b_{forget}\right) \tag{2}$$

$$output_i=\sigma\left(W_{output}\cdot\left[T_i,L_{i-1},C_{i-1}\right]+b_{output}\right) \tag{3}$$

其中，输入门 $input_i$；遗忘门 $forget_i$；输出门 $output_i$；sigmoid 激活函数 σ；门控单元权重矩阵 W；偏置 b；L_{i-1} 表示 LSTM 单元在上一个时刻的隐藏状态；C_{i-1} 则指的是 LSTM 单元在上一个时刻所存储的记忆

信息。此外，T 表示在第 i 时刻的输入向量，这亦是 BERT 层最终输出的向量。

LSTM 通过其“门”结构，能够有效地管理特征信息的遗失，从而解决长距离依赖的问题。然而，单向 LSTM 网络仅能捕捉到过去的特征，无法获取未来数据特征。因此，本文选择使用 BiLSTM 模型，通过双向特征结合的方式来处理时间序列。这种方法在考虑前向特征的同时，也结合了后向数据特征，从而避免了由于句子语义复杂而导致的历史特征的丢失。

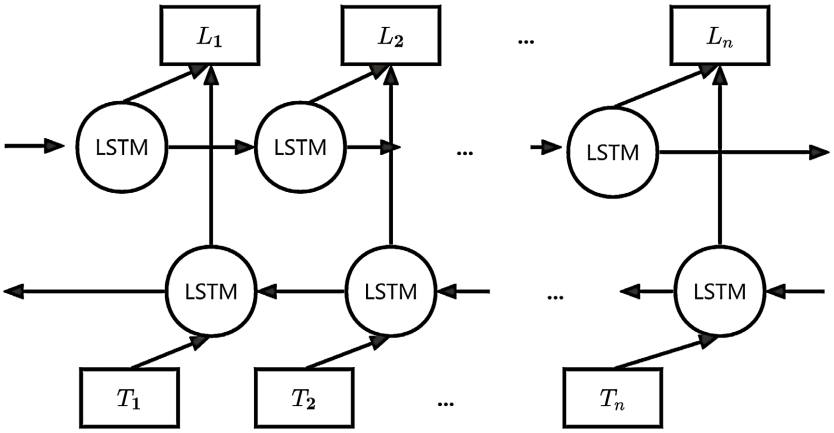

Figure 4. BiLSTM structure
图 4. BiLSTM 架构图

RoBERTa 预训练层的输出向量 $L=\{L_1,L_2,L_3,\cdots,L_n\}$ 传送到 BiLSTM 神经网络层，并继续传递到下一个神经网络层。BiLSTM 架构图，如图 4 所示。

通过上述的 BiLSTM 模型，虽然可以捕捉到电子病历文本整体的上下文特征，但是容易忽略文本局部特征。CNN 在提取局部特征方面表现出色。膨胀卷积神经网络(DCNN)是 CNN 的一种特殊变体，通过在卷积核中引入膨胀距离，扩展了感受野的范围，这一方式有助于获取更多的局部特征[10]。迭代膨胀卷积神经网络(IDCNN)由多层带有不同膨胀率的卷积神经网络构成，通过前一层的膨胀卷积输出，对当前层的特性向量进行计算。

IDCNN 计算方法，如公式(4)所示。

$$\tilde{D}_i=\sigma\left(H_{i-1}\cdot\tilde{D}_{i-1}\right) \tag{4}$$

其中 H_i 为第 i 层膨胀卷积神经网络，$\tilde{D}_i$ 为第 i 层卷积网络学习到的特征向量。

IDCNN 模型的局部特征处理能力相对 DCNN 模型来说更强。通过 IDCNN 模型解决 BiLSTM 存在的忽视局部特征的问题，提取 VTE 数据集全局特征，同时确保局部特征被完整保留。IDCNN 架构如图 5 所示。

通过训练 RoBERTa-BiLSTM-IDCNN 模型，能够为每个标签产生特定分数，最终输出分数最高的标签。因输出的得分不够精确，导致有时会出现标签位置错误的情况。

