结合全局信息增强的医学领域命名实体识别研究

要媛媛[1]，付　潇[2]，杨东瑛[1]，王洁宁[1]，郑　文[1, 3*]

（1. 太原理工大学 计算机科学与技术学院 (大数据学院)，晋中 030600；2. 中国船舶集团有限公司综合技术经济研究院，北京 100081；3. 长治医学院 山西省智能数据辅助诊疗工程研究中心，长治 046000）

摘要　中文医疗问诊文本中，由于口语化的不规则表达和专业术语的频繁出现，药物名称等实体难以被精准地识别出来。为了充分利用中文句子词间关系的重要作用，提出了一种用于增强全局信息的医学命名实体识别模型。模型利用注意力机制增强了词嵌入表征，并在使用双向长短时记忆网络的序列处理能力获取上下文信息的基础上，同时从两个方面丰富了句子的全局信息表示。其一是根据句法关系获取词语之间额外依赖关系构建了图卷积网络层用于丰富词间的依赖；其二是构建了辅助任务用于预测词间句法依赖关系的类别。在中文医疗问诊数据集上的实验结果表明，模型具有很好的竞争力，F1 值达到 94.54%。与其他模型相比，在药物和症状等实体类别的识别上取得了明显提高。在微博公开数据集上的实验也表明，模型具有通用领域的应用价值。

关键词　命名实体识别；医疗问诊；双向长短时记忆网络；注意力机制；图卷积网络

中图分类号　TP391　　**文献标志码**　A　　DOI　10.12178/1001-0548.2023064

Research on Named Entity Recognition in the Medical Domain with Global Information Augmentation

YAO Yuanyuan[1], FU Xiao[2], YANG Dongying[1], WANG Jiening[1], and ZHENG Wen[1, 3*]

(1. College of Computer Science and Technology (College of Data Science), Taiyuan University of Technology, Jinzhong 030600, China;
2. China Institute of Marine Technology & Economy, Beijing 100081, China;
3. Shanxi Engineering Research Centre for Intelligent Data Assisted Treatment, Changzhi Medical College, Changzhi 046000, China)

Abstract　Entities such as drug names are difficult to identify accurately in Chinese medical questioning texts due to the frequent occurrence of colloquial irregular expressions and jargon. To make full use of the important role of inter-word relations in Chinese sentences, a medical named entity recognition model for enhancing global information is proposed. The model enhances the word embedding representation using an attention mechanism and enriches the global information representation of sentences in two ways simultaneously, based on the use of the sequence processing capability of bidirectional long and short-term memory networks to obtain contextual information. Firstly, a graphical convolutional network layer was constructed to enrich inter-word dependencies based on syntactic relationships to obtain additional dependencies between words; secondly, an auxiliary task was constructed to predict the class of syntactic dependencies between words. Experimental results on the Chinese medical consultation dataset show that the model is very competitive, with an F1 value of 94.54%. Significant improvements were achieved in the recognition of entity classes such as drugs and symptoms compared to other models. Experiments on the Weibo public dataset also show that the model has general-domain applications.

Key words　attention mechanism; bidirectional long and short-term memory network; graph convolutional net- work; medical consultation; named entity recognition

“互联网+医疗”平台发展模式的快速推进，使医患交流拥有了更便捷、通畅的途径，患者通过网站或手机移动端享受异地在线问诊服务，从医学专家那里获取问诊解答。在线问诊需求增长迅速，但参与线上答疑的医生等人力投入行有限，问诊回复容易出现不及时的情况。由此加速了在线自动医

收稿日期：2023−03−06；修回日期：2023−10−31
基金项目：国家自然科学基金（11702289）；山西省关键核心技术和共性技术研发攻关专项项目（2020XXX013）
作者简介：要媛媛，主要从事医疗大数据方面的研究。
*通信作者 E-mail: zhengwen@tyut.edu.cn

疗问诊的发展，用人机对话辅助问诊过程，进一步提高在线诊疗的效率。

构建自动医疗问诊系统，需要从对话文本中抽取结构化文本信息用于构建知识库，因此医学领域中使用自然语言处理（Natural Language Processing, NLP）来进行实体识别和文本分类等任务。命名实体识别是医学文本挖掘的重要步骤之一，应用于医疗问诊中，旨在从患者的在线咨询文本中自动识别具有医学分析价值的实体，如药物、手术、疾病等，从而利用它们在医学知识库中搜索答案。

近年来，深度学习被广泛应用于图像分割、目标检测等任务中[1]，在医学领域中的应用和研究也取得了新的突破[2-4]。然而，在识别医学命名实体上仍然存在许多挑战。首先，在医疗问诊中，文本包含了很多简洁的对话，模型能够获取的信息十分有限，使得短句中的症状实体很难被识别。其次，受一些口语化表述的影响，医学词汇的多义现象会增加识别的困难，如“退烧”在不同的句子中所代指的信息是不同的，当与“吃”关联的时候，它指的是药品类别实体。此外，一些药品和疾病实体的名称是偏长的，同时名称中一般含有专业、偏僻的表述字，如“小儿氨酚黄那敏颗粒”，通过理解“颗粒”而将其判定为药品的可能性会更大。因此，增加词间的关联信息对判断实体类别有帮助。

