Modeling and Simulation 建模与仿真, 2024, 13(4), 4075-4089
Published Online July 2024 in Hans. https://www.hanspub.org/journal/mos
https://doi.org/10.12677/mos.2024.134370

Hans汉斯

ABS-HDL：基于BIASRU的中文医学命名实体识别模型

盛萱妍，邵　清

上海理工大学光电信息与计算机工程学院，上海

收稿日期：2024年6月3日；录用日期：2024年6月26日；发布日期：2024年7月3日

摘　要

中文医学命名实体识别旨在从中文非结构化医学文本中提取关键实体。针对模型训练时间长、传统字符向量处理方法容易忽视词边界等问题，提出了基于多头交互注意力的中文医学命名实体识别模型：ABS-HDL (ALBERT-BIASRU-SoftAttention-CRF Hybrid Deep Learning)。该方法首先使用ALBERT预训练模型分别获得词向量表示和字向量表示。其次，将字向量和词向量结合成一个字词向量矩阵。接着，本文提出了BIASRU语义提取层，通过将多头交互注意力融入到SRU中，实现了对字词向量矩阵特征的有效学习，并通过双向建模精确捕获序列上下文间的关系。此外，在软注意力机制权重分配层中，模型能够动态调整权重分配，增强了对实体边界的识别能力。最后，使用CRF解码层来优化标签序列的预测。实验结果表明，该模型在中文糖尿病数据集上与现有模型相比表现更好。

关键词

命名实体识别，ALBERT，简单循环单元，多头交互注意力，软注意力

ABS-HDL: Chinese Medical Named Entity Recognition Model Based on BIASRU

Xuanyan Sheng, Qing Shao

School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai

Received: Jun. 3rd, 2024; accepted: Jun. 26th, 2024; published: Jul. 3rd, 2024

Abstract

Chinese medical named entity recognition aims to extract key entities from unstructured Chinese

文章引用：盛萱妍，邵清．ABS-HDL：基于 BIASRU 的中文医学命名实体识别模型[J]．建模与仿真，2024，13(4)：4075-4089. DOI: 10.12677/mos.2024.134370

medical texts. Addressing issues such as the lengthy training time for models and the traditional character vector methods' tendency to overlook word boundaries, a Chinese medical named entity recognition model based on multi-head interactive attention is proposed: ABS-HDL (ALBERT-BIASRU-SoftAttention-CRF Hybrid Deep Learning). This method initially employs the ALBERT pre-trained model to obtain separate word vector and character vector representations. Subsequently, it combines these vectors into a unified character-word vector matrix. Furthermore, this paper introduces the BIASRU semantic extraction layer, which integrates multi-head interactive attention into the SRU, effectively learning the features of the character-word vector matrix and precisely capturing the relationships within the sequence context through bidirectional modeling. Moreover, in the soft attention mechanism weight allocation layer, the model dynamically adjusts the distribution of weights, enhancing the ability to recognize entity boundaries. Lastly, a CRF decoding layer is used to optimize the prediction of the label sequence. Experimental results demonstrate that this model performs better on a Chinese diabetes dataset compared to existing models.

Keywords

Named Entity Recognition, ALBERT, Simple Recurrent Unit, SoftAttention Mechanism

1. 引言

中文医学命名实体识别(Medical Named Entity Recognition, MNER) [1]旨在从临床文本或医学文献中准确提取出与医学相关的实体信息，包括疾病名称、药物名称、人体部位等。这一领域的研究和应用面临众多挑战，特别是医学文本中大量存在的专业术语和缩写，以及中文医学术语的多义性问题[2]，这些都大大增加了命名实体识别的难度。

鉴于中文医学实体结构的复杂性及其识别的高难度，近年来，众多学者开始探索先进技术在医学实体识别领域的应用前景[3]。命名实体识别技术随时间演进，主要经历了基于规则的方法、基于统计模型的方法以及基于深度学习的方法三个发展阶段。随着自然语言处理技术的不断进步，深度学习方法已逐步成为该领域的研究重心。特别是面对中文医学文本的多义性特征，王颖洁等人[4]的研究表明，深度学习模型能够有效学习上下文中的语义关联，这使得它们在处理医学长文本的命名实体识别任务上，相比于传统机器学习方法展现出了更加出色的适用性和效果。

近年来，大规模预训练模型被广泛应用[5]，Kalyan K S 等人[6]对基于 Transformer 架构的多种医学文本预训练模型进行了综合对比，包括 BERT、BioBERT、RoBERTa 等，其研究验证了这些预训练模型在医学自然语言处理任务中的卓越性能。特别是最近推出的 ChineseBERT [7]，作为 BERT 的一种优化版本，它通过在字符级别上进行分析，并将汉字的形态学特征与拼音特征结合起来，显示出在多项自然语言处理任务中的优异表现。

目前数据预训练的策略主要分为两类：粗粒度分析与细粒度分析。粗粒度分析主要将词汇转化为词向量，这些词向量捕捉了整个词汇的语义及其上下文信息。对于中文医学文本中频繁出现的药品名称、疾病名称等，词向量因其能够捕捉到较完整的语义信息而更为适用。例如：“肺炎”既可以指疾病本身，也可以指由于疾病引起的症状。词向量能够提供更丰富的上下文语义信息，解决上述问题。Ramos-Vargas R E 等[8]比较分析了 ELMo、Pooled Flair 和 Transformer 等在 BioNER 中的表现。结果表明，这些通用词

向量训练模型在医学命名实体识别任务中表现良好。Nath N 等[9]提出一种基于词向量的无监督方法，使用无需标注的临床文本，进行自动临床编码。与传统的实体识别模型相比，所提出的模型能够准确有效地提取实体。

相对而言，细粒度分析更侧重于将每个字符映射为字向量，通过这种方式，能够表示更细节的语言特性，如拼音、字形等。LEI S 等[10]提出了一种结合多特征嵌入和多网络融合模型(MFE-MNF)该模型嵌入了多粒度特征，即字符、词、偏旁部首和外部知识，扩展了字符的特征表示并定义了实体边界。Ding J 等[11]提出了一种针对汽车评论的中文 NER(JMCE-CNER)的联合多视角字符嵌入模型，从发音、偏旁部首和字形等多个视图提取更深层次的字符特征，生成多视图字符嵌入。实验结果表明，该模型效果较好。但是中文实体没有明确的词边界，学习字符级别的向量，虽然能嵌入中文特征，但是容易忽视词边界。

