基于局部增强的中文医疗命名实体识别模型

陈晶[1,2]，邢珂萱[2,3]，孟伟伦[3]，郭景峰[3]，冯建周[3]
（1.广东海洋大学数学与计算机学院，广东 湛江 524088；2.河北省计算机虚拟技术与系统集成重点实验室，河北 秦皇岛 066004；
3.燕山大学信息科学与工程学院，河北 秦皇岛 066004）

摘　要：医学实体的识别往往受到其相邻上下文的影响，目前的命名实体识别方法通常依赖于BiLSTM捕捉文本中的全局依赖关系，缺乏对字符之间局部依赖关系的建模。针对这一问题，提出了一种基于局部增强的中文医疗命名实体识别模型LENER。首先，LENER使用包括字音、字形和语义在内的多源信息来丰富底层字符表征。然后，结合相对位置编码对滑动窗口划分出的序列片段进行局部注意力计算，并通过非线性计算融合局部信息和BiLSTM得到的全局信息。最后，对识别出的实体头部和尾部进行组合，进而提取出实体。实验结果表明，LENER模型具有良好的实体识别能力，与其他模型相比，LENER模型的F_1值提升了0.5%~2.0%。

关键词：中文命名实体识别；上下文环境；注意力机制；多源信息；滑动窗口
中图分类号：TP391.1
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2024117

Chinese medical named entity recognition model based on local enhancement

CHEN Jing[1,2], XING Kexuan[2,3], MENG Weilun[3], GUO Jingfeng[3], FENG Jianzhou[3]
1. College of Mathematics and Computer Science, Guangdong Ocean University, Zhanjiang 524088, China
2. The Key Laboratory for Computer Virtual Technology and System Integration of Hebei Province, Qinhuangdao 066004, China
3. College of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China

Abstract: In the medical field, the recognition of medical entities is often influenced by their adjacent context, the current named entity recognition methods typically rely on BiLSTM to capture the global dependency relationships within text, lacking modeling of local dependencies between characters. To resolve this problem, a Chinese medical named entity recognition model LENER based on local enhancement was proposed. Firstly, the representation of characters was enriched by LENER utilizing multi-source information, including phonetic, graphic and semantic features. Secondly, relative position encoding was combined to perform local attention calculations on sequence segments divided by sliding windows, and local information was fused with global information obtained from BiLSTM through nonlinear computation. Finally, the recognized entity heads and tails were combined by LENER to extract the entities. The experimental results show that the LENER model has excellent entity recognition capabilities, and the F_1 value is improved by 0.5% to 2% compared with other models.

Keywords: Chinese named entity recognition, contextual environment, attention mechanism, multi-source information, sliding window

收稿日期：2024–03–11；修回日期：2024–06–06
通信作者：邢珂萱，xkx0513@163.com
基金项目：国家自然科学基金资助项目（No.62172352, No.42306218）；中央省部共建基金资助项目（No.226Z0102G, No.226Z0305G）；广东海洋大学科研启动基金资助项目（No.060302102304）
Foundation Items：The National Natural Science Foundation of China (No.62172352, No.42306218), Central Government Guides Local Science and Technology Development Fund Projects (No. 226Z0102G, No. 226Z0305G), Guangdong Ocean University Research Fund Project (No.060302102304)

0　引言

随着社会对医疗数据的收集、应用和管理能力的不断增强，产生了海量电子医疗数据。医疗命名实体识别（NER, name entity recognition）技术作为智能医疗系统和医疗知识图谱等任务的关键基础，为医疗领域的自动化和智能化提供了重要支持和保障。基于技术进行划分，现有的NER技术可以分为基于规则的方法、基于机器学习的方法以及基于深度学习的方法3类[1]。目前，深度学习已经成为NER的主流技术，模型通过自动学习高质量的特征来提高实体识别的准确率和效率。

在医疗领域，中文NER任务的目标是从非结构化的医疗文本中识别出疾病名称、症状、身体部位、药品等医疗实体[2]。尽管目前应用于垂直领域的NER方法通常结合多种信息源以提升识别效果，但是医疗行业的特殊性、专业性及隐私性等特点使得现有的基于深度学习的方法或ChatGPT模型难以获取规范、统一标注的外部知识进行训练。因此，本文选择从中文医疗文本内部的语言模式着手进行研究。

基于现有研究，受中文语法的限制，实体在句中的位置通常遵循特定规则，这些规则将实体与其上下文环境的关系分为以下3类。

1) 全局依赖关系：实体的标注受全局上下文环境的影响，这意味着序列中的每个字都可能对当前位置的标注产生影响。

2) 局部依赖关系：实体的标注显著受到相邻上下文环境的影响，这主要表现在实体的边界划分与分类通常受修饰词、数值指标或字符结构的影响。如图1所示，“[”和“]”代表实体的边界。在句子①中，修饰词“急性”与“肾衰竭”一同被标注为疾病类型的实体。在句子②中，由于存在提示性文字“检查”，因此后面的“血PH”“血钾”和“血渗透压”都被标注为医学检验项目类型的实体，而非身体部位类型的实体。

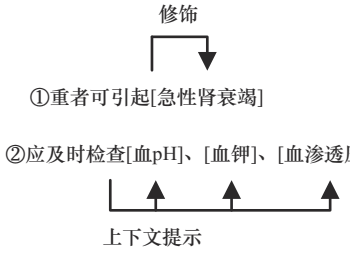

图1　受相邻上下文影响的实体边界划分及类型判定

3) 弱依赖关系：一些实体的识别不需要考虑其上下文关系，如人名、药品名称等类型的实体，它们本身具有的语义信息就满足识别需求。

目前，许多工作基于双向长短期记忆（BiLSTM, bi-directional long short-term memory）网络、Transformer等模型对文本的全局依赖关系进行编码。相比于Transformer，BiLSTM的网络结构具有时序性，因此更适用于NER任务。在BiLSTM模型中，当前时间步的输入与前一时刻的隐藏状态是相互独立的，这导致了上下文信息的缺失[3]，从而使模型难以直接建立字符与其相邻上下文之间的依赖关系。然而，在医疗文本中，这种局部依赖关系是不可忽略的。因此，仅依赖BiLSTM全局信息的编码模块容易忽视局部依赖对实体识别的影响，从而导致实体边界模糊和实体类别模糊的问题。

此外，在多个医疗数据集上进行统计后发现，不同类别实体所包含的字形结构在分布和组成上存在明显差异。如图2所示，微生物类（mic）的实体包含较多的“囝”与“艹”字形，这与微生物类主要描述真菌、细菌等原生生物相符；身体类（bod）的实体则包含较多的“月”字形，这与表征人体器官的字符通常包含“月”字形相符。因此，深入理解中文字形特点有助于更好地理解医疗实体的构成与关联关系。

