基于词典和多特征融合的中文医学命名实体识别

雷宇翔,廖 涛
安徽理工大学 计算机科学与工程学院,安徽 淮南 232001

摘 要:目的 针对现有方法中存在因分词导致级联错误从而影响实体识别效果,以及如何构建并融合高质量医学实体特征的问题,提出一个基于词典和多特征融合的中文医学命名实体识别模型。**方法** 该模型首先利用词典匹配机制和 Lattice 点阵结构来融合字符与医学词汇信息,利用字词的相对位置信息获取相对位置嵌入,并对汉字拼音进行编码得到拼音嵌入;然后提出一个融合 Transformer 模型来挖掘不同特征之间的互补性,以增强词汇信息并促进字词信息和拼音信息更好地融合;最后,将融合多特征的字符表示输入到条件随机场中来获得预测的标签。**结果** 在 CCKS-2019 和 Resume 数据集上的实验结果表明,该方法在多个指标上均得到了较好的提升。**结论** 避免了分词错误对命名实体识别效果造成的影响,通过融合 Transformer 模型更好地融合了多种医学实体特征,加强了模型识别词边界的能力,进而提高了模型识别医学实体的准确率,为后续构建医学知识图谱,实现智能化医学诊断提供了帮助。

关键词:中文命名实体识别;Lattice;多特征融合;Transformer
中图分类号:TP391.1;R-05 **文献标识码:**A doi:10.16055/j.issn.1672-058X.2026.0002.004

Chinese Medical Named Entity Recognition Based on Lexicon and Multi-feature Fusion

LEI Yuxiang, LIAO Tao
School of Computer Science and Engineering, Anhui University of Science and Technology, Huainan 232001, Anhui, China

Abstract: Objective A Chinese medical named entity recognition model based on lexicon and multi-feature fusion is proposed to address the problems in existing methods, including the cascading errors caused by word segmentation that affect the entity recognition effect and the issue of how to construct and fuse high-quality medical entity features. **Methods** Firstly, the model fused the information of characters and medical vocabulary with the lexicon matching mechanism and Lattice structure. It obtained relative position embeddings by using the relative position information of words and characters, and encoded Chinese character pinyin to get pinyin embeddings. Then, a fusion Transformer model was proposed to mine the complementarity between different features, so as to enhance the vocabulary information and promote better fusion of word and character information and pinyin information. Finally, the character representation fused with multiple features was input into a conditional random field to obtain predicted labels. **Results** Experimental results on the CCKS-2019 and Resume datasets demonstrated that the proposed method achieved notable improvements across multiple evaluation metrics. **Conclusion** The proposed method effectively avoids the negative impact of word segmentation errors on named entity recognition. It achieves efficient fusion of various medical entity features through the fusion Transformer model. As a result, it significantly enhanced the model's ability to recognize word boundaries, thereby improving the accuracy of medical entity recognition. This provides strong support for the subsequent construction of medical knowledge graphs and the realization of intelligent medical diagnosis.

Keywords: Chinese named entity recognition; Lattice; multi-feature fusion; Transformer

收稿日期:2024-07-01 **修回日期:**2024-10-09 **文章编号:**1672-058X(2026)02-0027-08
基金项目:国家自然科学基金面上项目(62076006)资助;安徽高校协同创新项目(GXXT-2021-008)资助.
作者简介:雷宇翔(1999—),男,江苏无锡人,硕士研究生,从事自然语言处理研究.
通信作者:廖涛(1977—),男,安徽淮南人,博士,副教授,从事信息抽取研究. Email:tliao@aust.edu.cn.
引用格式:雷宇翔,廖涛. 基于词典和多特征融合的中文医学命名实体识别[J]. 重庆工商大学学报(自然科学版),2026,43(2):27-34.
LEI Yuxiang, LIAO Tao. Chinese medical named entity recognition based on lexicon and multi-feature fusion[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2026, 43(2): 27-34.

命名实体识别(Named Entity Recognition, NER)是自然语言处理(Natural Language Processing, NLP)领域的一个关键任务,在信息抽取[1]、问答系统[2]、机器翻译[3]等多个NLP任务中都发挥着至关重要的作用。随着医疗信息化水平的不断进步,在临床报告、电子病历、医学文献和医学新闻中所包含的医学数据也随之增多,然而这些数据大多是非结构化的文本数据,如何通过NER技术对电子医疗文本进行规范化和标准化,对于医疗领域的专业研究有着重要的意义。

面向中文医学领域的命名实体识别[4](Chinese Medical Named Entity Recognition, CMNER)需要从医学领域数据文本中识别包括药品、身体器官、症状、疾病、治疗方法等医疗领域特有的实体,由于其复杂的语言结构,研究人员需要获取丰富的词边界信息才能有效地识别医学实体。然而,与英文NER任务获取词信息的方法不同,中文文本没有明显的分隔符来直接获取词级信息完成实体识别。通过引入中文分词(Chinese Word Segmentation, CWS)工具进行分词获取词级信息是一种可行的方法,但是分词结果的错误也会影响下游NER模型识别实体边界的性能[5]。因此,有不少研究者将研究重心转向基于字符级别的中文NER方法。如Meng等[6]将汉字视为图像,利用卷积神经网络提取汉字字形信息,并将字形信息引入到字符表示中;Yin等[7]使用卷积神经网络(Convolutional Neural Network, CNN)提取部首特征,然后将其与字符特征连接起来作为最终的字符表示;Zhang等[8]使用拼音映射表获取拼音特征,并将其融入字符向量表征中,拼音信息作为额外的特征输入到模型中有效地补充语法分析的不足。然而仅依靠字符信息的模型无法充分捕获相邻字符之间的语义信息和上下文关系,最终影响NER模型区分不同实体类型的能力。于是如何避免分词错误对命名实体识别任务产生影响的同时获取词信息,并融合多级别字信息,构建高质量医学实体特征,成为当前中文NER研究者研究的重点。

