基于 ERNIE-Bi-GRU-Attention 的医疗实体关系抽取模型

姚洁仪[1,2] 王春亮[1,2]
YAO Jieyi WANG Chunliang

摘 要 医疗信息文本信息处理存在文本长、专业术语多、实体间关系复杂等问题，因此，提出一种基于 ERNIE-Bi-GRU-Attention 的医疗实体关系抽取模型。首先通过预训练模型 ERNIE 使向量获得丰富的语义信息和医疗先验知识，解决医疗专业术语问题；其次通过 Bi-GRU-Attention 进行语句编码，捕获有效上下文信息，有利于关系抽取；然后使用经典 CRF 输出实体标签；将实体标签特征和语句编码向量特征拼接进行一阶和二阶特征融合；最后通过分类器获得最终关系标签输出。通过在医疗数据集上验证，结果表明与其他模型相比，使用此模型医疗实体关系抽取的效果有所提升。

关键词 信息抽取；关系抽取；预训练；BiGRU；医疗文本

doi：10.3969/j.issn.1672-9528.2024.02.047

0 引言

人工智能技术的快速发展和信息数字化趋势对医疗行业产生了巨大的影响。大量的医疗文本数据积累，包括电子病历、医学文献、临床试验报告等，其中包含了丰富的患者病历信息、疾病诊断和治疗方案等关键内容。对医疗文本数据进行分析和挖掘至关重要，它有助于提升医疗服务质量、推动临床决策，并有效管理疾病。然而，医疗文本数据的规模庞大、结构复杂以及存在大量的非结构化文本使得人工分析和利用变得困难。因此，利用人工智能和自然语言处理等技术自动抽取医疗信息已成为当前的重要研究问题[1-3]。关系抽取是信息抽取的重要子任务，其任务旨在确定文本中提取的两个实体之间的关系。由于医疗领域的复杂性和数据的噪音，医疗信息抽取面临许多挑战，如医疗文本长、语句结构复杂、实体关系交错等，对此，本文提出了一种基于 ERNIE-BiGRU-attention 的医疗文本关系抽取模型。本模型基于预训练模型 ERNIE 训练词向量，以解决医疗实体专业词汇多，语义难理解的问题，再使用 Bi-GRU 模型捕获上下文信息，且使用 attention 机制分配权重，更好关注于关键信息，加强词向量语义信息，采用 CRF 方法获得实体标签向量，与词向量一同输入分类器进行关系预测。分类器采用了 MLP 和 Biaffine 结合的方法，不仅能够挖掘实体浅表关系还能挖掘潜在联系，有利于解决医疗实体关系复杂的问题。

1. 水电工程智能视觉检测湖北省重点实验室 湖北宜昌 443000
2. 三峡大学计算机与信息学院 湖北宜昌 443000

1 相关工作

近年来，关系抽取的研究工作可以分为抽取式和生成式。其中抽取式分为基于标注、片段、填表和阅读理解的模型，生成式分为传统 seq2seq 和预训练模型。本文主要研究抽取式关系抽取模型。

基于标注的模型，通常使用二分标注序列确定实体头尾或者实体间的位置关系。文献 [4] 提出了一种标注模式，使用 BIEOS 表示实体的开始以及结束，同时使用数字 1、2 来标注头尾实体，模型标签中融入了关系类型，但此模型无法解决三元组重叠问题。文献 [5] 提出将关系中的实体分为主体和客体，先抽取主体后再抽取与之对应的关系与客体，解决了三元组重叠问题，但存在主体提取出现偏差则影响后续客体和关系提取的问题，以及当关系类别较多时关系冗余的问题。文献 [6] 提出了预测关系机制，先筛选出有用关系集合，可以解决文献 [5] 中关系冗余造成的效率低下问题。文献 [7] 也是基于文献 [5] 的基础上，引进 Biaffine 模型，设计双向关系提取模型，降低由于主体抽取错误带来的后续任务影响。

基于片段的模型，即将一个文本分为多个 span 片段，再根据片段进行关系抽取。文献 [8] 提出了标准的基于片段抽取关系的方法，先列出文本所有的片段进行向量编码，再进行判断每个片段是否为实体及其类别，最后再将实体两两配对进行关系预测。文献 [9] 在文献 [8] 的基础上使用了 Bert 预训练模型来训练文本，使 token 有更丰富的向量表征，同时限制了 span 的长度，降低片段数量过多带来的噪声影响。文献 [10] 提出了一个基于 pipeline 框架的关系抽取模型，使用了两个编码器分别提供文本信息给实体识别和关系抽取，提