可见，在命名实体识别任务中，由于实体在句中是相对稀疏的词级单元，以词为基础的信息增强具有重要意义，词间丰富的信息对于句意的理解和实体的挖掘都是十分重要的。为了增强在中文语境下医疗问诊文本实体的预测效果，本文提出了一种通过双向长短时记忆网络（Bi-directional Long Short-Term Memory, Bi-LSTM）结合图卷积网络（Graph Convolutional Network, GCN）实现全局信息增强的医学领域命名实体识别模型 Glo-MNER（Global Information Augmented Medical Named Entity Recognition）。此外，模型中构建辅助任务用于预测句法间的关系类别：一方面，考虑到中文里部分药品等实体的生僻性，对于词的字级表征，使用注意力机制[4]获取更值得关注的部分；对于文本序列表征，使用 Bi-LSTM 提取句子特征[5]，提高模型记忆长序列的能力。另一方面，依据中文句法中词语的有向修饰或者依赖关系，利用 GCN[6]提取显式的词间关联，同时将预测词间关系的类别作为额外的训练任务用来进一步丰富句子的全局信息。在医疗问诊和通用领域数据集中的实验结果表明，本文模型与之前的研究方法相比具有明显的优势。

1 相关工作

命名实体识别（Named Entity recognition, NER）是 NLP 中常见任务之一，也是构建结构化文本数据的重要方法之一。传统的基于规则和词典的实体识别方法依赖于人工制定的语义和结合词典的语法、句法规则。文献 [7] 通过制定正则表达式构建 FASTUS 名称识别系统，将识别任务分为识别短语、识别模式、融合事件 3 部分。文献 [8] 针对语音输入提出了使用 Brill 规则的 NER 系统，表明自动规则推理可以作为基于隐马尔可夫模型的 NE 方法的可行替代方案。基于机器学习的 NER 方法，使实体抽取任务转化为分类问题。因此一些有监督的分类方法被用于处理 NER 任务，文献 [9] 使用决策树算法，利用不同的特征子集来训练多个独立的分类器，实现了多语言 NER 系统；在预测实体标签时，语言的序列特征和相邻词语的标签位置关系也是值得关注的，文献 [10] 将隐马尔可夫模型应用到了 NER 任务，用于识别姓名、日期等常见实体；文献 [11] 通过特征归纳的方法来提高准确性并减少特征数量，并基于条件随机场（Conditional random field, CRF）来关注前后标签的顺序问题。但是，由于特殊领域的文本和词典的独特性，多数基于人工规则和机器学习的实体识别方法难以跨领域使用。

随着深度学习在 NLP 方向上不断取得突破，许多研究提出了基于神经网络的 NER 方法。最早由文献 [12] 提出通过单向 LSTM 解决命名实体识别问题的神经网络。之后在 NER 任务中，由于文本的语言连贯特性，一些研究致力于丰富端到端的序列表示。文献 [13] 利用 Bi-LSTM、卷积神经网络（Convolutional Neural Networks, CNN）和 CRF 的组合构建神经网络来丰富模型的特征表示能力，同时使用了词的字符级向量表示来进行序列表征。在中文领域，考虑到汉语的分词问题，文献 [14] 提出了一种格结构的 LSTM 模型，使这种独立于分词的方式提供了显式的词序列信息且避免了分词造成的错误传播；在这一基础上，文献 [15] 将词典信息整合到字符表示中，验证了该方法与其他序列架构结合可以快速应用于 NER 任务；文献 [16] 进一步将字符和单词序列视为两种不同的模态，考虑了单词和字符格结构上的密集交互，通过设计跨

格的注意力模块捕获不同模态细粒度相关性。除此之外，将句子结构或上下文的依赖信息应用于 NER 任务也是可行的，文献 [17] 证实了句子的语法结构对 NER 具有积极的影响。

医学领域的 NER 主要是为了抽取疾病、医疗操作等关键实体，为构建结构化的医疗知识库提供核心要素，同时为智能辅助诊疗分析等智能化场景应用提供基础支撑。为了实现自动挖掘医学领域关键实体，早期的研究主要通过对规则的统一定义以及应用机器学习算法来完成。文献 [18] 通过句法和词汇模式用于无监督医学命名实体识别。文献 [19] 制定规则用于处理时间表达式的识别和归一化，通过规则与机器学习相结合，用于临床文本的信息提取。循环神经网络的出现使得深度学习在医学文本处理上得到了更广泛的应用。在嵌入增强方面，文献 [20] 使用 LSTM 和 CNN 构建组合的字符级表征，来增强医学文本词的表示。在文本表征方面，Bi-LSTM 结构在处理医学信息依然能起到重要作用[21]。近年来基于 Transformer 的双向编码器表示 (Bidirectional Encoder Representations from Transformers, BERT) 被广泛应用在各个领域，被证明具有有效分类中文领域医学类文本的能力[22]。除了文本序列学习方面的研究工作外，通过汉字词典和字形增强中文领域的医学命名实体识别也能获得很好的模型表现[23]。针对数据集特点设置特征词用于注释也能够提高医学领域的 NER 效果[24]。本文中，为了解决医学领域文本中的一些实体抽取困难的问题，通过增加句法结构信息以及进一步丰富词的字符级向量表示来提升模型在中文医学领域的命名实体识别效果。

