高技术通讯 2024 年 第 34 卷 第 10 期:1058-1069

doi:10.3772/j.issn.1002-0470.2024.10.005

# 引入词汇信息的中文医学命名识别模型研究①

陈 晶②* 孙亚轩** *** 邢珂萱** ***

(*广东海洋大学数学与计算机学院 湛江 524088)

(**燕山大学信息科学与工程学院 秦皇岛 066004)

(***河北省虚拟技术与系统集成重点实验室 秦皇岛 066004)

**摘 要** 医学领域文本存在大量的专业词汇,相比于通用领域更容易出现分词错误和未登录词的问题,其结果会导致上下文语义缺失,并影响命名实体识别(NER)的准确率。为了解决上述问题,本文提出了引入词汇信息的基于门控循环单元的中文医学命名实体识别模型 WI-NER。首先,基于中文医学数据集的特点,描述了中文医学领域的命名实体识别的任务定义、实体位置和实体类别标签,并将模型在嵌入层对匹配专业词的字符进行特征嵌入与向量融合;其次,在上下文编码层添加词汇门控单元,利用循环神经网络的记忆与遗忘机制,自动提取实体识别所需的特征,并通过引入词汇信息和先验知识,实现了中文医学命名实体识别效果的提升;最后,对本模型在 3 个数据集上进行了实验验证,结果表明,本文提出的中文医学命名实体识别模型在准确率方面优于基线模型,达到了预期的医学领域特性。

**关键词** 中文医学命名识别; 先验知识; 嵌入层; 门控单元; 词汇信息

医学文献文本数据包含基因、药物、疾病和蛋白质等多个子领域的实体,而命名实体识别(named entities recognition,NER)的主要任务是识别命名实体的文本范围,将其分类为预定义的类别,它是问答系统、翻译系统和知识图谱的基础[1]。目前,针对医学知识图谱的研究还处于探索阶段,并且不同医学子领域的数据都具有专业性强、关系分布不平衡和结构复杂等特点,这使得知识图谱技术在医疗领域普遍存在效率低、限制多和拓展性差等问题。

研究者们尝试利用机器学习和统计学算法研究命名实体识别和实体抽取的问题,常用的方法包括支持向量机[2]、隐马尔科夫模型[3]、条件随机场[4]和最大熵[5]等。文献[6]基于支持向量机模型进行生物医学命名实体识别,其研究中引入了词缓存、无监督训练等方法,达到了提高训练效果的目的。文献[7]提出了一种最大熵算法,并在 Medline 数据集中对该算法进行了实验验证,结果表明该算法准确率和召回率都得到显著提升。近年来,机器学习中的相关研究成果被广泛地应用在实体抽取任务中。文献[8]提出了双向长短期记忆-条件随机场(bidirectional long short-term memory-conditional random field,BiLSTM-CRF)模型,通过双向长短期记忆网络模型捕获序列依赖信息,并利用条件随机场结构考虑全局序列标注信息。文献[9]提出 Stack-LSTM(stack long short-term memory)模型,该模型将若干实体词集中处理,并考虑了词与词之间的相关性。文献[10]提出 Lattice LSTM (lattice long short-term memory)方法,该方法使用 Lattice 结构来表征单词,并将词语信息有效地整合到基于字符的 BiLSTM-CRF 模型中,有效地缓解了中文分词错误的问题。

① 国家自然科学基金(62172352,61871465,42306218),中央政府引导地方科技发展基金(226Z0102G,226Z0305G),河北省自然科学基金(2022203028)和广东海洋大学科研启动基金(060302102304)资助项目。

② 女,1976 年生,博士,教授;研究方向:社交网络分析,知识建模与知识表示;联系人,E-mail: xychenjing@gdou.edu.cn。

(收稿日期:2023-10-11)

由大语料训练的预训练模型在中文序列标记方面表现出色,应用基于 Transformer 的双向编码器表示(bidirectional encoder representations from transformers,BERT)[11]提取字符特征的模型在中文命名实体识别等方面优于基于静态嵌入的方法。Transformer[12]是 Google 公司为解决 SequencetoSequence[13]模型问题而提出的多头自注意力机制模型,并在机器翻译任务中取得了重大的突破。文献[14]提出了 PLTE(porous lattice transformer encoder)模型,通过引入一种新的多孔机制增强局部建模能力,并使用轴共享结构替换全连接,简化了 Transformer 架构。文献[15]提出了 FLAT(flat lattice transformer)模型,利用 Transformer 中的自注意力机制捕获长距离依赖,设计了一种相对位置编码。该模型解决了潜在词汇冲突,并利用图形处理器(graphic processing unit,GPU)的并行计算能力,显著地提升了运行效率。在引入词汇信息(word information,WI)的方式上,Lattice LSTM[12]和 FLAT Transformer[15]都属于设计了一个动态框架来兼容词汇输入。文献[16]提出了 LR-CNN(lexicon rethinking convolutional neural network)模型,采用卷积神经网络(convolutional neural network,CNN)对字符特征进行编码,同时采取注意力机制融入词汇信息,有效地处理了词汇信息冲突问题。文献[17]提出协作图神经网络(collaborative graph network,CGN)模型,并构建了基于协作的图网络和 3 种不同的建图方式;通过生成对抗网络(generative adversarial network,GAN)进行特征提取,将基于字符的上下文表征与图网络表征加权融合。文献[18]提出基于 lexicon 信息的图网络(lexicon-based graph network,LGN)模型,该模型将每一个字符作为节点,匹配到的词汇信息构成边,并通过图结构实现局部信息的聚合。文献[19]提出了 WC-LSTM 模型,该模型将每个字符为结尾的词汇信息进行固定编码表示,并进行并行化处理。文献[20]提出了 Multi-digraph 模型,该模型基于多图结构显示建模字符和词典的交互,结合长短期记忆网络(long short-term memory,LSTM)和条件随机场(conditional random field,CRF)解决了词典匹配冲突的问题。文献[21]提出了一种在嵌入层引入词汇信息的方法 Soft-lexicon,并对每个字符依次获取各个位置对应所有词汇集合后再进行编码,有效地降低了信息损失的效率。