高抽取性能，且在关系抽取模型的输入中融入了实体信息，但此模型仍存在 pipeline 模型误差传递的问题，相较于文献 [8-9] 先抽取主体再匹配客体和关系的方法也多了许多数据噪声。

基于填表的方法，即对每个关系创建一个表，表项用来表示实体对是否具有这个关系，再根据关系表进行关系抽取。文献 [11] 将 token 对构建矩阵，且提出了新的实体间链接标注来构建关系表，此模型使用单步联合抽取模型，解决了多步产生的暴露偏差和误差传递问题，但是同时新的标注方法也带来了实体关系冗余的问题，效率较低。文献 [12] 提出了一种基于全局特征的关系表填充关系抽取模型，全局特征是指实体对和实体对之间的联系以及关系和关系之间的联系，将其进行关联建模，此方法减少了许多冗余信息，提升了模型性能。

基于阅读理解的方法，该方法是文献 [13] 提出的，即把实体关系抽取任务看作是多轮问答任务，将实体类别和关系类别作为问题查询模板，在文本中使用 MRC 模型提取实体和关系作为回答，从而完成实体关系抽取任务。

2 方法

本文提出了一个基于 ERNIE-BiGRU-attention 的医疗文本关系抽取模型，模型如图 1 所示。该模型由预训练编码模块、命名实体识别模块、关系抽取模块三个模块组成。预训练编码模块是基于预训练模型从文本中学习词向量表示，通过 Bi-GRU 层进行编码得到向量序列，经过 attention 层进行权重分配，再将向量输入于命名实体识别模块，命名实体识别模块采用 CRF 序列标注方法获得实体标签向量，与 Bi-GRU 得到的向量序列一同作为关系抽取模块的输入。关系抽取模块中使用了 MLP 和 Biaffine 分类器进行预测，最终关系得分由两者一同决定。

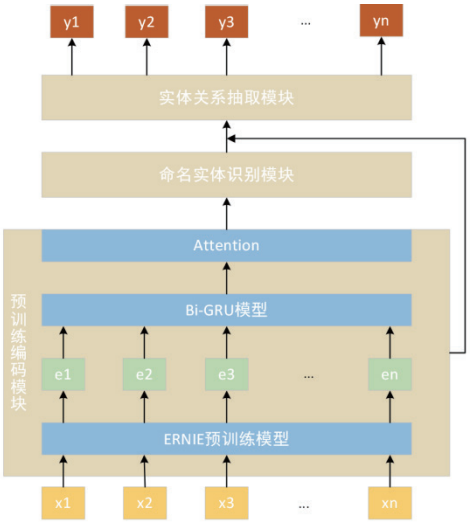

图 1 模型框架

2.1 预训练编码模块

2.1.1 ERNIE 预训练模型

该模块使用 ERNIE[14] 预训练模型得到词嵌入向量。ERNIE 是在 Bert[15] 基础上改进之后的预训练模型，主要改变了 Bert 的掩码策略。Bert 的策略是在预训练的过程中随机 mask 掉一个句子中的文字，是字符级别，导致无法充分理解句子的语法结构和其中的语义信息，缺乏全局建模信息能力。而 ERNIE 是短语级别和实体级别的掩码机制，它能够将一个由多个字组成的短语或者实体作为一个统一单元去进行 mask 和预测，如此丰富了语义信息，能够在预训练中潜在学习到实体的先验知识和长语义依赖。ERNIE 模型由多层双向 transformer 编码器构建而成。transformer[16] 中的自注意力机制计算句子中任意两个词的联系，即注意力分布，着重关注与此词相关的信息，捕获关键特征来理解整体句义。本文模型将文本输入至 ERNIE 预训练模型内，经过训练后得到表示词向量序列$X_i = \{x_1, x_2, ...x_n\}$。为了进一步提取文本特征与上下文信息，将得到的词嵌入向量序列输入至 Bi-GRU 进行编码。