2 医学领域命名实体识别模型

本文构建的医学领域命名实体识别模型结构如图 1 所示，模型的架构描述如下。

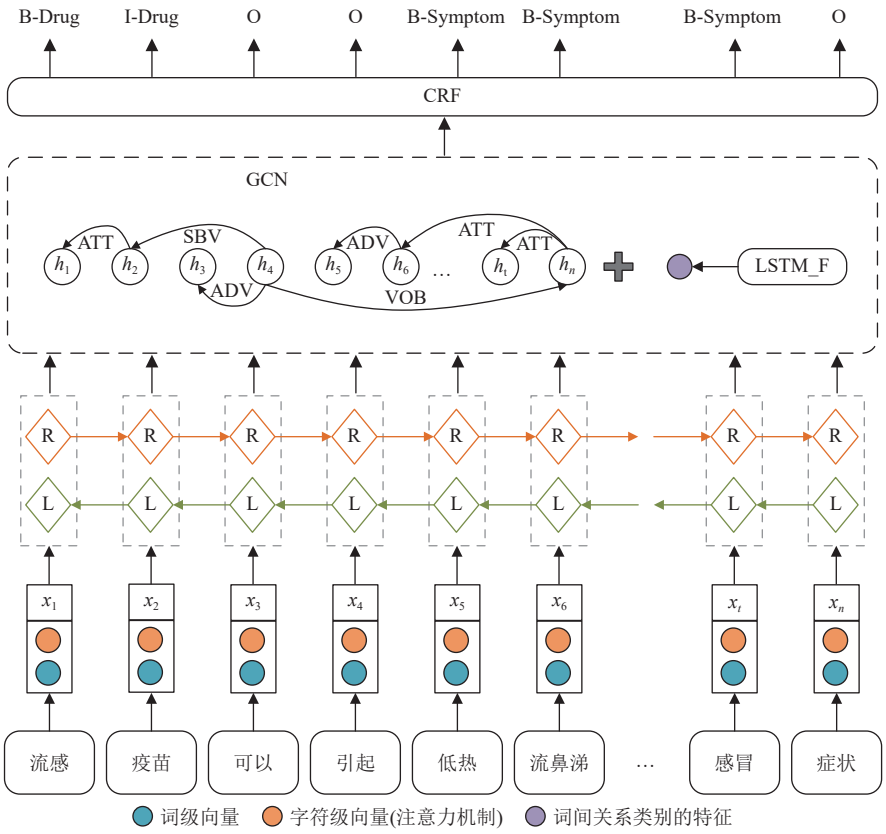

图 1 Glo-MNER 模型架构

1）词向量表征增加了引入注意力机制的字符级特征；2）句子序列特征通过 Bi-LSTM 学习；3）使用图网络表示词语间的句法联系，并额外增加词间关系类别的辅助学习任务；4）通过 CRF 层输出最终标签。

2.1 丰富字符级表征

由于问诊的文本涉及到口语化和非正式的语言环境，而部分实体名称又具有专业性特点，词的重要性要更突出，因此，有必要进一步丰富词的信息。本文在词级向量的基础上，增加了基于注意力

机制的字符级向量来丰富词语的含义表示。

基于注意力机制的字符向量表示，使得最终的词表示包含内部的注意力特征，借此缓解模型中由分词导致的错误传播影响。注意力机制可以关注一个词内字的注意力焦点，尤其是对于名称较长的实体，一些专业化表述是更值得关注的部分。因此输入层向量是按照词向量、字符向量进行拼接得到。

给定一个输入$\boldsymbol{X}=(\mathrm{x}_1,\mathrm{x}_2,\cdots,\mathrm{x}_n)\in\mathbb{R}^{n*d}$，初始化权重矩阵$\boldsymbol{W_K}$和$\boldsymbol{W_V}$，同时使用$\boldsymbol{Q}\in\mathbb{R}^{1*d}$软选择包含重要特征的相关信息，注意力信息的计算为：

$$\boldsymbol{K}=\boldsymbol{X}\boldsymbol{W_K},\ \boldsymbol{V}=\boldsymbol{X}\boldsymbol{W_V} \tag{1}$$

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V} \tag{2}$$

式中，$\boldsymbol{W_K}$、$\boldsymbol{W_V}$是可学习参数。最终的词向量表示为$\boldsymbol{x}_i=\left[\boldsymbol{x}_i^{\mathrm{word}};\boldsymbol{x}_i^{\mathrm{attention}}\right]$。

2.2 获取长序列特征

循环神经网络（Recurrent Neural Network, RNN）是在传统前馈神经网络基础上，针对获取序列前后相关性而做出的改进，具有处理可变长度序列的能力。RNN 模型运用了门控的机制来存储输入向量，并充分利用了序列特性，该 RNN 记忆信息称为隐藏状态，它使 RNN 能够预测输入数据序列中的下一个输入是什么，RNN 通过连接隐含层之间的节点，可以动态地学习序列特征。给定输入序列$(\mathrm{x}_1,\mathrm{x}_2,\cdots,\mathrm{x}_n)$，在$t$时刻，RNN 更新记忆信息$h_t$为：

$$h_t=\mathrm{f}(\boldsymbol{W}_x x_t+\boldsymbol{W}_h h_{t-1}+b_t) \tag{3}$$

式中，f为非线性激活函数；$\boldsymbol{W}_x$和$\boldsymbol{W}_h$是模型权重矩阵；b_t是偏差。

传统的循环神经网络模型可以传递词之间的语义信息，但无法捕捉长距离的语义连接。而长短期记忆（Long Short-Term Memory, LSTM）扩展了 RNN 的记忆信息，通过引入门控机制和记忆单元克服了梯度消失的问题，这种扩展具有获取长时间记忆信息的能力。