NER 方法大多是通过结合其他信息源,并利用深度学习或 ChatGPT 模型来提高识别效果。文献[22]在零样本和少样本设定下,全面分析了 InstructGPT 在基础的结构化抽取任务、命名实体识别上的表现。文献[23]针对标注样本稀缺给基于文本的小样本命名实体识别方法带来的问题,提出了一种融合多模态数据的小样本命名实体识别模型,该模型有效地解决了由文本与图像蕴含语义信息粒度不一致导致的模态对齐效果不佳的问题。文献[24]构建了非小细胞肺癌的知识图谱以弥补传统医学知识在数据挖掘和疾病分析方面的不足,并对抽取的 7 类医学实体和 2 万余条实体关系进行验证。文献[25]提出了一种集成字符与词汇信息的中医药古文献命名实体识别的 Lattice LSTM 模型,并对疾病、证候、方剂、症状和药材 5 类实体进行抽取和实验验证。文献[26]针对中文命名实体识别(Chinese named entities recognition,CNER)的现有文献无法有效捕捉汉语的复杂性和模糊性的问题,提出了一种双通道注意力增强中文命名实体识别(dual-channel attention enhancement for Chinese named entity recognition,DAE-NER)模型。该模型通过调整语义特征向量的权值,重构语义权值,减轻上下文语义差异的干扰;为了增强汉语文本中不同粒度的语义表征,在字符和句子 2 个层面设计了注意力增强模块。通过在不同数据集中进行实验,验证了所提出的 DAE-NER 可以有效地提高 CNER 的整体性能[26]。

从关注词语特征和分类结果的角度出发,文献[27]提出了一种基于多元数据嵌入的命名实体模型 MECT(multi-metadata embedding based cross-transformer),该模型融合 Cross-Transformer 提高了中文实体识别的性能。实验结果表明,MECT 可以更好地捕捉汉字的语义信息。文献[28]对生物医学中的文本实体进行识别,并将它们规范化为标准化的疾病词汇,提出了一种新的神经网络模型。该模型将端到端的疾病识别转换为一个动作序列的预

测任务,并通过不同的数据集验证了模型的效果。

文献[29]通过分析文本特征,基于扩展先前次优路径的识别方法,提出了一种新颖的嵌套方法命名实体识别,该方法在每个节点上维护一组隐藏状态并有选择地利用它们为每个级别的识别构建不同的势函数。文献[30]将字符偏旁部首特征与基于字典方法相融合,对 NER BERT 模型进行微调,取得了很好的实验效果。文献[31]以触发词和关系词作为识别实体词的关键特征词,并基于古籍文本的特征,构建 Bert-BiLSTM-MHA-CRF 模型,该模型融合句法特征和 Bert-BiLSTM-MHA-CRF 模型实现了对古籍文本深层次、细粒度的命名实体识别的目的。文献[32]针对法律文书命名实体识别的序列标注模型仅关注字符信息无法获得语义和词语的上下文信息问题,提出了一个融合外部信息特征并对边界限制的司法命名实体识别模型 SBENER(semantic and boundary enhance named entity recognition)。该模型将司法盗窃罪的信息融入字符序列以增强语义特征,解决了序列标注模型丢失词语信息及缺少边界限制的问题,效果优于其他基线模型。

通过对目前的文献分析可知,针对词汇和字符信息特征进行编码时,仅对嵌入层做优化,使得准确率的提升效果有限;由于医疗行业具有特殊性、专业性及隐私性等特点,基于深度学习的方法或 ChatGPT 模型存在研究过程中很难获取到具有规范、统一标注的外部知识的问题。

为了解决上述问题,本文以分析中文医学文本数据集的特点为基础,提出了引入词汇信息(WI)的基于门控循环单元(gated recurrent unit,GRU)的中文医学命名实体识别模型 WI-NER,本文的贡献如下。

(1) 针对中文医学数据集的领域特点进行分析,提出了一种引入词汇信息的基于门控循环单元的中文医学命名实体识别模型 WI-NER;

(2) 该模型以门控循环单元为基础,设计了一种字-词混合粒度的门控循环单元网络,并将词汇信息融入到字粒度的循环神经网络中,达到了提升模型性能的目的。

## 1 WI-NER 模型的总体框架

本文将中文医学领域的命名实体识别任务定义为:对于语料库中的各个句子实例 $s = [c_1, c_2, \cdots, c_n]$,$c_i$ 表示句子中的第 $i$ 个字符,预测序列中每个元素的实体位置标签 $t_{\text{pos}}$ 及实体类别标签 $t_{\text{type}}$,其中,实体位置标签集 $T_{\text{pos}} = \{\text{B, M, E, S, O}\}$,含义如表 1 所示;实体类别标签集 $T_{\text{type}} = \{\text{dis, sym, bod, pro, equ, dru, ite, dep, mic}\}$,含义如表 2 所示。将位置标签集 $T_{\text{pos}}$ 与类别标签集 $T_{\text{type}}$ 结合,可得预测标签集 $T = \{\text{O, B-dis, M-dis, E-dis, S-dis, B-sym}, \cdots, \text{S-mic}\}$。最后输出预测的标签序列 $Y = [y_1, \cdots, y_n]$,$y_i$ 表示句子中第 $i$ 个字符对应的预测标签,即句子实例 $s$ 到预测标签集 $T$ 的映射。

**表 1 实体位置标签**

| 标签类型 | 含义 | 说明 |
|---|---|---|
| B | Begin | 实体的起始位置 |
| M | Middle | 实体的内部位置 |
| E | End | 实体的结束位置 |
| S | Single | 实体仅为当前字 |
| O | Other | 不属于任何实体 |

**表 2 实体类别标签**

| 标签类型 | 含义 | 说明 |
|---|---|---|
| dis | disease | 疾病 |
| sym | symptom | 临床表现 |
| bod | body | 身体 |
| pro | procedure | 医疗程序 |
| equ | equipment | 医疗设备 |
| dru | drug | 药物 |
| ite | item | 医学检验项目 |
| dep | department | 科室 |
| mic | microbes | 微生物类 |

本文提出的 WI-NER 模型由嵌入层、编码层和解码层组成,总体框架如图 1 所示。WI-NER 模型的处理过程如下:

(1) 在嵌入层通过 BERT 和动态构建的医学实体特征向量,将序列中每个元素从字符表示转换成向量表示;

(2) 将序列的向量表示输入到编码器中,利用引入词汇信息的双向门控循环单元,自动提取特征并输出隐藏向量;

(3) 使用 CRF 解码器预测序列中每个字符对应的实体位置和实体类型标签。

## 2　WI-NER 模型的工作流程

### 2.1　嵌入模块

嵌入模块将句子实例 $s$ 中的每个字符 $c_i$ 表示为字向量形式 $\boldsymbol{x}_i$, 对当前字符依次获取各个位置标签对应所有词汇集合,构建动态词典特征并将其添加到每个字符的表示中,然后再进行编码嵌入表示,具体过程如图 2 所示。