为了解决上述问题,本文提出基于词典和多特征融合的中文医学命名实体识别模型。该模型结合词典匹配机制和Lattice点阵结构来融合字符与医学词汇信息,以此避免中文分词产生的级联错误。然后利用字词的相对位置编码获取相对位置嵌入,有效地捕获相邻字符之间的上下文关系,并对汉字拼音进行编码得到拼音嵌入,以此获得更加丰富的字符语义信息来帮助模型识别医学实体边界。最后提出一个融合Transformer模型,利用其内部注意力机制来挖掘不同特征之间的互补性,以促进字词特征和拼音特征更好地融合。

1 相关工作

命名实体识别[9]的发展从早期基于词典与规则的方法,到基于机器学习的方法,再发展为基于深度学习的方法。早期学者为了让命名实体识别模型具有更强的解释性和灵活性,他们预先构建词典和规则,以适应不同的自然语言文本。如陈曙东等[10]在序列建模前通过提取动态词典匹配的语义来增强命名实体的识别效果,但该方法需要该领域专家耗费大量人工成本定义规则,且其可移植性也并不可观。后来部分学者将机器学习方法引入命名实体识别研究当中,他们先进行数据收集和数据标注,然后从标注好的文本数据中提取特征,该方法具有较强的可移植性。如Liu等[11]首先建立一个医学领域的数据集,然后利用条件随机场[12](Conditional Random Field, CRF)研究了不同类型特征在中文医疗命名实体识别任务中的作用,该方法无须依赖特征模板,但需花费较大人工成本标注数据集。

目前,大部分学者采用基于深度学习的方法来提取文本语义信息,该方法相较于前两种方法能利用不同的神经网络代替词典和规则来获取更广泛的上下文信息。其中,基于长短期记忆网络(Long Short-Term Memory, LSTM)的模型[13-15]是目前应用最广泛的方法。该方法使用LSTM作为编码器捕获文本特征并使用CRF作为解码器预测标签,不仅能学习序列关系还能避免长距离依赖问题,但LSTM的循环结构会增加内存访问模式的复杂性,使GPU在处理复杂的内存访问模式时受到限制。为了提高整体效率,Yan等[16]提出TENER模型将基于注意力机制的Transformer模型引入命名实体识别任务中,用于对字符级特征和单词级特征进行建模。Transformer模型不采用循环和卷积,而是使用注意力机制计算输入文本序列中每个元素的相关性,并通过赋予它们权重来实现对序列的上下文理解。

然而中文医学文本通常具有复杂的语言结构,其中包含大量的术语和医学专业名词。直接对中文医学文本进行中文分词,产生的错误可能会造成误差传播。为此,不少研究者尝试通过词典来引入外部医学知识,并结合深度学习中的神经网络,来增强对非结构化医学文本中医学实体的识别。Zhang等[17]设计了一种基于词典的中文NER模型Lattice-LSTM,其采用Lattice点阵表示外部词典中的字和词,通过连接线表示字词间联系以捕捉字词之间的语义关系,避免了分词错误造成的影响并有效提升了命名实体识别的性能。在Lattice-LSTM模型的基础上,Zhao等[18]引入MKGCN(Multi-Modal Knowledge Graph Convolutional Network)[19]来提取词典和知识图谱中的词汇信息,并采用对抗训练来增强模型的鲁棒性。考虑到部首信息对中文医学实体识别产生的影响,Gui等[20]在嵌入层中将词典信息以及通过CNN获取的部首信息进行整合,并设计了一个级联的Transformer网络更好地整合这两种类型的语义信息。然而,大部分改进方法中构建的额外特征比较简单,都是将LSTM或CNN与自注意力机制相结合来融合不同的特征。如何获取并且高效融合医学实

体特征成为本文的研究目标。

综上所述,为了避免分词产生的级联错误带来的影响,并获取高质量的医学实体特征,本文设计了一个基于词典和多特征融合的中文医学命名实体识别模型。该模型结合词典匹配机制和 Lattice 点阵结构来融合字符与医学词汇信息,利用字词的相对位置编码获取相对位置嵌入,有效地捕获相邻字符之间的上下文关系,并对汉字拼音进行编码得到拼音嵌入以获得更加丰富的字符语义信息。然后提出一个融合 Transformer 模型,利用其内部的注意力机制来挖掘不同特征之间的互补性,以增强词汇信息并促进字词特征和拼音特征的更好融合。实验结果表明,本文模型避免了分词错误对命名实体识别效果造成的影响,减少了计算时间和成本,通过融合 Transformer 模型更好地融合了多种不同特征,提高了实体识别的准确性。

2　模型框架

本文提出的基于词典和多特征融合的中文医学命名实体识别模型如图 1 所示,共分为 3 层:多特征嵌入层、多特征融合编码层以及标签解码层。

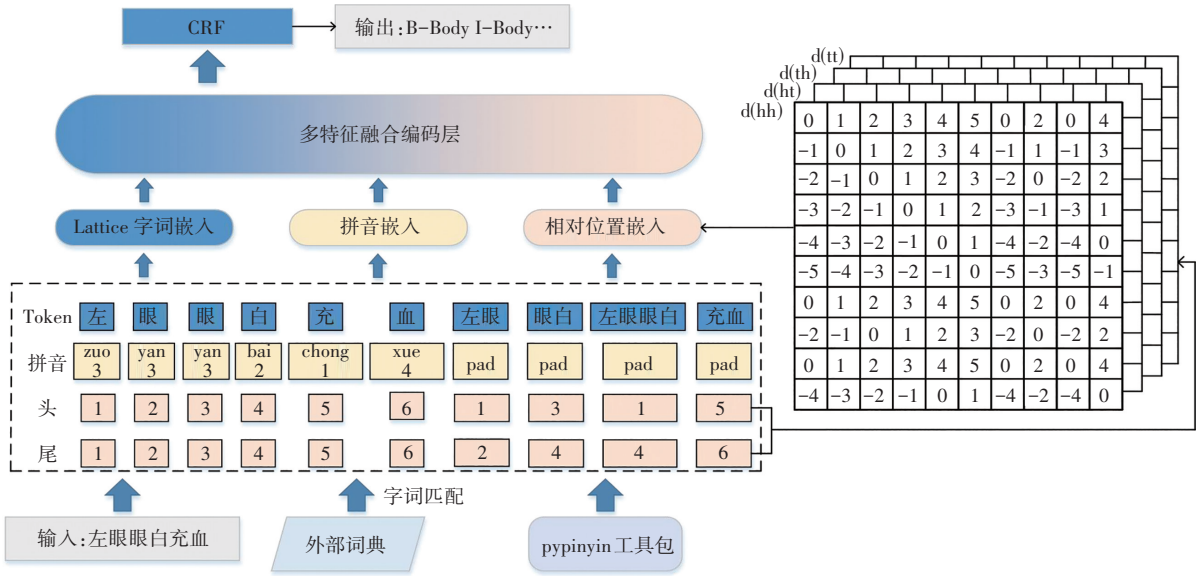

图 1　基于词典和多特征融合的中文医学命名实体识别模型

Fig. 1　Chinese medical named entity recognition model based on lexicon and multi-feature fusion

