融合词向量及词属性推理的中文电子病历实体识别方法*

武学鸿

（1 中南大学计算机学院 长沙 410083
2 湖南科创信息技术股份有限公司 长沙 410205）

杨 峰

（中南大学计算机学院 长沙 410083）

李建华

（湖南科创信息技术股份有限公司 长沙 410205）

徐 倩

（中南大学湘雅医学院 长沙 410003）

〔摘要〕 阐述基于 BiLSTM－CRF 基准模型，融合词向量及词属性推理机制实现中文电子病历命名实体识别的方法，包括医学语料库构建与词向量训练、融合词属性推理机制等，分析实验结果，指出医学领域词向量及词属性推理机制的引入有助于提升中文电子病历命名实体识别效果。

〔关键词〕 领域词向量；词属性推理机制；中文电子病历；命名实体识别；自然语言处理
〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2022.07.007

A Method for Entity Recognition of Chinese Electronic Medical Record Based on Word Vector Combined with Word Attribute Reasoning

WU Xuehong, *1School of Computer Science and Engineering*, *Central South University*, *Changsha 410083*, *2Research Institute*, *Hunan Creator Information Technology Co. Ltd.*, *Changsha 410205*, *China*; *YANG Feng*, *School of Computer Science and Engineering*, *Central South University*, *Changsha 410083*, *China*; *LI Jianhua Research Institute*, *Hunan Creator Information Technology Co. Ltd.*, *Changsha 410205*, *China*; *XU Qian*, *Xiangya School of Medicine*, *Central South University*, *Changsha 410003*, *China*

〔**Abstract**〕 The paper introduces the method of combining word vector and word attribute reasoning to realize Named Entity Recognition (NER) of Chinese Electronic Medical Record (EMR) based on BiLSTM－CRF benchmark model, including medical corpus construction, word vector training, integration of word attribute reasoning, etc., analyzes the experimental results, and points out that the introduction of word vector and word attribute reasoning in medical field can improve NER effect of Chinese EMR.

〔**Keywords**〕 domain word vector; word attribute reasoning; Chinese Electronic Medical Record (EMR); Named Entity Recognition (NER); Natural Language Processing (NLP)

1 引言

中文电子病历的命名实体识别研究最初主要是基于词典和规则的方法[1－4]，形成了医疗领域 3 个代表性通用工具，即 MedLEE、MedKA 和 cTAKES[5]。目

〔修回日期〕 2021－10－11
〔作者简介〕 武学鸿，工程师，发表论文 7 篇。
〔基金项目〕 国际科技创新合作基地项目“湖南省人工智能与医学大数据国际科技创新合作基地”（项目编号：2019CB1007）。

前最常用的是基于机器学习的研究方法[6-10]。Wang Y、Yu Z 和 Chen L 等[11]通过基于规则和条件随机场的算法对病历数据进行实体识别。Lei J、Tang B 和 Lu X 等[9]对比分析条件随机场、支持向量机、最大熵和结构支持向量机等不同算法在中文电子病历命名实体识别中的效果。于楠、王普和翁壮等[12]提出基于多特征融合的中文电子病历命名实体识别方法，多特征融合的条件随机场模型在仅选取基本特征、上下文窗口为 7 时识别效果最好。中文命名实体识别研究方法常将词向量、字向量或者二者的简单结合作为分类模型的输入特征[13-14]。本文构建医学语料库来训练医学专业词向量，同时设计词属性推理模型，以词语中的每个字作为输入得到词语的词属性向量，将其作为词语的内部特征，最后将二者结合作为模型的输入特征，见图 1。