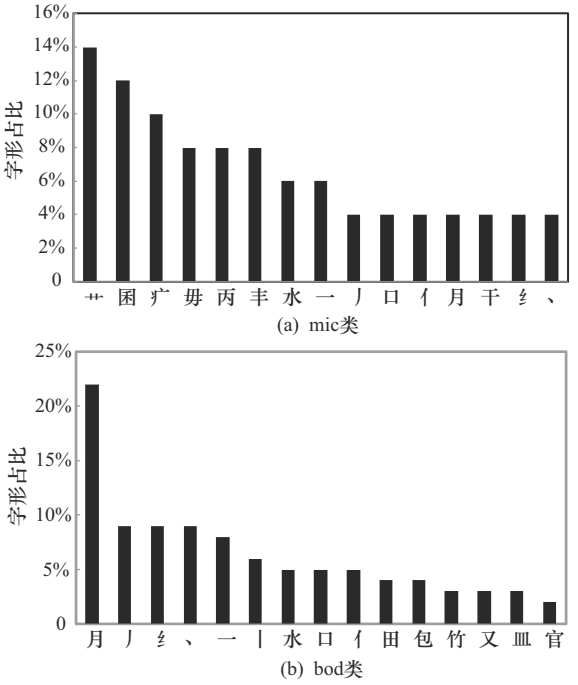

图2　CMeEE数据集中字形分布

为了解决上述问题，本文提出了基于局部增强的中文医疗命名实体识别模型LENER（local enhancing-based name entity recognition）。LENER 由基于 BiLSTM 的全局模块和基于滑动窗口注意力机制的局部注意力模块组成，在编码结束后通过非线性计算来平衡全局信息和局部信息对实体识别带来的影响。本文的主要贡献如下。

1) 使用包括字音、字形和语义在内的多源特征进行字符增强。

2) 针对医疗实体与上下文之间存在的局部依赖关系，本文将序列切分为多个片段，并通过注意力机制在片段上进行局部依赖的研究，从而弥补 BiLSTM 模型在局部建模方面的不足，解决实体边界划分模糊的问题。

3) 在多个公开中文数据集上对本文模型进行评估，证明了本文模型的优点和有效性。

1　相关工作

医疗领域的NER方法与通用领域的方法大体相似。传统的医疗NER通常采用基于字典和规则的方法，这种方法通过使用预定义的规则来识别医疗实体。虽然这种方法可以达到较高的精度，但是受限于手动定义的规则，容易引发OOV（out of vocabulary）问题。近年来，深度学习技术在NER领域的研究取得了显著进展，受到了研究人员的广泛关注。医疗NER在深度学习领域的研究可以分为基于单神经网络、基于多任务学习、基于迁移模型和基于混合模型的方法[4]。当前的NER模型主要通过整合多种神经网络并进行微调来实现，其中常用的神经网络包括卷积神经网络（CNN, convolutional neural network）[5]、长短期记忆（LSTM, long short-term memory）网络[6]以及Transformer[7]。曹依依等[8]提出了基于CNN的医疗NER模型，该模型采用迭代扩张卷积作为提取特征的编码器。Dang等[9]对BiLSTM-CRF模型进行微调，以识别各种类型的命名实体，如基因和蛋白质。Yu等[10]采用多层BiLSTM和双仿射结构计算每个跨度的分数，并选择排名top-k的实体进行提取。Zhang等[11]提出了WB（well-behaved）-Transformer模型，该模型分别对中文电子病历中的字符和对应的词进行编码，降低了错误词边界错误的影响。Yan等[12]提出了TENER模型，对Transformer的编码器结构进行优化，以便捕获字符之间的距离和方向信息。

提升NER性能的一个趋势是在模型中引入外部知识。Gong等[13]提出了一种基于领域词典与条件随机场（CRF, conditional random field）的预标注−精确标注双层标注模型，通过一次预标注和二次精确标注的形式，将人工构建的准确性和机器学习的自动性结合起来。Zhang等[14]提出了Lattice LSTM模型，针对当前字符，查询词典中以该字符结束的所有词汇信息，并通过构造有向无环图的方式进行信息融合。Li等[15]设计了一种平面的Flat-Lattice Transformer结构，基于Transformer融合词汇信息的动态结构，并重新设计了位置编码来融合Lattice结构。Ma等[16]提出了一种在嵌入层利用词汇信息的方法，保留识别过程中匹配到的所有词，并采用静态词频作为权重对词进行聚合。Gui等[17]引入了一个具有全局语义性的基于词典知识的图神经网络，以解决基于循环神经网络模型在NER方面的局限性。Yang等[18]对文本和标签分别进行编码，并设计了基于注意力机制的语义融合模块，通过标签知识显式增强文本表示。Li等[19]将NER视为一种机器阅读理解任务，并将标签形式化为提取问题。Liu等[20]充分利用BERT模型的序列建模能力，把词汇信息融入BERT底层的编码过程中。Geng等[21]提出了一种基于检索的方法来解决NER问题，同时利用了动态改变的词典信息和知识的语义信息。

另一个趋势是利用文本内部的结构信息。Wu等[22]提出了基于部首流和语义流的双流模型MECT，考虑汉字是象形文字，并且其结构通常携带了一定语义信息，因此该模型在Flat-Lattice Transformer的基础上结合部首信息以增强语义理解。Xuan等[23]提出了一种CNN结构，专门用于获取字形信息和相邻图之间的交互信息。Gu等[24]提出了RICON模型，该模型通过规律感知模块整合实体内部规律，通过规律诊断模块利用上下文信息来验证挖掘出的内部规律的可靠性。Li等[25]发现NER任务中的边界检测和类型预测2个子任务可以通过共享信息进行交互增强，因此设计了模块化交互网络（MIN），该网络基于双仿射交互机制增强边界信息和类型信息。Chen等[26]在实体识别过程中，除了使用图注意力网络层捕捉词之间的依赖关系外，还采用Star-Transformer来近似建模远程依赖关系，以应

对实体在文本中的稀疏性问题。Li 等[27]在编码器中引入了句法图结构，并借助图卷积网络（GCN, graph convolutional network）进行增强，从而提升了模型对嵌套实体和不连续实体的提取能力。

2　LENER 模型

LENER 模型结构如图 3 所示，主要由嵌入层、上下文编码层和解码层组成。嵌入层结合字音、字形和字义等多源信息对字符进行表征。上下文编码层包括全局模块和局部注意力模块，前者专注于提取字符在全局上下文中的依赖信息，而后者侧重于字符在相邻上下文片段中的局部依赖关系。解码层包括实体头索引分类层、实体尾索引分类层以及实体匹配算法层。

2.1　嵌入层

汉字的字音、字形和语义 3 种特征之间相互关联，为实体的识别提供了一定线索。对于实体而言，构成不同类别实体的字符在字形结构上往往存在差异。例如，身体类的实体常带有“月”字旁，而疾病类的实体常带有“疒”字旁，不同的字形结构隐含着不同的语义信息。同时，在一些实体中存在音译词，由于文本来源不同，同一个词可能有不同的字形表征方式，引入汉字的字音可以减少“同音不同字”对语义造成的干扰。为此，本文提出了融合字音、字形、语义的特征嵌入方案，进而增强汉字字符的表征能力。