2.1　多特征嵌入层

2.1.1　Lattice 字词嵌入

为了解决中文分词错误对下游序列标注等任务产生的影响,本文使用词典匹配方法来获取文本中包含的词汇信息,为模型提供更多中文词语的语义和边界信息,从而提高实体识别效果。具体来说,对于给定的一段输入文本序列和一个医学词典,通过将输入文本序列与词典进行匹配,得到一组匹配的单词,如图 2 所示,该图是由字符和所有匹配的词汇组成的图结构。

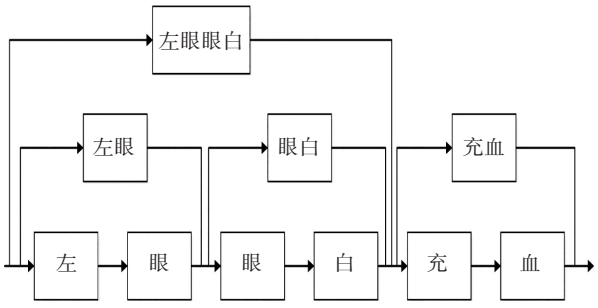

图 2　字词匹配图

Fig. 2　Word matching graph

然而由于 Transformer 编码器不能直接对图结构数据进行建模,因此受到扁平化 Lattice Transformer 模型 FLAT[21] 的启发,本节对 Lattice 点阵结构进行类似的平铺转换操作。如图 3 所示,首先通过将匹配后的词汇添加到字符序列的末尾来构造平面晶格,然后利用 token 来标记平面点阵中的字符和词汇,最后为每个 token 分配头和尾的位置,即第一个字符和最后一个字符的位置,如果一个 token 只包含一个字符,则头和尾的位置是相同的。

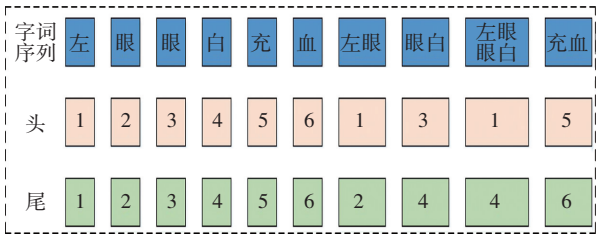

图 3　扁平化的 Lattice 字词结构图

Fig. 3　Flattened Lattice word structure diagram

由于 Lattice 由字符和匹配的词汇组成,因此为了得到 Lattice 字词嵌入,需要初始化字符嵌入和词嵌入。首先使用预训练模型 Bert[22],通过预训练好的字符向量将每个字符映射到相应的字符嵌入。然后,随机初始化匹配词的嵌入,经过将词嵌入和字符嵌入拼接组成后,得到 Lattice 字词嵌入 $\boldsymbol{E}_{\mathrm{L}}$。通过该方法获取的字

词嵌入能有效避免分词错误带来的影响。

2.1.2 拼音嵌入

每个汉字都可以由汉语拼音进行表达,已有研究者对汉字的拼音特征进行提取来帮助命名实体识别。受此启发,本模型对中文医学文本进行分析,文本中许多医学实体都包含相同的汉字,如炎“yán”、肿“zhǒng”、血“xuè”等。通过对这些频繁出现在医学命名实体结尾的汉字进行拼音特征提取,来解决医学文本中实体边界的模糊性问题。

为了获取每个单词的拼音特征,首先利用 Pypinyin 工具包得到输入医学文本中每个汉字的拼音特征;然后为每个汉字构建一个由拼音字母和声调组成的 27 维向量,其中前 26 位对应字母表中 26 个字母,字母表中存在组成该汉字拼音字母的位置设为 1,其他位置设为 0,如“血”的拼音为“xuè”,则在该 27 维向量的第 24、21、5 这 3 个位置上分别设为 1,最后一位为汉字的声调,“阴平、阳平、上声、去声”4 个声调分别由数字 1 至 4 来表示,则“血”的第 27 位为 4;最后根据对应的拼音序列 $\boldsymbol{X}^{\mathrm{py}}=(\boldsymbol{x}_0,\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n)$ 得到输入文本的拼音特征 $\boldsymbol{E}_{\mathrm{py}}$,其计算如式(1)所示:

$$\boldsymbol{E}_{\mathrm{py}}=\boldsymbol{F}^{\mathrm{py}}(\boldsymbol{X}^{\mathrm{py}})=(\boldsymbol{e}_0,\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_n) \tag{1}$$

其中,$\boldsymbol{F}^{\mathrm{py}}$ 为矩阵所对应映射函数,$\boldsymbol{x}_n$ 表示输入医学文本中第 n 个汉字对应的拼音,$\boldsymbol{e}_n$ 表示输入医学文本中第 n 个汉字的拼音向量。