2.1.2 Bi-GRU 模型

循环门单元 [17]（gated recurrent unit，GRU）是长短期记忆网络 [18]（long short-term memory，LSTM）的一种结构变体模型。LSTM 网络通过门控机制缓解循环神经网络 RNN 中存在的梯度消失问题，将短期记忆与长期记忆结合起来。为了记忆长期记忆，每个 LSTM 单元有 3 个控制门：遗忘门、输入门、输出门，控制当前信息的引入是从历史信息中选择还是从当前时刻的信息中选择。GRU 比 LSTM 少了一个门控单元，只有更新门和重置门两个门控单元，参数量相较 LSTM 而言大大减少，但效率比 LSTM 更高。GRU 结构如图 2 所示，具体函数如下：

$$R_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \quad (1)$$

$$Z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \quad (2)$$

$$\tilde{h}_t = \tanh(W \cdot [R_t \odot h_{t-1}, x_t]) \quad (3)$$

$$h_t = Z_t \odot \tilde{h}_t + (1 - Z_t) \odot h_{t-1} \quad (4)$$

式中：x_t 表示输入序列中第 t 个时间步的词向量，h_{t-1} 表示前一时间的隐状态，Z_t 表示更新门，R_t 表示重置门，$\tilde{h}_t$表示候选隐状态，h_t 表示当前时间步的隐状态，$\boldsymbol{W}$ 表示权重矩阵，σ 表示 sigmoid 函数。

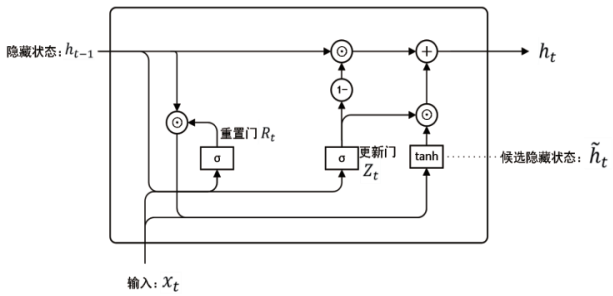

图 2 GRU 结构图

双向 GRU（bidirectional gated recurrent unit，Bi-GRU）由两个信息传递相反的 GRU 循环层构成，双向 GRU 能捕捉到两个方向上的信息，能很好地利用上下文进行语义分析，使模型含有更丰富的向量信息，提升模型分类效果。双向连接方式如图 3 所示。

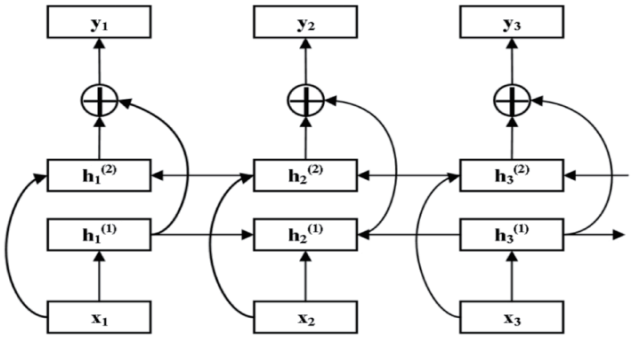

图 3 双向 GRU 示意图

最终输出的 Y 是将两层输出进行连接，函数为：

$$y_t = h_t^{(1)} \oplus h_t^{(2)} \tag{5}$$

2.1.3 attention 机制

attention 层将对 Bi-GRU 输出的特征向量 y_t 进行权重分配。虽然 Bi-GRU 整合了句子上下文信息，具有序列全局性，但仍存在长距离信息弱化问题，attention[19] 机制作用是审视整个序列，计算其中的注意力分布，从中间抓住重点信息。本文引入注意力机制来增加与命名实体相关的语义特征权重，为接下来的命名实体识别任务加强局部特征提取效果。权重分配计算位：

$$\alpha_{i,j} = \frac{\exp\left(\text{score}(x_i,x_j)\right)}{\sum_{k=1}^{n}\exp(\text{score}(x_i,x_k))} \tag{6}$$