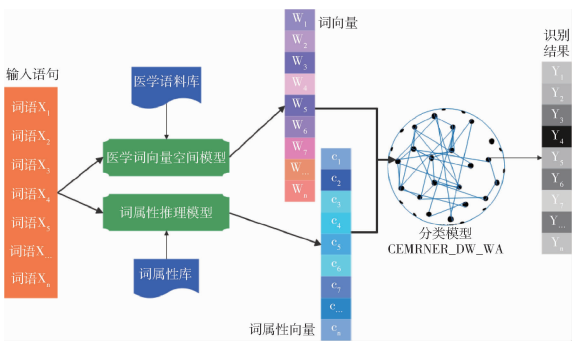

图 1 融合词向量及词属性推理的中文电子病历命名实体识别方法

2 医学语料库构建与词向量训练

2.1 构建医学语料库

首先构建丰富的医学语料库，通过收集及预处理获取语料库内容，见表 1。

表 1 医学语料库

语料库组成	数据量	数据来源
线上文本集	70 万条	健康信息网
官方文本集：词条型	9 万条	卫生官网
官方文本集：资料型	20 本	卫生官网
电子病历文本集	1 000 篇	医院临床

2.2 词向量训练

2.2.1 相似词分析法 采用 skip-gram 方法依次将语料分割为单个词、初始化向量、二次抽样、使用词典查找方法和正则表达式来修正、训练并保存 skip-gram 隐含层节点参数，从而生成医学专业领域词向量。采用欧式距离的倒数表示词语之间的相似度，欧式距离越小则词语之间语义越相关。假设有两个词语 A、B，向量分别为 a（x_{11}，x_{12}，…，x_{1n}），b（x_{21}，x_{22}，…，x_{2n}），则 A 与 B 的欧氏距离 d_{ab}的计算公式为：

$$d_{ab} = \sqrt{\sum_{k=1}^{n}(x_{1k} - x_{2k})^2} \qquad \text{公式 1}$$

A、B 之间的相似度 sim_{AB}计算公式为：

$$\text{sim}_{AB} = \frac{1}{d_{ab}} = \frac{1}{\sqrt{\sum_{k=1}^{n}(x_{1k} - x_{2k})^2}} \qquad \text{公式 2}$$

本文列举了部分典型词语，分别计算其在通用领域和专业领域词向量空间模型中排名前 3 的最相似词，见表 2、表 3。对比表 2 及表 3，如“发烧”，在前者中最相似词分别是“发热”“燃烧”“发光”，而在后者中则是“发热”“感冒”“高热”；又如“拍片”在前者中最相似词分别是“拍电影”“拍视频”“照相”，而在后者中则是“CT”“造影”“B 超”。基于常识分析来看，医学领域词向量模型中推荐的相似词更加准确。

表 2 基于通用领域词向量空间中最相似词

词语	相似词及相似度
发烧	发热 0.45，燃烧 0.25，发光 0.20
进食	吃饭 0.74，进餐 0.54，食物 0.26
拍片	拍电影 0.47，拍视频 0.43，照相 0.36
头部	上部 0.63，尾部 0.59，头上 0.41
化疗	癌症 0.53，放疗 0.47，肿瘤 0.31

表 3 基于医学领域词向量空间中最相似词

词语	相似词及相似度
发烧	发热 0.81，感冒 0.74，高热 0.53
进食	饮食 0.73，吃饭 0.70，食纳 0.64
拍片	CT 0.49，造影 0.38，B 超 0.23
头部	头面部 0.74，面部 0.60，头上 0.53
化疗	放疗 0.65，治疗 0.30，手术 0.11

2.2.2 类比推理法 除相似词分析外，类比推理也是一种评价词向量空间模型的有效方法。该方法给出一对有语义联系的词语对 A 与 B，然后给定一个词语 C，可以计算得到与 C 有同类语义联系的词语 D，见表 4。从表 4 中结果可以看出两种词向量空间模型对词语的语义特征提取存在差别。如“紧张”一词，参考“升高 - 降低”词语对的反义性，专业领域词向量模型推理出为“松软”，而通用领域词向量模型推理结果为“放松”。

表 4 类比推理结果

词语对	通用领域词向量	专业领域词向量
头部 - 头痛 腹部 - ?	腹胀	腹痛
升高 - 降低 紧张 - ?	放松	松软
呼吸 - 正常 脉搏 - ?	正常	平稳
血压 - 高血压 血糖 - ?	糖尿病	高血糖