2.1.3 相对位置嵌入

为了更有效地提取实体边界信息,本模型首先通过 FLAT 模型在 Transformer 编码器基础上改进信息获取的方法,来获取字词在句子中的相对位置嵌入。FLAT 模型为文本中的每个字符和词汇都设置了头位置与尾位置,因此本文根据字词序列中每个 token 的头尾信息来计算 4 种字词相对位置信息,相对位置如式(2)—式(5)所示:

$$d_{ij}^{(hh)}=\mathrm{head}[i]-\mathrm{head}[j] \tag{2}$$

$$d_{ij}^{(ht)}=\mathrm{head}[i]-\mathrm{tail}[j] \tag{3}$$

$$d_{ij}^{(th)}=\mathrm{tail}[i]-\mathrm{head}[j] \tag{4}$$

$$d_{ij}^{(tt)}=\mathrm{tail}[i]-\mathrm{tail}[j] \tag{5}$$

其中,$\mathrm{head}[i]$ 和 $\mathrm{tail}[i]$ 表示 token 的起始位置和结尾位置,$d_{ij}^{(hh)}$ 表示字词序列中索引为 i 的 token 的起始位置到索引为 j 的 token 的起始位置之间的距离。同理,$d_{ij}^{(ht)}$、$d_{ij}^{(th)}$、$d_{ij}^{(tt)}$ 分别表示每对 token 起始位置和结尾之间的距离,以及两者结尾位置之间的距离。然后用一个非线性变换来计算每对 token 之间的 4 种相对位置嵌入 $\boldsymbol{E}_{\mathrm{P}}$,其计算如式(6)所示:

$$\boldsymbol{E}_{\mathrm{P}}=\mathrm{ReLU}(W_r(P_{d_{ij}^{(hh)}}\oplus P_{d_{ij}^{(ht)}}\oplus P_{d_{ij}^{(th)}}\oplus P_{d_{ij}^{(tt)}})) \tag{6}$$

其中,W_r 表示可学习参数,$\oplus$表示串联运算符,$P_{d_{ij}^{(ht)}}$、$P_{d_{ij}^{(th)}}$、$P_{d_{ij}^{(hh)}}$、$P_{d_{ij}^{(tt)}}$ 表示 4 种相对位置编码计算,如式(7)所示:

$$P_d^{(2k)}=\sin(d/10\ 000^{2k/d_{\mathrm{model}}})$$
$$P_d^{(2k+1)}=\cos(d/10\ 000^{2k/d_{\mathrm{model}}}) \tag{7}$$

其中,k 表示位置编码的维度索引,d_{model} 表示输入维度的大小。这种基于正弦的位置编码方法使 Transformer 能够对字符的位置和每两个字符之间的距离进行建模。利用字词的相对位置编码获取相对位置嵌入,有效识别了字符与匹配词之间的位置关系,提高了模型识别词边界的能力。

2.2 多特征融合编码层

如图 4 所示,在该层中设计了一个融合 Transformer 模型来实现多特征融合编码,该模型由一个 Lattice Transformer 和一个拼音 Transformer 组成。

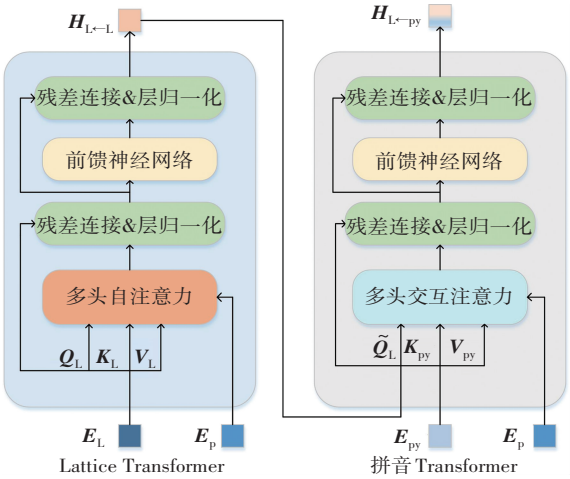

图 4 融合 Transformer 模型图

Fig. 4 Architecture of the fused Transformer model

2.2.1 Lattice Transformer

本文设计了一个 Lattice Transformer 来将字符-词点阵结构中的词汇信息集成到字符表示中,利用多头自注意力机制对字符与匹配词之间的密集交互进行建模,并引入字词的相对位置信息来提高模型识别词边界的能力。

为了计算 Lattice 字词嵌入的自注意力,本文对 Lattice 字词嵌入 $\boldsymbol{E}_{\mathrm{L}}$ 和相对位置嵌入 $\boldsymbol{E}_{\mathrm{P}}$ 执行线性映射来获得矩阵 $\boldsymbol{Q}_{\mathrm{L}}$、$\boldsymbol{K}_{\mathrm{L}}$ 和 $\boldsymbol{V}_{\mathrm{L}}$。其计算如式(8)—式(10)所示:

$$\mathrm{SA}_{\mathrm{L}\leftarrow\mathrm{L}}(\boldsymbol{E}_{\mathrm{L}},\boldsymbol{E}_{\mathrm{P}})=\mathrm{softmax}(A_{\mathrm{L}\leftarrow\mathrm{L}})\boldsymbol{V}_{\mathrm{L}} \tag{8}$$

$$A_{\mathrm{L}\leftarrow\mathrm{L},ij}=(Q_{\mathrm{L},i}+u_{\mathrm{L}})^{\mathrm{T}}K_{\mathrm{L},j}+(Q_{\mathrm{L},i}+v_{\mathrm{L}})^{\mathrm{T}}E_{\mathrm{P}_{ij}}W_{\mathrm{L}}^{\mathrm{P}} \tag{9}$$

$$\boldsymbol{Q}_{\mathrm{L}},\boldsymbol{K}_{\mathrm{L}},\boldsymbol{V}_{\mathrm{L}}=\boldsymbol{E}_{\mathrm{L}}[W_{\mathrm{L}}^{\mathrm{Q}},W_{\mathrm{L}}^{\mathrm{K}},W_{\mathrm{L}}^{\mathrm{V}}] \tag{10}$$