本文引入了先验知识与词汇信息相结合的方式进行嵌入模块的实现,是个体在进行学习、推理或理解新信息之前已经具备的知识或经验。先验知识可以来源于个人之前的学习、实践或观察,是机器在学习新知之前已经存储在大脑中的语言知识和科学知识等,有助于理解和解释新的信息、知识。

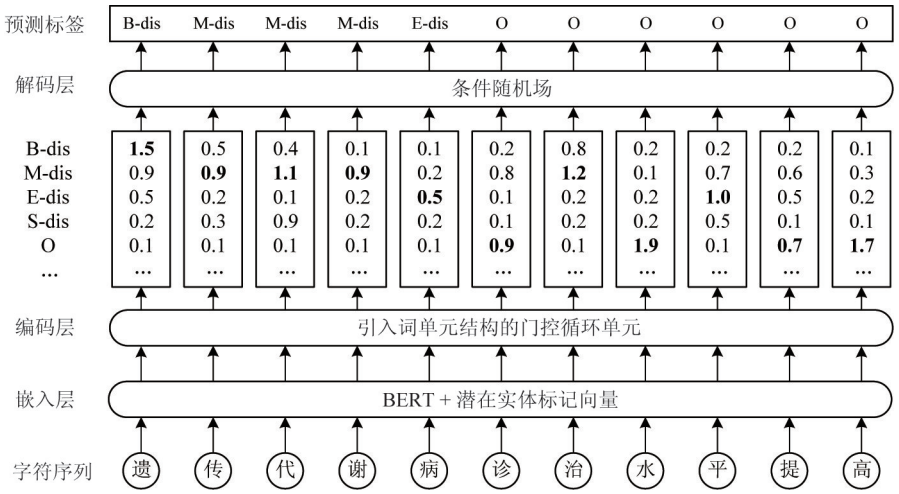


**图 1　WI-NER 模型的总体框架**

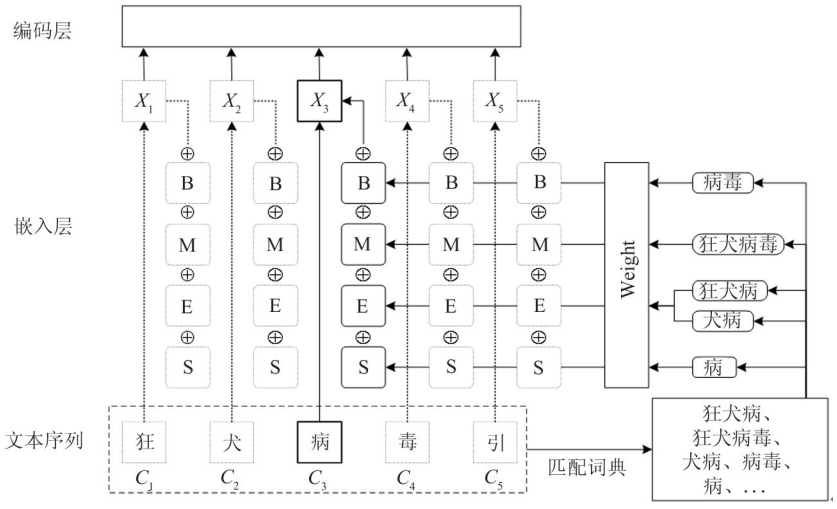


**图 2　嵌入模块**

不同领域的语料库会有明显的不同特点,例如在医学领域的语料库中“病”“药”等字作为实体尾部的概率远远高于其他领域的语料库,也就是语料库的先验知识。为了引入语料库的先验知识,提前遍历与输入语句相关的所有语料 $S$,统计各个字符在所有实体中被各标签标记的次数矩阵 $\boldsymbol{z}_i(t_{\text{pos}},$

$t_{type}$)，获得医学实体字频矩阵 $\boldsymbol{p}(t_{pos}, t_{type})$，进一步构建语句的先验知识嵌入矩阵 $\boldsymbol{k}(t_{pos}, t_{type})$，并对其进行降维处理，获得输入语句的潜在实体标记向量 $\boldsymbol{k}_i$。该向量体现了字符在潜在实体中不同位置与类别的分布概率，增强了字符的实体类别特征。

将输入序列记为 $s = [c_1, c_2, \cdots, c_n]$，首先将序列 $s$ 的每个字符 $c_i$ 映射到一个密集向量 $\boldsymbol{x}_i$，如式(1)所示。

$$\boldsymbol{x}_i = e^c(c_i) \tag{1}$$

其中 $e^c()$ 是字向量查找表中从字符到向量表示的映射操作。

然后开始引入词信息，对于字符序列 $s$，其中匹配到的单词记为 $w_{j,k} = \{c_j, \cdots, c_k\}$，将匹配到的词根据该字符的位置分类到集合{B, M, E, S}中，如式(2)所示。

$$\begin{cases} B(c_i) = \{w_{j,k} \mid j = i < k\} \\ M(c_i) = \{w_{j,k} \mid j < i < k\} \\ E(c_i) = \{w_{j,k} \mid j < i = k\} \\ S(c_i) = \{w_{j,k} \mid j = i = k\} \end{cases} \tag{2}$$

字符 $c_i$ 在位置 $t_{pos}$ 的实体类别特征的信效度 $v_i(t_{pos})$ 由该字符的样本数量和离散程度决定，样本数量与离散程度越高，其实体类别特征越可信并有效，即字频 $z_i$ 和潜在实体标记向量 $\boldsymbol{p}_i$ 中元素的方差 $\sigma_i(t_{pos})$，如式(3)所示。其中 sigmoid 函数可使字频实数映射到(0,1)区间内，以便神经网络收敛和反向传播时求导。

$$v_i(t_{pos}) = \text{sigmoid}(z_i) \times \sigma_i(t_{pos}) \tag{3}$$

依据信效度 $v_i(t_{pos})$ 将潜在实体标记向量 $\boldsymbol{k}_i$ 嵌入到原字符向量 $\boldsymbol{x}_i$ 中，如式(4)和式(5)所示。

$$\boldsymbol{k}_i(t_{pos}) = \tanh(\boldsymbol{p}_i(t_{pos}) \times t_{pos}(c_i) \times v_i(t_{pos})) \tag{4}$$

$$\boldsymbol{x}_i = [\boldsymbol{x}_i; \boldsymbol{k}_i(B); \boldsymbol{k}_i(M); \boldsymbol{k}_i(E); \boldsymbol{k}_i(S)] \tag{5}$$