对于输入序列 $S=\{c_1,c_2,c_3,\cdots,c_n\}$，其中，$c_i$ 表示 S 中的第 i 个字符，该序列中的 n 个字符通过预训练模型 RoBERTa 映射为密集向量形式，得到原始的字符语义嵌入 $\boldsymbol{x}_s^i$，如式(1)所示。

$$\boldsymbol{x}_s^i=\mathrm{RoBERTa}(c_i) \tag{1}$$

汉字的字音包括拼音和音调两部分，其中音调包括阴平、阳平、上声及去声，分别用数字 1、2、3、4 表示。通过字音查找表获取输入序列中字符的拼音和音调，从而得到字音序列 $\mathrm{PY}=\{w_p^1,w_p^2,\cdots,w_p^n\}$，其中，$w_p^i$ 代表单个字符的拼音与音调信息。

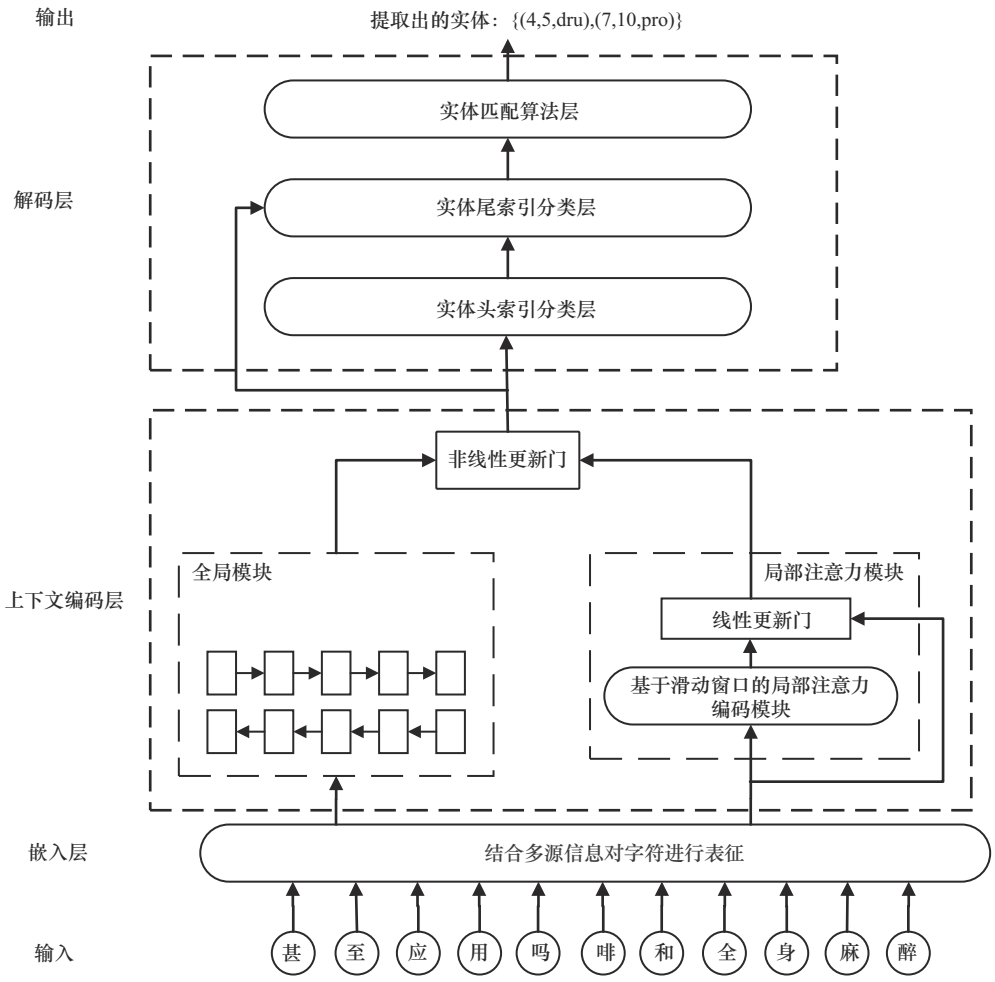

图 3　LENER 模型结构

例如，当c_i=“病”时，w_p^i = ['b','i','n','g','2']。

汉字的字形包含笔画、部件和整字三级结构单位。在NER任务中，汉字的字形特征通常基于一级部件进行拆解。然而，二级部件和三级部件也包含丰富的字符语义和字音信息。因此，本文采用递归拆解方案来拆解字形，拆解过程如图4所示。通过字形查找表对汉字的字形进行迭代提取，直至全部汉字拆解完毕或达到最大字形结构数量。通过递归拆解可以得到合成部件、基础部件和笔画3种粒度的字形，将这3种粒度的字形特征进行拼接，得到字形序列CZ = { w_c^1,w_c^2,⋯,w_c^n}。例如，对“病”字进行递归拆解，得到w_c^i= ['疒'，'丙'，'丶'，'厂'，'冫'，'一'，'人'，'冂'，'丶'，'一'，'丿'，'冫'，'一'，'丿'，'乀'，'丨'，'乙']。

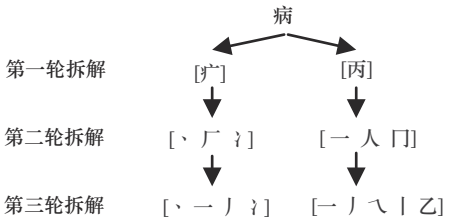

图4　汉字的递归拆解过程

如图5所示，在获取汉字的字形序列CZ和字音序列PY后，首先，分别通过2个多层一维卷积层处理CZ和PY，以提取汉字的字形和字音特征。然后，通过最大池化和全连接层分别获取汉字的字音嵌入$\boldsymbol{x}_p$和字形嵌入$\boldsymbol{x}_c$。最后，采用拼接方式将字符的语义嵌入、字形嵌入和字音嵌入进行融合，得到经过多源信息加强的字符嵌入$\boldsymbol{X}$ = { $\boldsymbol{x}_1$,$\boldsymbol{x}_2$,⋯,$\boldsymbol{x}_n$}。

2.2　上下文编码层

在NER中，实体的识别效果通常与其所处的上下文环境密切相关。在医疗文本中，局部上下文环境（如修饰词、提示词等）对实体的类型判定和边界划分具有直接影响。因此，本文在语义编码层设计了全局模块和局部注意力模块，分别用于提取实体与上下文之间的全局和局部依赖关系。在解码阶段，将非线性融合后的全局嵌入和局部嵌入输入分类器中，并通过实体匹配算法对分类出的头索引和尾索引进行匹配。

2.2.1　全局模块

全局模块负责对序列中所有字符间的依赖关系进行建模，该层常用的架构包括CNN、Transformer和BiLSTM网络。考虑文本的方向信息对实体识别的重要影响，本文采用具有天然时序性的单层BiLSTM网络进行全局编码。前向LSTM如式(2)所示。

$$\begin{bmatrix} i_t \\ f_t \\ o_t \\ \tilde{c}_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} \left(\boldsymbol{W} \begin{bmatrix} x_t^c \\ h_{t-1} \end{bmatrix} + b \right) \tag{2}$$

其中，σ表示逐元素的Sigmoid函数，$\boldsymbol{W}$为权重矩阵，b为偏置项，i_t、f_t和o_t分别为输入门、遗忘门和输出门，$\tilde{c}_t$为候选记忆状态，tanh是神经元激活函数，x_t^c和h_{t-1}分别为t时刻的输入变量和$t-1$时刻的隐藏层变量。后向LSTM以与前向LSTM相反

