

基于残差注意力网络的医疗命名实体识别方法

王维欣^{1,2}, 徐国愚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10月18日; 录用日期: 2024年11月15日; 发布日期: 2024年11月27日

摘要

针对临床医疗记录中的复杂语义实体和长短距离依赖关系识别准确率低的难题, 文章提出了一种双向语义与残差注意力网络的医疗文本命名实体识别方法。利用BERT-wwm预训练模型捕捉语义特征, 结合双向门控循环单元BiGRU用于处理复杂长程语义关联; 增加残差连接的注意力Attention结构, 保障专注于关键信息的同时, 不会丢失捕捉到的整体序列特征; 条件随机场CRF负责最后的序列标注预测, 对前序多层神经网络抽取的特征序列进行最优路径解码。实验结果表明, 通过本方法能够有效提升命名实体识别的准确率。

关键词

残差注意力网络, 医疗命名实体识别, 双向门控循环单元, 长短距离依赖

Medical Named Entity Recognition Method Based on Residual Attention Networks

Weixin Wang^{1,2}, Guoyu Xu¹

¹Computer and Information Engineering College, Henan University of Economics and Law, Zhengzhou Henan

²Information Center, Zhejiang Hospital, Hangzhou Zhejiang

Received: Oct. 18th, 2024; accepted: Nov. 15th, 2024; published: Nov. 27th, 2024

Abstract

Aiming at the challenge of low recognition accuracy for complex semantic entities and long- and short-range dependencies in clinical medical records, this paper proposes a medical text named entity recognition method that integrates bidirectional semantics with a residual attention network.

文章引用: 王维欣, 徐国愚. 基于残差注意力网络的医疗命名实体识别方法[J]. 计算机科学与应用, 2024, 14(11): 119-126. DOI: 10.12677/csa.2024.1411221

The method leverages the BERT-wwm pre-trained model to capture semantic features and combines it with a Bidirectional Gated Recurrent Unit (BiGRU) to handle complex long-range semantic associations. An Attention mechanism with residual connections is added to ensure focus on key information while preserving the overall sequence characteristics captured. A Conditional Random Field (CRF) is responsible for the final sequence labeling prediction, performing optimal path decoding on the feature sequences extracted by the preceding multi-layer neural networks. Experimental results demonstrate that this approach can effectively improve the accuracy of named entity recognition.

Keywords

Residual Attention Networks, Named Entity Recognition, BiGRU, Long- and Short-Range Dependencies

Copyright © 2024 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

随着医院信息化水平的提高,电子病历系统为医生提供了结构化记录和成熟模板,极大地简化了病历书写的过程。然而,这种便利性也带来了新的问题和挑战。一方面,一些医生过分依赖电子病历模板的导入、复制和粘贴功能,导致病历内容的雷同,忽视了患者个体差异,往往不能真实、客观地反映患者的病情变化,增加了医疗安全风险。另外,由于医生工作忙碌或对系统的依赖,会出现病历书写上的疏漏或错误,影响病历的准确性和完整性。

病历质量控制的核心在于保证记录信息的真实性、完整性及时效性。目前,电子病历系统的质控主要集中在完整性校验上。虽然能够在一定程度上确保病历的格式规范,但难以深入到病历内容的质量控制。比如,对于病历文书的内涵、临床合理性及专业性的评估,当前的电子系统尚不能完全胜任,对终末电子病历的内涵质控仍然需大量人工参与。当前深度学习和自然语义处理技术在各个领域取得了快速发展,将命名实体识别(Named Entity Recognition, NER)技术应用于病历质控,有望实现质控效能与精确度的双重提升。

2. 问题分析

医疗文本命名实体识别的核心目标是在非结构化病历中,精准检测并分类具有特定意义的命名实体,如疾病、症状、药物、治疗方案及检验检查等[1]。这在病历质量控制中具有重要应用,能够自动提取病历中的关键医学信息,帮助发现错漏信息、不规范记录,进而提高医疗文书的专业性和临床合理性。传统的机器学习方法为了优化模型效果,通常需要大量的特征工程工作,研究者需定义和构建多种特征,如词性标注、词语拼音转换、词根或词干提取、笔画信息分析等,还需利用词典等资源构造特征[2]。相比之下,深度学习,尤其是基于神经网络的方法(如双向长短期记忆网络 BiLSTM),能够实现端到端的学习,自动捕捉并学习深层次、复杂的语义特征,显著减少了特征工程的工作量[3]。目前, BERT-BiLSTM-CRF 模型当前已被广泛应用,但在实际医疗场景,病历文本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,识别病历文本实体需结合上下文多维度语义提取文本特征,导致模型出现长距离依赖问题,在某些情况下

