第 39 卷 第 5 期
2024 年 10 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol. 39 No. 5
Oct. 2024

DOI:10.13364/j.issn.1672-6510.20230180
数字出版日期：2024–06–21；数字出版网址：http://link.cnki.net/urlid/12.1355.N.20240621.1013.002

基于 RBAC 模型的中文医疗命名实体识别

张 斌，赵婷婷，张碧霞，陈亚瑞，王 媛
（天津科技大学人工智能学院，天津 300457）

摘 要：中文医疗命名实体识别旨在从非结构化数据中抽取结构化实体，目前的主流研究都使用了大量的训练数据。针对中文医疗命名实体识别训练数据匮乏的问题，提出了基于联合分词的 RBAC（RoBERTa-BiGRU-Attention-CRF）模型和基于语义搜索的命名实体识别数据增强方法。首先利用预训练模型和双向门控循环单元（BiGRU）提取文本的深度双向语义表示，再将该语义表示分别送入分词模块和命名实体识别模块。分词模块利用条件随机场（CRF）得到分词信息。命名实体识别模块利用 BiGRU 与多头注意力得到混合语义表示，再送入 CRF 得到命名实体识别的标签序列。在 CCKS2019 中文电子病历数据集上的实验结果表明，该方法在数据量较少的情况下 F_1 达到 90.5%，证明了该方法的有效性。

关键词：多任务学习；预训练模型；双向门控循环单元；多头注意力；条件随机场；数据增强
中图分类号：TP391 **文献标志码**：A **文章编号**：1672-6510（2024）05-0056-07

Chinese Medical Named Entity Recognition Based on RBAC Model

ZHANG Bin，ZHAO Tingting，ZHANG Bixia，CHEN Yarui，WANG Yuan
（College of Artificial Intelligence，Tianjin University of Science & Technology，Tianjin 300457，China）

Abstract：Chinese medical named entity recognition aims to extract structured entities from unstructured data. Current mainstream research uses a large amount of training data. Aiming at the problem of lack of training data for Chinese medical named entity recognition，a RoBERTa-BiGRU-Attention-CRF（RBAC）model based on joint segmentation and a novel data enhancement method for named entity recognition based on semantic search are proposed in this article. Specifically，the pretrained model and the Bidirectional Gated Recurrent Unit（BiGRU）are first used to extract the deep bidirectional semantic representation of the text，and then the semantic representation is sent to the word segmentation module and the named entity recognition module respectively. The word segmentation module uses conditional random fields（CRF）to obtain word segmentation information. The named entity recognition module uses BiGRU and multi-head attention to obtain a mixed semantic representation，and then is sent to CRF to obtain the tag sequence for named entity recognition. Experimental results on the CCKS2019 Chinese electronic medical record datasets showed that the F_1 of this method reached 90.5% when the amount of data was small，thus proving the effectiveness of this method.

Key words：multi-task learning；pretrained model；BiGRU；multi-head attention；CRF；data enhancement

引文格式：
张斌，赵婷婷，张碧霞，等. 基于 RBAC 模型的中文医疗命名实体识别[J]. 天津科技大学学报，2024，39（5）：56–62.
ZHANG B，ZHAO T T，ZHANG B X，et al. Chinese medical named entity recognition based on RBAC model[J]. Journal of Tianjin university of science & technology，2024，39（5）：56–62.

随着医学技术的不断发展和应用，医疗数据的挖掘与分析变得越来越重要，其中命名实体识别（named entity recognition，NER）技术在医学领域中具有广泛的应用价值。医疗命名实体识别旨在从文本

收稿日期：2023–09–30；**修回日期**：2024–02–28
基金项目：国家自然科学基金项目（61976156）；天津市企业科技特派员项目（20YDTPJC00560）
作者简介：张 斌（1999—），男，河南周口人，硕士研究生；**通信作者**：赵婷婷，教授，tingting@tust.edu.cn

中提取医学专业术语、患者身体状况、药品信息、手术操作方法等相关实体，为医学研究、临床决策、个性化医疗等提供数据支持。然而，中文医学领域的文本具有很强的复杂性和非结构性，且许多医学场景的数据匮乏，这给命名实体识别带来了挑战。因此，设计高效、准确的医疗命名实体识别系统对促进医疗信息化建设和提升医学诊疗水平具有重要意义。