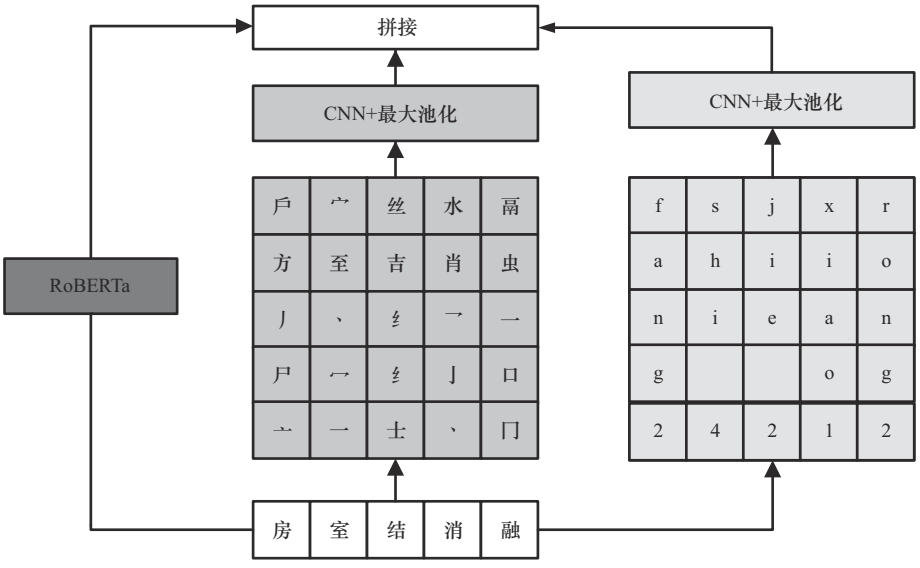

图5　多源信息增强的字符嵌入

的顺序对序列进行建模，编码后得到基于全局依赖的隐藏序列 $\mathbf{HG}=\{\boldsymbol{h}_G^1,\boldsymbol{h}_G^2,\cdots,\boldsymbol{h}_G^n\}$，其中，$\boldsymbol{h}_G^i$ 表示第 i 个字符的全局依赖向量。

2.2.2　局部注意力模块

为了减少全局信息对局部信息的干扰，本文构建了以当前时间步输入的字符为中心、固定大小为 $2m+1$ 的滑动窗口，利用这些滑动窗口将序列划分为多个序列片段后再进行处理。每个滑动窗口内部被视为与当前时间步输入直接相关的局部上下文环境，局部注意力模块示意如图6所示，其中每个圆圈代表一个字符，通过不同深度的颜色来表示不同字符在经过注意力机制处理后的信息强度，颜色越深代表嵌入后的字符信息在经局部注意力计算后增强得越多。

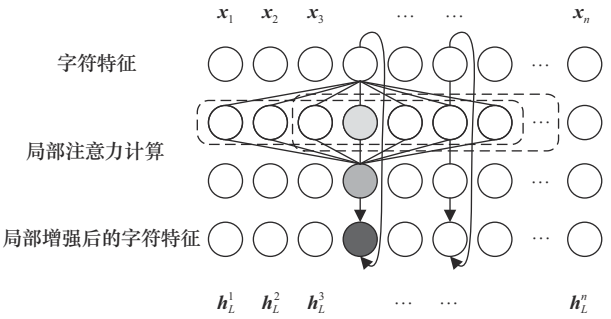

图6　局部注意力模块示意

本文通过线性注意力机制计算窗口内所有字符对当前时间步输入的注意力权重。考虑注意力机制是无向运算，而方向性的缺失会对实体识别造成一定影响，因此LENER模型采用了NEZHA[28-29]中的函数式相对位置编码来表征字符间的相对位置，以确保模型能够充分捕捉字符间的相对位置信息，具体如式(3)和式(4)所示。

$$\mathbf{pos}_{ij}[2k]=\sin\left(\frac{j-i}{10\,000^{\frac{2k}{d_z}}}\right) \tag{3}$$

$$\mathbf{pos}_{ij}[2k+1]=\cos\left(\frac{j-i}{10\,000^{\frac{2k}{d_z}}}\right) \tag{4}$$

其中，$\mathbf{pos}_{ij}$ 代表2个字符索引为 i 和 j 时的相对位置编码，d_z 为隐藏层维度，$j-i$ 代表2个索引的距离及方向，$2k$ 和 $2k+1$ 分别代表偶维度和奇维度，d_z 代表嵌入向量维度。位置编码的每个维度都由正弦函数决定，在模型训练期间是固定的。

LENER模型在计算注意力时，除了保留关键的语义交互外，还将key中的绝对位置编码替换为上述的相对位置编码，同时去除原本的缩放系数 $\sqrt{d_k}$，进一步锐化注意力，降低噪声干扰。之后对value进行加权时再次引入字符间的相对位置，得到字符的局部依赖关系，具体如式(5)~式(7)所示。

$$\boldsymbol{e}_{ij}=\boldsymbol{x}_i\boldsymbol{W}^q(\boldsymbol{x}_j\boldsymbol{W}^k+\mathbf{pos}_{ij})^{\mathrm{T}} \tag{5}$$

$$\boldsymbol{\alpha}_{ij}=\frac{\exp(\boldsymbol{e}_{ij})}{\sum_{r=1}^{2m+1}\exp(\boldsymbol{e}_{ir})} \tag{6}$$

$$\boldsymbol{h}_{\mathrm{La}}^i=\sum_{j=1}^{2m+1}\boldsymbol{\alpha}_{ij}(\boldsymbol{x}_j\boldsymbol{W}^v+\mathbf{pos}_{ij}) \tag{7}$$

其中，$\boldsymbol{W}^q$、$\boldsymbol{W}^k$、$\boldsymbol{W}^v$ 为可训练参数，$\boldsymbol{x}_i$ 为嵌入层的字符表征，$\boldsymbol{e}_{ij}$ 为分数矩阵，$\boldsymbol{a}_{ij}$ 为注意力权重向量，$\boldsymbol{h}_{\mathrm{La}}^i$ 代表字符 i 在经局部注意力计算后得到的局部依赖向量。

为了提高模型的泛化性能，在获取局部向量 $\boldsymbol{h}_{\mathrm{La}}^i$ 后，将其与预训练模型得到的原始语义特征向量 $\boldsymbol{x}_s$ 进行线性运算，从而综合利用不同特征维度之间的信息。运算后得到基于局部上下文信息的隐藏序列 $\mathbf{HL}=\{\boldsymbol{h}_L^1,\boldsymbol{h}_L^2,\cdots,\boldsymbol{h}_L^n\}$，其中，$\boldsymbol{h}_L^i$ 为第 i 个字符的局部依赖向量，具体如式(8)所示。

$$\boldsymbol{h}_L^i=\mu_1\boldsymbol{h}_{\mathrm{La}}^i+\mu_2\boldsymbol{x}_s^i \tag{8}$$

