基于批数据过采样的中医临床记录四诊描述抽取方法

王亚强[1,2,3†]，**李凯伦**[1,2,3]，**蒋永光**[4]，**舒红平**[1,3]
[1]成都信息工程大学软件工程学院
[2]成都信息工程大学数据科学与工程研究所
[3]软件自动生成与智能服务四川省重点实验室
[4]成都中医药大学基础医学院
[†]通讯作者：yaqwang@cuit.edu.cn

摘要

中医临床记录四诊描述抽取对中医临床辨证论治的提质增效具有重要的应用价值，然而该抽取任务尚有待探索，类别分布不均衡是该任务的关键挑战之一。本文围绕该任务展开研究，构建了中医临床四诊描述抽取语料库；基于无标注中医临床记录微调通用预训练语言模型实现领域适应；利用小规模标注数据，采用批数据过采样算法，实现中医临床记录四诊描述抽取模型的训练。实验结果表明，本文提出方法的总体性能均优于对比方法，与对比方法的最优结果相比，本文提出的方法将少见类别的抽取性能F1值平均提升了2.13%。

关键词： 四诊描述抽取；类别分布不均衡；批数据过采样；临床记录；中医

Four Diagnostic Description Extraction in Clinical Records of Traditional Chinese Medicine with Batch Data Oversampling

Yaqiang Wang[1,2,3†], **Kailun Li**[1,2,3], **Yongguang Jiang**[4], **Hongping Shu**[2,3]
[1]College of Software Engineering, Chengdu University of Information Technology
[2]Institute for Data Science and Engineering, Chengdu University of Information Technology
[3]Sichuan Key Laboratory of Software Automatic Generation and Intelligent Service
[4]Department of Preclinical Medicine, Chengdu University of Traditional Chinese Medicine
[†]Corresponding author：yaqwang@cuit.edu.cn

Abstract

Four diagnostic description extraction in clinical records of traditional Chinese medicine (TCM) has important application value in improving the quality and efficiency of TCM clinical syndrome differentiation and treatment. However, the extraction task is yet to be explored, and imbalanced class distribution is one of the key challenges of this task. As a first exploration of this task, we firstly constructed a TCM clinical four diagnostic description extraction corpus and then solved the domain adaptation by fine-tuning the general domain pre-trained language model based on unlabeled TCM clinical records. At last, we trained our proposed four diagnostic description extraction model by utilizing a small labeled dataset through a well-designed batch data oversampling algorithm. The experimental results show that the performance of the proposed method in this paper is better than that of the compared methods, and the proposed method improves the extraction performance F1 score of the rare class by 2.13% on average.

Keywords: Four diagnostic description extraction , Imbalanced class distribution , Batch data oversampling , Clinical records , Traditional Chinese medicine

1 引言

辨证论治，又称辨证施治，是中医特有的一种对疾病研究、处理、认知和治疗的基本原则与方法(印会河, 2005)。辨证是论治的前提和依据，四诊（即望、闻、问、切）信息是中医专家综合分析病人的病情，认知疾病，最终辨清证型的重要参考(李红岩 et al., 2022)。快速、准确地获取中医临床记录中的四诊信息，对提升中医专家辨证和诊疗的效率与质量以及为中医临床辅助辨证提供更丰富的医学语义信息具有重要的价值(屈丹丹 et al., 2021)。

在四诊信息中，局部的、具体的疾病、症状、脉象、舌质等实体信息的抽取已有广泛研究。Wang等人(2012)基于条件随机场等统计序列标注模型首次尝试从中医临床记录中抽取症状信息。肖瑞等人(2020)围绕中医临床记录中的疾病和症状信息抽取，采用深度学习模型展开研究。然而，面向全局的、叙述性的中医临床记录四诊描述的抽取还未见相关报道。