最后，将嵌入了先验词汇信息的字向量输出到编码层。

### 2.2 编码模块

本文在门控循环单元的基础上引入词汇信息的门控循环单元(word information GRU, WI-GRU)，使得预训练模型中的词汇信息可以融入到基于字符编码的训练网络中。编码层 WI-GRU 的整体结构如图 3所示。其中，矩形代表字向量和词向量，圆角矩形中 $C$ 和 $W$ 分别代表 WI-GRU 中的字单元和词单元，圆形 $h$ 代表单元输出向量和传递隐状态。

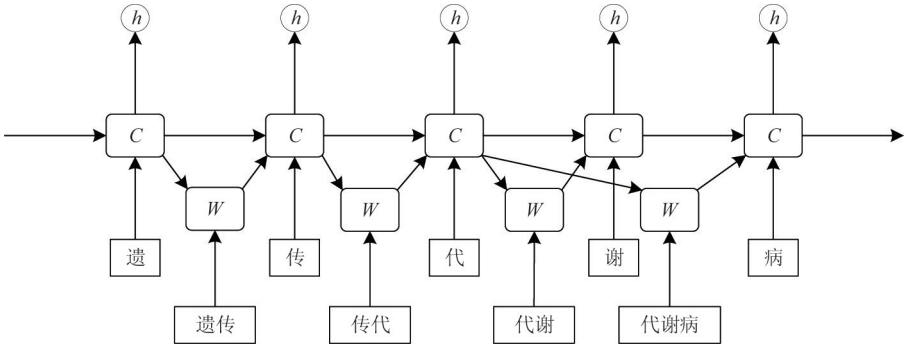


**图 3 编码层 WI-GRU 整体结构**

WI-GRU 结构中的普通字单元使用字向量和上一个单元传递的隐状态，训练重置门和更新门。其中，重置门用来控制单元生成候选隐状态时上一个隐状态传递的程度；更新门用来控制单元生成真正隐状态时上一个隐状态和当前候选隐状态的权重，并将隐状态输出到下一个单元。

图 4 是 WI-GRU 中普通字单元的内部结构和运算流程，其中 $\boldsymbol{x}_i$ 是作为输入的字向量，$h_{i-1}$ 和 $h_i$ 分别是上一个字单元传递来的隐状态和这一个字单元的隐状态，$R_i$ 和 $Z_i$ 分别是被训练的重置门和更新门，$\tilde{h}_i$ 为单元的候选隐状态，重置门和更新门的计算方式如式(6)和(7)所示，隐状态的计算方式如式(8)和(9)所示。

$$R_i = \text{sigmoid}\left(W_r \begin{bmatrix} \boldsymbol{x}_i \\ h_{i-1} \end{bmatrix} + b_r\right) \tag{6}$$

$$Z_i = \text{sigmoid}\left(W_z\begin{bmatrix}\boldsymbol{x}_i\\ h_{i-1}\end{bmatrix}+b_z\right) \tag{7}$$

$$\tilde{h}_i = \tanh\left(W_h\begin{bmatrix}\boldsymbol{x}_i\\ R_i\odot h_{i-1}\end{bmatrix}+b_h\right) \tag{8}$$

$$h_i = Z_i\odot h_{i-1} + (1-Z_i)\odot \tilde{h}_i \tag{9}$$

式中的“⊙”表示哈达玛积运算,即同阶矩阵之间的“按元素乘”,设矩阵 $\boldsymbol{A},\boldsymbol{B}\in C^{m\times n}$ 且 $\boldsymbol{A}=\{a_{ij}\}$,$\boldsymbol{B}=\{b_{ij}\}$,则矩阵 $\boldsymbol{A}$ 与矩阵 $\boldsymbol{B}$ 的哈达玛积如式(10)所示。

$$\boldsymbol{A}\odot\boldsymbol{B}=\begin{bmatrix}a_{11}b_{11} & \cdots & a_{1n}b_{1n}\\ \vdots & \ddots & \vdots\\ a_{m1}b_{m1} & \cdots & a_{mn}b_{mn}\end{bmatrix} \tag{10}$$

式中的 sigmoid( ) 和 tanh( ) 都是连续可导的激活函数,可以将神经网络非线性化,定义域都为 $R$,值域分别为(0,1)与(-1,1),计算公式如式(11)和式(12)所示。

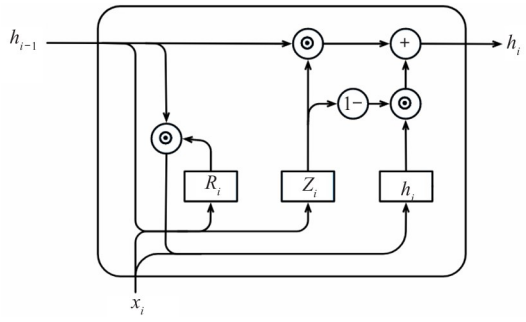


**图 4　WI-GRU 中的普通字单元**

$$\text{sigmoid}(x) = \frac{1}{1+e^{-x}} \tag{11}$$

$$\tanh(x) = \frac{\sinh x}{\cosh x} = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{12}$$

对于字符序列 $s=[c_1,\cdots,c_n]$,其中匹配到的单词记为 $w_{j,k}=[c_j,\cdots,c_k]$,先将单词与预训练的单词词典进行匹配,得到单词的嵌入表示 $\boldsymbol{x}^w_{j,k}$,如式(13)所示。

$$\boldsymbol{x}^w_{j,k} = e^w(w_{j,k}) \tag{13}$$

其中,$e^w(\,)$ 表示与单词向量词典匹配的操作。以 GRU 为基础,WI-GRU 添加了以词向量为输入的词汇门控单元。以该词的首字单元输出的隐状态作为上一个隐状态,训练重置门和更新门。其中,重置门用来控制单元生成候选隐状态时首字单元的隐状态传递的程度;更新门用来控制单元生成真正隐状态时首字隐状态和该词候选隐状态的权重,然后将生成的隐状态输出到该词的尾字单元。

本文中词单元与普通字单元不同之处在于:词单元只在匹配到词时被构建,以词向量和该词的首字单元的隐状态作为输入,最终生成的隐状态只输出到该词的尾字单元;而普通字单元在字向量从嵌入层传到编码层时即被构建,以字向量和前一个字的隐状态作为输入,最终生成的隐状态会被输出到下一个字单元和解码层。