其中，μ_1 和 μ_2 为超参数。

2.3　解码

本文通过门控机制将语义编码层的全局依赖信息和局部依赖信息融合，得到最终的字符嵌入特征 $\boldsymbol{h}_i$，具体如式(9)所示。

$$\boldsymbol{h}_i=\sigma(\boldsymbol{h}_L^i)\odot\boldsymbol{h}_G^i \tag{9}$$

其中，σ 表示Sigmoid激活函数，$\odot$ 表示element-wise乘法运算。

与基于序列标注的NER任务和基于跨度提取的NER任务不同，本文将医疗命名实体的识别视为抽取实体头部和尾部的问题，在抽取的同时为其赋予类型。在抽取结束后，将抽取出的实体头部和尾部按照一定规则进行组合，从而提取实体。

将嵌入特征直接输入2个全连接层进行分类。首先通过全连接层Start_Linear得到 p_{sc}^i，p_{sc}^i 表示每个字符开始一个 c 类实体的概率。然后将嵌入特征与实体的开始位置信息 s_i 进行拼接，随后送入另一个全连接层End_Linear进行分类。最终得到 p_{ec}^i，p_{ec}^i 表示每个字符结束一个 c 类实体的概率。

$$p_{\mathrm{sc}}^i=\boldsymbol{W}^s\boldsymbol{h}_i+b^s \tag{10}$$

$$p_{\mathrm{ec}}^i=\boldsymbol{W}^e[\boldsymbol{h}_i;s_i]+b^e \tag{11}$$

其中，$\boldsymbol{W}^s$、$\boldsymbol{W}^e$、b^s 和 b^e 为可训练参数，s_i 表征实

体开始位的 one-hot 编码，在验证与测试时，s_i = softmax(p_{sc}^i)。

在提取序列的头概率列表和尾概率列表后，通过 argmax 分别提取实体的头列表 $start_t$、尾列表 end_t、头概率列表 $start_p$ 和尾概率列表 end_p。然后遍历文本序列，选取概率最高的实体头以及与实体头类别相同且概率最高的实体尾进行组合，从而提取实体。实体匹配算法描述如算法 1 所示。

算法 1　实体匹配算法

输入　$start_t$,$start_p$,end_t,end_p

输出　跨度列表 L

1)设置参数 state 为 1，待提取的实体尾字符序号 e_l 及对应的概率 e_p、待提取的实体头字符序号 s_l 及对应的概率 s_p、类别 s_t 均置为空
2)for $i \leftarrow 0$ to n do
3) if $start_t^i$ 为 0
4) if end_t^i 为 0: 执行 continue 语句
5) end if
6) if end_t^i 不为 0
7) if 参数 state 为 2
8) if s_t 中元素与 end_t^i 中元素相同，则将 i 赋值给 e_l，end_p^i 赋值给 e_p，并置 state 为 3
9) end if
10) end if
11) if 参数 state 为 3 then:
12) if s_t 中的元素与 end_t^i 中的元素相同并且 $end_p^i > e_p$，则将 i 赋值给 e_l，将 end_p^i 赋值给 e_p
13) end if
14) end if
15) end if
16) end if
17) if $start_l^i$ 不为 0
18) if 参数 state 为 1
19) 将 i 赋值给 s_l，将 $start_p^i$ 赋值给 s_p，将 $start_t^i$ 赋值给 s_t 并置 state 为 2
20) end if
21) if 参数 state 为 2
22) if $start_p^i > s_p$，则将 i 赋值给 s_l，将 $start_t^i$ 赋值给 s_t，将 $start_p^i$ 赋值给 s_p
23) end if
24) end if
25) if 参数 state 为 3
26) 将 [s_l,e_l,s_t] 加入 L 中
27) 将 i 赋值给 s_l，将 $start_t^i$ 赋值给 s_t，将 $start_p^i$ 赋值给 s_p 并将 state 置为 2
28) end if
29) end if
30) end for

2.4　损失函数

考虑医疗 NER 属于多分类问题，且医疗数据集中存在类别不均衡的现象，本文选取 focal loss 进行优化。定义模型的实体头部损失函数及实体尾部损失函数分别为

$$FL_{start} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}\varphi_{ij}(1 - P_s^{ij})^{\gamma}\log(P_s^{ij}) \quad (12)$$

$$FL_{end} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}\varphi_{ij}(1 - P_e^{ij})^{\gamma}\log(P_e^{ij}) \quad (13)$$

其中，C 代表实体的类别数；P_s^{ij} 为 p_s^i 经 softmax 函数得到的结果，其代表模型对第 i 个样本属于第 j 个类别的预测概率，P_e^{ij} 同理；φ_{ij} 代表平衡因子，用于调整不同类别的权重；γ 表示困难样本的系数，较大的 γ 值会降低易分类样本的权重，使模型更关注困难样本。最终模型的总损失为

$$FL_{total} = \lambda_1 FL_{start} + \lambda_2 FL_{end} \quad (14)$$

其中，λ_1 和 λ_2 为超参数，头部损失 FL_{start} 和尾部损失 FL_{end} 分别基于 λ_1 和 λ_2 加权。

3　实验

3.1　数据集

为了验证本文所提方法的有效性和广泛性，本文在 4 个公开数据集上进行了实验，包括医疗数据集 CMeEE、cMedQANER 和通用数据集 Resume、Weibo。表 1 和表 2 展示了这些数据集的统计信息。

CMeEE 数据集是一个标准的医学 NER 数据集，包括疾病（dis）、临床表现（sym）、药物（dru）、医疗设备（equ）、医疗程序（pro）、身体（bod）、医学检验项目（ite）、微生物类（mic）和科室（dep）9 类实体。该数据集包含训练集数据 15 000 条，验证集数据 5 000 条，测试集数据 3 000 条。

cMedQANER 数据集来源于 ChineseBLUE1.0，是来自问答论坛的中文医疗 NER 数据集，包含 11 类实体。该数据集包含训练集数据 1 673 条，验证集数据 175 条，测试集数据 215 条。

表1 医疗数据集实体数量统计

CMeEE			cMedQANER		
类别	训练集/个	测试集/个	类别	训练集/个	测试集/个
dis	15 843	4 935	dis	3 897	432
sym	12 269	4 130	cro	691	78
dru	3 930	1 440	sym	2 260	231
equ	888	238	bod	2 518	238
pro	6 332	2 057	tre	1 067	145
bod	17 697	5 883	time	212	32
ite	2 581	923	drug	539	62
mic	1 908	584	fea	311	28
dep	348	110	phy	376	45
—	—	—	test	486	49
—	—	—	dep	146	10

表2 通用数据集实体数量统计

Resume			Weibo		
类别	训练集/个	测试集/个	类别	训练集/个	测试集/个
NAME	949	112	PER	1 196	249
CONT	260	28	LOC	87	48
LOC	47	6	GPE	189	48
RACE	115	14	ORG	182	22
PRO	149	16	—	—	—
EDU	607	87	—	—	—
ORG	4 599	552	—	—	—
TITLE	2 650	324	—	—	—

Resume数据集是根据新浪财经网关于上市公司的高级经理人的简历摘要数据经过筛选、过滤和人工标注生成的。该数据集包括人名（NAME）、国籍（CONT）、地名（LOC）、种族（RACE）、专业（PRO）、教育背景（EDU）、机构（ORG）、职称（TITLE）8类实体。该数据集包含训练集数据3 821条，验证集数据463条，测试集数据477条。