可能仍有优化空间。

本文根据医疗文本语料的特点以及现有的命名实体识别模型的不足, 提出了双向语义与残差注意力网络的医疗文本命名实体识别模型(BERT-www-BiGRU-Residual Attention-CRF, BBRAC)。通过结合多种深度学习模型, 提高解决临床医学领域中复杂的语义歧义和长距离依赖关系的能力。

3. 基于双向语义与残差注意力的命名实体识别模型

病历质控实体识别模型 BBRAC 主要包括两个部分: BERT-www 预训练层、BiGRU-Residual Attention-CRF 神经网络层。整体结构, 如图 1 所示。

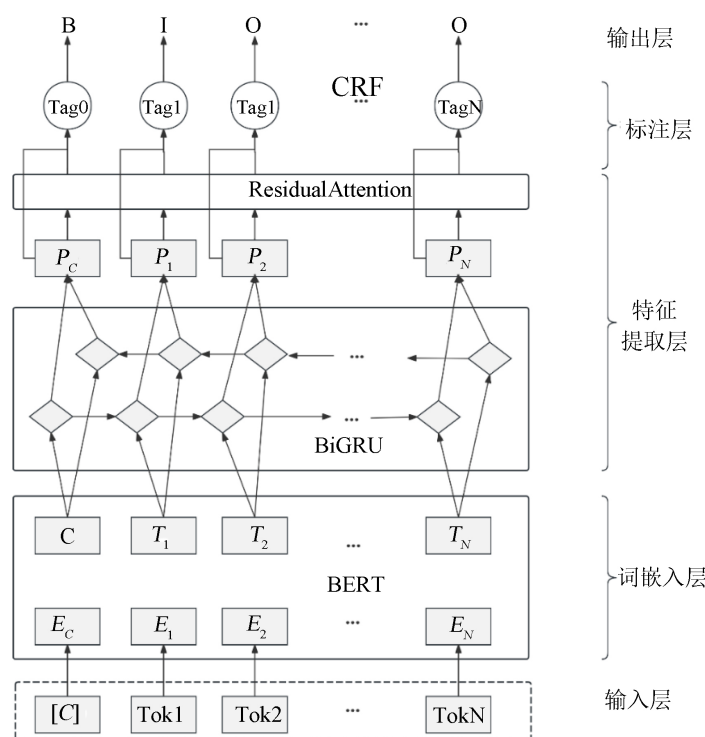


Figure 1. EMTBA model structure
图 1. EMTBA 模型结构图

3.1. BERT-www 预训练层

为了解决临床医学领域命名实体的语义歧义问题, 本文采用了预训练模型策略。BERT 模型采用了多层的 Transformer 的 encoder 结构, 基于多个相同的编码器堆叠而成。每个 Transformer 层都包括两个多头自注意力机制和前馈神经网络子层, 在预训练阶段采用无监督学习的方法, 从大量文本语料中通过两种预训练任务学习获得先进的语言表示[4][5]。在具体应用中, 只需要对预先训练好的模型进行少量微调, 便可以适应特定任务, 无需为了适应一个新的特定数据集而从零开始训练一个全新的模型, 显著扩大了模型的应用范围并加强了对不同类型数据的处理能力。BERT 通过引入位置嵌入(Position Embeddings), 增强了模型对上下文的敏感度, 从而有效处理了多义词在不同上下文中的语义。BERT 模型, 如图 2 所示。

BERT-www 是 BERT 模型的一种改进模型, 在预训练过程中创新性得采用了一种整词掩码策略(Whole Word Masking), 取得了良好的效果。研究发现, 扩大预训练模型的规模往往能够提升其在下游任

务上的性能, BERT-wwm 采用了比标准 BERT 更大的训练数据集, 达到了 5.4 B 字符级别, 从而能够生成包含丰富上下文依赖关系的高质量词向量表示, 这些表示不仅蕴含了单词本身的语义信息, 还反映了它们在整个句子中的语境关联。