如图 2 所示，LSTM 模型单元由 3 个门组成：遗忘门、输入门和输出门。遗忘门决定保留或删除现有的信息，输入门指定新信息添加到内存的程度，输出门控制单元中的现有值是否有助于输出。

基于h_{t-1}和x_t的值，遗忘门通常用 sigmoid 函数决定需要从 LSTM 存储器中删除多少信息。遗忘门的输出f_t是一个介于 0 和 1 之间的值，其中 0 表示完全摆脱已学习的信息，1 表示保留全部。该输出的计算如下：

$$f_t=\sigma\left(\boldsymbol{W}_{fh}h_{t-1}+\boldsymbol{W}_{fx}x_t+b_f\right) \tag{4}$$

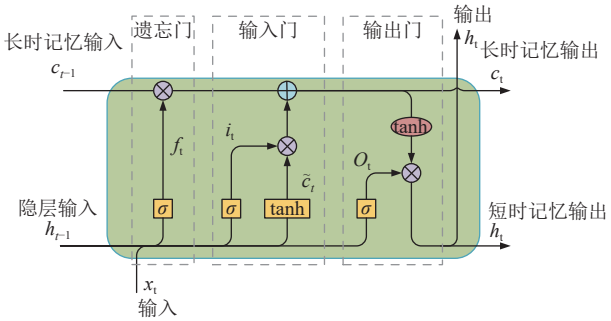

图 2 LSTM 单元结构

输入门决定是否将新的学习信息添加到 LSTM 的存储器中，它由 sigmoid 层和 tanh 层组成。其中i_t决定哪些信息需要更新，计算方式如式 (5)，$\tilde{c}_t$则表示将被添加到 LSTM 存储器中的候选值向量，如式 (6)。

$$i_t=\sigma(\boldsymbol{W}_{ih}h_{t-1}+\boldsymbol{W}_{ix}x_t+b_i) \tag{5}$$

$$\tilde{c}_t=\tanh(\boldsymbol{W}_{\tilde{c}h}h_{t-1}+\boldsymbol{W}_{\tilde{c}x}x_t+b_{\tilde{c}}) \tag{6}$$

当前的记忆单元c_t由遗忘门f_t对旧的记忆信息c_{t-1}选择性删除后，增加新的候选信息，即输入门提供的候选向量，计算过程为：

$$c_t=f_t\cdot c_{t-1}+i_t\cdot\tilde{c}_t \tag{7}$$

LSTM 网络的输出门o_t使用 sigmoid 层来决定 LSTM 存储器中的哪一部分对输出是有贡献的，计算方式见式 (8)。h_t是最终的记忆信息，使用 tanh 函数实现对记忆单元c_t从−1 到 1 的映射，如式 (9) 所示。

$$o_t=\sigma(\boldsymbol{W}_{oh}h_{t-1}+\boldsymbol{W}_{ox}x_t+b_o) \tag{8}$$

$$h_t=o_t\tanh(c_t) \tag{9}$$

考虑到中文文本数据中，词的含义不只局限于单向的依赖传递，因此，本文使用 Bi-LSTM 学习每一时刻的上下文信息。Bi-LSTM 具有双向的 LSTM 单元，它平等对待所有输入词，捕获每一个词的上下文表征。本文将拼接的多维词向量序列作为输入，Bi-LSTM 的输出即为整句话的向量表示。

然而，Bi-LSTM 的序列处理特性决定，虽然它能够表达长距离依赖，但它对位置上相联系的词

更敏感，即它表达局部上下文信息的能力更强，这样会忽略一些远距离而在语义上相关联的词信息表达。

2.3 词间关系表示

在中文句子中，语法关系有助于理解句子的成分和语义，依存句法分析旨在找出句子的核心语义，它显式地描述了词之间的依赖关系。如图 3 所示，将“引起”作为核心词，加强了“感冒”（症状）和“气管炎”（症状）等的联系，此外，“抗感冒”（药物类别）和“药物”的联系也是值得关注的。因此，本文使用词语间有向的依赖关系[25]构建图卷积网络。

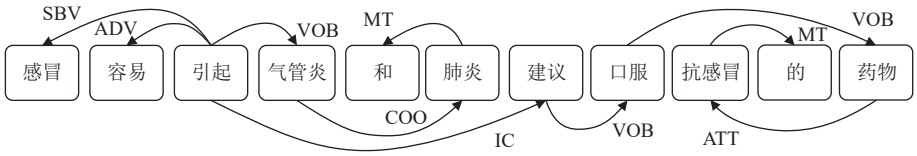

图 3 词依赖图（曲线箭头表示句子成分之间的依赖关系）

树是一种特殊的图，而依存分析树实际是一个完全图的子图。本文使用 DDParser（Baidu Dependency Parser）作为依存分析树初始化工具，DDParser 产出的依存关系如图 4 所示。图中箭头表示一条依存关系三元组的连接，POS 表示词性，Relation 表示关系，To 表示关系方向，句子中所有的关系三元组共同构成了依存句法分析树。然后用图结构表示依存分析树：$\boldsymbol{G}=(\boldsymbol{V},\boldsymbol{E})$，其中，$\boldsymbol{V}$表示句子中节点集合，由上一步产出的词向量表示，$\boldsymbol{E}$表示边集合，由依存分析树各个词语的关系表示，利用依存句法分析树中各个词语之间的概率值对图中边进行初始化。最终句子的依存分析树$\boldsymbol{G}$被表示为$\boldsymbol{n}\times\boldsymbol{n}$邻接矩阵$\boldsymbol{K}$，其中$\boldsymbol{A}_{ij}$表示节点$i$通过$\boldsymbol{G}$中的单个依存路径连接到节点$\boldsymbol{j}$。