早期的 NER 方法主要基于规则，此类方法利用手工设计的规则和模式从文本中识别命名实体。Kim 等[1]提出了一种自动生成规则的方法，用 Brill 规则推理方法直接捕获一组简单规则中的语义信息用于 NER 任务。基于规则的方法在多数场景效果不佳，也很难从一个应用场景移植到另一个场景。为了克服上述问题，基于统计学习的 NER 方法逐渐被研究者们所关注。张华平等[2]提出了一种基于角色标注的中国人名自动识别方法，使用经典的动态规划算法（维特比算法）对切词结果进行角色标注，使用模式最大匹配实现中国人名的识别，能够有效提升模型的识别和泛化能力。随着深度学习的兴起，许多研究者开始探索深度学习与 NER 结合的可能性。Huang 等[3]提出了基于 BiLSTM-CRF（bidirectional long short-term memory-conditional random fields）的 NER 方法，该方法将 NER 建模成序列标注任务，使用 BiLSTM（Bi-directional long short-term memory）提取文本特征，再用条件随机场（conditional random fields，CRF）计算整个标记序列的联合概率分布。BiLSTM-CRF 模型考虑了标签序列之间的关系，对词嵌入的依赖性较小，具有较强的鲁棒性，已经成为 NER 领域的经典方法，为基于预训练模型的 NER 的发展奠定了基础。自 BERT（bidirectional encoder representation from Transformers）[4]提出以来，基于预训练模型的 NER 方法成为主流。谢腾等[5]将 BERT 与 BiLSTM-CRF 结合应用在中文 NER 中，其优势在于充分发挥了 BERT 在文本特征提取方面的卓越能力，能够深入学习词级别、句法结构以及上下文中的语义信息特征。这使 BERT-BiLSTM-CRF 相较于传统模型在 NER 任务上取得更为卓越的效果。然而，目前主流的深度学习模型往往是在拥有大量训练数据的场景下表现优秀，在训练数据较少的场景表现往往较差。

本研究基于深度学习技术优化医疗命名实体识别模型，旨在提高其在训练数据有限的场景下的准确性和泛化性，为医学研究和临床诊疗提供更可靠的数据支持。

本研究的贡献主要包含以下 4 点：

（1）在中文医疗领域构建了一个分词与命名实体识别的多任务联合训练框架。这一框架显著提升了模型识别中文医疗实体边界的能力，尤其在数据匮乏的情况下，仍能使模型在中文医疗命名实体识别任务上表现出色。

（2）在命名实体识别任务中，使用混合语义向量进行解码，使模型能够同时关注到全局与局部的语义特征，进一步提升模型在医疗场景中的识别能力与泛化能力。

（3）提出了一个全新的数据增强方法，利用语义搜索（semantic search）[6]的方式对 THUOCL（THU open Chinese lexicon）[7]中的医疗词典进行检索，找到训练集中每个与词典中实体最接近的实体进行替换，从而完成中文医疗 NER 的数据增强。

（4）在 CCKS2019 中文电子病历数据集上进行了大量的实验，其表现显著优于主流模型的方法，证明了本文模型的有效性。

1　RBAC 模型

通过序列标注的方法解决中文医疗命名实体识别问题。RoBERTa-BiGRU-Attention-CRF 模型由 4 个部分组成，分别是向量表示层、特征提取层、辅助分词模块、实体识别模块，具体模型结构如图1所示。

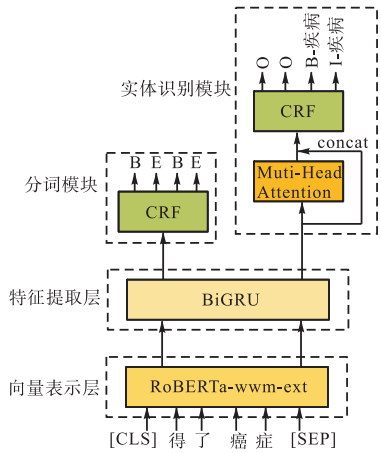

图 1　RoBERTa-BiGRU-Attention-CRF 模型结构
Fig. 1　Structure of RoBERTa-BiGRU-Attention-CRF model

向量表示层为 RoBERTa-wwm-ext[8]（RoBERTa with whole word masking-extended）预训练模型，用于将输入文本转换为一组向量表示。特征提取层为