$$score(x_i,x_j) = \frac{W_a x_i x_j}{|x_i||x_j|} \tag{7}$$

式中：W_a 是权重参数。

2.2 命名实体识别模块

本文实体识别任务采用 BIO 标注方法 [20]，即为每个字符标注为“B-X”“I-X”或“O”，其中，“B-X”表示这个字符是在 X 类型实体的开头，“I-X”表示这个字符是在 X 类型实体的中间位置，“O”表示不属于任何实体。经过编码模块处理后的信息可能出现不合理的标注情况，如两个“B”相邻等，没有考虑到相邻标签的相互影响，因此，本文引入 CRF 模型来进行约束，确保标签信息是有效合理的。CRF 条件随机场 [21] 是一种概率无向图模型，它与隐马尔可夫模型和马尔可夫随机场等模型密切相关，它具有一些独特的特点，CRF 不仅考虑输入序列的特征，还考虑不同标签之间的相互依赖关系，因此能够更准确地捕捉标签序列的结构信息。CRF 的基本思想是给定输入序列条件下，通过最大化标签序列的条件概率来选择最佳标签序列。将编码模块的输出作为 CRF 层的输入，使用维特比算法解码输出得分最大的序列为标签预测序列，计算公式为：

$$P(y|x) = \frac{\exp(s(x,y))}{\sum_{y'}\exp(s(x,y'))} \tag{8}$$

$$s = \sum_{i=0}^{n} A_{y_i,y_{i+1}} + \sum_{i=1}^{n} P_{i,y_i} \tag{9}$$

式中：x 为输入序列，y 为标签序列，P 是初始得分矩阵，A 是转换得分矩阵，s 为最后得分。

2.3 实体关系抽取模块

实体关系抽取，即预测两个实体间是否具有某种预定义关系。也可视为预测关系标签任务。本文使用 Biaffine 分类器 [22] 和 MLP 分类器相结合的方法进行实体间的关系预测。每个关系都有一个主体 subject 和一个客体 object，主体作为关系的主动方，客体作为关系的被动方，即关系三元组为（subject、relationship、object）。又因每个实体都有可能成为主体或客体，且当其作为主体和作为客体时，所处于的上下文语境以及语法都是不相同的，所以需要使用两个 MLP 分类器进行主客体分离计算。

$$s_i = MLP_1(h_i) \tag{10}$$

$$o_j = MLP_2(h_j) \tag{11}$$

该模块的输入分为两部分，一部分是 Bi-GRU-attention 编码模块的输出向量，其中携带了上下文相关语义信息，有助于关系抽取，另一部分是命名实体识别得到的标签向量，将两个向量拼接后作为实体关系抽取模块的输入。

得到了主客体特征表示后，将主体和客体进行一阶融合特征提取，本模型采用将主体和客体的隐藏层向量直接相加再通过 MLP 分类器提取特征。

$$r'_{ij} = W_{r'}(s_i + o_j) + b_{r'} \tag{12}$$

考虑到有些词语之间具有深层关系，本模型使用 Biaffine 分类器提取二阶融合特征。 Biaffine 双仿射变换具有更多地参数，能更好的学习实体信息，加强信息交互能力和数据融合程度。Biaffine 计算如下。

$$r''_{i,j} = s_i^T U o_j + W_{r''}[s_i; o_j] + b_{r''} \tag{13}$$

式中：U、W 和 b 为可训练参数。最终关系预测分数由以上两者共同决定计算：

$$r_{i,j} = Softmax\left((r'_{i,j} + r''_{i,j})\right) \tag{14}$$

本文将从 $r_{i,j}$ 中选择预测分数最高的标签作为预测结果，采用交叉熵计算 loss。

3 实验与分析

3.1 数据集

为测试本模型性能，使用中文医学文本实体与关系抽取数据集 CMeIE。CMeIE 是在 CHIP2020 中发布（Guan 等人 [23]）的基于 schema 的中文医学信息抽取数据集。它包含儿科训练语料和百种常见疾病训练语料，儿科训练语料来源于 518 种儿科疾病，百种常见疾病训练语料来源于 109 种常见疾病。其中训练集数据 14 339 条，验证集数据 3585 条，测试集数据

4482 条，关系数为 53 条，包含 10 种同义词子关系，43 种其他子关系。

3.2 实验参数

本文模型的实验环境为 NVIDIA GeForce RTX 3090，64 GB 显存。深度学习框架为 PyTorch，网络参数优化器采用 adam 优化器，经过实验最后参数设置如表 1 所示。在训练过程中，为了防止过拟合情况的发生，采用了“early stop”策略，在每个训练轮次结束后，记录下当前模型在验证集上的损失值以及 F_1 分数，当损失或 F_1 分数在指定轮数过后依旧没有得到提升，则停止训练，并记录最优情况下的训练轮数。最后将训练集和验证集合并作为新的训练集，按照之前记录的最优轮数进行训练。