其中,$W_{\mathrm{L}}^{\mathrm{Q}}$、$W_{\mathrm{L}}^{\mathrm{K}}$、$W_{\mathrm{L}}^{\mathrm{V}}$、$W_{\mathrm{L}}^{\mathrm{P}}$、$u_{\mathrm{L}}$、$v_{\mathrm{L}}$ 均为可学习参数,$\mathrm{SA}_{\mathrm{L}\leftarrow\mathrm{L}}$ 用于计算 Lattice 字词嵌入 $\boldsymbol{E}_{\mathrm{L}}$ 和相对位置嵌入 $\boldsymbol{E}_{\mathrm{P}}$ 的相关性。然后将多个头部的输出与其线性映射进行连接来计算多头自注意力 MSA,其计算如公式(11)所示:

$$\mathrm{MSA}(\boldsymbol{E}_{\mathrm{L}},\boldsymbol{E}_{\mathrm{P}})=[\mathrm{SA}_1(\boldsymbol{E}_{\mathrm{L}},\boldsymbol{E}_{\mathrm{P}}),\cdots,\mathrm{SA}_h(\boldsymbol{E}_{\mathrm{L}},\boldsymbol{E}_{\mathrm{P}})]W_{\mathrm{L}}^{\mathrm{O}} \tag{11}$$

其中,$W_{\mathrm{L}}^{\mathrm{O}}$ 为可学习参数。接着将多头自注意力的输出经过残差连接以及层归一化操作后,得到中间层的输出 $\boldsymbol{Z}$,并将其输入到前馈神经网络中进行非线性变换。最后,

利用另一轮的残差连接和层归一化来计算该 Transformer 的输出 $\boldsymbol{H}_{\mathrm{L}\leftarrow\mathrm{L}}$,其计算如式(12)、式(13)所示:

$$\boldsymbol{H}_{\mathrm{L}\leftarrow\mathrm{L}}=\mathrm{LayerNorm}(\mathrm{FFN}(\boldsymbol{Z})+\boldsymbol{Z}) \tag{12}$$

$$\boldsymbol{Z}=\mathrm{LayerNorm}(\mathrm{MSA}(\boldsymbol{E}_{\mathrm{L}},\boldsymbol{E}_{\mathrm{P}})+\boldsymbol{E}_{\mathrm{L}}) \tag{13}$$

其中,LayerNorm 代表层归一化操作,FFN 表示前馈神经网络用于处理和转换自注意力机制中编码的信息。

2.2.2　拼音 Transformer

为了将拼音信息融入字符表示中,本模型采用类似 2.2.1 小节中将词汇信息融入字符表示的方法,在构造的拼音 Transformer 中,利用包含相对位置信息的交叉注意力机制对 Lattice 字词特征和拼音特征之间的密集交互进行建模,交叉注意力的计算如式(14)—式(17)所示:

$$\mathrm{CA}_{\mathrm{L}\leftarrow\mathrm{py}}(\boldsymbol{H}_{\mathrm{L}\leftarrow\mathrm{L}},\boldsymbol{E}_{\mathrm{py}},\boldsymbol{E}_{\mathrm{P}})=\mathrm{softmax}(A_{\mathrm{L}\leftarrow\mathrm{py}})\boldsymbol{V}_{\mathrm{py}} \tag{14}$$

$$A_{\mathrm{L}\leftarrow\mathrm{py},ij}=(\tilde{Q}_{\mathrm{L},i}+v_{\mathrm{py}})^{\mathrm{T}}E_{P_{ij}}W_{\mathrm{py}}^{\mathrm{P}}+(\tilde{Q}_{\mathrm{L},i}+u_{\mathrm{py}})^{\mathrm{T}}K_{\mathrm{py},j} \tag{15}$$

$$\tilde{\boldsymbol{Q}}_{\mathrm{L}}=\boldsymbol{H}_{\mathrm{L}\leftarrow\mathrm{L}}W_{\mathrm{L}\leftarrow\mathrm{L}}^{\mathrm{Q}} \tag{16}$$

$$\boldsymbol{K}_{\mathrm{py}},\boldsymbol{V}_{\mathrm{py}}=\boldsymbol{E}_{\mathrm{py}}[W_{\mathrm{py}}^{\mathrm{K}},W_{\mathrm{py}}^{\mathrm{V}}] \tag{17}$$

其中,$W_{\mathrm{L}\leftarrow\mathrm{L}}^{\mathrm{Q}}$、$W_{\mathrm{py}}^{\mathrm{K}}$、$W_{\mathrm{py}}^{\mathrm{V}}$、$W_{\mathrm{py}}^{\mathrm{P}}$、$u_{\mathrm{py}}$、$v_{\mathrm{py}}$ 均为可学习参数,$\widetilde{\boldsymbol{Q}}_{\mathrm{L}}$ 为将 Lattice 编码器的输出 $\boldsymbol{H}_{\mathrm{L}\leftarrow\mathrm{L}}$ 进行线性映射得到的矩阵,$\boldsymbol{K}_{\mathrm{py}}$、$\boldsymbol{V}_{\mathrm{py}}$ 为将拼音向量矩阵 $\boldsymbol{E}_{\mathrm{py}}$ 进行映射所得到的矩阵。随后进行与 Lattice 变压器中类似的残差连接、层归一化以及在前馈神经网络中进行非线性变换等操作,得到更新后的 Lattice-拼音感知字符表示,即为拼音编码器的输出 $\boldsymbol{H}_{\mathrm{L}\leftarrow\mathrm{py}}$。

2.3　标签解码层

在标签解码层中,使用 CRF 作为模型的解码器来预测输出标签,对于输入医学文本序列 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_n\}$,则输出预测标签序列为 $\boldsymbol{Y}=\{y_1,y_2,\cdots,y_n\}$,其计算如式(18)—式(19)所示:

$$P(\boldsymbol{Y}\mid\boldsymbol{X})=\frac{\exp\left(\sum_i s(y_i,y_{i-1})\right)}{\sum_{Y'}\exp\left(\sum_i s(y'_i,y'_{i-1})\right)} \tag{18}$$

$$s(y_i,y_{i-1})=W^{y_i}r_i+T_{y_{i-1}y_i} \tag{19}$$

其中,Y'为所有可能标签的集合,W^{y_i} 和 $T_{y_{i-1}y_i}$ 皆为可训练参数。最后采用 Viterbi 算法计算最优序列,训练损失计算如式(20)所示:

$$L=\sum_{i=1}^{n}\log(P(\boldsymbol{Y}\mid\boldsymbol{X}))+\lambda\|\theta\| \tag{20}$$