2.2.3 结果分析 综上所述专业领域词向量中词语的语义特征比通用领域词向量更接近词语在医疗语境的含义。

3 融合词属性推理机制

词属性推理机制是指根据词语中的每个字推理出其属性分类信息，再与词语向量结合形成融合后的文本特征输入到分类模型中去。词属性推理机制本质上属于分类任务，为了尽可能地提取到词语中单个字与多个字的信息，使用多卷积核的卷积神经网络模型来实现词属性推理模型，见图 2。

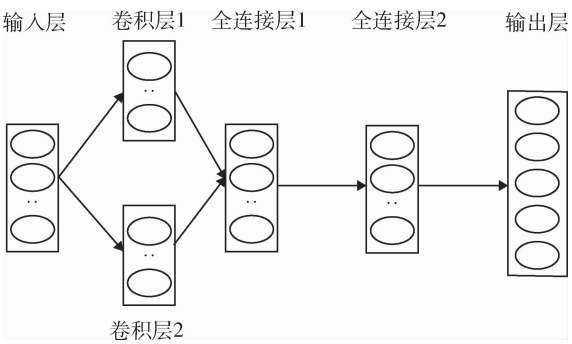

图 2 词属性推理模型结构设计

词语的字向量输入词属性推理模型后，经过卷积层的输出为：

$$c_1 = \text{conv}(f_1, X, \text{‘valid’}) + b_1 \quad \text{公式 3}$$

$$c_2 = \text{conv}(f_2, X, \text{‘valid’}) + b_2 \quad \text{公式 4}$$

式中 conv 表示二维的卷积操作运算，f_1、f_2 表示两个不同大小的卷积核，X 表示输入词语的字向量（x_1，x_2……x_i），“valid”是本文选择的卷积操作处理边界问题的一种方式，b_1、b_2 分别表示两个卷积层的偏置量。将 c_1、c_2 进行连接操作后输入到下一层的全连接层中：

$$c = \text{concat}(c_1, c_2) \quad \text{公式 5}$$

$$m = c \times w_1 + b_3 \quad \text{公式 6}$$

式中 concat 操作表示将两个卷积层结果连接起来。式中 w_1 是全连接层 1 的权重矩阵，b_3 是偏置量。

$$m' = m \times \text{Bernoulli}(p) \quad \text{公式 7}$$

$$n = m' \times w_2 + b_4 \quad \text{公式 8}$$

以上公式中 Bernoulli 函数是为了生成概率向量，也就是随机生成一个只包含 0、1 的向量，其中 p 表示生成过程中的概率。将全连接层 1 的输出结果与概率向量相乘，相当于随机丢弃了其中占比为 p 的信息，这种做法也被称为 dorpout 处理。w_2 是全连接层 2 的权重矩阵，b_4 是偏置量。模型最后一层对全连接层 2 的输出结果进行 softmax 操作，结果作为模型的最终输出，计算公式如下：

$$\bar{y} = \text{softmax}(n) \quad \text{公式 9}$$

4 实验结果分析

4.1 分析工具

为了分析词向量以及词属性推理机制对实验结果的影响，设计 3 个基于 BiLSTM - CRF[15] 的扩展模型：模型 CEMRNER 使用通用领域词向量，模型 CEMNER_ WA 使用通用领域词向量并结合词属性推理机制，模型 CEMRNER_ DW_ WA 使用专业领域词向量并结合词属性推理机制。3 个模型都基于 tensorflow 框架开发实现。

4.2 实验数据来源及格式

实验数据来自于临床电子病历，采用 eHost 工具对 1 000 份电子病历标注并审核确认，标注的实体类

别有：疾病、症状、药品、部位、治疗以及检查。每份电子病历的标注结果都会输出对应的 xml 文件。Annotations 标签为根标签，其属性 textSource 记录了病历文件名称，每个实体信息由 annotation 以及 classMention 两个标签描述，其中 annotation 标签中的 span 子标签描述该实体在病历中的位置，spannedText 为实体内容，classMention 标签中的 metionClass 子标签指明了当前实体类别。标注完成后，将实验数据按照 8：1：1 的比例随机划分出训练集、验证集以及测试集。