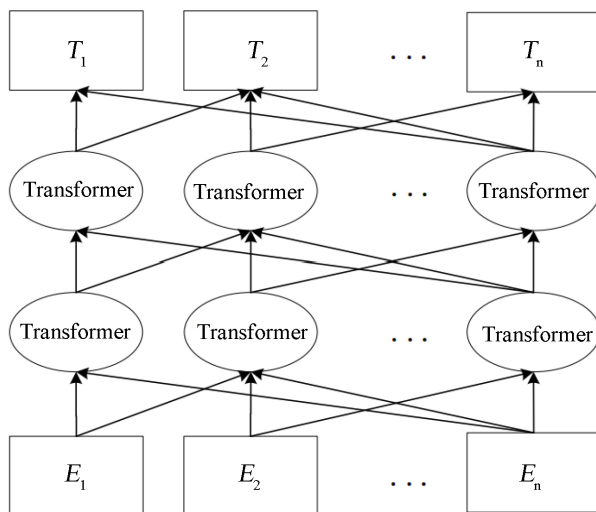


Figure 2. BERT network structure
图 2. BERT 网络结构

3.2. BiGRU-Residual Attention-CRF 神经网络层

GRU 通过训练过程可以学到应该记忆哪些信息和遗忘哪些信息, 可以更好的捕捉到较长距离的依赖关系[6]。BiGRU 由前向 GRU 和后向 GRU 组成, 前向 GRU 经过多个时间步的运算输出一个结果向量, 后向 GRU 同样经过运算输出一个结果向量, 将两个结果向量拼接即得到最后包含双向语义信息的输出结果。此外, BiGRU 通过更新门和重置门机制门控机制设计使得模型能够更好地处理长距离依赖问题, 尤其对于长程语义关联有着良好的建模效果。

但由于 GRU 是一种单向的模型, 它从左到右(从前到后)处理序列数据。这意味着只能访问到当前时刻及之前的上下文信息, 而不能获取之后的信息。因此, 研究者提出了由前向 GRU 和后向 GRU 组成的 BiGRU, 前向 GRU 经过多个时间步的运算输出一个结果向量, 后向 GRU 同样经过运算输出一个结果向量, 将两个结果向量拼接即得到最后包含双向语义信息的输出结果。不同于 GRU 仅利用过去时刻的信息来进行预测, BiGRU 同时利用了过去时刻和未来时刻的状态信息, 结合正反两个方向的 GRU 模型处理输入数据, 并共同确定最终的输出, 使得结果更加准确。

双向 GRU 内部结构图对于第 t 时刻的输入 x_t , 正向 GRU 和反向 GRU 得到的隐状态分别为 $\vec{h}_t$ 和 $\bar{h}_t$, 则 BiGRU 在第 t 时刻的隐状态 h_t 由 $\vec{h}_t$ 和 $\bar{h}_t$ 拼接而得, 具体的计算过程如公式(1-3)所示。

$$\vec{h}_t = \overrightarrow{GRU}(x_t, \vec{h}_{t-1}) \quad (1)$$

$$\bar{h}_t = \overleftarrow{GRU}(x_t, \bar{h}_{t-1}) \quad (2)$$

$$h_t = [\vec{h}_t, \bar{h}_t] \quad (3)$$

注意力机制(Attention Mechanism)通过对不同特征分配权重, 使得模型能够集中关注于对当前输出数据有用的特征信息, 从而提高模型的准确率[7]。注意力机制, 如图 3 所示。

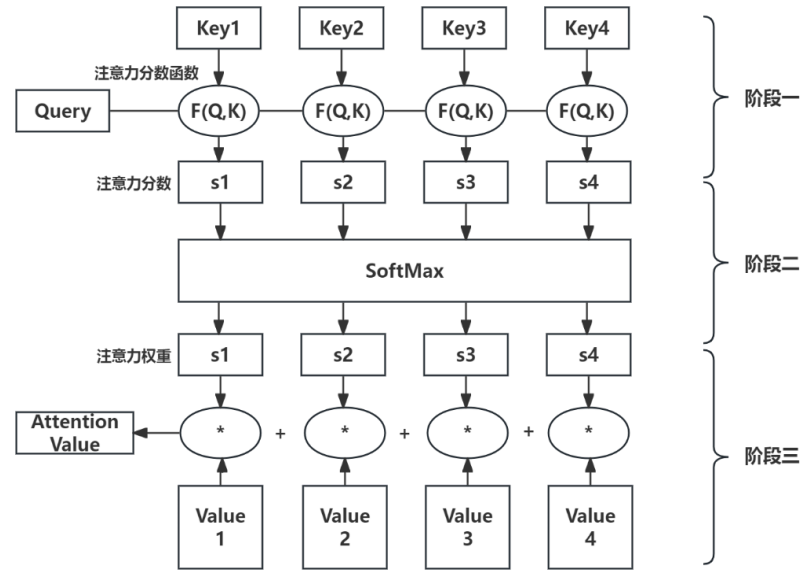


Figure 3. Schematic diagram of Attention

图 3. Attention 的原理图

Attention 的原理分为三步：1、对 query 与 key 进行相似度计算，获得注意力分数；2、归一化注意力分数，获得直接可用的注意力权重；3、将权重与 value 进行加权求和，获得最终输出。