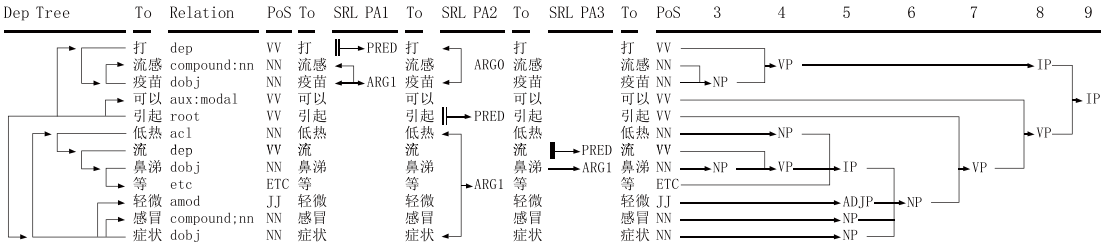

图 4 依存句法分析示例

图卷积网络在图结构上对相邻节点特征进行卷积，每一个图卷积层通过邻居节点的属性映射来导入新的节点嵌入。用来表达文本类句法关系时，节点表示词的特征，边代表词之间的句法关系，本文使用每个节点的传出边构建有向的图结构。

单层 GCN 只能捕获节点i的一跳邻居节点信息，通过$\boldsymbol{K}$层 GCN 可以捕获节点i的$\boldsymbol{K}$跳邻居节点信息。GCN 模型正向传播时通过捕获到的词语之间句法依存关系，找出与核心词相关的依存关系构成句法关系三元组进一步提升实体识别准确率，同时模型训练过程中通过反向传播不断更新邻接矩阵$\boldsymbol{K}$的权重值，纠正 DDParser 的初始化权重修正依存分析树中节点之间的关系概率值，形成正向闭环。

$$\overrightarrow{h}_v^{k+1}=\mathrm{ReLU}\left(\sum_{u\in\overrightarrow{N}(v)}\left(\overrightarrow{W}^k h_u^k+\overrightarrow{b}^k\right)\right) \tag{10}$$

式中，$\boldsymbol{W}$是由节点和边构成的权重矩阵；h_u^k表示第k层节点u的嵌入向量；$\overrightarrow{N}(v)$表示节点 v 所有最近邻居节点的集合（包括v本身）；$\overrightarrow{b}^k$是偏差项。

2.4 增强全局信息的辅助任务

虽然使用 GCN 在一定程度弥补了模型在全局信息捕获上的不足，但远距离的词间依赖信息依然是较少的。同时，一些句法关系类别对语义关系的理解有进一步的补充作用，如图 3 中，“抗感冒”和“药物”的定中关系（Attribute, ATT）是常见的。

因此，使用 LSTM 将预测实体间依赖关系的类别作为二分类的辅助任务，将其隐藏特征与 GCN 层的输出结果相加，计算公式如下：

$$H_p=\mathrm{LSTM}_p(x_i) \tag{11}$$

$$H=\boldsymbol{W}_p\sum_p^n\boldsymbol{H}_p+\boldsymbol{W}_{\mathrm{GCN}}\boldsymbol{H}_{\mathrm{GCN}} \tag{12}$$

式中，H作为最终 CRF 层的输入；$\boldsymbol{H}_p$表示不同的

句法关系类别；$\boldsymbol{W}$是可学习权重矩阵。

由于中文句法关系种类较多，为了降低模型复杂性，对实体的常见关系种类进行统计，并保留了五类，见表 1。

表 1 实体对应句法关系的占比 %

标签	VOB 动宾	ATT 定中	SBV 主谓	HED 核心	COO 并列
药物名称	43.22	11.43	9.8	5.47	10.13
药物类别	56	19.96	10.26	1.76	5.23
医疗操作	26.21	15.49	9.8	13.4	4.25
症状	24.8	18.20	12.33	15.06	8.77
医学检验	33.82	14.99	12.44	10.85	6.24

2.5 标签标注

在序列标注方面，考虑到少量的分词错误和标注的连续特性（如 B-Drug 必须在实体的第一个标注位，它后面是 I-Drug，不能是 I-Operation），本文依然采用 NER 任务常用的解码方法 CRF 来提供额外的标签转移特征。

对于序列$\boldsymbol{H}=(\mathrm{h}_1,\mathrm{h}_2,\cdots,\mathrm{h}_n)$，在经过编码后，需要计算得到每个词对应每个标签上的分数。通过计算节点的分数和节点之间的转移分数得到输出序列$\boldsymbol{Y}=(\mathrm{y}_1,\mathrm{y}_2,\cdots,\mathrm{y}_n)$的分数为：

$$\mathrm{score}(\boldsymbol{Y})=\sum_{i=1}^{n}q_i+\sum_{i=2}^{n}(\boldsymbol{A}[i-1,i]) \tag{13}$$

式中，q_i表示每个词h_i对应的标签得分的分布；转移矩阵$\boldsymbol{A}$中$\boldsymbol{A}[i-1,i]$表示标签序列y_{i-1}到y_i的转移分数。之后将序列$\boldsymbol{Y}$的分数进行归一化，得到标注结果的概率为：

$$p(\boldsymbol{Y}|\boldsymbol{H})=\frac{q^{\mathrm{score}(\boldsymbol{Y})}}{\sum_{\tilde{Y}}q^{\mathrm{score}(\tilde{\boldsymbol{Y}})}} \tag{14}$$