BiGRU（bidirectional gated recurrent unit）[9]，用于进一步从向量表示层的输出中提取更深层次的语义信息，这一层的输出会被分别送入辅助分词模块和实体识别模块。辅助分词模块接收特征提取层的输出，通过 CRF 得到辅助分词任务的标签序列。实体识别模块首先接收特征提取层的输出作为实体识别模块的输入，通过多头注意力捕获句子中词与词之间的依赖关系，再将特征提取层的输出与多头注意力的输出拼接，得到包含全局特征与局部特征的混合语义向量，最后将混合语义向量送入 CRF，得到中文医疗命名实体识别的标签序列。

1.1　向量表示层

向量表示层的作用是将输入文本转换为一组向量表示，向量表示包含了输入字符在高维空间的深度双向语义信息，这些信息会被用于文本分类、序列标注等任务。

本研究选择了 RoBERTa-wwm-ext 作为向量表示层的编码器。RoBERTa-wwm-ext 是一个由 BERT 改进得到的基于 Transformer 的双向语言模型。与 BERT 相比，RoBERTa-wwm-ext 在预训练时删除了 NSP（next sentence prediction）任务，将 BERT 的静态 MASK 策略替换为动态 MASK 策略，并在训练数据规模和训练步数上进行一些调整，以进一步提升模型的性能和鲁棒性。此外，RoBERTa-wwm-ext 采用全词掩码（whole word masking）的方式，通过对输入文本中的词进行整体掩码的方式进行训练，使得模型能够更好地处理不同粒度的中文任务（表 1）。RoBERTa-wwm-ext 模型已在多项自然语言处理任务中表现优异[10]，是目前在中文场景下最强大、应用最广泛的语言模型之一。

表 1　全词掩码

Tab. 1　Whole word masking

说明	样例
原始文本	使用语言模型来预测下一个词的 probability。
分词文本	使用 语言 模型 来 预测 下 一个 词 的 probability 。
原始 Mask 输入	使 用 语 言 [MASK] 型 来 [MASK] 测 下 一 个 词 的 pro [MASK] ##lity 。
全词 Mask 输入	使 用 语 言 [MASK] [MASK] 来 [MASK] [MASK] 下 一 个 词 的 [MASK] [MASK] [MASK] 。

向量表示层的输入是一个 token 序列，该序列是由语料库中的文本句子生成的。在将数据送入向量表示层之前，需要先根据指定的最大序列长度对它们进行填充处理。接着将序列转换为向量形式，这些向量将作为输入传递给向量表示层。本研究使用的数据集是 CCKS2019 中文电子病历数据集。对于给定的一组电子病历纯文本，将这一组数据中的每条电子病历纯文本通过预训练模型 RoBERTa-wwm-ext 得到对应的向量表示，如式（1）—式（3）所示。

$$\boldsymbol{D}=\{t_1,\cdots,t_N\} \tag{1}$$

$$\boldsymbol{t}_i=\{c_{i1},\cdots,c_{iM}\} \tag{2}$$

$$\boldsymbol{c}_{ij}=\{e_{ij1},\cdots,e_{ijW}\} \tag{3}$$

其中：$\boldsymbol{D}$ 是 N 维向量，表示包含 N 条电子病历纯文本的一组数据；$\boldsymbol{t}_i$ 是 M 维向量，表示 $\boldsymbol{D}$ 中的第 i 条电子病历文本，这里 M 为第 i 条电子病历文本经过填充或截断处理后的长度；$\boldsymbol{c}_{ij}$ 是 W 维向量，表示第 i 条电子病历文本中第 j 个字的向量表示。通过向量表示层得到的向量表示将会传入特征提取层作为输入。

1.2　特征提取层

目前常用的特征提取层编码器大多是序列模型，例如循环神经网络（RNN）。RNN 可以建立序列数据之间的联系，能够将序列的当前状态与先前状态相结合，从而有效地预测下一个状态。虽然 RNN 能够处理序列数据，但容易产生梯度爆炸和梯度消失的问题，不能很好地学习长距离依赖的信息。为了解决上述问题，Hochreiter 等[11]提出了 BiLSTM，BiLSTM 结构的门控机制在很大程度上克服了梯度爆炸和梯度消失以及长期依赖的问题，可以轻松地处理长序列数据，并且能够捕捉到序列中的上下文信息，从而在自然语言处理、语音识别等任务中取得了较好的成绩。