对于相同维度 d 下的 query 和 key，直接计算两者的内积作为注意力分数，其计算公式如式(4)所示。

$$s(q, k_i) = \frac{q \cdot k_i}{\sqrt{d}} \quad (4)$$

其中， q 表示 query， k_i 表示第 i 个 key，而分母 $\sqrt{d}$ 是缩放因子，用于调节内积结果。

$$a(q_i, k_s) = \frac{\exp(s(q_i, k_s))}{\sum_{i=1}^N \exp(s(q_i, k_i))} \quad (5)$$

其中， q_i 表示当前时刻的 query， k_s 表示当前的第 s 个 key， N 表示 key 的总个数。

最后，对每个 key 所对应的权重和 value 进行加权求和，得到最终的输出，其计算式如公式(6)所示。

$$Attention(q_i, K, V) = \sum_s^m a(q_i, k_s) v_s \quad (6)$$

残差网络模型结构如图 4 所示：

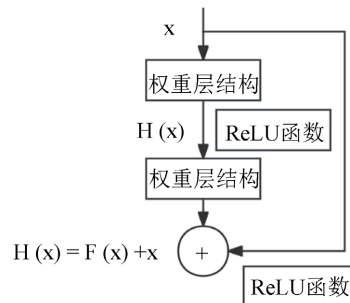


Figure 4. Residual network model structure

图 4. 残差网络模型结构

x 为输入, $H(x)$ 为特征的输出, $F(x)$ 为残差, 其表达式如公式(7)所示。

$$F(x) = H(x) - x \quad (7)$$

特征信息 X 可以直接与后面的层相互连接, 这样使得后面的层可以学习到残差值。残差结构通过恒等映射来增大模型的深度, 其基本的运算如公式(8)所示。

$$x_L = x_l + \sum_{i=1}^{L-1} F(x_i, w_i) \quad (8)$$

其中 x_L 为第 L 层深度单元特征信息的表示。当残差值等于 0 时, 残差网络相当于进行恒等映射, 使得模型的精度不会受到影响。事实上, 由于数据的复杂性和多样性, 残差值通常不为 0, 这意味着模型在不断堆叠层的过程中学习到了更多的特征信息。

最后输出层采用条件随机场(CRF)进行序列标注预测, 负责对前序多层神经网络抽取的特征序列进行最优路径解码。CRF 模型是一种判别式模型[8], 通常应用于序列标注任务[9]。序列标注的核心目标是为句子中的每个词语分配一个合适的标签, 从而形成一个完整的标注序列。

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某省级三甲医院脱敏后的 1000 份电子病历, 结合卫健委《病历书写基本规范(试行)》《浙江省病历质量管理规定》《浙江省住院病历质量检查评分表(2021 版)》的相关内容, 总结归纳出了需要标记的 8 种实体类型。实验采用 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。

Table 1. Entity type number statistics of medical record dataset

表 1. 病历数据集的实体类型数量统计

实体类型	实体示例
疾病(Dis)	2351
身体部位(Bod)	6753
药物(Dru)	9817
手术(Ope)	305
其他治疗(Tre)	3908
症状(Sym)	11,572
检验(Itc)	4580
检查(Pro)	3891

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 F1 值, 如公式(9-11)所示。

$$P = \frac{T_p}{T_p + F_p} \times 100\% \quad (9)$$

$$R = \frac{T_p}{T_p + F_N} \times 100\% \quad (10)$$

$$F_1 = \frac{2PR}{P + R} \times 100\% \quad (11)$$

4.3. 实验结果

实验一：模型综合性能比较实验

从表 2 所示, 四个命名实体识别模型中, BiGRU-Attention 由于参数较少, 在训练和推理时比 BiLSTM 版本更快, 但是其精确率最低, 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; BERT-BiLSTM-CRF 模型更擅长处理长期依赖问题, 因此对于需要复杂上下文理解的任务表现更好, 但是训练和推理速度较慢; FastText-CNN 模型精确率最低; 相对而言, 使用残差注意力网络的 BBRAC 模型在准确率、召回率和 F1 值上均有所提升, 表明模型具备更强的全局上下文信息提取能力。这是因为残差注意力网络保留了整体序列的特征, 同时缓解深层神经网络训练过程中可能的梯度消失、爆炸问题以及模型退化问题, 提高了序列标注任务的性能, 尤其是在需要捕捉上下文信息、长距离依赖关系以及优化整个序列标签分配的任务。