图 5 是 WI-GRU 中普通字单元的内部结构和运算流程,其中 $\boldsymbol{x}^w_{j,k}$ 是作为输入的词向量,$h_j$ 是该词的首字单元输出的隐状态,$R^{cw}_{j,k}$ 和 $Z^{cw}_{j,k}$ 分别是该词单元的重置门和更新门,$\tilde{h}^w_{j,k}$ 是词单元的候选隐状态,重置门与更新门的计算方式如式(14)和(15)所示,候选隐状态和隐状态的计算方式如式(16)和(17)所示。

$$R^{cw}_{j,k} = \text{sigmoid}\left(W^{cw}_r\begin{bmatrix}\boldsymbol{x}^w_{j,k}\\ h_j\end{bmatrix}+b^{cw}_r\right) \tag{14}$$

$$Z^{cw}_{j,k} = \text{sigmoid}\left(W^{cw}_z\begin{bmatrix}\boldsymbol{x}^w_{j,k}\\ h_j\end{bmatrix}+b^{cw}_z\right) \tag{15}$$

$$\tilde{h}^w_{j,k} = \tanh\left(W^{cw}_h\begin{bmatrix}\boldsymbol{x}^w_{j,k}\\ R^{cw}_{j,k}\odot h_j\end{bmatrix}+b^{cw}_h\right) \tag{16}$$

$$h^w_{j,k} = Z^{cw}_{j,k}\odot h_j + (1-Z^{cw}_{j,k})\odot \tilde{h}^w_{j,k} \tag{17}$$

生成所有以字符 $c_k$ 为分词结尾的词单元 $w_{j,k}$,分配权重,进而更新该字符单元的输出向量。一个字单元可能同时是多个词单元的末字单元,因此,除

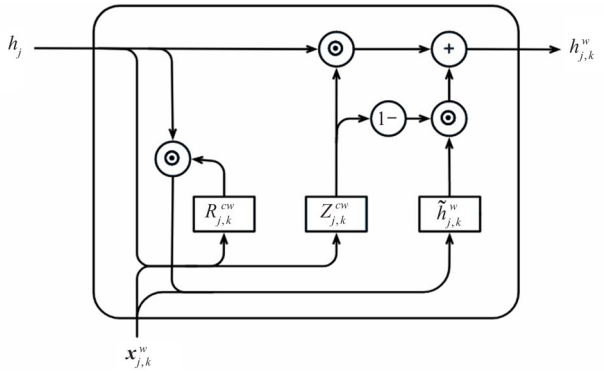


**图 5　WI-GRU 中的词单元**

了前字单元的隐状态外，还会接收多个词单元的隐状态，所以要为每个词单元设置传入更新门 $Z_{j,k}^{wc}$。通过词单元的传入更新门和该末字单元的更新门，来控制该末字单元最终生成隐状态时前一个字单元隐状态的权重 $\alpha_{k-1}^{c}$、该字单元的候选隐状态权重 $\alpha_k^c$ 和若干词单元隐状态的权重 $\alpha_{j,k}^{w}$。

图 6 为 WI-GRU 中普通字单元的内部结构和运算流程，其中 $\boldsymbol{x}_k$ 是作为输入的字向量，$h_{k-1}$ 是上一个字单元传递来的隐状态，$h_{j,k}^{w}$ 代表若干个以该字为末尾的词单元传递的隐状态，$h_k$ 是这个字单元输出的隐状态，$R_k$ 是被训练的重置门，$Z_k$ 是被训练的字更新门，$Z_{j,k}^{wc}$ 是被训练的控制各个词单元隐状态传输的词更新门，$\tilde{h}_k$ 为单元的候选隐状态，$\alpha$ 是各个隐状态的权重。

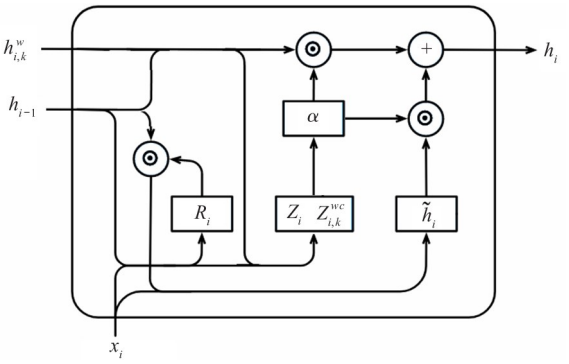


**图 6　WI-GRU 中匹配词的末字单元**

末字单元的重置门和字更新门与普通字单元一致，词更新门的计算方式如式(18)所示，各个隐状态权重的计算方式如式(19)、(20)、(21)所示，该字单元隐状态的计算方式如式(22)所示。

$$Z_{j,k}^{wc} = \text{sigmoid}\left(W_z^{wc}\begin{bmatrix} x_k \\ h_{j,k}^{w} \end{bmatrix} + b_z^{wc}\right) \tag{18}$$

$$\alpha_{k-1}^{c} = \frac{\exp(1 - Z_k^c)}{\exp(Z_k^c) + \exp(1 - Z_k^c) + \sum \exp(Z_{j,k}^c)} \tag{19}$$

$$\alpha_k^c = \frac{\exp(Z_k^c)}{\exp(Z_k^c) + \exp(1 - Z_k^c) + \sum \exp(Z_{j,k}^c)} \tag{20}$$

$$\alpha_{j,k}^{w} = \frac{\exp(Z_{j,k}^c)}{\exp(Z_k^c) + \exp(1 - Z_k^c) + \sum \exp(Z_{j,k}^c)} \tag{21}$$

$$h_k = \alpha_{k-1}^{c} \odot h_{k-1} + \alpha_k^c \odot \tilde{h}_k^{c} + \sum (\alpha_{j,k}^{w} \odot h_{j,k}^{w}) \tag{22}$$

最后，将所有字符单元的隐向量序列 $H = [h_1, \cdots h_n]$ 输出到解码模块。

### 2.3　解码模块

条件随机场 CRF 是一个序列标注算法，在序列标注任务中一个词的标签常与其上下文存在关系，因此，基于条件随机场考虑输入观测序列，计算整体的联合概率分布，对比概率分布数值并选择最大者，利用标签之间的约束条件选出最优输出。