然而，近年来 BiGRU 逐渐受到越来越多的关注。它采用更加简洁的门控机制，具有更高的计算效率，并在一些任务中取得了与 BiLSTM 相当甚至更好的效果。在处理某些长序列任务时，由于双向循环结构和门控机制的作用，BiGRU 比 BiLSTM 收敛速度更快，并且能够更好地捕获序列中的关键信息。此外，BiGRU 处理稀疏序列数据的能力很强，通常需要的参数更少，训练时间更短。基于上述优点，选择 BiGRU 作为特征提取层的编码器。

BiGRU 是门控循环单元（GRU）的一种扩展形式，是由两个方向的 GRU 组合形成的，每个 GRU 由更新门和重置门组成，门结构可以选择保存上下文信息解决 RNN 梯度消失或梯度爆炸的问题。GRU 结

构如图 2 所示。

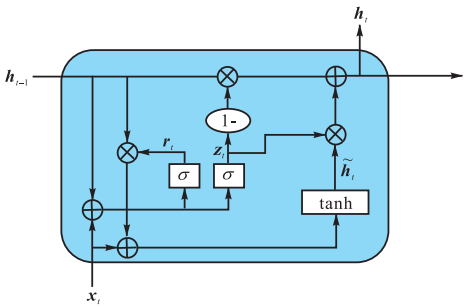

图 2　GRU 结构

Fig. 2　GRU structure

对于时间 t，GRU 单元状态计算公式为

$$\boldsymbol{r}_t=\sigma(\boldsymbol{\omega}_r\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_r) \tag{4}$$

$$\boldsymbol{z}_t=\sigma(\boldsymbol{\omega}_z\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_z) \tag{5}$$

$$\tilde{\boldsymbol{h}}_t=\tanh(\boldsymbol{\omega}_h\cdot[\boldsymbol{r}_t\times\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_h) \tag{6}$$

$$\boldsymbol{h}_t=(1-\boldsymbol{z}_t)\times\boldsymbol{h}_{t-1}+\boldsymbol{z}_t\times\tilde{\boldsymbol{h}}_t \tag{7}$$

其中：σ 表示 Sigmoid 函数，$\cdot$ 表示点积，$\boldsymbol{\omega}_r$、$\boldsymbol{\omega}_z$、$\boldsymbol{\omega}_h$ 为权重矩阵，$\boldsymbol{b}_r$、$\boldsymbol{b}_z$、$\boldsymbol{b}_h$ 为偏置参数。$\boldsymbol{x}_t$ 为时间步为 t 时的输入向量，$\boldsymbol{h}_t$ 为隐藏状态，也是输出向量，包含前 t 个时间的所有有效信息。$\boldsymbol{z}_t$ 为更新门，用于控制前一个单元隐藏层输出对当前单元状态的影响。更新门值越大，前一个单元隐藏层输出对当前单元状态的影响越大。$\boldsymbol{r}_t$ 为重置门，用于控制 $\boldsymbol{h}_{t-1}$ 对 $\tilde{\boldsymbol{h}}_t$ 的重要性。重置门的值越小，忽略前一个单元隐藏层信息的程度就越大。$\tilde{\boldsymbol{h}}_t$ 为在当前单元中需要更新的信息。两个门都捕获序列长度依赖性。在性能等价于 LSTM 时，GRU 的结构比 LSTM 简单，训练速度更快。

本研究采用的 BiGRU 网络由前向 GRU 和后向 GRU 组成。前向 GRU 的隐藏层表示为 $\vec{\boldsymbol{h}}_t$，后向 GRU 的隐藏层表示为 $\overleftarrow{\boldsymbol{h}}_t$。通过式(4)—式(7)得到单向 GRU 在时间 t 的隐藏层输出，如式(6)—式(7)所示。BiGRU 在时间 t 的隐藏状态输出，通过前向 GRU 单元和后向 GRU 单元的隐藏层输出拼接，计算公式为

$$\vec{\boldsymbol{h}}_t=\mathrm{GRU}(\boldsymbol{x}_t,\vec{\boldsymbol{h}}_{t-1}) \tag{8}$$

$$\overleftarrow{\boldsymbol{h}}_t=\mathrm{GRU}(\boldsymbol{x}_t,\overleftarrow{\boldsymbol{h}}_{t-1}) \tag{9}$$

$$\boldsymbol{h}_t=[\vec{\boldsymbol{h}}_t,\overleftarrow{\boldsymbol{h}}_t] \tag{10}$$

