

基于膨胀卷积神经网络的中文医疗命名实体识别研究^{*}

胡海洋 赵从朴 马 琰 姜会珍 张 晶

(中国医学科学院 北京协和医学院 北京协和医院疑难重症及罕见病国家重点实验室 信息中心 北京 100730)

朱卫国

(1 中国医学科学院 北京协和医学院 北京协和医院疑难重症及罕见病国家重点实验室 信息中心 北京 100730

2 中国医学科学院 北京协和医学院 北京协和医院全科医学科 北京 100730)

〔摘要〕 介绍国内外电子病历命名实体识别研究现状, 阐述基于注意力机制和膨胀卷积神经网络的电子病历命名实体识别模型构建方法、结构以及实验环境、具体设计和结果, 实验表明该模型具有更好的命名实体识别效果。

〔关键词〕 命名实体识别; 膨胀卷积神经网络; 字嵌入; 自然语言处理

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2021.09.008

Study on Chinese Medical Named Entity Recognition Based on the Dilated Convolutional Neural Network HU Haiyang, ZHAO Congpu, MA Lian, JIANG Huizhen, ZHANG Jing, Department of Information Center, State Key Laboratory of Complex Severe and Rare Diseases, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100730, China; ZHU Weiguo, 1Department of Information Center, State Key Laboratory of Complex Severe and Rare Diseases, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100730, 2Department of Primary Care and Family Medicine, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100730, China

〔Abstract〕 The paper introduces the study status of Named Entity Recognition (NER) in Electronic Medical Records (EMR) at home and abroad, expounds the building method and structure of the NER model of EMR based on the attention mechanism and dilated Convolutional Neural Network (CNN) and the experimental environment, design and results. The experiment shows that the model has better effect for NER.

〔Keywords〕 Named Entity Recognition (NER); dilated Convolutional Neural Network (CNN); word embedding; Natural Language Processing (NLP)

〔收稿日期〕 2021-03-29

〔作者简介〕 胡海洋, 初级职称; 通讯作者: 朱卫国, 博士, 副教授, 副主任医师。

〔基金项目〕 中国医学科学院医学与健康科技创新工程“医学大数据信息采集和分析评估”(项目编号: 2016-I2M-2-004); 科技部国家重点研发计划“数字诊疗装备研发”重点专项(项目编号: 2018YFC0116905)。

1 引言

在医疗信息化领域运用自然语言处理 (Natural Language Processing, NLP) 技术自动抽取和分析电子病历文本中的信息逐渐成为研究热点, 其中医学命名实体识别 (Named Entity Recognition, NER) 是最基本和最重要的医学自然语言处理任务。现有模型大多使用循环神经网络 (Recurrent Neural Network, RNN), 计算速度较慢且较难达到实际应用要求。针对上述问题, 本文以临床医疗知识发现为导向, 在医学自然语言处理平台基础上, 提出基于多头自注意力机制的膨胀卷积神经网络模型, 通过字嵌入和位置嵌入算法减轻模型对分词效果依赖; 使用膨胀卷积神经网络模型进行训练, 开展医疗命名实体识别及阴阳性判断研究。

2 研究现状

2.1 国外

由于欧美国家医疗信息化发展较早, 英文电子病历命名实体识别研究已进行多年^[1], 早期研究通常基于总结归纳的方法, 即利用词典或者规则对医学实体进行映射和匹配。其中 IBM 的 MedKAT^[2] 利用国际疾病分类肿瘤学分册 (ICD - O) 实例化临床文本病理报告中的病症信息, 由梅奥诊所搭建的 eTAKES 平台^[3] 利用医疗领域术语词典抽取病历信息并建立映射编码, 但基于规则和词典系统高度依赖于专家归纳, 系统对新实体的抽取效果较差、效率较低。发展至今电子病历命名实体识别研究主要采用机器学习算法, 目前很多机器学习算法已应用于 NER, 包括隐马尔可夫模型 (Hidden Markov Model, HMM)^[4], 最大熵模型 (Maximum Entropy, ME)^[5], 条件随机场 (Conditional Random Fields, CRF)^[6] 等。虽然基于机器学习方法能取得较好效果, 但是对特征选取的要求较高, 需要引入大量人工构建的特征, 人工选取出反映实体信息特性的特征集合。Bhatia P、Celikkaya B 和 Khalilia M^[7] 对医疗文本命名实体识别任务提出一种新型的端到端