中医临床记录中的四诊描述不仅包含局部的、具体的实体修饰信息。如图1所示，实体的“有”或“无”、时间的“长”或“短”、情况的“重”或“轻”等，还蕴含着实体之间的关联信息，如实体之间的因果关系、并列关系等。因此，中医临床记录四诊描述抽取的结果将形成对实体信息抽取研究的补充，为下游任务提供更丰富的医学语义信息。

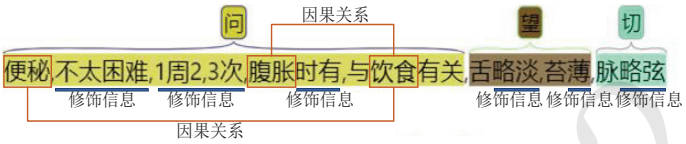

图 1. 中医临床记录中的四诊描述包含的修饰和关联语义信息

与中医临床记录中的实体信息抽取任务不同，中医临床记录四诊描述抽取任务具有其特殊性。首先，与实体的字面值相比，四诊描述的文本长度通常较长，会带来更强的稀疏性。在本文的实验数据集中，每段四诊描述平均包含12个字[1]。此外，如图2所示，通过对不同的四诊描述进行计数发现，四诊描述呈现长尾分布。

其次，由于中医专家的临床实践习惯不同，使得四诊描述天然存在类别分布不均衡的问题。一般地，望诊、问诊、切诊被中医专家更广泛地在临床实践中使用，而闻诊的使用相对较少。基于本文实验数据统计发现（如图3所示），中医临床记录中包含望诊和切诊描述的实例数量少于包含问诊描述的实例数量，而包含闻诊描述的实例数量相较于其他三诊描述格外稀少。

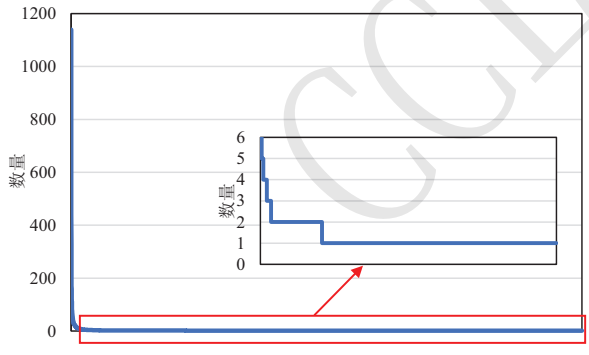

图 2. 不同的中医四诊描述的计数结果

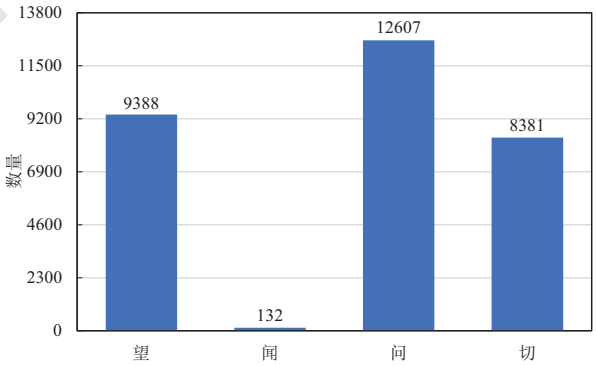

图 3. 中医临床记录中的四诊描述分类计数

因此，本文首次开展了中医临床记录四诊描述抽取任务的探索，针对中医临床记录中的四诊描述的特点，在最新的模型框架基础上，设计并验证了相应的改进策略与算法，取得以下具体成果：

首先，本文将中医临床记录四诊描述抽取定义为基于字的序列标注任务，采用广泛使用的“BIO”标注模式(Irsoy and Cardie, 2014)，提出基于BERT+BiLSTM+CRF(Dai et al., 2019)的中医临床记录四诊描述序列标注模型。在该模型中，利用BERT的动态上下文语义嵌入

[1]本文将中医临床记录中的标点符号也均视为字

学习能力和多头注意力机制(Vaswani et al., 2017)，实现对中医临床记录四诊描述的文本语义信息增强，进而在数据稀疏条件下，保证四诊信息抽取的性能。