$q^{\mathrm{score}(\tilde{Y})}$表示所有可能序列对应的分数的指数。

3 实 验

3.1 实验数据及标注模式

本文使用来自中国计算语言学大会 (CCL 2021) 举办的第一届智能对话诊疗评测比赛[1] 的公开文本数据，共计 97 522 条，按照 8∶1∶1 的比例划分训练集、验证集和测试集。本实验采用了命名实体识别标注模式中的 BIO 标注模式，O 表示非医学实体词， B、I 分别表示医学实体的首部、非首部。数据中的实体类别和数量统计见表 2。

表 2 实验数据

标签	实体	训练集	验证集	测试集
Drug	药物名称	5 065	593	659
Drug_Category	药物类别	4 318	522	599
Operation	医疗操作	1 530	164	179
Symptom	症状	26 156	3 121	3 290
Medical_Examination	医学检验	7 854	963	1 077

3.2 实验环境及评价指标

本文的硬件环境为 Xeon(R) Silver 4114 @ 2.20 GHz，操作系统使用 CentOS release 7.6.1810（64 位），开发环境为 python3.6。

在词特征提取过程中，使用了 2 层的 Bi-LSTM 和 2 层的 GCN，利用 LSTM 预测关系类别的辅助任务隐层数设置为 1。

对于训练参数，训练进行了 50 个 epoch，batch size 大小设置为 10。在训练过程中，使用学习率为 5.0×10^{-4} 的 Adam 优化器进行参数优化，学习率的衰减率为 0.05，为了避免过拟合现象，采用 L2 正则化，lambda 参数设置为 1.0×10^{-8}。

本文使用 NER 主要评价指标：精确率（Precision, P）、召回率（Recall, R）、综合指标（F1-score, F1）评估模型性能。

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

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

$$\mathrm{F1}=\frac{2PR}{\mathrm{P}+\mathrm{R}} \tag{17}$$

式中，TP表示预测正确的五类医学实体；FP表示将无关词预测为医学实体；FN表示将医学实体预测为非该实体。

此外，为了验证模型在不同实体上的突出表现，本文同时使用了宏平均 F1（MacroF1）来评估模型性能：

$$\mathrm{MacroF1}=\frac{1}{n}\sum_{i=1}^{n}\mathrm{F}_1^i \tag{18}$$

式中，F_1^i表示第i类实体的 F1 值。

本文使用 UAS 和 LAS 评估依存分析树质量，UAS 表示已正确分配的关系三原组中的词语所占

1 https://tianchi.aliyun.com/dataset/dataDetail?dataId=95414

的百分比；LAS 表示已正确分配的三元组关系占三元组关系总数的百分比。评估公式如下：

$$\text{UAS} = \frac{\text{已正确分配关系的三元组的词语数量}}{\text{所有关系三元组的词语数量}} \quad (19)$$

$$\text{LAS} = \frac{\text{已正确分配关系的三元组的数量}}{\text{所有关系三元组数量}} \quad (20)$$

3.3 实验结果与分析

3.3.1 与以往研究比较

为了验证本文提出的模型在医疗问诊文本命名实体识别的效果，实验中对比了本文提出的模型与近年来用于中文命名实体识别任务的神经网络模型。其中，BERT 通过加载预训练模型配合微调的方法，在多项 NLP 任务中取得了出众表现，它的优势在于下游任务的微调时对训练数据集的需求量小，但构建预训练模型需要大量的文本数据，实验中使用由谷歌提供的 BERT-base-Chinese 预训练模型，并使用 CRF 进行解码标注；Lattice-LSTM[14]、SoftLexicon[15] 和 DCSAN[16] 都是基于字典和深度学习的模型，可有效解决中文分词错误，但需要构建对应的词典，在后两个模型的对比实验中，均使用了双向字符级的向量嵌入。表 3 展示了各模型在医学领域数据中的结果。

可以看出，总体上 Glo-MNER 在医疗问诊数据集上的表现相较其他模型有明显的提升，微平均F1值达到了 94.31%，而宏平均F1值也表现出了明显优势，说明模型在各类实体上的识别效果是出色的。虽然没有构建专门的医学领域词典，但 3 种基于词典的方法在进行这类词级任务时表现良好。由于训练数据偏少，BERT-CRF 模型取得了一定的优势，但对于识别难度相对大的实体类并没有起到关键作用，根据宏平均F1值可以看出，DCSAN 在某些实体上的识别效果突出，而本文模型的宏平均值处于稳定且较好的水平。

表 3 不同实体识别模型实验结果比较 %

Model	P	R	F1	MacroF1
Lattice-LSTM	90.84	90.34	90.59	90.28
SoftLexicon(LSTM)	91.25	92.28	91.76	91.84
BERT-CRF	91.60	93.01	92.30	91.92
DCSAN	91.52	92.70	92.11	92.51
Bi-LSTM-CRF	86.19	90.11	88.11	88.30
Glo-MNER	**94.67**	**93.95**	**94.31**	**94.54**

3.3.2 全局信息增强的效果

通过人工对比 DDParser 初始化产出的依存关系三元组与训练了 N 次后产出的依存关系三元组，分别为：DDParser 初始化依存树、模型迭代 50 次、100 次和 500 次之后的依存句法分析树。发现模型训练能够有效纠正初始化过程中的部分错误标注以及概率预测不准确等问题。实验评估如表 4 所示。