1.3　命名实体识别与分词联合训练

为了使模型能够在有限的数据场景下得到更好的识别效果，需要模型能够尽可能多地利用有限数据中的信息。本文提出了一种将 NER 任务和中文分词(Chinese word segmentation，CWS)任务进行联合训练的多任务学习框架。在该方法中，将特征提取层输出的向量表示分别送入一个分词模块和一个命名实体识别模块进行处理。其中，分词模块的主要作用是提供更准确的分词结果，使模型能够更好地学习到中文的边界信息，以帮助命名实体识别模块准确识别命名实体的边界。

1.3.1　CRF

NER 任务的解码器中最常用的是 Softmax 和 CRF，但 Softmax 不能考虑全局的标签之间的约束信息，因此在本文方法的 CWS 模块和 NER 模块使用了 CRF 而非 Softmax 进行解码。

给定一个输入序列 $X=(x_1,x_2,\cdots,x_n)$ 和对应的标签 $Y=(y_1,y_2,\cdots,y_n)$，通过式(11)计算标签序列的得分。

$$S(X,Y)=\sum_{i=1}^{n}\boldsymbol{P}_{i,y_i}+\sum_{j=0}^{n+1}\boldsymbol{A}_{y_j,y_{j+1}} \tag{11}$$

其中：$\boldsymbol{A}$ 表示转移得分矩阵，$\boldsymbol{A}_{y_j,y_{j+1}}$ 表示从标签 y_i 转移到标签 y_{j+1} 的转移得分。y_0 表示句子开始标签，y_{n+1} 表示句子终止标签，k 表示标签种类数，$\boldsymbol{A}\in R^{(k+2)\times(k+2)}$。输出层的得分矩阵为 $\boldsymbol{P}\in R^{n\times k}$，矩阵元素 P_{i,y_i} 表示第 i 个词在第 y_i 个标签下输出的得分。

1.3.2　多头注意力

本文模型在特征提取层使用了 BiGRU，目的是捕捉句子序列的上下文特征。然而，句子中每个字符的语义信息对 NER 任务的贡献不同。文本中有大量的无用信息，导致信息冗余。BiGRU 模型很难从句子序列中捕获重要信息。因此，在 BiGRU 网络捕获上下文特征后，使用了多头注意力(multi-head attention，MHA)[12]进一步捕获重要信息。它通过为不同位置的输入信息分配不同的权重，使模型能够更关注重要的信息。多头注意力的计算如图 3 所示。

多头注意力将注意力矩阵拆分为多个头，每个注意力的计算公式为

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{12}$$

其中：$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 为输入的查询、键、值矩阵，它们通常表示为 3 个不同的线性变换，d_k 为键矩阵的维度。多头自注意力分别计算每个头(head，用符号 h 表示)的注意力，最后再将每个头的输出进行拼接。计算公式为

$$\mathrm{MHA}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Concat}(h_1,\cdots,h_n)\boldsymbol{W}^{\mathrm{O}} \tag{13}$$

其中：$h_i=\mathrm{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^{Q},\boldsymbol{K}\boldsymbol{W}_i^{K},\boldsymbol{V}\boldsymbol{W}_i^{V})$ 是第 i 个注意

力头，Concat 是拼接操作，$\boldsymbol{W}^{O}$ 是最终输出的权重矩阵。通过使用多头注意力机制，模型能够同时关注输入序列的不同部分，从而可以更准确地捕捉序列之间的依赖关系。

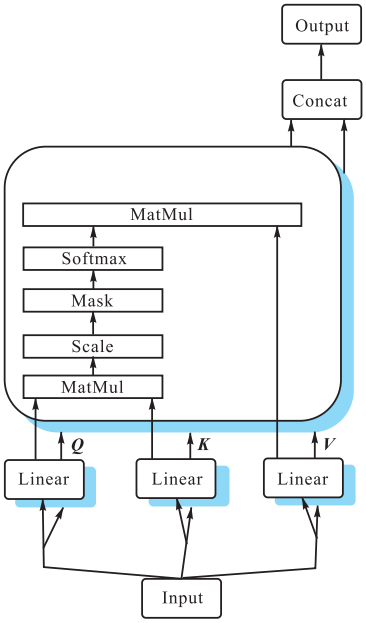

图 3　多头注意力结构

Fig. 3　Structure of multi-head attention

1.3.3　联合训练

本文提出了一种 NER 与 CWS 联合训练的框架，通过采用损失函数加权的方法进行联合训练，同时兼顾了分词模块和命名实体识别模块的性能表现，以达到最优的多任务学习效果。