其中,λ 为正则化超参数,θ 为可训练参数集。在实验中,损失函数由 Adam 优化器进行优化。基于词典和多特征融合的中文医学命名实体识别模型的算法流程如图 5 所示。先引入外部词典,对输入的数据集文本进行字词匹配,利用 FLAT 对字词进行平铺转换为点阵结构,利用点阵结构获取字词相对位置信息,使用拼音工具包获取字符拼音;然后设计一个融合 Transformer 对上述多特征进行融合;最后使用 CRF 对预测标签进行解码操作。

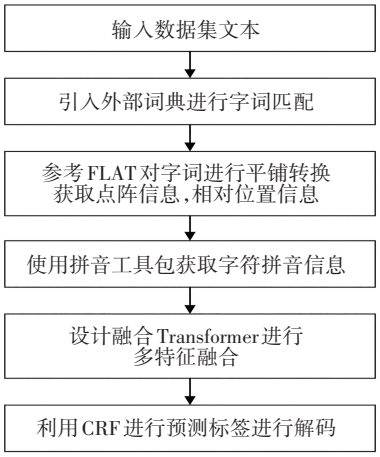

图 5　基于词典和多特征融合的中文医学命名实体识别模型的算法流程图

Fig. 5　Algorithm flow chart of Chinese medical named entity recognition model based on lexicon and multi-feature fusion

3　实验与结果分析

3.1　数据集

为了评估该模型的实验效果,本文在两个公共数据集上进行实验。为了证明该模型对医学文本的适用性,首先使用中国知识图谱与语义计算大会 CCKS-2019 发布的中文电子病历实体数据集进行实验。为了验证模型的泛化能力,在 Resume 简历数据集上进行重复实验来证明该模型在通用领域的适用性,数据集的具体情况如表 1 所示。

表 1　实验中使用的公开中文数据集

Table 1　The public Chinese dataset used in the experiment

数据集	实体类型	数据集划分	句子数
CCKS-2019	身体部位、药物、疾病、影像检查、手术、实验室检查	训练集	1 000
		验证集	79
		测试集	300
Resume	国籍、教育背景、人名、地址、组织名、专业、民族、职称	训练集	3 821
		验证集	463
		测试集	477

对于数据集的预处理操作,本文使用 BIO 规则对实体类型进行标注,其中 B 代表实体的开头(Begin)、I 代表内部(Inside),而 O 代表实体之外的位置(Outside)。

3.2　评价指标

为了评估模型的有效性,本次实验采用准确率 S_{P}、召回率 S_{R} 以及 F1 值 S_{F1} 这 3 个评价指标,这些指标的具体计算方式如式(21)—式(23)所示:

$$S_{\mathrm{P}}=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FP}}}\times100\% \tag{21}$$

$$S_{\mathrm{R}}=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FN}}}\times100\% \tag{22}$$

$$S_{F1}=\frac{2S_P S_R}{S_P+S_R}\times 100\% \tag{23}$$

其中,N_{TP} 表示将正样本预测为正类的个数,N_{FN} 表示将正样本预测为负类的个数,N_{FP} 表示将负样本预测为正类的个数。

3.3 实验参数设置

本次实验使用 Pytorch 库构建模型,使用 SGD 作为优化器,字符嵌入通过预训练的嵌入进行初始化,字符、拼音和相对位置的嵌入维数设置为 50 并随机初始化。实验中所使用的其他超参数可见表 2。

表 2 超参数设置

Table 2 Hyperparameter settings

超参数	数 值
初始学习率	0.001
Batch_size	10
epoch	50
Transformer	6
多头注意力机制	8
Dropout	0.2

3.4 对比实验

为了验证所提出的融合方法优于当前大多采用基本串联操作融合多种特征的方法,本文通过设计两种主流的融合方法与本模型方法进行比较。

如图 6 所示,模型 A 通过门控机制整合多粒度特征,并输入到双向长短期记忆网络(Bi-directional Long Short-Term Memory, BiLSTM)中进行上下文编码;模型 B 先将字词嵌入和拼音嵌入进行连接,再输入到 Transformer 中进行编码。将这两个模型在 CCKS-2019 和 Resume 数据集中与本文模型进行比较,比较结果如表 3 所示。在医学领域数据集中,本文模型在 S_P、S_R、S_{F1} 上分别达到了 84.83%、86.17%、85.49%;在通用 Resume 据集上,3 个评价指标分别达到了 96.06%、95.88%、95.97%。与仅使用交互门控+BiLSTM 进行特征融合的方法对比,各项指标均得到明显提升,其中 S_{F1} 在两个数据集上分别获得了 0.71% 和 0.35% 的增益。与仅使用连接操作+自注意力机制进行特征融合的方法对比,S_{F1} 在两个数据集上分别获得了 0.45% 和 0.21% 的增益。实验结果表明,与主流的基本连接操作相比,通过融合 Transformer 对多粒度特征之间的密集交互进行建模,可以获得更好的融合结果。

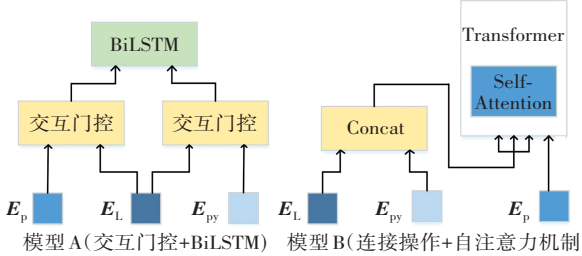

图 6 主流多特征融合方法图

Fig. 6 Diagram of mainstream multi-feature fusion methods

表 3 多特征融合方法结果对比

Table 3 Comparison of results using different multi-feature fusion methods /%

模 型	CCKS-2019			Resume		
	S_P	S_R	S_{F1}	S_P	S_R	S_{F1}
模型 A(交互门控+BiLSTM)	84.29	85.28	84.78	95.83	95.42	95.62
模型 B(连接操作+自注意力机制)	84.45	85.63	85.04	95.92	95.61	95.76
本文模型	**84.83**	**86.17**	**85.49**	**96.06**	**95.88**	**95.97**