其次，采用在无标注的中医临床记录数据上微调通用领域BERT的方法(Gururangan et al., 2020)，验证BERT在进行领域适应后对中医临床记录四诊描述序列标注性能的影响。实验结果发现，该方法有助于提升中医临床记录四诊描述的整体标注性能。对于各类描述的标注结果来说，该方法对“I-望”、“I-闻”、“I-问”、“I-切”等标签的标注有更积极的促进作用。

第三，提出基于批数据过采样的模型训练算法，提升模型对少见的四诊描述类别的标注性能。该算法在基于小批量梯度下降算法（Mini-Batch Gradient Descent，MBGD）(Ruder, 2016)的中医临床记录四诊描述序列标注模型训练框架基础上，通过过采样包含少见类别的数据实例，实现在每轮随机划分的批量数据中，策略性地增加对少见类别数据的学习关注。该方法在实现序列标注模型对常见四诊描述类别的标注性能提升的基础上，大幅提升了少见类别的标注性能。

实验结果表明，本文提出的基于批数据过采样的中医临床记录四诊描述抽取方法的效果优于HMM(Rabiner, 1989)、CRF(Lafferty et al., 2001)、BiLSTM和BiLSTM+CRF(Lample et al., 2016)等对比模型。与对比模型在本文任务上的最佳性能相比，本文方法的标注性能F1值平均提升了1.37%。特别地，本文方法大幅提升了少见类别的标注性能F1值，“B-闻”和“I-闻”标签的F1值分别达到了62.22%和61.54%，相比最佳的对比方法平均有2.13%的提升。

2 相关工作

2.1 中医临床记录信息抽取

中医临床记录信息抽取是近年来中医信息化领域广泛研究的课题。Zhang等人(2022)综述了从2010年至今中医文本信息抽取的相关工作，中医临床记录信息抽取是其中的重要任务之一。目前，中医临床记录信息抽取主要针对疾病、症状、体征、诊断、方剂、药物等局部的、具体的实体信息抽取任务展开，针对包含丰富语言学和临床语义信息的中医临床记录四诊描述抽取的研究甚少。因此，本文开展了中医临床记录四诊描述抽取任务的探索研究。

与一般领域的信息抽取任务相同，中医临床记录信息抽取通常采用序列标注方法实现(Wang et al., 2014)。该类方法将抽取任务转换为序列标注任务，通过对中医临床记录中的基本语义单元进行分类实现对连续的基本语义单元构成的目标类别信息的抽取。其中，语义单元一般为中文字，分类标签通常由待抽取的信息分别定义的BIO标签形成，B表示待抽取的语义单元在待抽取信息的开始位置，I表示待抽取的语义单元在待抽取信息的中间和结束位置，O表示非待抽取的语义单元(Irsoy and Cardie, 2014)。作为初步探索工作，本文沿用了该语义单元和分类标签定义方法。

2.2 序列标注模型

HMM、CRF是被广泛使用的统计序列标注模型，在训练数据规模不大的情况下，因模型复杂度相对较低，它们通常能够取得与深度序列标注模型相当的标注性能(Nasar et al., 2021)。作为中医临床记录四诊描述抽取任务的初探，本文在自建数据集上验证了HMM和CRF的性能，并将它们作为基线模型与目前被更广泛应用的深度序列标注模型BiLSTM+CRF进行比较。

当前，深度序列标注模型在各项信息抽取任务（包括中医临床信息抽取任务）上都取得了优秀的性能，BiLSTM+CRF是其中的代表(Lample et al., 2016)。因此，本文将其作为SOTA基线模型应用于中医临床记录四诊描述抽取任务。此外，BERT能够基于上下文信息，利用多头注意力机制，获取当前待标注语义单元的多角度的丰富的语义信息，动态地形成该语义单元的词嵌入，从而提升下游预测任务模型的性能。因此，本文采用BERT+BiLSTM+CRF来解决中医临床记录四诊描述抽取任务由于数据稀疏带来的语义模糊问题。