Transformer 模型, 采用共享编码器分别解码方式提高性能。Weegar R、Pérez A 和 Dalianis H 等^[8] 采用集成学习方法, 针对西班牙和瑞典 4 个不同病历数据集分别使用支持向量机 (Support Vector Machine, SVM) 算法、CRF 和 Perceptron 方法训练模型。

2.2 国内

在中文电子病历实体识别领域, 由于中文语言特点和电子病历结构不同, 国内研究技术发展历程也不同。如许源、葛艳秋和王强等^[9] 利用 CRF 与 RUTA 规则相结合的医学实体识别方法; Cai X、Dong S 和 Hu J^[10] 提出一种基于自我注意机制的 LSTM - CRF 模型, 将词性信息纳入深度学习模型中以提高中文实体边界检测准确性。Wang H、Zhang W 和 Zeng Q 等^[11] 采用一种规则和 CRF 相结合的模型, 针对肝癌病例构建 1 个包含 115 份病历的语料库。目前部分深度学习模型, 如 RNN 和卷积神经网络 (Convolutional Neural Network, CNN), 在中文自然语言处理任务中得到广泛运用, Yin M、Mou C 和 Xiong K 等^[12] 使用 CNN 提取汉字部首级特征并使用自注意机制捕获字符之间依赖关系, 取得较好效果。

3 资料与方法

3.1 研究数据集

本研究所用电子病历来自北京协和医院普通内科共 685 份不明原因发热患者, 病历内容包括入院、出院、病程记录 3 部分。入院记录包括患者主诉、现病史、既往史、体格检查等信息。病程记录是医生针对患者病情和诊疗过程所进行的连续性记录, 内容包括患者病情变化、复查重要结果及临床意义、上级病房意见、会诊意见、医师分析讨论意见、诊疗措施及效果、医嘱变化及原因、向患者和其近亲属通知的重要问题等。出院记录包括患者出院时的诊断结论和医嘱。所有数据均进行脱敏处理。

3.2 协和医学自然语言处理平台 (PUMCH - mNLP)

提供医疗场景下病历文本分析及挖掘的核心工

具，旨在帮助医院高效处理病历数据。通过该平台可将非结构化的医学文本转化为含有关键医学信息的结构化数据，以 json 格式输出或存储于数据库中。

3.3 数据实体标注

许源、葛艳秋和王强等^[9]将入院记录中的命名实体类型分为 5 类，即临床症状、危险因素、检验检查、治疗、药物。夏光辉、李军莲和邢宝坤等^[13]构建的电子病历语料库中实体类型分为疾病、症状、检查和治疗 4 种。本研究借鉴上述语料库构建经验，将电子病历中的实体分为症状、疾病、检验项目、检查项目、用药和过敏史 5 种类型，其中症状、疾病、用药、过敏史的类型区分阴阳性。参考统一医学语言系统（Unified Medical Language System, UMLS）对概念类型的界定方法，本研究定义的命名实体遵循实体中不重叠、不嵌套和不加标点符号（如逗号、句号等）的原则。语料标注工作初期利用 PUMCH - mNLP 标注，在具有医学专业背景相关人员协助修订下完成标注。

4 基于膨胀卷积神经网络的电子病历命名实体识别模型

4.1 概述

电子病历中文本序列 $T = (t_1, t_2, \dots, t_n)$ ， t_i 为该序列中第 i 个字符，需要生成对应输出序列 $\text{Output} = o_1, o_2, \dots, o_n, o_i \in L, L$ 为所有标签集合，最后通过解码算法从 Output 中解码出实体。模型共由嵌入层、注意力层、卷积层、输出层 4 部分组成，嵌入层使用字嵌入和位置嵌入算法，解决分词算法错误及注意力层句子间依赖性，卷积层对注意力层输出进行卷积操作，输出层由两层全联接层组成，采用切片式算法进行实体分割，见图 1。