近年来，注意力方法逐渐被应用于实体识别模型中，这些方法可以有效提高中文实体识别模型的性能。Li Y 等人[12]，使用空间注意力弥补 BiLSTM 在有效特征提取方面的不足，从而提升医学实体的识别效果；Li D [13]等将自注意力方法嵌入到 BERT 模型的 Transformer 计算框架中，更好地表达单词之间的长距离句法依存关系。於张闲等[14]在 BERT-BiLSTM 的基础上引入软注意力机制，通过软注意力机制计算每个词的权重，突出实体边界值，减少无关词的干扰，从而提高识别精度。近期不少学者开始使用 FLAT 作为词汇增强方法，将字向量与字向量同时建模，可以使每个字符与任意潜在词语交互，例如：“血”的自包含词语有“验血”、“血糖”等。谢靖等[15]以 Flat-lattice Transformer (FLAT)结构为微调模型，验证了 FLAT 在中医专业领域中文献命名实体识别工作上的有效性。但是这种词汇增强方法在训练大规模文本时显著增加了计算成本与内存消耗。

综合上述研究可以看出，虽然深度学习模型在中文医学实体识别领域已经得到了广泛应用，但仍面临着一些不可忽视的挑战。主要问题包括：① 在学习字符级向量时，可能会导致语义信息和边界信息的丢失；② 在学习词级向量时，有可能会忽略中文组合词。③ 传统的字词联合学习方法计算成本较高。为了克服这些难题，本研究提出了一种基于 BIASRU 的中文医学实体识别模型，其创新之处和主要贡献概述如下：

(1) 针对 FLAT 联合字词建模计算成本过大的问题，利用轻量级预训练模型 ALBERT，生成嵌入层字向量与词向量。针对字符向量容易忽视词边界、词向量容易忽视实体嵌套的缺点，通过 Lexicon Matching 技术，连接不同长度字符向量和词向量，使模型实现字词联合建模。

(2) 针对字符向量易于忽略词边界、词向量易于忽略实体嵌套结构的问题，将 Inter-Attention 内嵌至 SRU 中，联合模型两个不同长度的字符和词序列。使字符序列能够融合词边界和语义信息。并且搭建双向循环的模型提高长文本处理能力。该创新模块称为 BIASRU (Bi-directional Simple Recurrent Unit with Built Inter-Attention, BiASRU)。

2. 模型设计

2.1. 模型结构

基于 BIASRU 中文医学实体识别模型结构如图 1 所示。包括 ALBERT 预训练层、字词向量融合层、BiASRU 语义提取层、软注意力机制权重分配层、CRF 解码层六个模块。

首先，ALBERT 预训练层通过无监督方式提取丰富的语义特征，分别获得字向量表示与词向量表示；其次，将两个向量序列通过融合层输出向量矩阵；然后，BiASRU++语义提取层将字符特征与词典信息融合、学习上下文特征；接着，软注意力机制权重分配层通过关注全局上下文信息进行实体重要边界特征的捕获；最后，CRF 解码层对软注意力机制权重分配层的输出进行邻近标签关系预测，获得一个最佳预测序列，完成中文医学实体识别任务。

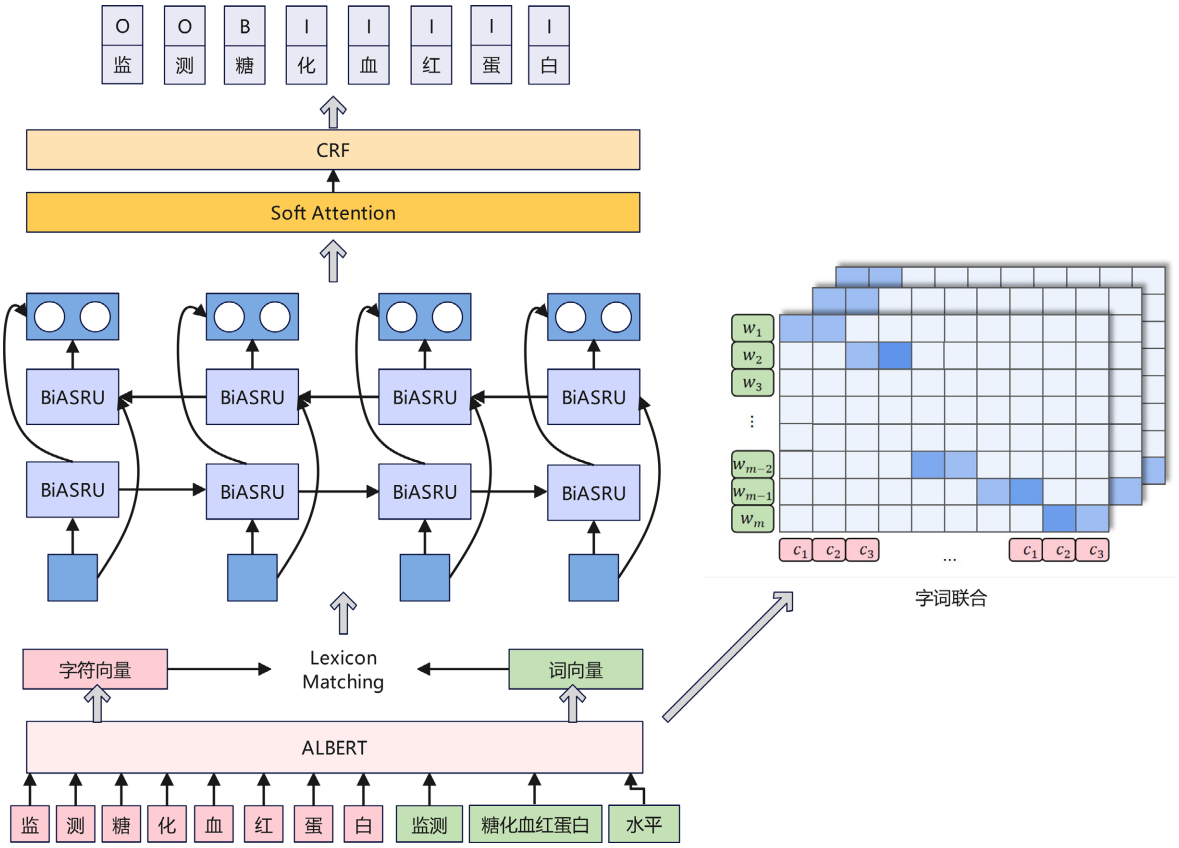


Figure 1. Overall structure of model
图 1. 模型整体架构

2.2. ALBERT 预训练层

ALBERT [16]是一个比传统 BERT 架构小得多的轻量级 BERT 架构，致力于提高模型的效率和性能，通过采用嵌入参数分解(Factorized Embedding Parameterization)和跨层参数共享(Cross-layer Parameter Sharing)参数共享的方法，在显著减少 BERT 架构参数数量的同时，又不严重损害模型性能。