推理层的条件随机场(CRF)能够通过在标签之间引入约束关系来辅助纠正预测过程中的错误。输入模型的序列为 $X=\{X_1,X_2,X_3,\cdots,X_n\}$，对应的标签序列为 $Y=\{Y_1,Y_2,Y_3,\cdots,Y_n\}$，其得分公式如公式(5)所示。

$$score(X,Y)=\sum_{i=1}^{n}\tilde{W}_{Y_i,Y_{i+1}}+P_{i+1,Y_{i+1}} \tag{5}$$

$\tilde{W}$ 代表得分矩阵，Y_i 到 Y_{i+1} 的转移得分值记作 $\tilde{W}_{Y_i,Y_{i+1}}$，而 P 则是指上一层的得分向量。第 $i+1$ 层的标签 Y_{i+1} 对应的得分值为 $P_{i+1,Y_{i+1}}$。Y 标签序列的概率计算如公式(6)所示。

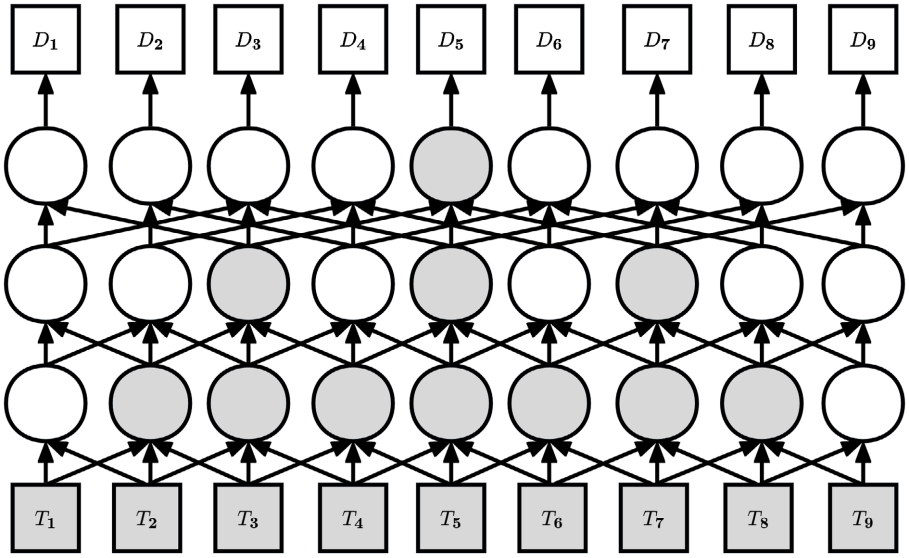

Figure 5. IDCNN structure
图 5. IDCNN 架构

$$P(Y \mid X) = \frac{e^{score(X,Y)}}{\sum_{\tilde{Y} \in Y_X} e^{score(X,\tilde{Y})}} \tag{6}$$

$\tilde{Y}$ 涵盖了所有可能的标签组合序列。在解码阶段，为了确定最优的标签序列，采用了维特比算法。维特比算法被用于寻找最优路径，其具体的计算过程如公式(7)所示。

$$Y^* = \text{argmax}\left\{\ln\left(P(Y \mid X)\right)\right\} \tag{7}$$

经上述系列步骤，CRF 有效修正预测中的错误，提高识别的准确性。

因此，序列 X 输入模型后经过 RoBERTa-BiLSTM-IDCNN-CRF 模型的优化处理，得到准确的标签序列 Y^*。

4. 实验结果及分析

4.1. 实验数据

本文构建的实验数据来源于浙江省某省级三甲医院提供的 2000 份 VTE 住院患者电子病历。在临床医生的专业指导下，结合 2023 年发布的《肺血栓栓塞症诊治与预防指南》[11]关于静脉血栓栓塞症预防的相关内容，总结归纳出了需要标记的 8 种实体类型。本文采用 BIOES 标注法，“B”表示该字符为实体开头；“I”表示该字符为实体结尾；“O”表示该字符为非实体；“E”表示该字符为实体尾部；“S”表示单个字符为一个实体。标注结果如表 1，随机选取 70%实体数据集为训练数据，20%为测试数据，其余 10%为验证数据。