4.2 嵌入层

针对常用的中文分词算法可能出现的错误，将临床文本序列输入表示为字向量嵌入（Word Embeddings）和位置嵌入（Position Embeddings）叠加组成。将序列中每 1 个位置编码为 1 个向量并将其

融入字向量中，字向量嵌入记录每个字的信息，位置嵌入记录字的顺序信息，见图 2。

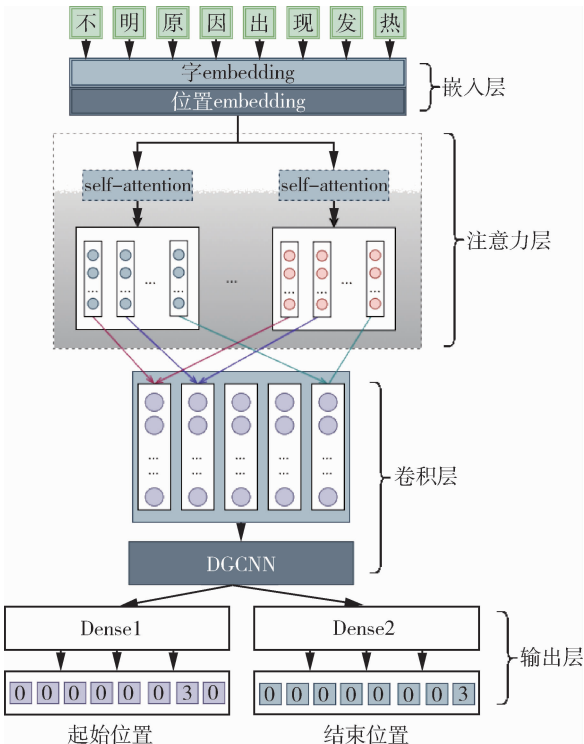


图 1 模型总结构

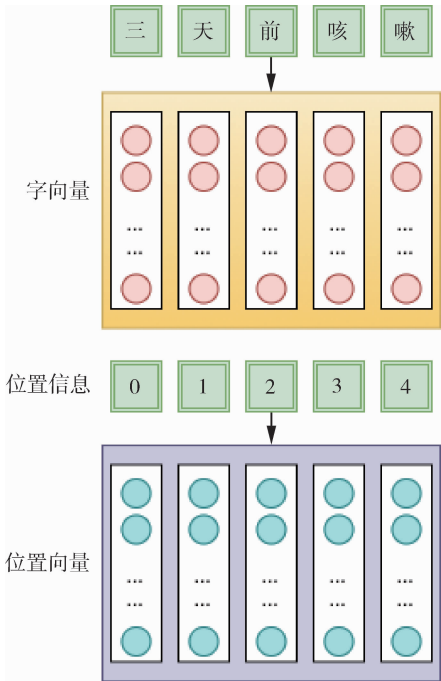


图 2 字向量、位置向量形成过程

4.3 多头自注意力层

注意力机制 (Attention Mechanism) 可以被高度抽象为模型。注意力机制最初应用于数字图像处理领域, 后逐渐应用于自然语言处理领域。注意力函数可以看作 1 个查询 Query (Q) 到一系列键 Key (K) 再到值 Value (V) 对的映射, 通过计算查询和键的相似度再乘以相应值, 求和后得到最终参数, 具体计算公式为:

$$\text{Attention}(Q, S) = \sum_{i=1}^{l_k} \text{Similarity}(Q, K) * V$$

计算自注意力时, 通常取 Query = Key = Value, 用实测点程序计算相似性, 便于内部程序变大, 用缩放点积的程序计算得到的点积, 缩小 $\sqrt{d_k}$ 倍 ($\sqrt{d_k}$ 为向量维度), 公式为:

$$\text{Attention}(Q, S, V) = \text{solfmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