Table 2. Performance comparison of different named entity recognition models

表 2. 不同命名实体识别模型效果对比

模型	P	R	F1
BiGRU-Attention	0.75	0.80	0.77
FastText-CNN	0.71	0.77	0.74
BERT-BiLSTM-CRF	0.76	0.81	0.78
BBRAC	0.80	0.82	0.81

实验二：消融实验

如表 3 所示, 通过消融实验来检验所提出模型模块的有效性, 可以得出以下结论: (1) BiGRU 模块和 Residual Attention 模块对于本文模型框架的重要性是不可忽视的; (2) 相比之下, BiGRU 模块的重要性更为突出。

Table 3. Ablation study results comparison

表 3. 消融实验结果对比

模型	P	R	F1
BERT	0.65	0.60	0.62
BERT-BiGRU	0.75	0.69	0.72
BERT-BiGRU-Residual Attention	0.78	0.75	0.76
BBRAC	0.80	0.82	0.81

结果显示, 该模型在各个医疗实体类别上均表现出了较强的适应性和区分能力, 总体略优于其他模型效果。通过对各实体类别识别准确率和 F1 值等关键指标进行逐一分析, 证实了该模型能够有效地挖掘医疗文本中蕴含的丰富语义信息, 并基于此进行精准且全面的实体标注, 同时也进一步验证了其在解决医疗领域复杂实体识别问题中的有效性和可靠性。

5. 结束语

在本研究中, 针对实际医疗文本中因语义复杂性和长短距离依赖导致命名实体识别准确率低的问题, 提出了一种结合双向语义与残差注意力机制的命名实体识别方法, 通过整合先进的语义提取技术, 有效提升了实体识别的精确度。通过 BERT-wwm 预训练模型与 BiGRU 网络的结合, 本模型成功处理了医疗文本中多维语义的复杂性和长距离依赖关系。同时, 加入的残差注意力网络确保模型能够专注于关键信息的提取而不丢失整体语义结构的特征, 最终利用条件随机场进行序列标注预测, 取得了理想的实验效果。实验结果表明, 提出的模型在准确率、召回率及 F1 值等评价指标上均优于传统方法, 展现了其在命名实体识别任务中的强大适应性。未来的研究可以继续优化模型结构, 探索更多医疗场景下的实体识别应用, 并结合更丰富的语料库, 以进一步提升模型的实用性与泛化能力。

参考文献

- [1] 杜晋华, 尹浩, 冯嵩. 中文电子病历命名实体识别的研究与进展[J]. 电子学报, 2022, 50(12): 3030-3053.
- [2] 赵铁军, 许木璠, 陈安东. 自然语言处理研究综述[J/OL]. 新疆师范大学学报(哲学社会科学版), 1-23. <https://doi.org/10.14100/j.cnki.65-1039/g4.20230804.001>, 2024-10-15.
- [3] Radford, A., Narasimhan, K., Salimans, T., *et al.* (2018) Improving Language Understanding by Generative Pre-Training. Preprint, OpenAI.
- [4] 杨飞洪, 王序文, 李姣. 基于 BERT-TextCNN 模型的临床试验筛选短文本分类方法[J]. 中华医学图书情报杂志, 2021, 30(1): 54-59.
- [5] Li, Y., Ding, Z., Ma, Z., Wu, Y., Wang, Y., Zhang, R., *et al.* (2023) Few-Shot Relation Classification Based on the BERT Model, Hybrid Attention and Fusion Networks. *Applied Intelligence*, **53**, 21448-21464. <https://doi.org/10.1007/s10489-023-04634-0>
- [6] Radford, A., Wu, J., Child, R., *et al.* (2019) Language Models Are Unsupervised Multitask Learners. Preprint, OpenAI.
- [7] Zhu, E., Sheng, Q., Yang, H., Liu, Y., Cai, T. and Li, J. (2023) A Unified Framework of Medical Information Annotation and Extraction for Chinese Clinical Text. *Artificial Intelligence in Medicine*, **142**, Article 102573. <https://doi.org/10.1016/j.artmed.2023.102573>
- [8] Liu, C. and Yang, S. (2022) Using Text Mining to Establish Knowledge Graph from Accident/incident Reports in Risk Assessment. *Expert Systems with Applications*, **207**, 117991. <https://doi.org/10.1016/j.eswa.2022.117991>
- [9] 郁圣卫, 卢奇, 陈文亮. 基于领域情感词典特征表示的细粒度意见挖掘[J]. 中文信息学报, 2019, 33(2): 112-121.