表 4 依存句法分析树质量评估 %

Method	UAS	LAS
DDParser	90.31	89.06
GCN迭代50次	94.80	90.68
GCN迭代100次	95.72	92.88
GCN迭代500次	96.23	93.45

为了验证模型增加的 GCN 模型和辅助任务在医学 NER 任务上的作用，本文针对字符表征、GCN 和辅助学习任务进行了消融实验，如表 5 所示。

表 5 消融实验 %

Model	P	R	F1
Glo-MNER	**94.67**	**93.95**	**94.31**
-attention	91.44	93.62	92.51
-GCN	92.72	90.15	91.42
-parser	92.11	91.30	91.70

当不附加额外的字符级嵌入进行训练时，F1值降低了 1.8 个百分点，模型的精确率下降较明显，表明在减少嵌入表征信息后，模型的查准率下降较多。同样，在去掉图卷积网络层关于句法结构的信息特征以及句法依赖类别的辅助学习任务时，比 Glo-MNER 分别降低了 2.89 个百分点和 2.61 个百分点，表明句法结构对于全局信息的表征具有重要作用。由此说明，丰富的词嵌入表示和基于全局信息增强的图卷积网络，以及辅助任务对模型提升在医疗问诊实体识别任务上都起到了明显的提升作用，本文提出的 Glo-MNER 模型效果可以达到最大化。

3.3.3 模型对不同实体的识别效果

为了探究模型宏平均 F1 的提升因素，分析了模型对每一类实体的识别效果。表 6 展示了各个模型在五类实体中的 F1 表现。

对于药物（Drug）和药物类别（Drug_Category）实体，在其他模型差别不大的情况下，Glo-MNER 取得了明显的提升，说明一些药物和药物类别在句子中，需要进一步的全局特征提取来获取，无论是词典方法还是多头注意力机制，对于部分实体的学习是困难的。

表 6 模型在每类实体上的表现（F1 分数）

%

Model	Drug	Drug_Category	Operation	Symptom	Medical_Examination
Bi-LSTM-CRF	86.14	88.52	88.66	87.76	90.42
Lattice-LSTM	88.96	90.31	89.01	90.79	92.31
BERT-CRF	90.58	91.51	91.10	92.43	93.98
SoftLexicon(LSTM)	90.40	91.73	91.58	91.75	93.61
DCSAN	90.87	91.80	93.51	91.81	**94.54**
Glo-MNER	**92.72**	**94.59**	**96.59**	**94.96**	93.83

由于医疗操作（Operation）和症状（Symptom）在句子中作为句法关系的核心词（Harley Ellis Devereaux, HED）的频率都很高，且大多数句子为短句，这种情况下通过上下文获取到的信息非常少，因此句法关系辅助任务可以起到很好的作用，模型在这两类实体的提升效果都十分明显。此外，症状在文本中并列出现是较常见的，因此并列结构（Chief Operating Officer, COO）也能对预测起到积极的作用。对于医学检验（Medical_Examination），由于实体多为动词，且比较常规化（如“检查”“化验”），各类方法的效果都相对较好且不具备明显差异。针对不同类别实体的实验结果表明，本文方法在识别专业性语义特征明显，在文本句法关系上有显著特点且较难识别的实体时具有突出的优势。

3.3.4 模型在通用领域数据集上的表现

为了进一步验证本文模型在命名实体识别任务上的效果，在公开数据集 Weibo NER[1] 上做对比实验，它来自社交媒体网站新浪微博中的文本数据，同样具有口语化和简写等中文表达特点。从表 7 中可以看到，本文模型的表现与它在医疗领域数据集中的观察结果是相似的，表明本文模型在不同领域也具有很好的竞争力。

表 7 在微博数据集上的表现

%

Model	P	R	F1
Peng和Dredze[26]	63.33	39.18	48.41
He和Sun[27]	61.68	48.82	54.50
SoftLexicon(LSTM)	-	-	58.12
FLAT[28]	-	-	63.42
Glo-MNER	**71.78**	**64.69**	**68.05**

4 结束语

本文针对医疗问诊文本中一些实体的识别难点，通过增强全局信息表征来提高模型的识别效果。首先，本文使用额外的字符向量特征抽取和图卷积网络层来进一步丰富词的表征，其次，增加了基于句法关系类别预测的辅助学习任务，并证明增强全局信息表征对于命名实体识别任务是有帮助的。在医学领域数据集上的实验表明，丰富的单词表征和全局信息表征改进了句意理解和实体预测。特别地，对于药物类别等实体，本文模型相较于对比模型获得了明显的相对提升。并且，在公开数据集上的实验结果证明了此模型在通用领域的命名实体识别任务上同样是有优势的。

在真实的医学文本中，存在许多噪声和错误，同时面临着实时性和跨领域跨语言适应性的限制因此模型的泛化效果有待进一步研究。此外，本文在词内信息的表示上采用了简单的拼接操作，没有考虑类似 DCSAN 模型中的各个特征表示之间的交互表示。在今后的工作中，将进一步考虑利用多模态交互增强词的多维度特征融合。

1 https://github.com/hltcoe/golden-horse

参 考 文 献

[1] ZHAO L, ZHI L Q, ZHAO C, et al. Fire-YOLO: A small target object detection method for fire inspection[J]. Sustainability, 2022, 14(9): 4930-4943.