多头自注意力是对自注意力的改进。首先, 使用不同矩阵对 Q 、 K 、 V 进行线性变换, 重复 h 次; 其次, 执行注意力函数来生成平行于 Q 、 K 和 V 的输出值, 每次线性变换都可以得到; 最后, 对输出值进行线性变换, 所得值为多头自注意力的结果。多头自注意力机制通过 h 次自注意力计算, 可以从不同维度中学习相关信息, 提高其连续抽取能力。本研究利用多头自注意力机制在中文病历文本内部结构中获取字符间的长距离依赖关系, 具体公式如下:

$$\text{MultiH}(Q, K, H) \text{ Concat} = (h_1, h_2, \dots, h_n) W^0$$

4.4 基于膨胀卷积神经网络的卷积层

4.4.1 卷积神经网络 一种经典的深度神经网络结构, 是多层感知机 (Multilayer Perceptron, MLP) 的变体模型。CNN 从生物学概念演变而来, 是近年来发展迅速的有效识别方法。与传统方法相比其具有权重共享和训练数据自动特征提取的优点^[14]。卷积运算是 CNN 重要特征。通过特征映射组织卷积层中的每个单元连接到前一层的局部感受野, 卷积运

算利用卷积核和局部感受野进行, 然后通过非线性运算获得激活函数的输出, 得到特征值。一般神经网络中的卷积运算以二维形式出现, 然而因为自然语言处理任务的特殊性, 卷积运算在自然语言处理任务中以一维卷积的形式出现。对于字符序列 T , 使用一维卷积核 W 卷积, 结果 c 可由以下公式得出^[15]:

$$c_t = W_c \oplus x_{t \pm k}$$

4.4.2 膨胀卷积神经网络 传统卷积神经网络对序列建模受到卷积核大小的限制, 但在自然语言处理中文字通常具有依赖性, 如果要抓取距离较远的依赖关系必须线性地堆叠模型层数。为抓取文字更长的依赖关系, 相对于 BiLSTM 长距离依赖性, 对膨胀卷积神经网络局部单元进行卷积编码可以更好地实现对医疗实体的识别。一般情况下 CNN 滤波器通过在输入矩阵的区域上连续滑动以执行卷积运算, 该区域普遍是连续的。然而膨胀卷积神经网络在滤波器中加入扩张宽度, 使得输入矩阵不再连续且每次进行卷积运算时都会跳过膨胀宽度中的输入数据。在扩张卷积运算过程中, 滤波器获得的输入矩阵上数据较多, 但滤波器本身尺寸不变而是扩展其感受域, 看起来像是在“膨胀”, 因此称之为膨胀卷积神经网络。与一般 CNN 相比膨胀卷积神经网络可以在不进行池操作的情况下获得更大感知域, 减少信息丢失。本文用膨胀卷积神经网络代替传统卷积神经网络。膨胀卷积工作机制是跳过与中心直接相邻的 $p-1$ 个输入, 并连接到中心和与中心相邻的第 p 个输入, 其中 p 为膨胀卷积速率。在感受域逐渐扩展、层数不断增加的过程中, 神经网络参数呈线性增长, 感受域呈指数增加。本文使用膨胀率依次为 1、2、5、1 的 4 层膨胀卷积神经网络, 对注意力层输出的向量进行卷积, 自动提取文本中特征, 输出相应特征向量。仅经过 3 步膨胀扩展后, 感受域就已扩展到输入矩阵中的所有数据, 其中每层参数数量相同, 可有效减少训练时的参数, 加快训练速度, 见图 3。