在本文实验中使用跨层参数共享，从而减少了整体的参数数量。同时保持模型的表现。参数共享过程如式(1)所示：

$$h_1 = TransformerLayer\left(h_{i-1}, E\right) \tag{1}$$

其中，E 是需要共享的参数，h 表示每层输出向量。通过跨层参数共享，P 参数可以被多个层共享使用，从而减少总体参数数量。

经动态学习文本向量表示后，得到字序列 $C=\left(C_0, C_1, C_2, \cdots, C_n\right)$ 作为字向量的输出，同时训练词向量，得到词序列 $W=\left(W_0, W_1, W_2, \cdots, W_m\right)$ 序列。然后将两个序列联合输入到字词向量融合层，经词汇匹配(Lexicon Matching)输出一个二维向量矩阵 V。字词融合过程如图 2 所示。

经字词融合后，ALBERT 预训练层输出三个序列，分别为字符向量序列 C、词向量序列 W 和二维向量矩阵 V，由 BIASRU 进行二次语义特征捕捉。

2.3. BIASRU 语义提取层

本文在传统简单循环单元 SRU (Simple Recurrent Unit, SRU) [17]的基础上，提出了一种改进模块—

BIASRU。SRU 作为 LSTM 的一种轻量级替代品，其设计摒弃了对上一时间步状态输出的严格依赖，从而在保持高效建模能力的同时，具备更快的并行处理能力。这一特性不仅提升了模型在处理复杂文本时的性能，而且显著缩短了模型的训练时间。SRU 结构如图 3 所示。

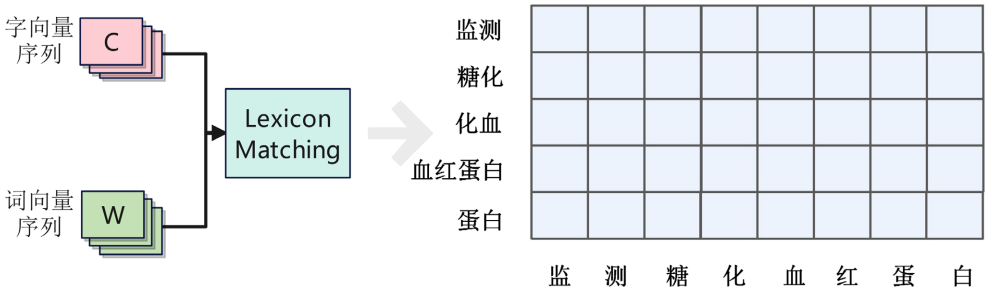


Figure 2. Character-word vector fusion
图 2. 字词向量融合

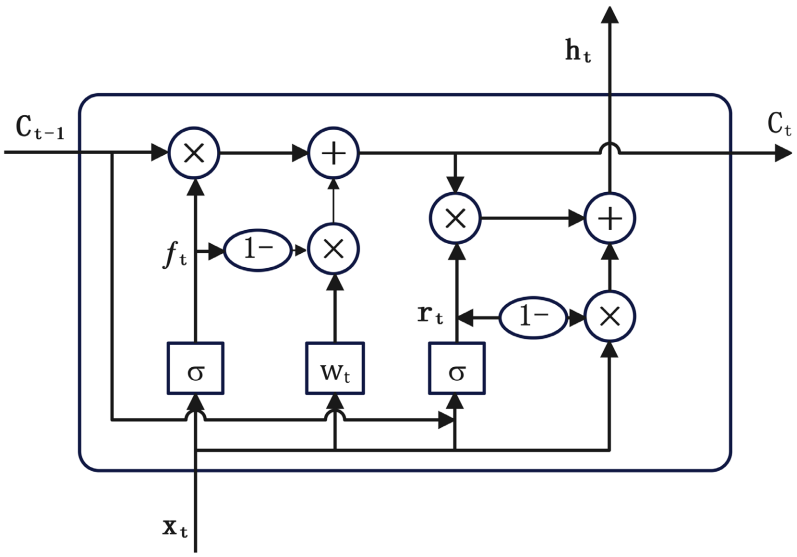


Figure 3. Structure of SRU
图 3. SRU 模型结构

SRU 计算过程如式(1)~(4)所示：

$$f_t = \sigma\left(W_f x_t + v_f \odot c_{t-1} + b_f\right) \tag{2}$$

$$r_t = \sigma\left(W_r x_t + v_r \odot c_{t-1} + b_r\right) \tag{3}$$

$$c_t = f_t \odot c_{t-1} + \left(1 - f_t\right) \odot \left(W x_t\right) \tag{4}$$

$$c_t = f_t \odot c_{t-1} + \left(1 - f_t\right) \odot \left(W x_t\right) \tag{5}$$

其中 σ 表示 Sigmoid 激活函数；$\odot$ 代表矩阵元素乘法；W_f、W、W_r、v_f、v_r 为可学习权重矩阵；b_f、b_r 为可学习权重值。由式(4)可知，SRU 不再依赖上一个时刻的输出 h，可实现并行化处理。

本文针对 SRU 做出改进：

(1) 将多头交互注意力(Multi-Head Inter-Attention)内嵌至 SRU 中。

交互注意力可处理处理输入序列 Q 另一个相关序列 K 间的交互关系，本文将单一的交互注意力扩展

为并行的多个注意力机制，称为多头交互注意力，每个“头”都独立地关注输入的不同部分。这种方法可以捕获数据的多种表示和特征，提高模型的表达能力。相关架构如图 4 所示。