模型的解码模块即为 CRF 结构，将编码层输出到隐向量序列 $H = [h_1, \cdots h_n]$ 送入解码层，通过 Softmax 计算第 $i$ 个字符 $c_i$ 与第 $k$ 个预测标签 $t_k \in T$ 一致的概率，如式(23)所示。

$$p(k \mid c_i^s) = \frac{\exp(W_k h_i + b_k)}{\sum \exp(W_j h_i + b_j)_{j=1}^{|T|}} \tag{23}$$

上式 $W_k$ 与 $b_k$ 是预测标签 $t_k$ 的可训练参数，字符 $c_i$ 与各个标签一致的概率可设为 $p(c_i^s) = [p_1, \cdots, p_{|T|}]$，对应的预测标签 $\text{argmax}[p_1, \cdots, p_{|T|}]$，最后解码层输出标签序列 $Y = [y_1, \cdots, y_n]$。

为了得到最优模型，需要计算预测标签序列 $Y$ 与样本真实标签序列 $Q = [q_1, \cdots, q_n]$ 的损失值 Loss，在模型中反向传播将其最小化，以实现模型所有参数 $W$ 的优化，损失值 Loss 的计算如式(24)所示。

$$Loss = -\frac{1}{n}\sum_{i=1}^{n} \log p(y_i \mid q_i) \tag{24}$$

## 3　WI-NER 模型算法描述

依据所提出的 WI-NER 识别模型框架，该模型的计算流程描述如算法 1 所示。

**算法 1**　WI-NER 模型算法

**输入**：文本序列 sentence，实体位置标签集 $T_{\text{pos}}$，实体类别标签集 $T_{\text{type}}$

**输出**：实体标签序列

BEGIN：

For $c_i$ in sentence do：#字粒度的嵌入层

　for ($t_{\text{pos}}$, $t_{\text{type}}$) in ($T_{\text{pos}}$, $T_{\text{type}}$) do：

　　$k_i(t_{\text{pos}}, t_{\text{type}})$　#根据式(4)，生成先验知识嵌入矩阵

$k_i = k_i(t_{pos}, t_{type})$.flatten( )　#将嵌入矩阵降维得到嵌入向量

$x_i = [e^c(c_i); k_i]$ #将字向量查找表中的字向量与嵌入向量拼接

end for

end for

for $c_k$ in sentence do: #字-词混合粒度的编码层

计算字单元的候选隐向量

for $w_{j,k}$ in words do: #遍历所有以 $c_k$ 为末字的词单元

根据式(13)查找表获得词的嵌入表示

根据式(15)计算词单元的传入更新门

根据式(17)计算词单元的隐向量

根据式(18)计算词单元的传出更新门

end for

根据式(22)计算各个隐向量的权重并计算最终隐向量

end for

根据式(23)解码层中计算实体标签概率

根据式(24)计算损失值

loss.backward( ) #计算梯度下降

optimizer.step( ) #迭代更新所有参数,将损失值最小化

END

## 4　实验结果与分析

### 4.1　实验环境和评价指标

使用中文医学实体抽取数据集 CMeEE、智能对话诊疗数据集 IMCS 和中药说明书实体识别数据集 TCM 作为本研究的测试语料库。

(1) CMeEE 语料来自中文医学领域的自然语言文献,例如医学教材、医学百科、入院记录和检验报告等,这些文本中蕴含大量医学专业知识和医学术语。

(2) IMCS 语料来自医患对话文本,其标注了对话中医疗相关的 5 类命名实体(症状、药品名、药物类别、检查、操作)。

(3) TCM 的语料来自中药药品说明书,有 1 000 份数据用于训练,句子最大长度超过 3 000,按照 13 类定义的实体进行标注。

本文采用严格 *Micro-F*1 值作为模型的评价指标,先通过总体样本统计 TP(预测正确的正例)、TN(预测正确的负例)、FP(预测错误的正例)、FN(预测错误的负例),计算查准率(precision)和查全率(recall),并最终计算出 *F*1 值。查准率、查全率和 *F*1 值的计算描述如式(25)、(26)和(27)所示。

$$Precision = \frac{\sum TP_i}{\sum (TP_i + FP_i)} \tag{25}$$

$$Recall = \frac{\sum TP_i}{\sum (TP_i + FN_i)} \tag{26}$$

$$Micro\text{-}F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{27}$$

与模型相关的超参数如表 3 所示,在嵌入模块使用预训练模型 BERT-base 对输入实例进行编码嵌入,将训练次数设为 100,使用随机梯度下降方法进行优化,在训练时使用网格搜索法结合交叉验证来确定最优参数,指定参数空间子集为:句子最大长度 $max_length \in \{200, 2\,000, 5\,000\}$,学习率 $lr \in \{0.001, 0.01, 0.1\}$。

**表 3　超参数**

| 参数名称 | 参数表示 | 值 |
|---|---|---|
| 句子长度 | *max_length* | 5 000 |
| 词向量维度 | *d_c* | 768 |
| 隐藏层维度 | *d_h* | 200 |
| 正则项系数 | $\lambda$ | 1e-8 |
| 学习率 | *lr* | 0.01 |
| 学习衰减率 | *lr_decay* | 0.05 |

为了验证 WI-NER 模型的有效性,本文在 CMeEE、IMCS 和 TCM 数据集上与 BERT + MLP 模型[11]、BiLSTM 模型[8]、Lattice 模型[10]和 FLAT 模型[15]进行对比分析,实验结果如表 4 所示。

**表 4　实验结果(准确率与召回率/%)**

| | CMeEE | | IMCS | | TCM | |
|---|---|---|---|---|---|---|
| | *Precision* | *Recall* | *Precision* | *Recall* | *Precision* | *Recall* |
| BERT + MLP | 62.51 | 61.54 | 84.13 | 89.68 | 65.66 | 77.94 |
| BiLSTM | 61.17 | 54.19 | 88.31 | 86.36 | 76.54 | 72.36 |
| Lattice | 60.86 | 56.74 | 89.74 | 85.62 | 66.67 | 85.76 |
| FLAT | 59.78 | 59.91 | 92.52 | 88.76 | 67.71 | 78.95 |
| WI-NER | 65.38 | 62.33 | 92.18 | 90.27 | 74.18 | 84.24 |

基于实验结果中的查准率与查全率，计算 *F*1 值，并计算每个模型在 3 个数据集的平均 *F*1 值，结果如表 5 所示。

**表 5　*F*1 值实验结果/%**

| | CMeEE | IMCS | TCM | 平均 |
|---|---|---|---|---|
| BERT + MLP | 62.02 | 86.82 | 71.28 | 73.37 |
| BiLSTM | 57.47 | 87.33 | 74.39 | 73.06 |
| Lattice | 58.73 | 87.63 | 75.02 | 73.79 |
| FLAT | 59.85 | 90.60 | 72.90 | 74.45 |
| WI-NER | 63.82 | 91.22 | 78.89 | 77.97 |