Weibo数据集是一个为NER标注的微博信息语料库，包括人名（PER）、地名（LOC）、行政区（GPE）、组织机构（ORG）4类实体。该数据集主要基于2013年11月至2014年12月从微博上采样的1 890条信息进行标注，其中包括训练数据集1 350条，开发数据集270条，测试数据集270条。

3.2 对比实验分析

NER的评价指标包括精确率P、召回率R和F_1值。本文采取严格Micro-F_1作为评测指标，即要求预测出的实体的起始位、结束位和实体类型完全匹配才算预测正确。评价指标定义为

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{15}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{16}$$

$$F_1 = \frac{2PR}{P + R} \tag{17}$$

其中，TP为真正例，FP为假正例，FN为假负例。

由于医疗领域具有特殊性、专业性及隐私性等特点，目前专门针对医疗领域的NER模型及数据很少开源，因此本节选取了8个在通用领域表现较好的模型作为基线模型，具体介绍如下。

1) FLAT：基于Transformer融合了一种无损引入词汇信息的动态结构，支持并行化计算，可以大幅提升推断速度[15]。

2) Lattice LSTM：通过Lattice LSTM表示句子中的词组，将潜在的词汇信息融合到基于字符的LSTM-CRF中[14]。

3) MECT：提出了一种能够将字特征、词特征和部首特征结合的双流模型来提高NER方法的性能[22]。

4) RICON：设计了规律感知模块和规律诊断模块，规律感知模块用于捕捉每个跨度的内部规律，规律诊断模块用于定位实体的边界[24]。

5) BERT+CRF：将BERT预训练模型与CRF结合，实现BERT-CRF模型。

6) SoftLexicon（LSTM）+bichar：为避免设计复杂的模型结构，并便于迁移到其他序列标注框架，提出了一种在嵌入层简单利用词汇的方法[16]。

7) LEBERT：把词汇信息融入BERT底层的编码过程中，不需要包含词汇类型信息的词典，只需要普通的词向量即可[20]。

8) TURNER：提出了2种不确定抽样方法，通过检索的方法解决NER问题。这些方法既可以解决模糊识别问题，又可以将检测到的知识作为文本，在检测过程中允许动态更新知识库[21]。

本节在医疗数据集CMeEE、cMedQANER和通用数据集Weibo、Resume上分别进行实验，并将结果与现有的基线模型的结果进行比较，实验结果如表3~表6所示。通用数据集F_1值对比如图7所示从表3~表6和图7可以看出，LENER在医疗数据集和通用数据集上的效果都优于基线模型，证明了本文所提模型的有效性。

表3　CMeEE数据集实验结果

模型	P	R	F_1
FLAT	56.14%	47.63%	51.33%
Lattice LSTM	60.86%	56.74%	58.73%
MECT	60.36%	60.38%	60.37%
RICON	66.25%	64.89%	65.57%
BERT+CRF	63.33%	64.93%	65.78%
LENER	68.46%	64.04%	66.18%

表4　cMedQANER数据集实验结果

模型	P	R	F_1
FLAT	83.19%	79.40%	81.25%
Lattice LSTM	80.20%	77.27%	78.71%
MECT	82.63%	80.59%	81.60%
RICON	—	—	—
BERT+CRF	79.68%	82.04%	80.93%
LENER	86.80%	77.48%	81.88%

表5　Weibo数据集实验结果

模型	P	R	F_1
FLAT	—	—	60.32%
Lattice LSTM	53.04%	62.26%	58.79%
MECT	—	—	70.43%
LEBERT	—	—	70.75%
BERT+CRF	74.21%	72.49%	73.34%
SoftLexicon（LSTM）+bichar	59.08%	62.22%	61.42%
TURNER	—	—	71.22%
LENER	77.22%	75.79%	76.48%

表6　Resume数据集实验结果

模型	P	R	F_1
FLAT	—	—	95.54%
Lattice LSTM	94.81%	94.11%	94.46%
MECT	96.40%	95.39%	95.98%
LEBERT	—	—	96.08%
BERT+CRF	95.68%	96.19%	95.54%
SoftLexicon（LSTM）+bichar	95.71%	95.77%	95.74%
TURNER	—	—	96.36%
LENER	97.80%	97.63%	97.72%

在基线模型中，FLAT、Lattice LSTM、LEBERT、SoftLexicon（LSTM）+bichar都利用了字典或知识库等外部知识。然而，这种引入外部知识的方法对外部知识的质量要求较高。本文从文本的内部语言模式入手，有选择性地利用实体的全局上下文信息和局部上下文信息来辅助NER。与使用外部知识的基线模型相比，本文所提模型不需要高质量的外部知识，更适用于外部知识难以获取或质量不高的情况，实验结果也证实了本文提出的字符局部增强模型的有效性。

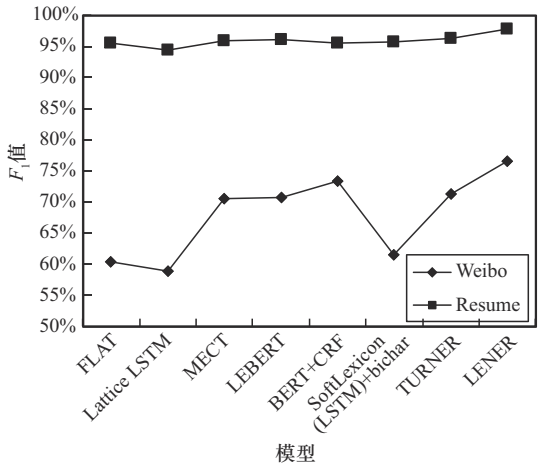

图7　通用数据集F_1值对比

此外，LENER模型在CMeEE数据集上的效果优于同样利用文本内部语言模式的RICON模型和MECT模型，表明综合考虑实体之外的上下文环境对提高实体识别的准确率是有效的。

3.3　实体边界划分效果实验

在全局模块的BiLSTM模型中，当前时间步输入与前一时刻隐藏状态之间缺乏直接交互，这导致了实体与其相邻文本之间的上下文信息缺失问题。由于该模型忽略了相邻上下文对实体的直接作用，进而影响实体的边界划分及分类的准确性。针对这一问题，本文引入了局部注意力计算单元，以补充和完善缺失的局部依赖关系。为进一步证明局部注意力计算单元的有效性，本节在CMeEE数据集上设置了消融实验，其中，模型1代表仅使用全局模块的模型，模型2代表完整的模型，并统计了下述3种情况出现的次数。

1) 情况1：预测实体与真实实体的类型不同，但左右边界一致。

2) 情况2：预测实体与真实实体的类型相同，左边界一致，但右边界不同。

3) 情况3：预测实体与真实实体的类型相同，右边界一致，但左边界不同。

实验结果如表7~表9所示，表8和表9中相差距离value定义为

$$\text{value} = \text{length}(\text{预测实体}) - \text{length}(\text{真实实体}) \quad (18)$$

表7 情况1出现次数

混淆类型	模型1/次	模型2/次
dis/sym	686	591
pro/ite	111	111
bod/ite	105	55
bod/dru	64	46
pro/equ	17	20
dis/bod	41	33