表 1 参数设置

名称	值
batch size	64
epoch	100
learning rate	2e-5
dropout	0.5

3.3 实验结果分析

在本文将模型与其他几个经典关系抽取模型进行比较。

Multi-head[24]：提出了一种多头注意力机制实体关系联合抽取方法，首次使用关系分数进行实体关系分类。

NovelTagging[25]：提出一种新的标注模式，将联合任务转化成标签问题。用 BIEOS 标签来指代实体的位置，并将关系类型融入标签中，且分出头尾实体。这种标注方法提升了模型性能。

Bert-CNN[26]：句子经过 Bert 和 Bi-LSTM 获取句子上下文语义表示，再将句子转换为二维化表示，通过 CNN 提取特征后，由 Biaffine 分类器和 MLP 分类器共同预测实体之间的关系。

表 2 中列出了本文模型与其他模型和基线模型在数据集上的结果。

表 2 实验结果

模型	F_1 值 /%	预测值 /%	召回率 /%
Multi-head	58.9	63.3	55.2
NovelTagging	25.6	51.4	17.1
Bert-CNN	59.3	64.4	58.0
Ours	61.8	66.1	57.8

观察到，本文模型在几乎所有的评估指标方面都优于对比模型。其原因在于，本文模型基于 ERNIE 预训练，学习了医疗领域先验知识，提升模型识别性能，且 Bi-GRU-attention 模块在处理长文本时捕捉了上下文重点信息，解决了长距离依赖信息丢失问题，以及在关系抽取时，将实体标签与上下文信息一同作为输入，丰富了向量中携带的特征信息，有效挖掘实体间内在联系，提高了关系抽取的准确性。

3.4 消融实验

为了验证本文模型中各个模块的贡献有效性，做了相关消融实验，分别去掉了 attention 层、Bi-GRU-attention 层，得到的性能指标结果如表 3 所示。其中可以看出 attention 层能对句中关键信息进行捕捉，提升局部特征。而 Bi-GRU 层能得到含有更丰富上下文的句子信息向量，使得模型具有全局特征，解决长距离依赖问题。

表 3 消融实验

Model	F_1/%	Precision/%	Recall/%
本文模型	61.8	66.1	57.8
-attention	60.9	65.6	57.0
-BiGRU-attention	60.3	65.9	55.8

4 结论

本文提出了一种基于 ERNIE-BiGRU-attention 的医疗文本关系抽取模型，主要解决医疗文本长、句式结构复杂、专业领域词汇多且实体关系交错等问题。本文模型通过 ERNIE 预训练模型学习，利用其多掩码机制获得更丰富的语义表示及学习先验知识，解决医疗领域专业实体词汇抽取问题。使用 Bi-GRU-attention 模型捕捉上下文信息且通过注意力分布计算权重分配，解决医疗文本长、容易丢失关键信息的问题。在进行关系抽取时，本文还使用 Biaffine 双仿射注意力机制，挖掘实体间潜在关系表征。然而医疗关系抽取中还存在一些问题，未来工作将对于嵌套实体和嵌套关系进行深入研究，继续提升本模型性能。

参考文献：

[1] 游新冬，赵明智，王星予，等 . 一种融合实体类别特征的医疗领域关系抽取方法 [J]. 北京信息科技大学学报（自然科学版），2022,37(6):19-25.

[2] ZHU Y, LI L, LU H et al. Extracting drug-drug interactions from texts with bioBert and multiple entity-aware attentions [EB/OL].(2020-06-01)[2023-10-06].https://pubmed.ncbi.nlm.nih.gov/32454243/.

[3] 赵丹丹，张俊朋，孟佳娜，等 . 基于预训练模型和混合神经网络的医疗实体关系抽取 [J]. 北京大学学报 (自然科学版), 2023, 59(1):65-75.

[4] ZHENG S, WANG F, BAO H, et al. Joint extraction of entities and relations based on a novel tagging scheme[EB/OL]. (2017-07-01)[2023-09-29].https://aclanthology.org/P17-1113/.

[5] WEI Z, SU J, WANG Y, et al. A novel cascade binary tagging framework for relational triple extraction[EB/OL]. (2020-07-30)[2023-09-21].https://aclanthology.org/2020.acl-main.136/.