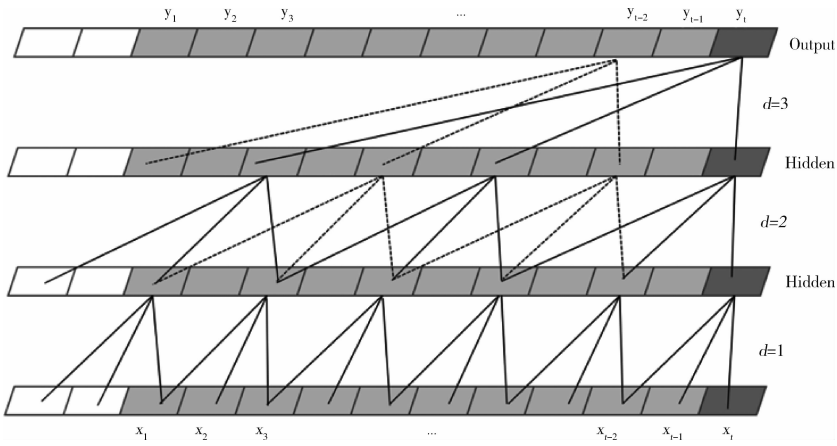


图 3 膨胀卷积神经网络原理

4.5 输出层

4.5.1 全连接层 (Full Connected Layer, FCL) 对于卷积层的输出 $A = (a_1, a_2, \dots, a_n)$, 分别使用两个以 sigmoid 函数作为激活函数的全连接层, 得到一个实体头部预测序列 $SP = (sp_1, sp_2, \dots, sp_n)$ 与一个实体尾部预测序列 $EP = (ep_1, ep_2, \dots, ep_n)$ 。其中 sp_i, ep_i 表示第 i 个位置的字符是某一特定实体的头部、尾部概率, 以概率值最大的类型作为位置标签。使用一种切片式解码算法对 S 和 E 进行处理, 将实体属性与阴阳性从原始文本 T 中提取出来。切片式解码算法利用预测得到的实体首部和尾部位置在文本上进行切片以获得实体。

4.5.2 损失函数 模型训练使用基于交叉熵的损失函数, 其形式为:

$$\text{Loss} = CE(SP, \text{TrueS}) + CE(EP, \text{TrueE})$$

其中 $CE(SP, \text{TrueS}) + CE(EP, \text{TrueE})$ 分别为首部位置预测、尾部位置预测与真实值的交叉熵, 将两个交叉熵相加作为模型最终的损失函数。

5 实验设计与分析

5.1 实验环境、评测标注

实验软件编程环境为 Python 3.5, 使用 TensorFlow 构建膨胀卷积神经网络模型进行训练和测试。所有实验均采用标准 Bakeoff 评分程序计算常用评

测指标: P (Precision, 准确率), R (Recall, 召回率), $F1$ (召回率和准确率的调和平均值)。计算公式如下:

$$P = TP / (TP + FP)$$
$$R = TP / (TP + FN)$$
$$F1 = 2 * P * R / (P + R)$$

5.2 实验设计

分别使用 BiLSTM、BiLSTM + CNN 和 BiLSTM + attention + CNN 3 种模型作为对比模型, 与命名实体识别中常用的双向长短时记忆网络模型做对比, 对于提出的膨胀卷积神经网络模型进行多种对比实验: 使用基本的双向长短时记忆网络模型 BiLSTM 进行训练; 利用 BiLSTM 算法对模型进行编码, 卷积神经网络解码 (BiLSTM + CNN); 通过双向长短时记忆网络模型加入自注意力机制, 卷积神经网络解码 (BiLSTM + Attention + CNN), 见表 1。

表 1 模型评价指标对比

模型	P	R	$F1$
BiLSTM	88.44	81.76	84.97
BiLSTM + CNN	91.00	81.04	85.73
BiLSTM + Attention + CNN	89.49	83.35	86.31
Attention + DGCNN	91.81	85.22	88.39

5.3 结果分析

通过对比 $F1$ 分数, 膨胀卷积神经网络模型比 BiLSTM 基线模型提高了 3.42%, 比 BiLSTM + CNN

提高了 2.66%，比 BiLSTM + Attention + CNN 提高了 2.04%。说明膨胀卷积神经网络模型在处理中文命名实体识别和阴阳性判断中具有较好优势。