[2] HOU F, ZHAO C, SU N L, et al. Quantitative assessment of interstitial lung disease based on RDNet convolutional network[C]//Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine. New York: IEEE, 2022: 1550-1553.

[3] LIU X, WU Y, CHEN Y, et al. Diagnosis of diabetic kidney disease in whole slide images via AI-driven quantification of pathological indicators[J]. Comput Biol Med, 2023, 166: 107470-107483.

[4] PENG M X, HOU F, CHENG Z X, et al. Prediction of cardiovascular disease risk based on major contributing features[J]. Scientific Reports, 2023, 13: 4778-4788.

[5] WANG M, ZHOU T, WANG H H, et al. Chinese power dispatching text entity recognition based on a double-layer BiLSTM and multi-feature fusion[J]. Energy Reports, 2022, 8: 980-987.

[6] LIU B S, LIU X L, REN H, et al. Text multi-label learning method based on label-aware attention and semantic dependency[J]. Multimedia Tools and Applications, 2022, 81(5): 7219-7237.
[7] APPELT D E, HOBBS J R, BEAR J, et al. SRI International FASTUS system: MUC-6 test results and analysis[C]//Proceedings of the 6th conference on Message understanding. New York: ACM, 1995: 237–248.
[8] KIM J H, WOODLAND P C. A rule-based named entity recognition system for speech input[C]//Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP 2000). [S. l.]: ISCA, 2000: 528-531.
[9] SZARVAS G, FARKAS R, KOCSOR A. A multilingual named entity recognition system using boosting and c4. 5 decision tree learning algorithms[C]//International Conference on Discovery Science. Heidelberg: Springer, 2006: 267-278.
[10] BIKEL D M, SCHWARTZ R, WEISCHEDEL R M. An algorithm that learns what's in a Name[J]. Machine Learning, 1999, 34(1): 211-231.
[11] MCCALLUM A, LI W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons[C]//Proceedings of the 7th conference on Natural language learning at HLT-NAACL 2003 - Volume 4. New York: ACM, 2003: 188–191.
[12] HAMMERTON J. Named entity recognition with long short-term memory[C]//Proceedings of the 7th conference on Natural language learning at HLT-NAACL 2003 - Volume 4. New York: ACM, 2003: 172–175.
[13] MA X, HOVY E. End-to-end sequence labeling via bi-directional lstm-cnns-crf[EB/OL]. [2023-01-23]. https://doi.org/10.48550/arXiv.1603.01354.
[14] ZHANG Y, YANG J. Chinese NER using lattice LSTM[EB/OL]. [2023-01-23]. https://arxiv.org/abs/1805.02023, 2018.
[15] MA R, PENG M, ZHANG Q, et al. Simplify the usage of lexicon in Chinese NER[EB/OL]. [2023-01-23]. https://arxiv.org/abs/1908.05969.
[16] ZHAO S, HU M H, CAI Z P, et al. Dynamic modeling cross- and self-lattice attention network for Chinese NER[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(16): 14515-14523.
[17] CETOLI A, BRAGAGLIA S, O'HARNEY A D, et al. Graph convolutional networks for named entity recognition[C]//Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories, Prague, Jan 23-24, 2018. Stroudsburg: ACL, 2018: 37-45.
[18] ZHANG S, ELHADAD N. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts[J]. J Biomed Inform, 2013, 46(6): 1088-1098.
[19] KOVACEVIC A, DEHGHAN A, FILANNINO M, et al. Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives[J]. J Am Med Inform Assoc, 2013, 20(5): 859-866.
[20] CHO M, HA J, PARK C, et al. Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognition[J]. J Biomed Inform, 2020, 103: 103381.
[21] WU H, JI J, TIAN H, et al. Chinese-named entity recognition from adverse drug event records: Radical embedding-combined dynamic embedding-based BERT in a bidirectional long short-term conditional random field (Bi-LSTM-CRF) model[J]. JMIR Med Inform, 2021, 9(12): e26407.
[22] ZHANG D, ZHANG H, WANG L, et al. Recognition of Chinese legal elements based on transfer learning and semantic relevance[EB/OL]. [2023-01-22]. https://10.1155/2022/1783260.
[23] YANG J, WANG H M, TANG Y T, et al. Incorporating lexicon and character glyph and morphological features into BiLSTM-CRF for Chinese medical NER[C]// Proceedings of the IEEE International Conference on Consumer Electronics and Computer Engineering. New York: IEEE, 2021: 12-17.
[24] ZHANG R Y, ZHAO P Y, GUO W Y, et al. Medical named entity recognition based on dilated convolutional neural network[J]. Cognitive Robotics, 2022, 2: 13-20.
[25] Zhang S, Wang L, Sun K, et al. A practical chinese dependency parser based on a large-scale dataset[EB/OL]. [2023-01-22].https://ar5iv.labs.arxiv.org/html/2009.00901.
[26] PENG N, DREDZE M. Improving named entity recognition for chinese social media with word segmentation representation learning[EB/OL]. [2023-01-22]. https://arxiv.org/abs/1603.00786.
[27] HE H F, SUN X. A unified model for cross-domain and semi-supervised named entity recognition in Chinese social media[C]//Proceedings of the Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. New York: ACM, 2017: 3216–3222.
[28] LI X, YAN H, QIU X, et al. FLAT: Chinese NER using flat-lattice transformer[J]. [EB/OL]. [2023-01-22]. https://arxiv.org/abs/2004.11795.

编 辑 税 红