对于分词任务，将其视作序列标注任务。分词模块接收特征提取层的输出作为输入，使用 CRF 作为解码器，得到分词的标签序列。

对于实体识别任务，也将其视作序列标注任务。实体识别模块首先接收特征提取层的输出 $\boldsymbol{O}_{\mathrm{BiGRU}}$ 作为输入，使用多头注意力机制获得包含注意力的向量表示 $\boldsymbol{O}_{\mathrm{MHA}}$，再将 $\boldsymbol{O}_{\mathrm{MHA}}$ 与实体识别模块的输入 $\boldsymbol{O}_{\mathrm{BiGRU}}$ 进行拼接，得到包含混合语义的向量表示，最后将混合语义向量送入 CRF 中进行解码，得到实体识别模块输出标签序列。

联合训练的损失函数为

$$L=(1-\lambda)L_{\mathrm{NER}}+\lambda L_{\mathrm{CWS}} \tag{14}$$

其中：L_{NER} 和 L_{CWS} 分别为 NER 模块和 CWS 模块的损失，L 为整个模型的总损失。$\lambda\in[0,1)$ 是控制 CWS 模块和 NER 模块比重的系数，当 λ 为 0 时，模型的损失为 NER 模块的损失，CWS 模块短路。

1.4　基于语义搜索的 NER 数据增强

为了增强模型在有限训练数据情况下的性能，提出了一种创新的利用语义搜索进行数据增强的方法。语义搜索通过理解搜索查询的内容提高搜索的准确性，被大量应用在搜索引擎上。与传统的搜索引擎只搜索基于词汇匹配的文档不同，语义搜索还可以搜索同义词。基于这个特性，利用语义搜索从清华大学整理的医学词典中搜索出与当前训练集中语义最接近的实体集合，使用搜索出来的实体集合替换原始训练数据集中的实体，从而达到有效数据增强的目的。与常规的数据增强方法[13]相比，本文的数据增强方法能够更有效地将关注点集中在实体上，而不是在一些对实体识别影响较小的字或词上进行过多关注。

2　实验与分析

2.1　数据集

使用 CCKS2019 中文电子病历数据集进行实验，训练集为 600 条数据，验证集和测试集分别为 200 条数据；包含 6 类实体，分别为疾病和诊断、检查、检验、手术、药物、解剖部位。THUOCL 是由清华大学自然语言处理与社会人文计算实验室整理推出的一套高质量的中文词库，词表来自主流网站的社会标签、搜索热词、医学词库、输入法词库等。在数据增强部分，将 THUOCL 中的医学词典作为外部词典进行语义搜索。

2.2　数据标注

由于本研究使用了多任务联合训练，因此需要针对不同的任务将数据处理成对应的格式。针对 NER 任务，采用 BIO 标记法标注每个字对应的标签。数据中的每个字将被标注成“B-X”“I-X”或“O”中的一个，其中 B 代表当前字是一个实体的开始，I 代表当前字是一个实体的中间部分，O 代表当前字不是实体的一部分，X 代表当前实体的类型。针对 CWS 任务，采用 BI 标记法标注每个字对应的标签。数据中的每个字将会被标注成“B”或“I”中的一个，其中 B 代表当前字是分词的开始，I 代表当前字是分词的一部分。NER 任务的标签是训练数据已经给定的，CWS 任务的标签通过百度的 ERNIE（enhanced representation through knowledge integration）3.0[14]进行标注得到。

2.3　评价指标

使用 F_1 分数作为评价指标。F_1 分数综合了模

型的精确率和召回率，可以更好地评估模型的综合性能。

精确率（precision，用符号 P 表示）表示模型预测的正例中真实正例的比例，公式为

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

召回率（recall，用符号 R 表示）表示真实正例中被模型预测为正例的比例，公式为

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

其中：N_{TP} 表示模型预测为正例且实际为正例的样本数，N_{FP} 表示模型预测为正例但实际为负例的样本数，N_{FN} 表示模型预测为负例但实际为正例的样本数。

F_1 分数综合了精确率和召回率，是精确率和召回率的调和平均数。公式为

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

2.4　实验结果

为了验证模型的有效性，使用本文模型与主流 NER 模型进行了对比，各个模型在 CCKS2019 中文电子病历数据集上的精确率、召回率、F_1 分数见表 2，本文模型 RBAC 在每个实体类别上的 F_1 分数见表 3。

表 2　各个模型在 CCKS2019 中文电子病历数据集上的精确率、召回率和 F_1 分数

Tab. 2　Precision, recall and F_1 score of various models on CCKS2019 Chinese electronic medical record datasets