BERT是利用通用领域大规模数据集训练得到的预训练模型(Devlin et al., 2018)，其生成的词嵌入携带的是通用语义信息。中医临床记录四诊描述抽取任务的待标注语义单元具有中医领域特殊含义，其上下文蕴含中医领域特殊语义。为更好地适应中医领域的特殊语义表达，借鉴Zhang等人(2020)方法的思想，本文利用中医临床记录数据在MC-BERT的基础上进行微调，以期获得能够更好地表达中医临床记录语义的预训练语言模型。

2.3 不均衡类别分布学习

数据采样是在不均衡类别分布学习中广泛采用的方法之一(刘树栋 and 张可, 2019)。该方法主要通过设计特殊的采样策略，如过采样、欠采样或过采样与欠采样融合，改变数据集的类别分布，达到数据集类别分布均衡的目标。其中，过采样算法是在数据有限条件下，更多被使用的数据采样方法。中医临床记录四诊描述存在类别分布不均衡问题，通常特定领域任务的数据规模有限，因此，本文将数据过采样方法应用到BERT+BiLSTM+CRF的模型训练过程。

BERT+BiLSTM+CRF的模型训练主要采用MBGD框架完成，该框架的参数学习过程的核心是基于每一组批数据估计梯度(Ruder, 2016)。类别分布不均衡会直接导致各组批数据中包含少见类别数据的可能性低，进而导致少见类别学习不充分。为了让模型在训练的过程中更多地关注少见类别，借鉴数据过采样方法(Lin et al., 2017; Shahee and Ananthakumar, 2018)，本文通过过采样少见类别数据，实现在每轮随机划分的批数据中，策略性地增加对少见类别数据的学习，进而达到模型在训练过程中充分学习少见类别数据的目标。

3 方法

3.1 任务定义

中医临床记录四诊描述抽取任务可归结为序列标注任务，因此任务可形式化定义为：给定一条中医临床记录$\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$，目标是训练一个序列分类器，该序列分类器将顺序地预测输入序列$\boldsymbol{x}$中第i个文字x_i对应的标签y_i。本文采用“BIO”标注模式，因此有y_i属于预定义的标签集合$\boldsymbol{L}$ = {O, B-望, I-望, B-闻, I-闻, B-问, I-问, B-切, I-切}。给定训练数据集，中医临床记录四诊描述抽取任务的模型优化目标为：

$$\hat{\boldsymbol{y}} = argmax_{\boldsymbol{y}}\, \mathrm{P}\left(\boldsymbol{y} \mid \boldsymbol{x}\right) \tag{1}$$