通过比较,发现引入字词信息和拼音信息可以为模型带来不同程度的性能提升,使用本文的融合 Transformer 进行多特征融合编码要优于目前基于主流多特征融合方法的模型,使模型能提取更全面的语义特征,有助于提高模型识别词边界的能力。另外将本文模型与近几年中文命名实体识别的主流模型进行对比以验证所提出模型的有效性。对比模型如下:Zhang 等[17]提出的 Lattice LSTM 模型对句子中的字符和词典识别的潜在词进行编码,将融合信息馈送到基于字符的 LSTM-CRF 中;Li 等[21]将 Lattice 结构转换为多跨度表示并引入位置编码融合字词信息;Ma 等[23]将词典信息融入字符表示并且能和不同的序列标注模型框架进行使用;Wu 等[24]提出的 BMECT 模型基于 FLAT 架构,并融合了 Lattice 字词信息以及偏旁部首信息的 MECT 模型;Kong 等[25]提出的 ACNN 模型构建了结合注意力机制的多层 CNN,充分利用 GPU 并行性来提高效率,为 CMNER 提供了一种新的方法。

如表 4 所示,列举了本文模型在 CCKS-2019 和 Resume 两个数据集上与主流模型进行性能比较的结果。发现本文提出的基于词典和多特征融合的中文医学命名实体识别模型相较于所列出的最优主流模型(ACNN),在 CCKS-2019 和 Resume 两个数据集上的 S_{F1} 值分别提高了 0.36% 和 0.3%,S_P 和 S_R 均有不同程度提升。这说明该实验通过融合 Transformer 融合字词信息、拼音信息和字词相对位置信息可以提取句子中的深层语义信息,更好地识别词边界进而提高命名实体识别的效果。

表 4 主流模型对比结果

Table 4 Comparison results of mainstream models /%

模 型	CCKS-2019			Resume		
	S_P	S_R	S_{F1}	S_P	S_R	S_{F1}
Lattice LSTM[17]	83.20	84.86	84.02	94.81	94.11	94.46
FLAT[21]	—	—	84.29	—	—	95.40
SoftLexicon[23]	84.21	85.10	84.65	95.30	95.77	95.53
BMECT[24]	84.96	85.29	85.12	95.71	96.23	95.97
ACNN[25]	83.07	**87.29**	85.13	94.91	**96.43**	95.67
本文模型	**84.83**	86.17	**85.49**	**96.06**	95.88	**95.97**

3.5　消融实验分析

在中文医学命名实体识别模型的训练过程中,为了验证引入词嵌入、字符拼音嵌入以及字词相对位置嵌入会对整体模型产生正面效果,本文通过去除不同成分来进行多特征融合编码层的消融实验。消融实验采用的数据集仍为 CCKS-2019 和 Resume,参数设置均相同。多特征融合编码层的消融实验如表 5 所示。其中模型 1 表示保留 Lattice Transformer 的同时移除拼音 Transformer;模型 2 表示保留拼音 Transformer 的同时移除 Lattice Transformer;模型 3 表示同时移除 Lattice Transformer 和拼音 Transformer,模型的实验效果等同于 BERT+CRF 的效果。

表 5　多特征融合编码层的消融实验结果

Table 5　Ablation results of multi-feature fusion coding layers　/%

模　型	CCKS-2019			Resume		
	S_P	S_R	S_{F1}	S_P	S_R	S_{F1}
模型 1(本文模型-拼音 Transformer)	84.50	84.11	84.30	95.56	95.31	95.43
模型 2(本文模型-Lattice Transformer)	84.39	84.04	84.21	95.33	95.26	95.29
模型 3(本文模型-Lattice& 拼音 Transformer)	83.87	84.12	83.99	94.28	94.47	94.37
本文模型	**84.83**	**86.17**	**85.49**	**96.06**	**95.88**	**95.97**

通过比较,当移除 Lattice Transformer 或拼音 Transformer,以及两者同时被移除时,S_P、S_R、S_{F1} 值均产生明显下降。综上所述,本文所提出的每个组件都可以很好地独立工作,并且当它们组合在一起时可以有效地提高模型的性能。

3.6　参数分析

为了研究融合 Transformer 对整体模型收敛速度的影响,在 CCKS-2019 数据集和 Resume 数据集上与模型 A(交互门控+BiLSTM)和模型 B(连接操作+自注意力机制)两种特征融合方法进行了比较,如图 7 和图 8 所示。