4.2. 评价指标

模型评估的评价指标包括准确率(Precision, P)、召回率(Recall, R)和 F_1 值，公式(8)~(10)。

$$P = \frac{T_p}{T_P + F_P} \times 100\% \tag{8}$$

$$R = \frac{T_p}{T_P + F_N} \times 100\% \tag{9}$$

$$F_1 = \frac{2PR}{P + R} \times 100\% \tag{10}$$

Table 1. Entity type number statistics of VTE dataset
表 1. VTE 数据集的实体类型数量统计

实体类型	实体示例
疾病(Dis)	2386
身体部位(Bod)	320
药物(Dru)	6816
手术(Ope)	1710
其他治疗(Tre)	2037
症状(Sym)	21,952
检验(Ite)	3672
检查(Pro)	2241

4.3. 模型参数

在构建 RBPIC 模型时进行命名实体识别实验前，需要设定的超参数包括句子的最大长度、学习率、丢弃率等。以下是实验中使用的超参数设置，如表 2 所示。

Table 2. NER model parameter design
表 2. 命名实体识别模型参数设计

参数名称	值
RoBERTa-embedding-size	768
Max seq length	250
batch size	12
transformer	12
LSTM_dim	64
Learning rate	1e−4
卷积核窗口大小	3
膨胀卷积块数	3
膨胀卷积距离	1, 1, 2
Dropout rate	0.4

4.4. 实验结果

实验一：不同下游模型在实体识别任务中的表现。

如表 3 所示，BiLSTM-CRF 模型与 IDCNN-CRF 模型相比，在准确率、召回率和 F1 值上均有所提升，表明 BiLSTM 模型具备更强的全局上下文信息提取能力。在 BiLSTM-Attention-CRF 模型提取能力略

有下降，原因在于 Attention 机制过于强调权重分配，导致在处理大量医学专业术语时复杂的语义使整体特征提取能力受到影响。BiLSTM-Serial-IDCNN-CRF 模型是在 BiLSTM-CRF 的基础上引入 IDCNN 模型进行串行连接，尽管此方法融合了两种模型的优点，增强了对局部特征的提取能力，但也因此增加了模型参数的计算量，从而延长了训练时间。相对而言，BiLSTM 和 IDCNN 神经网络并行连接而形成的 BiLSTM-Parallel-IDCNN-CRF 模型在特征提取能力上优于串行连接方式。这是因为疾病特征实体具有多维复杂的语义结构，直接采用串行连接会削弱 BiLSTM 获取全局特征的能力，而并行连接则更有效地保留了全局特征。因此，本文提出的并行连接的神经网络模型在特征提取方面具有明显的优势。

Table 3. Entity identification of different downstream models
表 3. 不同下游模型实体识别结果

下游模型	*P*	*R*	F_1
IDCNN-CRF	71.74%	67.91%	69.78%
BiLSTM-CRF	72.79%	73.48%	73.13%
BiLSTM-Attention-CRF	70.24%	69.52%	69.88%
BiLSTM-Serial-IDCNN-CRF	76.66%	76.84%	76.75%
BiLSTM-Parallel-IDCNN-CRF	80.32%	79.89%	80.10%

实验二：不同 BERT 预训练模型在实体识别任务中的表现。

如表 4 所示，本文提出的下游模型 BiLSTM-Parallel-IDCNN-CRF 作为基础架构，在此基础上对比四种不同的预训练模型。BERT 中文模型以字符为单位进行处理，在不同上下文中利用 Transformer 编码器动态生成向量表示，处理与上下文相关性强的任务。AlBert 是 BERT 模型的一种简化版本，通过减少参数数量和实现层间参数共享来达到精简的目的。由于其规模较小，它的性能不及标准的 BERT 预训练模型。BioBert 是一种特地为生物医学领域开发的 BERT 模型，在生物医学文献上进行额外预训练，部分实体与本文的 VTE 疾病匹配，因此可提升实体抽取的能力。RoBERTa 是对 BERT 模型的一种优化和增强版本，它利用了更大规模的数据集、更高的 batch 大小以及中文全词遮罩的技术，相比 BERT 有更好的性能表现。实验结果证明了本文提出 RBPIC 模型的优越性。