3.2 模型

本文以BERT+BiLSTM+CRF模型为基础实现中医临床记录四诊描述抽取，该模型的基础框架如图4所示。

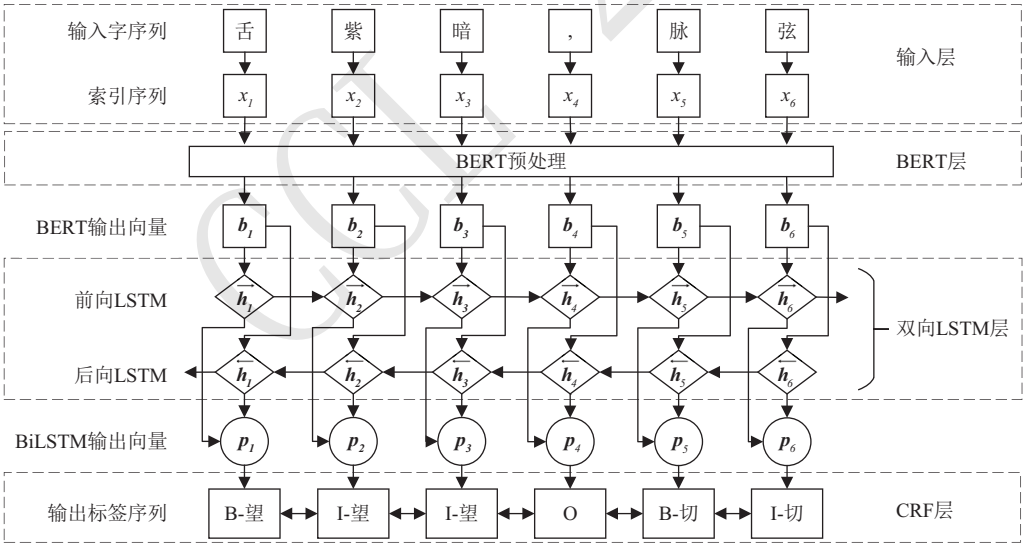

图 4. BERT+BiLSTM+CRF模型基础框架图

如图4的输入层所示，中医临床记录以字为基本单元进行切分，并将切分后的字替换为BERT词表中对应的索引值$x_1 \sim x_6$，形成索引序列。

输入字的索引序列经过图4的BERT层特征提取，得到包含丰富的上下文语义信息的字向量$\boldsymbol{b_1} \sim \boldsymbol{b_6}$。多头注意力机制是BERT模型最关键的部分。在BERT层中，注意力机制实质上是通过字序列的字与字之间的关联程度调整权重系数矩阵，从而获得字序列中所有的字在引入上下文信息后的语义表征向量，其计算公式如(2)所示。

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

其中，$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$为BERT的Embedding层输出的所有字向量经过不同的线性变换后得到的加权矩阵，d_k为字向量的维度。多头注意力机制从不同的角度学习输入序列中的上下文语义信息，均衡单一注意力机制可能产生的偏差，给字向量注入更多元的上下文语义信息，其公式如式(3)和式(4)所示。

$$MultiHead\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=Concat\left(\boldsymbol{head_1},\ldots,\boldsymbol{head_n}\right)\boldsymbol{W^O} \tag{3}$$

$$\boldsymbol{head_i}=Attention\left(\boldsymbol{Q}\boldsymbol{W_i^Q},\boldsymbol{K}\boldsymbol{W_i^K},\boldsymbol{V}\boldsymbol{W_i^V}\right) \tag{4}$$

其中，$\boldsymbol{W_i^Q}$、$\boldsymbol{W_i^K}$、$\boldsymbol{W_i^V}$为多头注意力机制中第$\boldsymbol{i}$个注意力头的$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$所对应的权重矩阵，$\boldsymbol{W^O}$为多头注意力拼接后的线性变换所需的权重矩阵。

在图4中，BiLSTM层的前向过程和后向过程的LSTM单元可以舍弃当前时刻输入字向量的无用信息，并将当前时刻输入字向量的有用信息传递到下一时刻的LSTM单元。然后，将双向过程中每个时刻对应的输出拼接，如公式(5)所示，得到包含长距离上下文信息的字向量$\boldsymbol{p_1}\sim\boldsymbol{p_6}$。

$$\boldsymbol{p}_t=\left[\overrightarrow{\boldsymbol{h}_t},\overleftarrow{\boldsymbol{h}_t}\right] \tag{5}$$

其中，$\overrightarrow{\boldsymbol{h}_t}$为前向过程的LSTM单元在时刻$t$的输出，$\overleftarrow{\boldsymbol{h}_t}$为后向过程的LSTM单元在时刻$t$的输出。

最后，在图4的CRF层中，CRF模型利用邻近标签之间的依赖关系对BiLSTM层输出的所有字向量进行解码，解码目标如式(6)所示，从而得到最优的预测序列。

$$\boldsymbol{Y}^*=\underset{\widetilde{\boldsymbol{Y}}\in\boldsymbol{Y_X}}{argmax}\ s\left(\boldsymbol{X},\widetilde{\boldsymbol{Y}}\right) \tag{6}$$