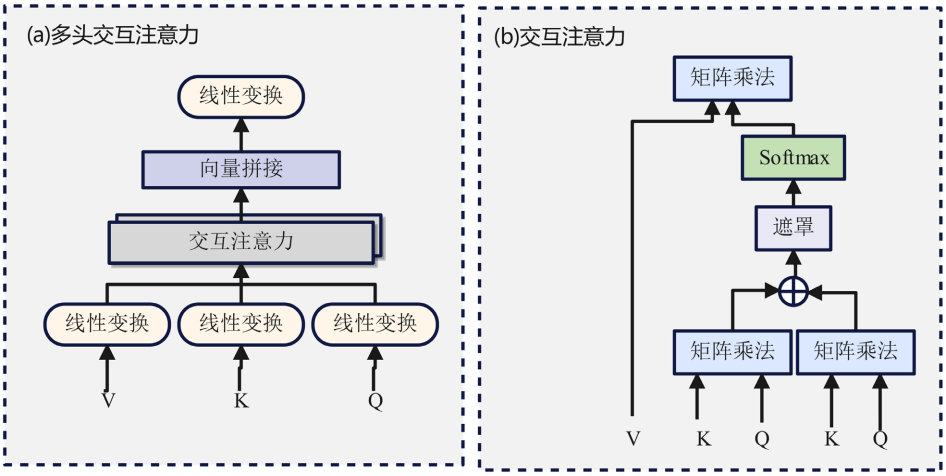


Figure 4. Multi-head Inter-Att
图 4. 多头交互注意力

左图为交互注意力结构，其中 Q 代表 ALBERT 输出的字符向量序列 C，K 代表 ALBERT 输出的词向量序列 W。Q 的长度为 q，K 的长度为 k。V 表示注意力权重矩阵。掩码机制用于字符和词的得分掩码，通过将序列中的空缺位置填充为 10^{-15}，确保在进行 Softmax 归一化时，这些位置的注意力权重接近于 0。交互注意力针对 Q 序列中的每一个位置生成一个表示，输出的 V 展示了 Q 受到 K 序列中各位置影响的程度。交互注意力计算过程如(6)~(8)所示：

$$[Q,K,V]=\left[X^{C}W_{q},X^{W}W_{k},X^{W}W_{v}\right] \tag{6}$$

$$InterAtt\left(A,V\right)=softmax\left(mask\left(A\right)\right)V \tag{7}$$

$$A_{ij}=\left(Q_{i}+u\right)^{T}K_{j}+\left(Q_{i}+v\right)^{T}R_{ij}^{*} \tag{8}$$

右图为多头交互注意力结构，它拓展了传统的单头注意力机制，通过使用多个注意力头来更好地捕捉不同子空间的特征表示。多头交互注意力的计算过程如式(9)~(10)所示：

$$Head^{(s)}=InterAtt\left(X^{C,(s)},X^{W,(s)}\right) \tag{9}$$

$$MultiHead\left(X^{C},X^{W}\right)=\left[Head^{(1)},\cdots,Head^{(l)}\right] \tag{10}$$

其中 l 是互注意力头的数量，$Head\left(s\right)$ 是第 s 个互注意力头在字符和词向量子空间上的输出结果。$X^{C,(s)}$，$X^{W,(s)}$ 分别是它们在各自子空间中的字符和词的向量表示。

接着将多头交互注意力内嵌至 SRU 结构中，提高模型的语义理解能力与泛化能力，模块称为 IASRU (Inter-Attention SRU)。其中序列循环计算过程，能够显著提高模型的并行处理能力，该过程如图 5 所示。

(2) 搭建双向循环的模型提高长文本处理能力。

考虑到文本数据的连续性，为了提高模型的语义提取能力，本文在 AISRU 基础上搭建了 BIASRU (Bidirectional ASRU, BiASRU)，通过结合一个正向的 AISRU 和一个反向的 AISRU，BIASRU 能够融入上下文信息，增强从序列中提取特征的能力。输出结果H_t是由正向 AISRU 和反向 AISRU 拼接而成。

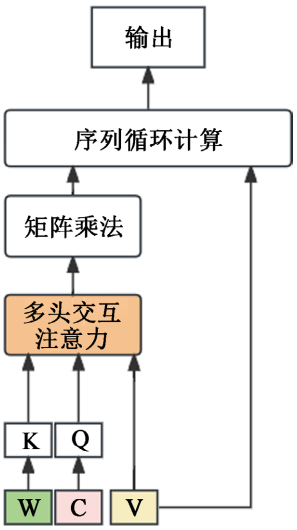


Figure 5. Structure of IASRU
图 5. IASRU 结构

图 6 展示了 BiASRU 的结构，其中正向层进行计算，并在每一步保存正向隐藏层的输出。然后在反向层进行计算，并在每一步保存隐藏层的输出。最后，将正向和反向层的输出合并以生成最终输出。

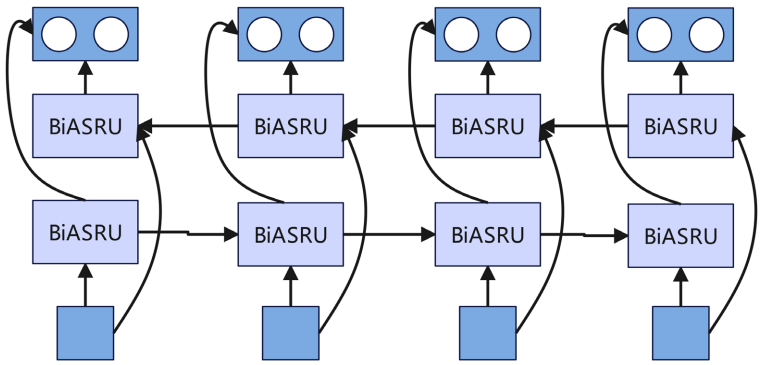


Figure 6. Structure of BIASRU model
图 6. BIASRU 模型结构

BiASRU 计算过程如式(11)所示：

$$H_t = \left[\overrightarrow{h_t}, \overleftarrow{h_t}\right] = \text{BIASRU}(C, W, V) \tag{11}$$

其中，C、W 为 ALBERT 层的输出向量序列、V 为字词向量融合层的输出矩阵，$\overrightarrow{h_t}$ 为 t 时刻 ASRU 正向输出，$\overleftarrow{h_t}$ 为 t 时刻 ASRU 反向输出，得到的输出 H 为结合了词信息的字符序列。

2.4. 软注意力机制权重分配层

BiASRU 层得到字符序列 H 作为软注意力权重分配层的输入，该层通过分配权重向量聚焦关键字符。例如，对于“微量白蛋白尿”这一术语，如果能够准确识别其作为医学实体的起始位置“微”和结束位置“尿”，将极大地提高实体识别的准确性。在这种情况下，起始和结束位置的字符权重相对更高，意味着它们更可能标识实体的边界。因此本文引入了软注意力机制，使模型在有效地处理长文本中的复杂信息的同时，能够精确捕捉实体关键字符，从而提高整体的实体识别精度。