对比各模型在各数据集的 *F*1 值，如图 7 所示。

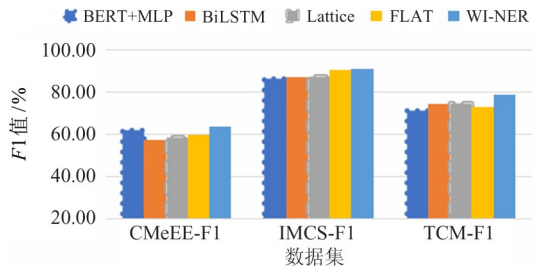


**图 7　实验对比与分析**

## 4.2　实验结果分析

对表 4 和表 5 的实验结果进行对比与分析可知以下几点。

（1）WI-NER 在 3 个数据集的 *F*1 值均高于基线模型，比基线模型中性能最好的 FLAT 模型平均高出 3.5%，验证了 WI-NER 模型能够更好地实现中文医学命名实体识别任务。

（2）从 3 个数据集下的实验结果可知，模型在 IMCS 数据集中准确率最高，在 CMeEE 数据集中准确率最低。说明在自然语言环境下，随着数据集中实体类别数量的增加，实现命名实体识别的难度也在增加。

（3）WI-NER 模型与基线模型中效果最好的 FLAT 模型相比，准确率更高。其原因是：WI-NER 模型不仅在基于字符编码的模型中融入了词汇信息，而且在嵌入模块中引入了医学领域的先验知识；由于 WI-NER 属于循环神经网络，在机器性能有限的情况下，WI-NER 模型可以处理更大长度的句子，与 FLAT 模型将长句子分割或截断、破坏句子语义不同。

为了验证在嵌入层引入先验知识和在编码层引入新的词粒度单元均有助于模型性能的提升，对模型进行了消融实验，消融实验结果如表 6 所示。其中，-WI 表示不在编码层引入词汇信息，取消词单元的计算，将普通的双向门控循环单元作为编码层，只计算字符单元之间的传递与输出；-Pattern 表示不在嵌入层引入先验知识，取消对数据中实体字符规律的统计，只将字符 $c_i$ 通过 BERT-base 映射到向量 $\boldsymbol{x}_i$。

**表 6　模型消融实验结果对比（准确率与召回率/%）**

| | CMeEE | | IMCS | | TCM | |
|---|---|---|---|---|---|---|
| | *Precision* | *Recall* | *Precision* | *Recall* | *Precision* | *Recall* |
| -WI | 60.13 | 59.54 | 83.35 | 89.62 | 68.62 | 83.19 |
| -Pattern | 63.58 | 61.86 | 89.24 | 86.29 | 71.57 | 82.72 |
| WI-NER | 65.38 | 62.33 | 92.18 | 90.27 | 74.18 | 84.24 |

通过实验结果中的准确率与召回率信息计算 *F*1 值，并计算每个模型在 3 个数据集的平均 *F*1 值，结果如表 7 所示。

**表 7　模型消融实验结果（*F*1 值/%）**

| | CMeEE | IMCS | TCM | 平均 |
|---|---|---|---|---|
| -WI | 59.83 | 86.37 | 75.21 | 73.80 |
| -Pattern | 62.71 | 87.74 | 76.74 | 75.73 |
| WI-NER | 63.82 | 91.22 | 78.89 | 77.97 |

根据表 6 和表 7 的消融实验结果可知：WI-NER 模型在各个数据集下的性能均优于-WI 模型，验证了基于字粒度的门控循环单元中加入的词粒度单元可以为编码层引入词汇信息，从而提升模型性能；WI-NER 模型在各个数据集下的性能均优于-Pattern 模型，验证了在嵌入层引入专业术语的实体字符分布规律作为先验知识，可以为模型性能带来

一定提升。

综合可知,WI-NER 模型在中文医学命名实体识别任务方面具有良好的优势,有效提高了识别效果的准确率。

## 5　结　论

本文提出了引入词汇信息的基于门控循环单元的中文医学命名实体识别模型 WI-NER,该模型充分利用了中文医学领域数据的特征和词汇信息。在嵌入层将字符转化为向量表示的过程中,对匹配专业词的字符进行特征嵌入与向量融合。在上下文编码层添加了词汇门控单元,使得模型可以同时对字符单元和词汇单元进行编码与信息传递。并分别在中文医学文献、医患对话文本和中药药品说明书的数据集上与多个模型进行了对比实验,分析了在命名实体识别过程中各功能模块的重要程度。实验表明,本文提出的 WI-NER 模型能利用中文医学领域数据的特征和词汇信息,更有效地识别文本中医学实体的位置与类别,提高中文医学命名实体识别的准确率。

本文提出的中文医学命名实体识别模型 WI-NER 在性能上均优于基线模型,但是这种序列标注方式无法解决实体嵌套问题,而且只引入了词汇信息和实体分布的先验知识,没有更加充分地利用中文医学实体的特点,比如字形、部首特点。因此,在接下来的工作中可以提出其他的标注方式来表达嵌套实体,并在编码层中考虑字形对实体类别的影响。

**参考文献**


[1] 李舒,朱雪娇. CCDSS 在慢性病管理中的应用进展[J]. 护理研究, 2018,32(21):3351-3353.

[2] CHEN P H, LIN C J, SCHOLKOPF B. A tutorial on v-support vector machines[J]. Applied Stochastic Models in Business and Industry, 2005,21:111-136.

[3] HU W, TIAN G, KANG Y, et al. Dual sticky hierarchical Dirichlet process hidden Markov model and its application to natural language description of motions[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018,40(10):2355-2373.

[4] LEE C, HWANG Y, OH H, et al. Fine-grained named entity recognition using conditional random fields for question answering[C]//Proceedings of Asia Information Retrieval Symposium. Berlin, Germany: Springer-Verlag, 2006:581-587.

[5] 李航. 统计学习方法[M]. 北京:清华大学出版社, 2012:80-87.