模型	精确率/%	召回率/%	F_1/%
BiLSTM + Softmax[15]	87.2	80.6	83.7
BiLSTM + CRF	88.3	80.8	84.3
BERT + CRF	85.7	79.2	82.2
BERT + BiLSTM + CRF	88.4	83.9	86.0
BERT + BiGRU + CRF	88.8	84.4	86.4
RoBERTa + BiLSM + CRF	88.4	84.4	86.3
RoBERTa + BiGRU + CRF	88.5	84.8	86.5
TemplateNER[16]	89.1	85.8	87.4
RBAC	89.3	86.1	87.5
LightNER[17]	90.2	87.5	88.8
RBAC + DA_2	90.5	87.6	89.0
EntLM[18]	91.3	87.4	89.3
RBAC + DA_4	92.7	88.4	90.5
RBAC + DA_5	91.1	87.3	89.2

注：MHA 指本文模型中的多头注意力，RBAC 为本文模型，DA_i 代表使用基于语义搜索的数据增强将数据增强到原数据的 i 倍。

在实验中，本文模型在该数据集上的精确率、召回率和 F_1 均取得了最优的结果，其中 F_1 分数高达 90.5%，显著优于当前主流模型。

表 3　RBAC 模型在每个实体类别上的 F_1 分数

Tab. 3　F_1 score of RBAC model on each entity category

实体	F_1/%			
	RBAC	RBAC + DA_2	RBAC + DA_4	RBAC + DA_5
Disease	85.43	87.46	89.17	88.45
Surgery	90.88	89.28	90.68	89.61
Medicine	95.32	95.44	96.25	95.21
Anatomy	84.65	85.35	86.06	85.27
Check	89.23	90.67	91.34	90.34
Inspection	89.18	89.03	90.89	90.32

注：Disease 为疾病和诊断，Surgery 为手术，Medicine 为药物，Anatomy 为解剖部位，Check 为检查，Inspection 为检验；RBAC 为本文模型，DA_i 代表使用基于语义搜索的数据增强将数据增强到原数据的 i 倍。

实验结果表明，使用基于语义搜索的数据增强非常有效。不同数据增强倍率下 RBAC 模型的实验结果如图 4 所示。当数据增强至原始数据的 4 倍时，训练出的模型已经达到了最优；当数据增强的倍数再继续增大时，模型效果开始下降。这可能是由于数据增强倍数过大导致新引入的数据噪声过多，从而影响了模型的效果。因此，在采用本文方法进行数据增强时，需要根据实际情况选择最合适的数据增强倍数。

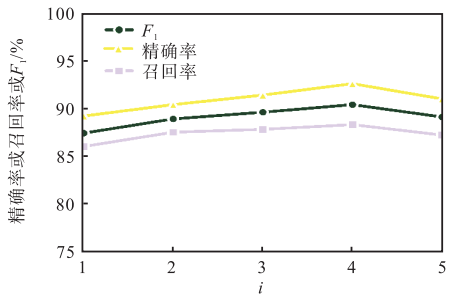

图 4　不同数据增强倍率下 RBAC 模型的实验结果

Fig. 4　Experimental results of RBAC model under different data enhancement factors

另一方面，实验结果还表明，本文方法在处理多样化的自然语言文本上具有很强的鲁棒性和可适应性。复杂实体的预测结果如图 5 所示。

sentence:，患者2015.11.27因“卵巢癌”于我院全麻上行经腹子宫+两侧附件切除+大网膜切除+部分肠壁转移病灶切除术。
entity_name:卵巢癌, entity_type:疾病和诊断, start_pos:15, end_pos:18
entity_name:经腹子宫+两侧附件切除+大网膜切除+部分肠壁转移病灶切除术, entity_type:手术, start_pos:26, end_pos:55

图 5　复杂实体的预测结果

Fig. 5　Prediction results for complex entities

本文模型对较长以及较复杂的实体也能精准识别。由于本文模型中使用了 BiGRU + MHA 的混合向量表示，因此模型综合了句子中的局部语义和全局语

义，从而能够精准地把长的、复杂的实体抽取出来。

综上所述，实验结果表明本文方法具有优异的性能和可泛化能力，为解决自然语言多样性和复杂性的问题提供了一种新的思路。

3 结　语

本研究设计了一套高效利用训练数据的联合训练框架，包含向量表示层 RoBERTa-wwm-ext、特征提取层 BiGRU、辅助分词模块和实体识别模块，同时提出了一个新颖的 NER 数据增强方法，即利用语义搜索的方式从已有医疗词典中检索出与训练数据中的实体最接近的实体集，再对原始数据集中的实体进行替换，有效提升了模型在训练数据匮乏场景下中文医疗命名实体识别中的表现。实验结果表明，本文模型在医疗领域的命名实体识别任务上具有出色的性能和实用性。