表8 情况2出现次数

value	模型1/次	模型2/次
–5	56	46
–4	70	29
–3	53	36
–2	211	131
–1	259	187
1	286	260
2	273	226
3	124	147
4	70	68

表9 情况3出现次数

value	模型1/次	模型2/次
–5	118	99
–4	139	146
–3	279	218
–2	340	263
–1	424	361
1	314	243
2	336	206
3	261	204
4	118	141

从表7可以看出，在加入局部计算模块后，预测实体与真实实体边界相同但类别不同的情况减少了。这一结果表明，局部计算模块注意到了由局部上下文不同而导致的语义细微差别。从表8和表9可以看出，在加入局部计算模块后，实体边界划分错误问题有所改善。在相差距离为–2、–1和1的情况下，边界划分错误的次数明显减少，说明注意相邻上下文与字符间的依赖关系是可行的。

同时，对比表8和表9可以发现，左边界正确、右边界错误这一情况的次数远少于右边界正确、左边界错误的次数。这可能是由于医疗实体的右边界常以某类关键词结尾，存在更明显的语言模式。

3.4 消融和滑动窗口选取实验

由于CBLUE 3.0的更新，天池平台对CMeEE数据集中的部分标注错误进行了修复。本节在天池平台最新发布的CMeEE_v2、cMedQANER、Weibo、Resume数据集上进行相关实验。

为了验证本文所提局部增强模型及多源信息融合的有效性，本文在上述2个数据集上进行消融实验。在本节实验中，模型1代表完整的模型，模型2代表去除了局部注意力模块的模型，模型3代表去除了多源信息嵌入的模型，模型4代表同时去除了局部注意力模块和多源信息嵌入的模型。实验结果如表10、表11和图8所示。

表10 CMeEE_v2消融实验结果

模型	P	R	F_1
模型1	77.76%	62.54%	69.32%
模型2	77.89%	62.25%	69.20%
模型3	77.30%	62.15%	68.90%
模型4	77.15%	62.08%	68.79%

表11 Weibo消融实验结果

模型	P	R	F_1
模型1	77.22%	75.79%	76.48%
模型2	76.20%	73.30%	74.72%
模型3	73.24%	78.04%	72.02%
模型4	74.56%	69.48%	71.93%

通过分析表10、表11与图8可以得到以下结论。

1) 将模型2的结果与模型4的结果进行比较后发现，引入包括字音、字形在内的多源特征后，F_1值有所提升，召回率在Weibo数据集上显著提高，这表明多源特征有助于模型识别出更多实体。

2) 将模型3与模型4的结果进行比较后发现，加入局部模块后F_1值有明显提升，说明充分理解字符与局部上下文和全局上下文之间的关系可以为模型提供更好的信息提取能力。

3) 将模型1与模型4的结果进行比较后发现，有效结合多源信息与局部增强模块能提升模型性能。

4) 将模型2与模型3的结果进行比较后发现，在医疗数据集上，相比于局部语义，多源信息对性能的影响更大。结合表8和表9的结果对比来看，可

能是由于医疗实体中存在如“XX 术”和“XX 菌”等固定搭配，因此医疗实体的尾字符中存在显著的字形分布。例如，疾病类型实体常以“病”“症”和“瘤”结尾，这 3 个字都包含偏旁部首“疒”，因此多源信息特征更易被模型学习。而在通用数据集上，这种固定搭配相对较少，局部信息的影响则更突出，这也说明了该模型具有较好的泛化性。

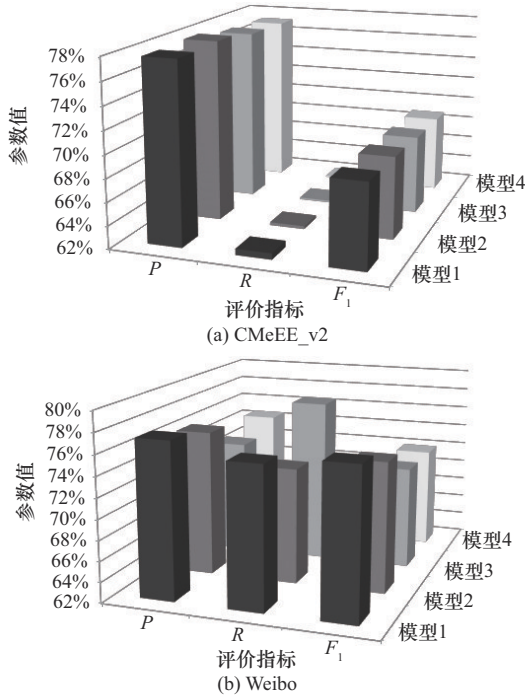

图 8　消融实验结果

此外，本文将每个滑动窗口的内部视为与当前时间步输入直接相关的局部上下文环境，因此局部上下文的选取范围与滑动窗口的选取范围保持一致。考虑实体间的距离会影响窗口覆盖的上下文长度和内部文本的语义差异，本文对实体间平均距离是否与滑动窗口大小的选取有关进行了实验，实验结果如表 12 所示。

从表 12 可以看出，在实体间平均距离的一定范围内浮动的滑动窗口的效果是最好的，证实了局部上下文的选取范围与实体间的平均距离有关。实体间平均距离定义为

$$\text{average\_dis}=\frac{\sum_{j=1}^{k}\sum_{i=1}^{n^j-1}(s_{i+1}-v_i)}{\sum_{j=1}^{k}(n^j-1)} \tag{19}$$

其中，k 代表句子数量，n^j 代表第 n 句话中包含的实体数量，s_{i+1} 代表第 i+1 个实体的头序号，v_i 代表第 i 个实体的尾序号。

3.5　跨语言命名实体识别实验

为验证模型的跨语言性，本节在 jnlbpa 数据集上进行实验。jnlbpa 数据集是基于 PubMed 数据库的英文数据集，实体类型为蛋白质、DNA、RNA、cell line 和细胞类型。本节选用了 3 种模型作为基线模型，分别为 BioBERT[30]、MINER[31]和 GRU，实验结果如表 13 所示。

表 13　jnlbpa 数据集实验结果

模型	P	R	F_1
BioBERT	72.68%	83.21%	77.59%
MINER	—	—	77.03%
GRU	64.35%	76.20%	69.77%
LENER	72.37%	82.13%	76.94%

表 12　滑动窗口选取

数据集	实体间平均距离	滑动窗口大小	P	R	F_1
CMeEE_V2	4.608 7	3	77.76%	62.54%	69.32%
		5	77.56%	62.42%	69.17%
		7	77.51%	62.27%	69.06%
cMedQANER	16.296 9	17	86.80%	77.48%	81.88%
		15	86.08%	76.52%	81.02%
		9	85.93%	76.44%	80.91%
Weibo	12.397 6	7	76.18%	74.93%	75.41%
		9	76.86%	76.02%	76.44%
		11	77.22%	75.79%	76.48%
		13	75.77%	73.30%	74.52%
		15	75.77%	74.11%	74.93%
Resume	2.060 8	3	97.80%	97.63%	97.72%
		5	97.63%	97.72%	97.67%
		7	97.80%	97.54%	97.67%