软注意力机制权重分配层的结构，如图 7 所示。

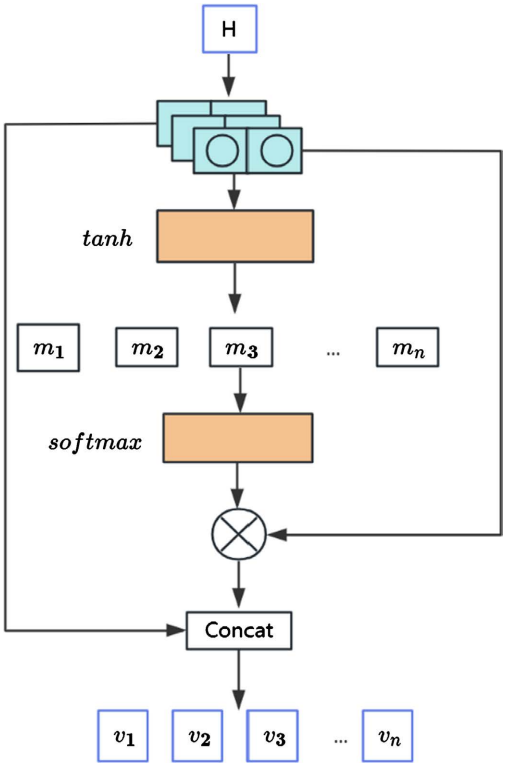


Figure 7. Structure of the Soft Attention
图 7. 软注意力机制权重分配层结构

将 BiASRU++输出的字符向量序列 H_i 综合起来，构成 $H = Concat\left(H_1, H_2, \cdots, H_n\right)$ 作为注意力层的输入，计算得出的权重与原始特征输出 H 求和后得到软注意力权重分配层输出 V'，计算过程如式(12)~(14)所示：

$$M = \tanh\left(W_A V + b_A\right) \tag{12}$$

$$\alpha = softmax\left(W^T M\right) \tag{13}$$

$$V' = H\alpha^T \tag{14}$$

式中，tanh 为非线性激活函数；W^T 和 W_A 表示可学习参数矩阵；b_A 为偏置值。α 矩阵表示句子中每个词的注意力得分。

2.5. CRF 解码层

理论上可以直接根据软注意力权重分配层的输出取最大标签得分进行序列标注，但若没有 CRF 解码层进行条件约束，可能会产生标注上的逻辑错误。CRF 解码层可以集成一些强制性的语言规则，例如特定类型的实体标签都以 B 开头，所以需要学习实体间的依赖关系来解码最优标签序列。

对于上一层输出 V' 对应的标签序列 $Y = \{y_1, y_2, \cdots, y_n\}$ 的预测得分 S 可由概率得分矩阵 P 和转移矩阵 A 相加计算得出：

$$S\left(V', y\right) = \sum_{i=1}^{n} P_{i,y_i} + \sum_{i=1}^{n+1} A_{y_i, y_{i+1}} \tag{15}$$

其中，n 是文本序列的字符长度，t 是预测的标签数，y_i 为第 i 个元素标签预测概率，A 是使用最大似然估计法来习得的概率转移矩阵参数。则标签序列 y 在 V 下的条件概率 P 为：

$$P\left(y|V'\right)=\frac{e^{S(V',y)}}{\sum_{\tilde{y}\in Y_X}S\left(V',\tilde{y}\right)} \tag{16}$$

其中，$\tilde{y}$ 表示真实标注序列。Y_X 表示所有可能的标签集合，接着为了最大化 $P\left(y|V\right)$ 定义损失函数：

$$\log\left(P\left(y|V'\right)\right)=S\left(V',y\right)-\log\left(\sum_{\tilde{y}\in Y_x}e^{S(V',\tilde{y})}\right) \tag{17}$$

接着解码具有最高概率的标签序列得到预测结果，具体如式(17)所示：

$$y^*=\arg\max_{\tilde{y}\in Y_X}S\left(V',\tilde{y}\right) \tag{18}$$

CRF 解码层对软注意力权重分配层的输出进行邻近标签关系预测，获得一个最佳预测序列 y*，完成中文医学实体识别任务。

3. 实验结果与分析

3.1. 数据集与评价标准

本研究中使用了阿里云天池实验室提供的中文糖尿病数据集作为训练和评估数据[18]。该数据集包含了大量的中文糖尿病相关文本样本，涵盖了疾病、药物、检查、症状等实体类型。中文糖尿病数据集包含 15 种医学实体类别。每个实体都用特定的标记进行了注释，以指示其所属的类别，15 种实体类型与划分规则如表 1 所示。

Table 1. Entity types and text segmentation
表 1. 实体类型与文本划分

	训练集	验证集	测试集
疾病相关			
疾病名称(Disease)	25,197	8399	8399
病因(Reason)	2849	950	950
临床表现(Symptom)	3166	1056	1055
检查方法(Test)	28,817	9607	9607
检查值(Test_value)	6400	2135	2135
治疗相关			
药品名称(Drug)	9944	3316	3316
频率(Frequency)	606	202	203
数量(Amount)	871	290	290
用药方法(Method)	604	201	202
治疗(Treatment)	894	299	230
手术(Operation)	493	164	164
副作用(Side Effect)	1050	352	351
常见实体			
持续时间(Duration)	6541	2182	2181

续表

部位(Anatomy)	16,864	5623	5623
程度(Leval)	1331	447	449
总数	105,348	35,116	35,119

本文实验将数据集进行数据预处理使实验数据更符合模型训练要求，删除数据集中多余的特殊字符、空格和标点符号，然后将数据集全部统一转换成 BIO 标注格式[19]。BIO 是基于词的标注方式，它将每个词进行标注，用于表示该词是否属于某个命名实体。其中 X 代表实体类别，B-X 表示实体 X 的开始，I-X 表示实体 X 的内部，O 表示非实体。

为了评估本文模型的有效性，使用了以下评价指标：

(1) 精确率(Precision)：精确率表示模型预测的正例中有多少是真正的正例。它的计算方式是将模型正确预测的正例数除以模型预测的正例总数。

(2) 召回率(Recall)：召回率衡量模型对真正正例的识别能力。它的计算方式是将模型正确预测的正例数除以真正的正例总数。

(3) F1 值：F1 值是精确率和召回率的调和平均值，综合了模型的准确性和完整性。它的计算方式是根据精确率和召回率计算得出。

3.2. 实验环境与参数选取

本研究的实验环境基于 PyTorch 1.10.0 深度学习框架搭建，全部实验均采用 Python 3.8.4 编程语言进行模型的构建与训练。实验运行在配备有两张 GTX 3090 显卡的服务器上，每张显卡具有 24 GB 的显存。服务器操作系统为 Ubuntu 20.04，配备 56 GB 机器内存和 8 核心的 Intel Xeon Processor CPU。在实验过程中，发现模块参数对于模型训练的精度有显著影响。模型规模和训练参数影响着模型的性能表现，经多次调参后得到最优参数，经多次调参后的最优参数如表 2 所示。