在公式(6)中，$\boldsymbol{Y_X}$表示所有可能的标注序列，$\boldsymbol{Y}^*$表示解码后获得最大评分的输出序列，s表示标注序列对应的分数函数。

3.3 模型训练方法

3.3.1 模型训练流程

如章节1中所述，中医临床记录四诊描述任务存在严重的类别分布不均衡问题，闻诊描述的数量远少于其它三诊描述的数量。直接利用具有该特点的训练数据对BERT+BiLSTM+CRF模型进行训练，将使模型对训练数据中较少训练数据对应的类别学习不充分，进而影响该类别的预测性能。为克服上述问题，本文设计了基于批数据过采样的小批量(mini-batch)梯度下降算法训练BERT+BiLSTM+CRF模型，以期在一定程度上缓解类别分布不均衡对中医临床四诊描述抽取模型性能的影响，算法训练模型的流程如图5所示。

在如图5的模型训练流程中，主要包含六个关键的处理步骤。

(1) 批数据过采样：在数据处理过程中，按批量大小M将训练数据集$\boldsymbol{D}$划分为包含$\lfloor|\boldsymbol{D}|/M\rfloor$个批量的批量集合$\boldsymbol{B}$。然后，使用批数据过采样的方式增加批量中闻诊信息的数量，生成批量集合$\boldsymbol{B'}$，用于模型训练，从而提高模型对于闻诊信息的抽取性能（此步骤将在3.3.3节中详细介绍）。

(2) 模型参数$\boldsymbol{\theta}_0$初始化：该步骤完成BERT+BiLSTM+CRF模型的初始化参数$\boldsymbol{\theta}_0$的设置。其中，BERT模型的参数是在一定规模的无标注中医临床记录数据上微调得到，该方法参见3.3.2节，BiLSTM模型和CRF模型的初始化参数为随机生成，服从均匀分布。

(3) 损失计算：该步骤将计算模型在当前批量包含的数据样本上的平均损失值。其中，$f_{\boldsymbol{\theta}_k}(\boldsymbol{x}'_i)$代表模型以当前批量中第$i$个数据样本$\boldsymbol{x}'_i$作为输入，且此时模型的参数为第$k$次迭代的参数$\boldsymbol{\theta}_k$。