由表13可以看出，虽然LENER模型的效果并非最佳，但其F_1值与其他模型相比差距并不显著。这一现象可能由两方面因素导致，一方面，汉语语法和英语语法存在差异，因此关注相邻上下文的方法并非最优解；另一方面，LENER采用了非线性融合方式对局部计算模块与全局模块进行整合，有效降低了局部模块引入的噪声，从而使LENER模型的F_1值与其他模型相差较小。

4 结束语

为了解决医疗领域中外部知识难以获取的问题，本文提出了一种基于局部增强的中文医疗命名实体识别模型LENER，该模型基于文本内部语言模式进行NER研究。首先通过观察医疗文本内部的语言模式发现，一些实体的局部上下文信息对实体的识别具有重要影响，因此，LENER模型选用滑动窗口对序列进行切分，在序列片段内通过注意力机制增强字符基于局部上下文的依赖关系。然后将局部特征与全局特征有选择性地进行融合，平衡全局依赖关系和局部依赖关系对最终实体识别的影响。最后为了丰富底层特征，在嵌入层对包括字音、字形和语义的多源信息进行融合。实验结果表明，与现有的基线模型相比，LENER在医疗数据集和通用数据集上的F_1值提升了0.5%~2.0%，并在一定程度上提升了命名实体的识别效果。未来的研究工作将进一步探讨如何有效地丰富局部上下文与字符间的交互信息，优化建模输入与其所处的上下文间的相互作用，以进一步提升命名实体识别的效果。

参考文献：

[1] LI J, SUN A X, HAN J L, et al. A survey on deep learning for named entity recognition[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(1): 50-70.

[2] ZHANG H, ZONG Y, CHANG B B, et al. Medical entity annotation standard for medical text processing[C]//Proceedings of the 19th Chinese National Conference on Computational Linguistics. Stroudsburg: ACL Press, 2020: 561-571.

[3] MELIS G, KOČISKÝ T, BLUNSOM P. Mogrifier LSTM[J]. arXiv Preprint, arXiv: 1909.01792, 2019.

[4] SONG B S, LI F, LIU Y S, et al. Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison[J]. Briefings in Bioinformatics, 2021, 22(6): 1-18.

[5] FUKUSHIMA K. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position[J]. Biological Cybernetics, 1980, 36(4): 193-202.

[6] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9: 1735-1780.

[7] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM Press, 2017: 6000-6010.

[8] 曹依依, 周应华, 申发海, 等. 基于CNN-CRF的中文电子病历命名实体识别研究[J]. 重庆邮电大学学报(自然科学版), 2019, 31(6): 869-875.
CAO Y Y, ZHOU Y H, SHEN F H, et al. Research on named entity recognition of Chinese electronic medical record based on CNN-CRF[J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2019, 31(6): 869-875.

[9] DANG T H, LE H Q, NGUYEN T M, et al. D3NER: biomedical named entity recognition using CRF-BiLSTM improved with fine-tuned embeddings of various linguistic information[J]. Bioinformatics, 2018, 34(20): 3539-3546.

[10] YU J T, BOHNET B, POESIO M. Named entity recognition as dependency parsing[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL Press, 2020: 6470-6476.

[11] ZHANG Z C, QIN X H, QIU Y L, et al. Well-behaved transformer for Chinese medical NER[C]//Proceedings of the 2021 3rd International Conference on Natural Language Processing (ICNLP). Piscataway: IEEE Press, 2021: 162-167.

[12] YAN H, DENG B C, LI X N, et al. TENER: adapting transformer encoder for named entity recognition[J]. arXiv Preprint, arXiv: 1911.04474, 2019.

[13] GONG L J, ZHANG Z F. Clinical named entity recognition from Chinese electronic medical records using a double-layer annotation model combining a domain dictionary with CRF[J]. Chinese Journal of Engineering, 2020, 42(4): 469-475.

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

[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. Stroudsburg: ACL Press, 2020: 6836-6842.

[16] MA R T, PENG M L, 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: ACL Press, 2020: 5951-5960.

[17] 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 (EMNLP-IJCNLP). Stroudsburg: ACL Press, 2019: 1040-1050.

[18] YANG P, CONG X, SUN Z Y, et al. Enhanced language representation with label knowledge for span extraction[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL Press, 2021: 4623-4635.

[19] LI X Y, FENG J R, MENG Y X, et al. A unified MRC framework for

named entity recognition[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL Press, 2020: 5849-5859.

[20] LIU W, FU X Y, ZHANG Y, et al. Lexicon enhanced Chinese sequence labeling using BERT adapter[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg: ACL Press, 2021: 5847-5858.

[21] GENG Z C, YAN H, YIN Z Y, et al. TURNER: the uncertainty-based retrieval framework for Chinese NER[J]. arXiv Preprint, arXiv: 2202.09022, 2022.

[22] WU S, SONG X N, FENG Z H. 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: ACL Press, 2021: 1529-1539.

[23] XUAN Z Y, BAO R, JIANG S Y. FGN: fusion glyph network for Chinese named entity recognition[C]//Knowledge Graph and Semantic Computing: Knowledge Graph and Cognitive Intelligence. Berlin: Springer, 2021: 28-40.

[24] GU Y J, QU X Y, WANG Z F, et al. Delving deep into regularity: a simple but effective method for Chinese named entity recognition[C]// Proceedings of the Findings of the Association for Computational Linguistics. Stroudsburg: ACL Press, 2022: 1863-1873.

[25] LI F, WANG Z, HUI S C, et al. Modularized interaction network for 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: ACL Press, 2021: 200-209.

[26] CHEN C, KONG F. Enhancing entity boundary detection for better 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: ACL Press, 2021: 20-25.

[27] LI F, LIN Z C, ZHANG M S, et al. A span-based model for joint overlapped and discontinuous 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: ACL Press, 2021: 4814-4828.

[28] WEI J Q, REN X Z, LI X G, et al. NEZHA: neural contextualized representation for Chinese language understanding[J]. arXiv Preprint, arXiv: 1909.00204, 2019.

[29] 沈传鑫, 王永杰, 熊鑫立. 基于图注意力网络的DNS隐蔽信道检测[J]. 信息网络安全, 2023, 23(1): 73-83.
SHEN C X, WANG Y J, XIONG X L. DNS covert channel detection based on graph attention network[J]. Netinfo Security, 2023, 23(1): 73-83.

[30] LEE J, YOON W, KIM S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics, 2020, 36(4): 1234-1240.

[31] WANG X, DOU S H, XIONG L M, et al. MINER: improving out-of-vocabulary named entity recognition from an information theoretic perspective[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL Press, 2022: 5590-5600.

[作者简介]

陈晶（1976– ），女，黑龙江哈尔滨人，博士，广东海洋大学教授、博士生导师，主要研究方向为社交网络分析和自然语言处理。

邢珂萱（1998– ），女，黑龙江大庆人，燕山大学硕士生，主要研究方向为自然语言处理。

孟伟伦（1998– ），男，河北衡水人，燕山大学博士生，主要研究方向为自然语言处理。

郭景峰（1962– ），男，黑龙江哈尔滨人，博士，燕山大学教授、博士生导师，主要研究方向为数据库理论及应用。

冯建周（1978– ），男，河北沧州人，博士，燕山大学副教授、硕士生导师，主要研究方向为自然语言处理与知识图谱。