Table 2. Training parameter settings
表 2. 模型参数设置

参数名称	参数值
批处理大小	16
初始学习率大小	3e−5
训练轮次	200 次
预热步数	10
词向量维度	128
优化器	RAdam [20]

BIASRU 隐藏层大小设置为 256，注意力头的数量为 8，维度大小为 32，软注意力机制维度大小为 512。为防止模型过拟合，神经网络节点随机失活概率设置为 0.5。

3.3. 实验结果分析

本文共设计了三组实验验证本文模型在中文医学命名实体识别中的效果。

实验一：对比探究基线模型与本文模型的实验效果。

实验二：比较近期不同模型在中文糖尿病数据集下的实验效果。

实验三：通过热力图可视化表示软注意力对模型的影响。

3.3.1. 实验一：基线模型对比分析

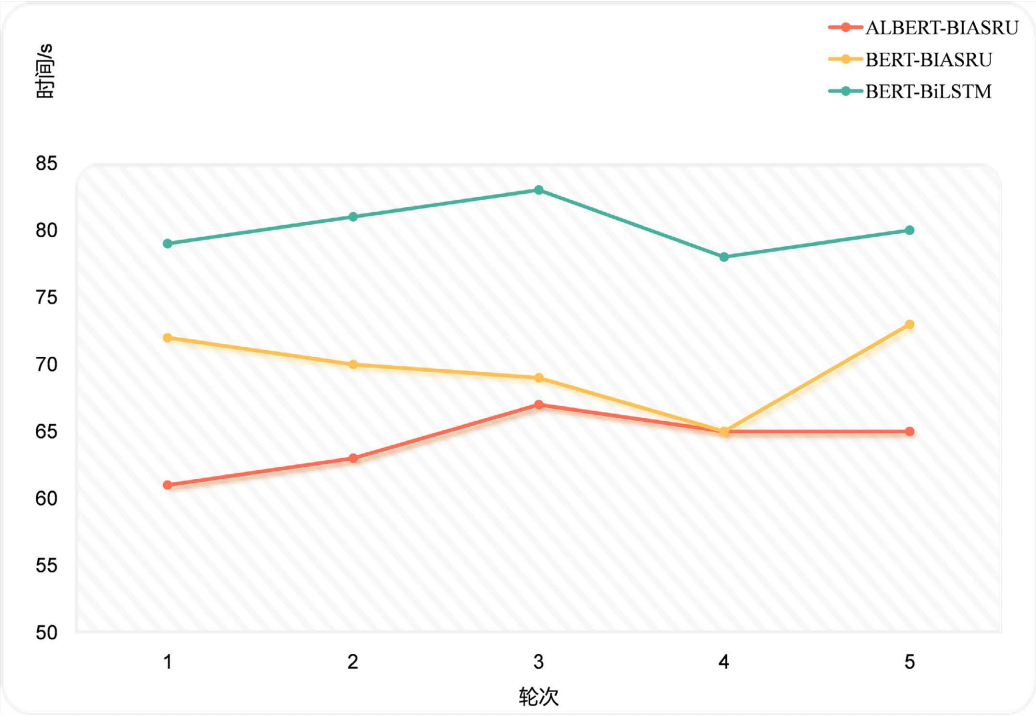
为证明本文模型的有效性，本实验将本文模型与当前主流方法 BERT-BiLSTM、BERT-BIASRU 进行比较。实验结果如表 3 所示，同时观察 BERT-BiLSTM、BERT-BiSRU 与本文模型在五个训练轮次中的时间消耗情况，具体表现如图 7 所示。

Table 3. Baseline model comparison
表 3. 基线模型对比

模型	F1 (%)	P (%)	R (%)
BERT-BiLSTM	90.87	90.69	90.72
BERT-BIASRU	91.75	91.54	91.89
ALBERT-BIASRU	92.83	92.56	92.65

注：ALBERT-BIASRU 为本文模型。

在本实验中，BERT 的输出维度被设定为 768。与此同时，本研究所提出的模型采用了 ALBERT 的参数共享机制，并将输出维度设置为 128。根据表 3 的结果，可以看出，相比于使用 BERT 的模型，采用 ALBERT 预训练的模型在 F1 分数上有了 0.98%的提升。这一结果表明，尽管 ALBERT 的参数数量少于 BERT，但其在性能上却能够实现更优的结果。



注：红线为本文模型。

Figure 8. Comparison of model training time
图 8. 模型训练时间对比

BERT-BIASRU 经过微调达到 91.75%的 F1 值，较 BERT-BiLSTM 的 F1 分数提高了 1.96%，说明了

BIASRU 模块的特征提取能力较优于 BiLSTM，这是因为 BIASRU 在 SRU 的基础上内嵌了多头交互注意力，为字符序列添加词向量的权重影响。双向建模的特点使模型同样能捕捉长距离特征。这些结果明确展示了 BIASRU 的强大功能和在处理复杂序列数据时的有效性。

图 8 显示在不同训练轮次中，BIASRU 的时间消耗普遍低于 BiLSTM。此外，采用了 ALBERT 预训练的模型时间损耗更低。这是因为相对于 LSTM，SRU 减少了复杂的门控机制，可以实现 GPU 并行计算，从而显著提升了训练速度并节省了时间成本。本文使用的 ALBERT 预训练层通过参数分解和跨层参数共享机制，结合 BIASRU 的并行计算能力，使得本文模型在时间消耗上达到了最低。