[6] KAZAMA J, MAKINO T, OHTA Y, et al. Tuning support vector machines for biomedical named entity recognition[C]//Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain. New York, USA:ACM, 2001:1-8

[7] LIN Y F, TSAI T H, CHOU W C,et al. A maximum entropy approach to biomedical named entity recognition[C]//Proceedings of International Conference on Data Mining in Bioinformatics. Berlin, Germany: Springer, 2004:56-61

[8] HUANG Z, XU W, YU K, et al. Bidirectional LSTM-CRF models for sequence tagging[EB/OL]. (2015-08-09)[2023-10-08]. http://arxiv.org/pdf/1508.01991.

[9] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition[EB/OL]. (2016-03-04)[2023-10-08]. http://arxiv.org/pdf/1603.01360.

[10] ZHANG Y, YANG J. Chinese NER using Lattice LSTM[EB/OL]. (2018-05-05)[2023-10-08]. http://arxiv.org/pdf/1805.02023.

[11] DEVLIN J, CHANG M W , LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2018-10-11)[2023-10-08]. http://arxiv.org/pdf/1810.04805.

[12] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. (2017-06-12)[2023-10-08]. http://arxiv.org/pdf/1706.03762.

[13] SUTSKEVER I, VINYALS O, LE Q V, et al. Sequence to sequence learning with neural networks[EB/OL]. (2014-09-10)[2023-10-08]. http://arxiv.org/pdf/1409.3215.

[14] XUE M, YU B, LIU T, et al. Porous lattice transformer encoder for Chinese NER[C]//Proceedings of the 28th

International Conference on Computational Linguistics. Barcelona, Spain: ACL, 2020:3831-3841.

[15] LI X N, YAN H, QIU X P, et al. FLAT: Chinese NER using flat-lattice transformer[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Seattle, USA: ACL, 2020:6836-6842

[16] GUI T, MA R, ZHANG Q, et al. CNN-based Chinese NER with lexicon rethinking[C]//The 28th International Joint Conference on Artificial Intelligence. Macau, China: IJCAI, 2019:4982-4988.

[17] SUI D B, CHEN Y B, LIU K, et al. Leverage lexical knowledge for Chinese named entity recognition via collaborative graph network[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China: EMNLP-IJCNLP, 2019:3821-3830.

[18] GUI T, ZOU Y C, ZHANG Q, et al. A lexicon-based graph neural network for Chinese NER[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Hong Kong, China: EMNLP-IJCNLP, 2019:3831-3841.

[19] LIU W, XU T, XU Q, et al. An encoding strategy based word-character LSTM for Chinese NER[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, USA: ACL, 2019: 2379-2389.

[20] DING R, XIE P, ZHANG X, et al. A neural multi-digraph model for Chinese NER with gazetteers[C]//Proceedings of the 57th Conference of the Association for Computational Linguistics. Florence, Italy: ACL, 2019: 1462-1467.

[21] MA R, PENG M, ZHANG Q, et al. Simplify the usage of lexicon in Chinese NER[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: ACL, 2020:5951-5960.

[22] 孙瑜,颜航,邱锡鹏,等. InstructGPT 在命名实体识别任务中的表现和挑战[J]. 中文信息学报, 2024,38(1): 74-85.

[23] 张天明,张杉,刘曦,等. 融合多模态数据的小样本命名实体识别方法[J]. 软件学报, 2024,35(3):1107-1124.

[24] 陈可,王俊. 基于自然语言处理的非小细胞肺癌知识图谱构建[J]. 中国卫生信息管理杂志, 2023,20(6): 1019-1025,1032.

[25] 曾江峰,庞雨静,高鹏钰,等. 基于 Lattice LSTM 的中医药古文献命名实体识别与应用研究[J]. 情报工程, 2023,9(5):112-122.

[26] LIU J X, SUN M Z, ZHANG W H, et al. DAE-NER: dual-channel attention enhancement for Chinese named entity recognition[J]. Computer Speech and Language, 2024,85:101581

[27] WU S, SONG X, FENG Z. MECT: multi-metadata embedding based cross-transformer for Chinese named entity recognition[EB/OL]. (2021-07-12)[2023-10-08]. http://arxiv.org/pdf/2107.05418.

[28] YIN X L, YUE Z, TAO Q, et al. A transition-based joint model for disease named entity recognition and normalization[J]. Bioinformatics, 2017, 33(15): 2363-2371.

[29] WANG Y R, SHINDO H, MATSUMOTO Y J, et al. Nested named entity recognition via explicitly excluding the influence of the best path[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Virtual: ACL, 2021:3547-3557.

[30] LI X Y, ZHANG H, ZHOU X H. Chinese clinical named entity recognition with variant neural structures based on BERT methods[J]. Journal of Biomedical Informatics, 2020,107(5):103422.

[31] 武帅,杨秀璋,何琳,等. 基于句法特征和 Bert-BiLSTM-MHA-CRF 的细粒度古籍实体识别研究[EB/J]. (2024-03-13)[2024-05-06]. 数据分析与知识发现, 2024:1-16.

[32] 张天宇,孙媛媛,杜文玉,等. 基于语义边界增强的司法命名实体识别[J]. 清华大学学报(自然科学版), 2024,64(5):749-759.

# Research on Chinese medical naming recognition model with vocabulary information

CHEN Jing[*], SUN Yaxuan[**,***], XING Kexuan[**,***]

([*]School of Electronics and Information Engineering, Guangdong Ocean University, Zhanjiang 524088)

([**]School of Information Science and Engineering, Yanshan University, Qinhuangdao 066004)

([***]Key Laboratory of Virtual Technology and System Integration, Yanshan University, Qinhuangdao 066004)

**Abstract**

There are a large number of specialized words in medical texts, which are more prone to word segmentation errors and unregistered words than in general fields, resulting in the loss of contextual semantics and affecting the accuracy of named entity recognition (NER). In order to solve the above problems, WI-NER, a Chinese medical named entity recognition model based on gated circulation unit with lexical information, is proposed in this paper. Firstly, on the basis of the characteristics of Chinese medical data set, the task definition, entity location and entity category label of named entity recognition in Chinese medical field are described. In addition, the model performs feature embedding and vector fusion on the characters matching professional words in the embedding layer. Secondly, a lexical gating unit is added to the context coding layer, and the features required for entity recognition are automatically extracted by using the memory and forgetting mechanism of recurrent neural networks. By introducing lexical information and prior knowledge, the recognition effect of Chinese medical named entities is improved. Finally, the model is verified by experiments on three datasets, and the results show that the accuracy of the Chinese medical named entity recognition model proposed in this paper is better than that of the baseline model, achieving the expected characteristics in the medical field.

**Key words**: Chinese medical naming recognition, prior knowledge, embedding layer, gated unit, vocabulary information