[6] ZHENG H, WEN R, CHEN X, et al. PRGC: potential

relation and global correspondence based joint relational triple extraction[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Online. Stroudsburg:Association for Computational Linguistics, 2021:6225-6235.

[7] REN F, ZHANG L, ZHAO X, et al. A simple but effective bidirectional extraction framework for relational triple extraction[EB/OL].(2021-12-09)[2023-09-05].https://arxiv.org/abs/2112.04940.

[8] DIXIT K, AL-ONAIZAN Y. Span-level model for relation extraction[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2019: 5308-5314.

[9] EBERTS M, ULGES A. Span-based joint entity and relation extraction with transformer pre-training[C]//European conference on artificial intelligence,european conference on artificial intelligence. Amsterdam:European Association for Artificial Intelligence, 2020:2006-2013.

[10]ZHONG Z, CHEN D. A frustratingly easy approach for entity and relation extraction[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online. Stroudsburg:Association for Computational Linguistics,2021:50-61.

[11] WANG Y, YU B, ZHANG Y, et al. TPlinker: single-stage joint extraction of entities and relations through token pair linking[C]//Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online). Stroudsburg:Association for Computational Linguistics, 2020:1572-1582.

[12] REN F, ZHANG L, YIN S, et al. A novel global feature-oriented relational triple extraction model based on table filling[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic. Stroudsburg:Association for Computational Linguistics, 2021:2646-2656.

[13] LI X, YIN F, SUN Z, et al. Entity-relation extraction as multi-turn question answering[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg:Association for Computational Linguistics, 2019:1340-1350.

[14]SUN Y,WANG S,LI Y,et al.ERNIE: enhanced representation through knowledge integration[EB/OL].(2019-01-19)[2023-09-19].https://arxiv.org/abs/1904.09223.

[15] LEE J S, HSIANG J. Patent classification by fine-tuning bert language model[J]. World patent information, 2020,61(6):51-54.

[16] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30:20-26.

[17] 温浩 , 何茜茹 , 王杰 , 等 . 基于 ERNIE-BiGRU 模型的摘要语步自动识别研究 [J]. 中文信息学报 ,2022,36(11):91-100.

[18] 蒋丽媛 , 吴亚东 , 张巍瀚 , 等 . 基于 BiLSTM-EPEA 模型的实体关系分类 [J]. 计算机时代 , 2023(5):46-50+56

[19] SHEN Y, HUANG X J. Attention-based convolutional neural network for semantic relation extraction[C]//Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. Osaka. Stroudsburg: Association for Computational Linguistics, 2016: 2526-2536.

[20] 张世豪 , 杜圣东 , 贾真 , 等 . 基于深度神经网络和 自注意力机制的医学实体关系抽取 [J]. 计算机科学 , 2021, 48(10): 77−84.

[21] 张华丽 , 康晓东 , 李博 , 等 . 结合注意力机制的 Bi-LSTM-CRF 中文电子病历命名实体识别 [J]. 计算机应用 , 2020, 40(S1): 98-102.

[22] YU J, BOHNET B, POESIO M. Named entity recognition as dependency parsing[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.Stroudsburg:Association for Computational Linguistics,2020:910-913.

[23] GUAN T, ZAN H, ZHOU X, et al. CMeIE: construction and evaluation of Chinese medical information extraction dataset[C]//Natural Language Processing and Chinese Computing: 9th CCF International Conference.Cham:Springer International Publishing, 2020: 270-282.

[24] BEKOULIS G, DELEU J, DEMEESTER T, et al. Joint entity recognition and relation extraction as a multi-head selection problem[J]. Expert systems with applications, 2018,114:34-45.

[25] ZHENG S, WANG F, BAO H, et al. Joint extraction of entities and relations based on a novel tagging scheme[J]. Proceedings of the 55th annual meeting of the association for computational linguistics ,2017(1):1227-1236.

[26] LI J, FEI H,LIU J,et al. Unified named entity recognition as word-word relation classification[EB/OL].(2022-06-28)[2023-10-11].https://ojs.aaai.org/index.php/AAAI/article/view/21344.

【作者简介】

姚洁仪（1998—），女，湖北武汉人，硕士，研究方向：自然语言处理。

（收稿日期：2023-11-30）