本文模型的 15 种实体识别结果如图 9 所示。

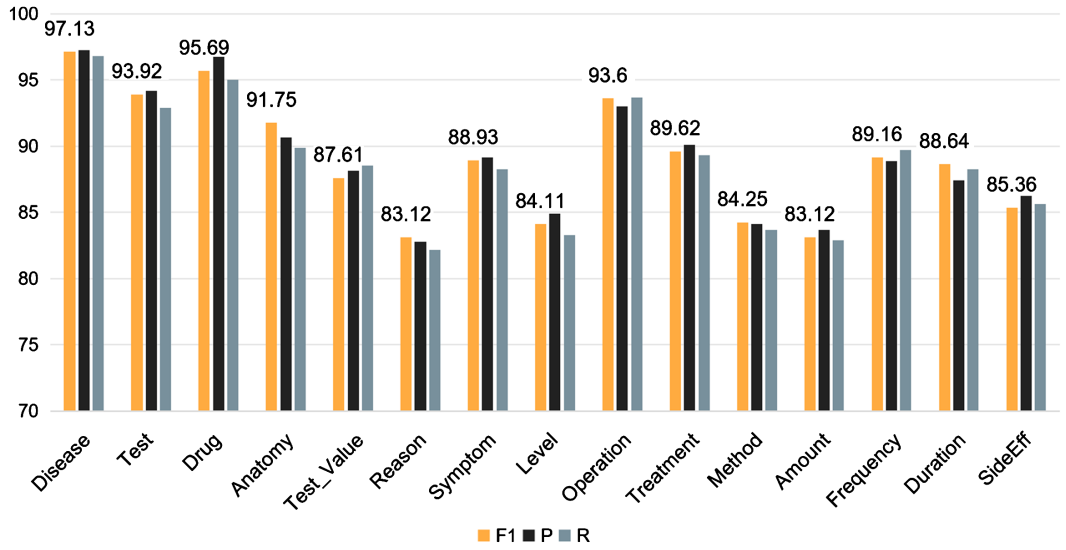


Figure 9. F1 values for 15 categories of entity
图 9. 15 类实体的 F1、P、R 值

医学文献类文本存在较多长句，长句的处理会增加模型的内存消耗，还导致训练过程中的梯度爆炸，从而影响模型的训练效率和效果。SRU 对门控机制进行了优化，控制信息的流动，减少了需要训练的参数量，同时保留了对长序列依赖的处理能力。同时 ALBERT 分解参数、参数共享的特点也在降低时间损耗的同时提高了训练效率。

根据图 9 所示，“症状”实体的 F1 值达到了 88.93%，这一结果表明，尽管医学文本中“症状”实体的识别准确率易受到“疾病”实体的影响，但本文提出的模型通过 BIASRU 的双向循环建模特性，有效提高了这两个实体之间的区分度，从而使“症状”实体的识别精度显著提升。“病因”实体的识别效率较低，经分析这可能是因为该实体在文本中的出现频率低，样本量少，导致数据不平衡，最终影响了模型的效果。后续可以通过同义词替换、句子重组等方式增加样本的多样性。

3.3.2. 实验二：性能对比分析

表 4 给出了各类模型的实验结果。

由表 4 可知，基于中文糖尿病文本训练的 RNN 模型，使用 Word2vec 来训练词向量表示，F1 值仅为 86.78%，但是由于 RNN 的递归性质，数据必须按序列顺序处理，因此医学类长文本的训练时间拉长。接着搭建 GRU-AT，使用 GRU 模型，然后由自注意力机制对 GRU 输出的隐状态进行权重计算，GRU-AT 模型的内部状态维度低，所以在捕捉复杂序列时表现不够出色，因此 F1 值仅为 88.52%。Fan Z [21]提出

的 BERT-BiDT-CRF 模型由两个 BiLSTM-CRF 模型和一个 CNN 模型组成。引入非目标场景数据集，提出句子级神经网络模型迁移学习。但由于该模型采用基础版本的 BERT 预训练，并不针对中文改进，F1 值为 85.88%。Wang Y 等[22]使用的 RoBERTa 模型：将单词替换成同义词实现数据扩充。这使模型一定程度上增强了泛化能力，该模型 F1 值达到了 91.27%，相较于基于 BERT-BiDT 的模型提升了 5.39%。韩普等[23]搭建的 BERT-BiLSTM-IDCNN-CRF 模型利用多任务学习构建粗粒度三分类任务以辅助实体识别任务，最后引入自注意力机制和 Highway 网络捕获全局重要信息并优化深层网络训练，达到了 92.28%的 F1 值。

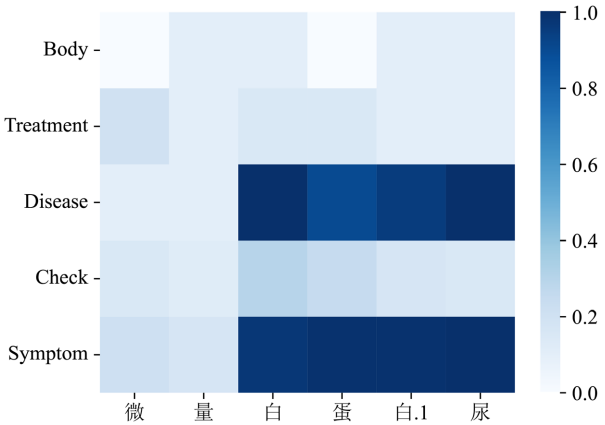
Table 4. Comparison of different model results
表 4. 各类模型对比结果

模型	F1 (%)	P (%)	R (%)
RNN	86.78	86.85	86.81
GRU-AT	88.52	88.54	88.65
BERT-BiDT-CRF [21]	85.88	86.10	85.65
RoBERTa [22]	91.27	91.18	91.36
BERT-BiLSTM-IDCNN-CRF [23]	92.28	92.28	92.56
ABS-HDL	92.83	92.56	92.65

然而，上述模型都是预训练词向量方法处理医学文本，对于中文医学类长文本来说，基于字向量的训练方法易忽视词边界，基于词向量的训练方法易忽视组合词。本文使用的 ALBERT 在预训练和微调中同时训练字向量和词向量，而 BIASRU 因为内嵌了多头交互注意力，可将词向量特征融合进字向量特征，生成一个多维向量矩阵，使模型在本文模型 F1 值达到 92.83%，在中文糖尿病数据集上达到了最优效果。

3.3.3. 实验三：软注意力机制可视化

中文糖尿病标注数据集中的文本存在很多长句，长句中非实体词比例较大，造成医学实体识别过程掺杂大量冗余噪声，而通过软注意力权重分配层，帮助模型聚焦关键位置，模型可以动态地分配权重，强调了边界位置的特征，减少中文医学文本的非实体词，降低因序列长度增加而导致的信息丢失。本节通过热力图的方式进一步说明软注意力机制对实体识别的影响。热力图直观展示了不同字符与实体类别之间的关联程度，其中颜色越深代表该字符与实体类别的联系越紧密，如图 10 所示。



注：“白.1”表示第 1 个与“白”重复字符。

Figure 10. Heatmap of SoftAttention weight distribution
图 10. 软注意力机制权重分配热力图

在图 10 中，以“微量白蛋白尿”为例，“白蛋白尿”对应实体类别“Disease”和“Symptom”，该实体被赋予了更大的关注度，并且“尿”的颜色更深，说明本文模型采用的软注意力机制对于和实体强相关的字形都赋予了更大的权重。软注意力权重分配层还整合了如边界特征的权重，以改善对实体边界的判别，能够更准确地识别出实体的开始和结束。