4.3 结果分析

对比分析模型 CEMRNER 与 CEMRNER_ DW 在测试集上 *F*1 值可知，结合词属性推理机制的模型 CEMRNER_ WA 较没有引入词属性推理机制的模型 CEMRNER 在总体水平上有所提升，*F*1 值由 0.85 提升到 0.88，从各实体类别来看整体识别效果更好。对比分析模型 CEMRNER_ WA 与 CEMRNER_ DW_ WA 在测试集上的表现可知，使用领域词向量的模型表现更优，总体 *F*1 值从 0.88 上升到 0.90。针对每类实体识别结果，CEMRNER_ DW_ WA 都表现出 1% ~2% 的提升。

5 结语

本文介绍了融合词向量及词属性推理的中文电子病历实体识别方法，以 BiLSTM - CRF 为基准模型分别实现 3 个对比模型。实验结果表明词属性推理机制的引入可以提升模型实体识别效果，同时也证明了专业领域词向量比通用领域词向量在医疗命名实体识别任务中表现更好。

参考文献

1 栗伟，赵大哲，李博，等. CRF 与规则相结合的医学病历实体识别 [J]. 计算机应用研究，2015，32（4）：1082 - 1086.

2 翟菊叶，陈春燕，张钰，等. 基于 CRF 与规则相结合的中文电子病历命名实体识别研究 [J]. 包头医学院学报，2017，33（11）：124 - 125.

3 Xu D，Zhang M，Zhao T，et al. Data - Driven Information Extraction from Chinese Electronic Medical Records [J]. Plos One，2015（10）：e01362708.

4 Cheng M，Li L，Ren Y，et al. A Hybrid Method to Extract Clinical Information from Chinese Electronic Medical Records [J]. IEEE Access，2019（7）：70624 - 70633.

5 杨锦锋，于秋滨，关毅，等. 电子病历命名实体识别和实体关系抽取研究综述 [J]. 自动化学报，2014，40（8）：1537 - 1562.

6 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 Medical Informatics and Decision Making，2019，19（2）：65.

7 Qiu J，Zhou Y，Wang Q，et al. Chinese Clinical Named Entity Recognition Using Residual Dilated Convolutional Neural Network with Conditional Random Field [J]. IEEE Transactions on Nanobioscience，2019，18（3）：306 - 315.

8 Qiu J，Wang Q，Zhou Y，et al. Fast and Accurate Recognition of Chinese Clinical Named Entities with Residual Dilated Convolutions [C]. Madrid：IEEE International Conference on Bioinformatics and Biomedicine - BIBM，2018.

9 Lei J，Tang B，Lu X，et al. A Comprehensive Study of Named Entity Recognition in Chinese Clinical Text [J]. Journal of the American Medical Informatics Association，2014，21（5）：808 - 814.

10 王博冉，林夏，朱晓东，等. Lattice LSTM 神经网络法中文医学文本命名实体识别模型研究 [J]. 中国卫生信息管理杂志，2019，16（1）：84 - 88.

11 Wang Y，Yu Z，Chen L，et al. Supervised Methods for Symptom Name Recognition in Free - text Clinical Records of Traditional Chinese Medicine：An Empirical Study [J]. Journal of Biomedical Informatics，2014（47）：91 - 104.

12 于楠，王普，翁壮，等. 基于多特征融合的中文电子病历命名实体识别 [J]. 北京生物医学工程，2018，37（3）：279 - 284.

13 Hu Q，Liu N，Wang J，et al. An Overlapping Sequence Tagging Mechanism for Symptoms and Details Extraction on Chinese Medical Records [J]. Computers & Electrical Engineering，2021（91）：107019.

14 Wen G，Chen H，Li H，et al. Cross Domains Adversarial Learning for Chinese Named Entity Recognition for Online Medical Consultation [J]. Journal of Biomedical Informatics，2020（112）：103608.

15 Huang Z，Xu W，Yu K. Bidirectional LSTM - CRF Models for Sequence Tagging [EB/OL]. [2020 - 11 - 16]. https：//www.doc88.com/p - 9929655991647.html.