6 结语

根据普通内科医生实际科研需求，建立基于自注意力机制的膨胀卷积神经网络命名实体识别模型，实现总体平均准确率 0.918、平均召回率 0.852、平均 F score 0.883。实验结果表明本文提出的膨胀卷积神经网络模型相较于传统 BiLSTM 方法，能更好地提高命名实体识别效果、加快计算速度，同时该模型对临床文本中实体阴阳性也有较好判别。本文仅针对不明原因发热电子病历数据进行命名实体识别与阴阳性识别研究，构建对应单病种自然语言提取模型，还需要通过更多其他疾病电子病历数据提取进行验证，同时在核磁共振报告、CT 报告、手术记录等文本提取中的应用也需要进一步研究。膨胀卷积神经网络算法虽然有较好识别效果，但仍有进一步改进空间，后续将寻找更高效算法进行优化。本文对自然语言识别模型的研究已应用于具有自主知识产权成果的协和医学自然语言处理平台，目前该平台已获得 1 项软件著作权。平台内集成病历结构化、分词标注、实体抽取、词频统计、词云统计等多项功能模块，目前已在北京协和医院妇产科、内科等临床科室得到应用，缓解了临床工作者科研数据统计压力，节省大量人力成本。

参考文献

- 1 Yetisgen M, Vanderwende L. Automatic Identification of Substance Abuse from Social History in Clinical Text [C]. Vienna: Conference on Artificial Intelligence in Medicine in Europe, 2017.
- 2 Coden A, Savova G, Sominsky I, et al. Automatically Extracting Cancer Disease Characteristics from Pathology Reports into a Disease Knowledge Representation Model [J]. Journal of Biomedical Informatics, 2009, 42 (5): 937 - 949.
- 3 Savova G K, Masanz J J, Ogren PV, et al. Mayo Clinical Text Analysis and Knowledge Extraction System (cTAKES):

- architecture, component evaluation and applications [J]. Journal of the American Medical Informatics Association Jamia, 2010, 17 (5): 507 - 513.
- 4 Rabiner L R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition [J]. Proceedings of the IEEE, 1989, 77 (2): 257 - 286.
- 5 Berger A L. A Maximum Entropy Approach to Natural Language Processing [J]. Computational Lingus, 1996, 22 (1): 39 - 71.
- 6 Charles Sutton. Andrew Mccallum. An Introduction to Conditional Random Fields [J]. Foundations & Trends in Machine Learning, 2010, 4 (4): 267 - 373.
- 7 Bhatia P, Celikkaya B, Khalilia M. Joint Entity Extraction and Assertion Detection for Clinical Text [C]. Florence: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019.
- 8 Weegar R, Pérez A, Dalianis H, et al. Ensembles for Clinical Entity Extraction [J]. Procesamiento del Lenguaje Natural, 2018 (60): 13 - 20.
- 9 Cai X, Dong S, Hu J. A Deep Learning Model Incorporating Part of Speech and Self - matching Attention for Named Entity Recognition of Chinese Electronic Medical Records [J]. BMC Med Inform Decision Makon, 2019 (65): 1472 - 1477.
- 10 许源, 葛艳秋, 王强, 等. 基于 CRF 与 RUTA 规则相结合的卒中入院记录医学实体识别及应用 [J]. 中山大学学报 (医学版), 2018, 39 (3): 455 - 462.
- 11 Wang H, Zhang W, Zeng Q, et al. Extracting Important Information from Chinese Operation Notes With Natural Language Processing Methods [J]. Journal of Biomedical Informatics, 2014 (48): 130 - 136.
- 12 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, Academic Press Inc., 2019 (98): 103289.
- 13 夏光辉, 李军莲, 邢宝坤, 等. 基于中文病例报告文献的医学诊疗命名实体识别研究 [J]. 医学信息学杂志, 2019, 40 (6): 54 - 59.
- 14 Lawrence S, Giles C L, Tsoi A C, et al. Face Recognition: aconvolutional neural - network approach [J]. IEEE Transactions on Neural Networks, 1997, 8 (1): 98 - 113.
- 15 王星, 李超, 陈吉. 基于膨胀卷积神经网络的中文分词方法 [J]. 中文信息学报, 2019, 33 (9): 24 - 30.