4. 结束语

针对传统方法训练时间长及字符向量处理中忽略词边界等问题，本文提出了基于 BIASRU 的中文医学实体识别模型，命名为 ABS-HDL。模型核心在于使用字词向量联合的方法，引入 ALBERT 预训练降低模型的时间消耗；还将多头交互注意力与 BiSRU 结合，给字符向量加入词向量边界特征；随后利用软注意力机制聚焦实体边界，进一步提高词边界处理能力；最终通过 CRF 解码层确保标注序列的合理性。实验结果表明，ABS-HDL 模型在中文糖尿病数据集上表现较好，实现了 92.83%的 F1 值，92.56%的准确率和 92.65%的召回率，充分证明了 ABS-HDL 在中文医学实体识别任务中的有效性。但本文的研究也存在不足，由于缺少部分实体训练样本，该实体准确率较低。今后会研究更有效的方法实现小样本的中文医学命名实体识别，进一步提升模型的准确率。

参考文献

[1] Pearson, C., Seliya, N. and Dave, R. (2021) Named Entity Recognition in Unstructured Medical Text Documents. https://doi.org/10.48550/arXiv.2110.15732

[2] Gao, Y., Wang, Y., Wang, P., *et al*. (2020) Medical Named Entity Extraction from Chinese Resident Admit Notes Using Character and Word Attention-Enhanced Neural Network. *International Journal of Environmental Research and Public Health*, **17**, Article 1614. https://doi.org/10.3390/ijerph17051614

[3] 郑强, 刘齐军, 王正华, 等. 生物医学命名实体识别的研究与进展[J]. 计算机应用研究, 2010, 27(3): 811-815. https://doi.org/110.3969/j.issn.1001-3695.2010.03.003

[4] 王颖洁, 张程烨, 白凤波, 等. 中文命名实体识别研究综述[J]. 计算机科学与探索, 2023, 17(2): 18. https://doi.org/110.3778/j.issn.1673-9418.2208028

[5] Kenton, J.D., Toutanova, M.W.C. and Bert, L.K. (2019) Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of naacL-HLT*, **1**, 2.

[6] Kalyan, K.S., Rajasekharan, A. and Sangeetha, S. (2021) AMMU—A Survey of Transformer-Based Biomedical Pretrained Language Models. DOI:10.48550/arXiv.2105.00827.

[7] Sun, Z., Li, X., Sun, X., *et al*. (2021) Chinesebert: Chinese Pretraining Enhanced by Glyph and Pinyin Information. arXiv preprint arXiv: 2106.16038. https://doi.org/10.18653/v1/2021.acl-long.161

[8] Ramos-Vargas, R.E., Román-Godínez, I. and Torres-Ramos, S. (2021) Comparing General and Specialized Word Embeddings for Biomedical Named Entity Recognition. *PeerJ Computer Science*, **7**, e384. https://doi.org/10.7717/peerj-cs.384

[9] Nath, N., Lee, S.H. and Lee, I. (2023) Application of Specialized Word Embeddings and Named Entity and Attribute Recognition to the Problem of Unsupervised Automated Clinical Coding. *Computers in Biology and Medicine*, **165**, Article ID: 107422. https://doi.org/10.1016/j.compbiomed.2023.107422

[10] Lei, S., Liu, B., Wang, Y., *et al*. (2023) Chinese Medical Named Entity Recognition Combined with Multi-Feature Embedding and Multi-Network Fusion. *Journal of Electronics and Information Technology*, **45**, 3032-3039.

[11] Ding, J., Xu, W., Wang, A., *et al*. (2023) Joint Multi-View Character Embedding Model for Named Entity Recognition of Chinese Car Reviews. *Neural Computing and Applications*, **35**, 14947-14962. https://doi.org/10.1007/s00521-023-08476-2

[12] Li, Y., Du, G., Xiang, Y., *et al*. (2020) Towards Chinese Clinical Named Entity Recognition by Dynamic Embedding Using Domain-Specific Knowledge. *Journal of Biomedical Informatics*, **106**, Article ID: 103435. https://doi.org/10.1016/j.jbi.2020.103435

[13] Li, D., Yan, L., Yang, J., *et al*. (2022) Dependency Syntax Guided BERT-BiLSTM-GAM-CRF for Chinese NER. *Expert Systems with Applications*, **196**, Article ID: 116682. https://doi.org/10.1016/j.eswa.2022.116682

[14] 於张闲，胡孔法. 基于 BERT-Att-biLSTM 模型的医学信息分类研究[J]. 计算机时代, 2020(3): 1-4.

[15] 谢靖，刘江峰，王东波. 古代中国医学文献的命名实体识别研究——以 Flat-lattice 增强的 SikuBERT 预训练模型为例[J]. 图书馆论坛, 2022, 42(10): 51-60.

[16] Lan, Z., Chen, M., Goodman, S., *et al*. (2020) ALBERT: A Lite BERT for Self-Supervised Learning of Language Representations. arXiv preprint arxiv:1909.11942.

[17] Lei, T. (2021) When Attention Meets Fast Recurrence: Training Language Models with Reduced Compute. arXiv preprint arXiv: 2102.12459, 2021. https://doi.org/10.18653/v1/2021.emnlp-main.602

[18] Aliyun. (year) A Labeled Chinese Dataset for Diabetes. https://tianchi.aliyun.com/competition/entrance/231687/information

[19] Ye, W., Li, B., Xie, R., *et al*. (2019) Exploiting Entity BIO Tag Embeddings and Multi-task Learning for Relation Extraction with Imbalanced Data. https://doi.org/10.18653/v1/P19-1130

[20] Liu, L., Jiang, H., He, P., *et al*. (2019) On the Variance of the Adaptive Learning Rate and Beyond. arXiv preprint arXiv:1908.03265, 2019.

[21] Fan, Z., He, X., Wang, L., *et al*. (2020) Research on Entity Relationship Extraction for Diabetes Medical Literature. 2020 *IEEE* 9*th Joint International Information Technology and Artificial Intelligence Conference* (*ITAIC*), Chongqing, 11 December 2020, 424-430.

[22] Wang, Y., Sun, Y., Ma, Z., *et al*. (2020) Named Entity Recognition in Chinese Medical Literature Using Pretraining-models. *Scientific Programming*, **2020**, 1-9. https://doi.org/10.1155/2020/8812754

[23] 韩普，顾亮，叶东宇，等. 基于多任务和迁移学习的中文医学文献实体识别研究[J]. 数据分析与知识发现，2023, 7(9): 136-145.