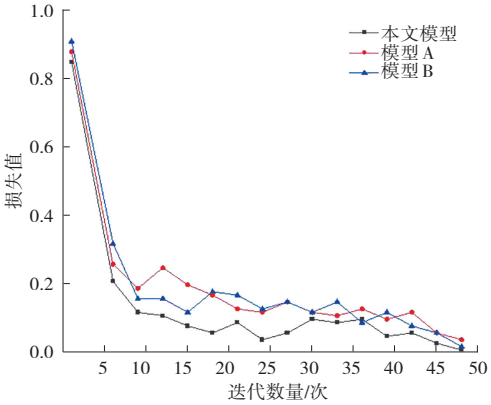

图 7　CCKS-2019 数据集的收敛速度对比

Fig. 7　Comparison of convergence speed of CCKS-2019 dataset

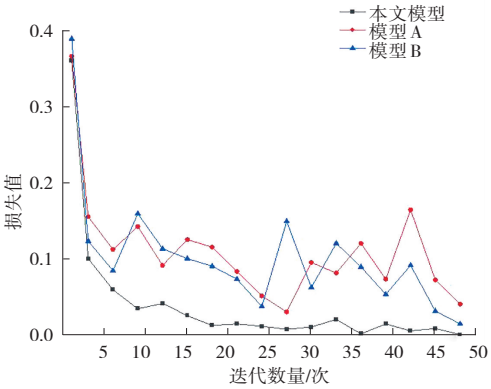

图 8　Resume 数据集的收敛速度对比

Fig. 8　Comparison of the convergence speed of the Resume dataset

通过以上参数分析可以看出:本文模型的损失值下降幅度最大且收敛速度最快,说明融合 Transformer 能高效地融合多种特征,更好地进行命名实体识别。

4　结束语

针对现有方法中存在的因分词导致级联错误从而影响实体识别效果,以及如何构建并融合高质量医学实体特征的问题,提出了基于词典和多特征融合的中文医学命名实体识别方法。该模型有效利用词汇和拼音信息,并通过设计的融合 Transformer 对多种特征进行融合来实现对中文医学文本的深入理解。将本文模型医学领域数据集和通用领域数据集与采用基本串联操作将多种特征结合起来的模型,以及最近比较主流的模型进行了比较。实验结果表明:文本模型在避免分词错误带来影响的同时还提高了模型识别的准确率。本次实验根据不同汉字有不同拼音信息这一性质帮助模型识别实体边界,但实际情况中存在同一汉字有不同读音或是不同汉字拼音却相同的情况。在未来工作中,将尝试引入其他与字符有关的特征表示来帮助拼音特征进行命名实体识别,例如引入计算机视觉方法帮助识别医学文本中的汉字字形特征。同时,尝试丰富中文医学语料库,丰富医学知识词典,在数据集中添加更多的训练数据以提高模型的泛化能力和对中文医学领域实体识别的数据支持。

参考文献(References):

[1]　崔博文, 金涛, 王建民. 自由文本电子病历信息抽取综述[J]. 计算机应用, 2021, 41(4): 1055-1063.
CUI Bo-wen, JIN Tao, WANG Jian-min. Overview of information extraction of free-text electronic medical records[J]. Journal of Computer Applications, 2021, 41(4): 1055-1063.

[2]　毛先领, 李晓明. 问答系统研究综述[J]. 计算机科学与探索, 2012, 6(3): 193-207.
MAO Xian-ling, LI Xiao-ming. A survey on question and answering systems[J]. Journal of Frontiers of Computer

Science & Technology, 2012, 6(3): 193-207.

[3] UGAWA A, TAMURA A, NINOMIYA T, et al. Neural machine translation incorporating named entity[C]// Proceedings of the 27th International Conference on Computational Linguistics. Kerrville: Association for Computational Linguistics, 2018: 3240-3250.

[4] NÉVÉOL A, DALIANIS H, VELUPILLAI S, et al. Clinical natural language processing in languages other than English: Opportunities and challenges[J]. Journal of Biomedical Semantics, 2018, 9: 12.

[5] LIU P, GUO Y, WANG F, et al. Chinese named entity recognition: The state of the art[J]. Neurocomputing, 2022, 473: 37-53.

[6] MENG Y, WU W, WANG F, et al. Glyce: Glyph-vectors for Chinese character representations[J]. Advances in Neural Information Processing Systems, 2019, 32: 2742-2753.

[7] 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, 2019, 98: 103289.

[8] ZHANG Y, LIU Y, ZHU J, et al. Learning Chinese word embeddings from stroke, structure and pinyin of characters[C]// Proceedings of the 28th ACM International Conference on Information and Knowledge Management. New York: ACM, 2019: 1011-1020.

[9] 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.
DU Jin-hua, YIN Hao, FENG Song. Research and development of named entity recognition in Chinese electronic medical record[J]. Acta Electronica Sinica, 2022, 50(12): 3030-3053.

[10] 陈曙东, 罗超, 欧阳小叶, 等. 基于动态词典匹配的语义增强中文命名实体识别算法[J]. 无线电工程, 2021, 51(7): 519-525.
CHEN Shu-dong, LUO Chao, OUYANG Xiao-ye, et al. A semantic-enhanced Chinese named entity recognition algorithm based on dynamic dictionary matching[J]. Radio Engineering, 2021, 51(7): 519-525.

[11] LIU K, HU Q, LIU J, et al. Named entity recognition in Chinese electronic medical records based on CRF[C]//Proceedings of the 14th Web Information Systems and Applications Conference. Piscataway: IEEE Press, 2017: 105-110.

[12] LAFFERTY J, MCCALLUM A, PEREIRA F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]//International Conference on Machine Learning. 2001: 282-289.

[13] XU K, ZHOU Z, HAO T, et al. A bidirectional LSTM and conditional random fields approach to medical named entity recognition[C]//Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2017. Cham Springer International Publishing, 2017: 355-365.

[14] 林昕, 朱小栋. 基于 Attention 机制的 LSTM 股价预测模型[J]. 重庆工商大学学报(自然科学版), 2022, 39(2): 75-82.
LIN Xin, ZHU Xiao-dong. Attention-mechanism-based LSTM model for stock price predicting[J]. Journal of Chongqing Technology and Business University (Natural Science Edition), 2022, 39(2): 75-82.

[15] SHI J, SUN M, SUN Z, et al. Multi-level semantic fusion network for Chinese medical named entity recognition[J]. Journal of Biomedical Informatics, 2022, 133: 104-144.

[16] YAN H, DENG B, LI X, et al. TENER: Adapting transformer encoder for named entity recognition[J]. Computation and Language, 2019, 1: 04474.

[17] ZHANG Y, YANG J. Chinese NER using lattice LSTM[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2018: 1554-1564.

[18] ZHAO S, CAI Z, CHEN H, et al. Adversarial training based lattice LSTM for Chinese clinical named entity recognition[J]. Journal of Biomedical Informatics, 2019, 99: 103290.

[19] XIONG Y, PENG H, XIANG Y, et al. Leveraging multi-source knowledge for Chinese clinical named entity recognition via relational graph convolutional network[J]. Journal of Biomedical Informatics, 2022, 128: 104035.

[20] GUI T, MA R, ZHANG Q, et al. CNN-based Chinese NER with lexicon rethinking[C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. 2019: 4982-4988.

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

[22] MA Z, ZHAO L, LI J, et al. SiBERT: A Siamese-based BERT network for Chinese medical entities alignment[J]. Methods, 2022, 205: 133-139.

[23] 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, PA, USA: ACL, 2020: 5951-5960.

[24] WU S, SONG X, FENG Z. MECT: Multi-metadata embedding based cross-transformer for Chinese named entity recognition[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg, PA, USA: ACL, 2021: 1529-1539.

[25] KONG J, ZHANG L, JIANG M, et al. Incorporating multi-level CNN and attention mechanism for Chinese clinical named entity recognition[J]. Journal of Biomedical Informatics, 2021, 116: 103737.

责任编辑:李翠薇