本文模型没有考虑嵌套命名实体[19]的场景，因此对于某些嵌套实体识别的效果可能欠佳。因此，未来研究的一个重要方向是进一步探索嵌套命名实体识别的技术手段，以提高模型在中文医疗嵌套命名实体识别任务中的性能和泛化能力。

参考文献：

[1] KIM J H，WOODLAND P C. A rule-based named entity recognition system for speech input[EB/OL]. [2023-06-01]. http://www.isca-archive.org/icslp_2000/kim00_icslp.pdf.

[2] 张华平，刘群. 基于角色标注的中国人名自动识别研究[J]. 计算机学报，2004(1)：85-91.

[3] HUANG Z，XU W，YU K. Bidirectional LSTM-CRF models for sequence tagging[EB/OL]. [2023-06-01]. https://doi.org/10.48550/arXiv.1508.01991.

[4] DEVLIN J，CHANG M W，LEE K，et al. BERT：pre-training of deep bidirectional transformers for language understanding[EB/OL]. [2023-06-01]. https://doi.org/10.48550/arXiv.1810.04805.

[5] 谢腾，杨俊安，刘辉. 基于 BERT-BiLSTM-CRF 模型的中文实体识别[J]. 计算机系统应用，2020，29(7)：48-55.

[6] REIMERS N，GUREVYCH I. Sentence-BERT：sentence embeddings using Siamese BERT-Networks[EB/OL]. [2023-06-01]. http://dx.doi.org/10.18653/v1/d19-1410.

[7] HAN S Y，ZHANG Y H，MA Y S，et al. THUOCL：Tsinghua open Chinese lexicon 2016[EB/OL]. [2023-06-01]. http://thuocl.thunlp.org/.

[8] CUI Y，CHE W，LIU T，et al. Pre-training with whole word masking for Chinese BERT[C]//IEEE. IEEE/ACM Transactions on Audio，Speech，and Language Processing. New York：IEEE，2021：3504-3514.

[9] DENG J，CHENG L，WANG Z. Self-attention-based BiGRU and capsule network for named entity recognition[EB/OL]. [2023-06-01]. https://doi.org/10.48550/arXiv.2002.00735.

[10] XU Z. RoBERTa-wwm-ext Fine-Tuning for Chinese text classification[EB/OL]. [2023-06-01]. https://doi.org/10.48550/arXiv.2103.00492.

[11] HOCHREITER S，SCHMIDHUBER J. Long short-term memory[J]. Neural computation，1997，9(8)：1735-1780.

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

[13] LI B，HOU Y，CHE W. Data augmentation approaches in natural language processing：a survey[EB/OL]. [2023-06-01]. http://dx.doi.org/10.1016/j.aiopen.2022.03.001.

[14] SUN Y，SHUOHUAN W，FENG S，et al. ERNIE 3.0：large-scale knowledge enhanced pre-training for language understanding and generation[EB/OL]. [2023-06-01]. https://doi.org/10.48550/arXiv.2112.12731.

[15] CUI L，ZHANG Y. Hierarchically-refined label attention network for sequence labeling[EB/OL]. [2023-06-01]. http://dx.doi.org/10.18653/v1/d19-1422.

[16] CUI L，WU Y，LIU J，et al. Template-based named entity recognition using BART[EB/OL]. [2023-06-01]. https://doi.org/10.48550/arXiv.2106.01760.

[17] CHEN X，LI L，DENG S，et al. LightNER：a lightweight tuning paradigm for low-resource NER via pluggable prompting[EB/OL]. [2023-06-01]. https://doi.org/10.48550/arXiv.2109.00720.

[18] MA R，ZHOU X，GUI T，et al. Template-free prompt tuning for few-shot NER[EB/OL]. [2023-06-01]. https://doi.org/10.48550/arXiv.2109.13532.

[19] 余诗媛，郭淑明，黄瑞阳，等. 嵌套命名实体识别研究进展[J]. 计算机科学，2021，48(S2)：1-10.

责任编辑：郎婧