Table 4. Different entity recognition situations of pre-trained models
表 4. 不同的预训练模型实体识别结果

模型	*P*	*R*	F_1
BERT-下游模型	81.34%	83.42%	82.37%
AlBert-下游模型	80.72%	82.25%	81.48%
BioBert-下游模型	83.23%	83.43%	83.33%
RoBERTa-下游模型(本文模型)	85.26%	84.57%	84.91%

5. 结束语

本文首先分析了命名实体识别在实际医疗场景，因中文电子病历文本复杂的语义关系，导致在 VTE 病历中出现疾病特征实体抽取困难、实体识别准确率低的问题。据此，本文提出了一种基于并行的神经网络命名实体识别模型。通过语言表征 RoBERTa 模型，更好地学习到病历实体中的特征信息。通过把双向长短期记忆网络和迭代膨胀卷积神经网络由传统的串行连接改为并行连接，改善了在串行连接过程中

双向长短期记忆网络获取全局特征能力减弱的问题，使并行连接的模型能从全局和局部两个方面获取文本信息，增强文本上下文特征的获取能力。实验证明，RoBERTa 模型比 BERT 模型实体识别准确率提高了 3.92%；并行连接的方式比串行连接的方式实体识别准确率提高了 3.66%。本文提出的 RBPIC 模型，提高了 VTE 疾病特征实体提取的准确率，对后续医疗领域实体识别的研究具有一定的参考价值。

参考文献

[1] 中华人民共和国国家卫生和计划生育委员会. 电子病历应用管理规范(试行) [J]. 中国实用乡村医生杂志, 2017, 24(6): 1-2.

[2] Syachrul, M.A.K., Bijaksana, M.A. and Huda, A.F. (2019) Person Entity Recognition for the Indonesian Qur'an Translation with the Approach Hidden Markov Model-Viterbi. *Procedia Computer Science*, **157**, 214-220. https://doi.org/10.1016/j.procs.2019.08.160

[3] Glabal Tone Communication Technology Co. Ltd. (2020) Korean Named-Entity Recognition Method Based on Maximum Entropy Model and Neural Network Model. Patent Application Approval Process (USPTO 20200302118).

[4] Imam, A.T., Alhroob, A. and Jumah, W. (2021) SVM Machine Learning Classifier to Automate the Extraction of SRS Elements. *International Journal of Advanced Computer Science and Applications*, **12**, 174-185. https://doi.org/10.14569/ijacsa.2021.0120322

[5] Liu, S., He, T. and Dai, J. (2021) A Survey of CRF Algorithm Based Knowledge Extraction of Elementary Mathematics in Chinese. *Mobile Networks and Applications*, **26**, 1891-1903. https://doi.org/10.1007/s11036-020-01725-x

[6] Goyal, A., Gupta, V. and Kumar, M. (2021) Recurrent Neural Network-Based Model for Named Entity Recognition with Improved Word Embeddings. *IETE Journal of Research*, **69**, 6970-6976. https://doi.org/10.1080/03772063.2021.2006805

[7] Chen, J.-Q., Zhu, Z.-C., Zhang, F., Zeng, K., Jiang, H.-Z. and Cheng, Z.-N. (2023) A BIGRU-Based Stacked Attention Network for Biomedical Named Entity Recognition with Chinese EMRs. *Studies in Health Technology and Informatics*, **308**, 757-767. https://doi.org/10.3233/SHTI230909

[8] 叶恩光, 张晓如, 张再跃, 等. 基于BERT和领域词典融合的中文电子病历命名实体识别[J]. 计算机与数字工程, 2024, 52(3): 746-750, 767.

[9] 周宇辉, 何志琴. 基于改进注意力机制的图像描述算法[J]. 智能计算机与应用, 2022, 12(2): 58-63.

[10] Yu, F. and Koltun, V. (2015) Multi-Scale Context Aggregation by Dilated Convolutions. https://doi.org/10.48550/arXiv.1511.07122

[11] 杜敏, 黄红娟, 徐西楚, 等. Wells 评分联合 D-二聚体水平对重症肺血栓栓塞患者近期预后不良的预测分析[J]. 中国急救复苏与灾害医学杂志, 2023, 18(12): 1610-1614.