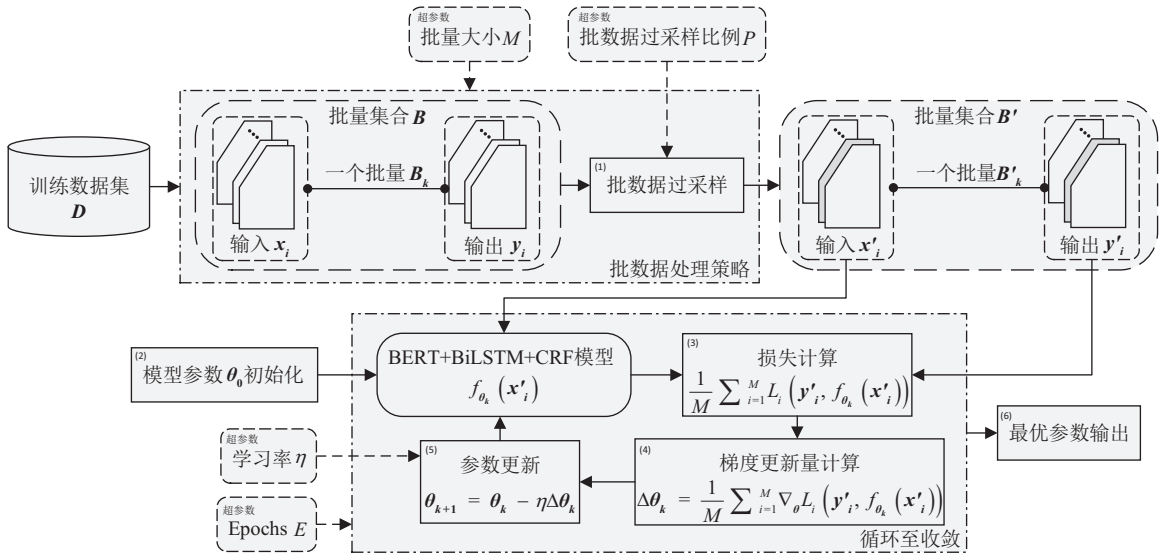

图 5. 基于批数据过采样的小批量梯度下降算法训练BERT+BiLSTM+CRF模型的流程图

(4) 梯度更新量计算：通过误差反向传播算法，当前轮批量$\boldsymbol{B}'_k$中所有数据样本计算梯度的平均值作为模型在第k轮迭代时的梯度更新量$\Delta\boldsymbol{\theta}_k$。

(5) 参数更新：基于当前轮迭代过程中的模型参数$\boldsymbol{\theta}_k$、梯度更新量$\Delta\boldsymbol{\theta}_k$和学习率$\eta$计算第k+1轮迭代过程的模型参数$\boldsymbol{\theta}_{k+1}$

(6) 最优参数输出：步骤(2)到步骤(5)循环执行$\lfloor|\boldsymbol{D}|/M\rfloor * E$轮（$E$为对数据集$\boldsymbol{D}$遍历的轮数），直到模型收敛，输出模型在收敛处的最优参数。

上述的步骤中，M、η、E均为模型训练过程中的超参，它们在本文实验中的取值参见4.3节。

算法 1： 批数据过采样算法

输入： 训练数据集 $\boldsymbol{D}$，闻诊信息数据集 $\boldsymbol{W}$，批数据过采样比例 P，批量大小 M，数据集洗牌函数 *shuffle*，数据集划分函数 *split*，数据样本随机移除函数 *remove*，数据样本随机选取函数 *select*，数据样本添加函数 *append*，批量添加函数 *add*

输出： 过采样闻诊信息后的批量集合 $\boldsymbol{B}'$

$\boldsymbol{S}$ = *shuffle*($\boldsymbol{D}$)　// 对训练数据集 $\boldsymbol{D}$ 进行洗牌操作

$\boldsymbol{B}$ = *split*($\boldsymbol{S}$, M)　// 将洗牌后得到的数据集 $\boldsymbol{S}$ 按批量大小 M 切分为批量集合 $\boldsymbol{B}$

$N = \lfloor|\boldsymbol{D}|/M\rfloor$　// 得到批量集合 $\boldsymbol{B}$ 中批量的数量 N

for $k = 1, 2, ..., N$ **do**

　　$\boldsymbol{B}_k$ = *remove*($\boldsymbol{B}_k$, P, M)　// 从批量 $\boldsymbol{B}_k$ 中随机移除$\lfloor P\times M\rfloor$个数据样本

　　$\boldsymbol{A}_k$ = *select*($\boldsymbol{W}$, P, M)　// 从闻诊信息数据集 $\boldsymbol{W}$ 中有放回地随机选取$\lfloor P\times M\rfloor$个数据样本

　　$\boldsymbol{B}'_k$ = *append*($\boldsymbol{B}_k$, $\boldsymbol{A}_k$)　// 将 $\boldsymbol{A}_k$ 加入批量 $\boldsymbol{B}_k$ 中，得到过采样闻诊信息后的批量 $\boldsymbol{B}'_k$

　　$\boldsymbol{B}'$ = *add*($\boldsymbol{B}'$, $\boldsymbol{B}'_k$)　// 将批量 $\boldsymbol{B}'_k$ 加入到批量集合 $\boldsymbol{B}'$中

end for

图 6. 批数据过采样算法

3.3.2 领域适应方法

为使通用领域的BERT预训练语言模型所生成的词嵌入携带更丰富的中医临床语义信息，使其更适用于中医临床记录四诊描述抽取任务，本文借鉴了关于特定领域BERT的领域适应方法的相关工作(Lee et al., 2020; Gururangan et al., 2020)的基本做法。

在Zhang等人(2020)提出的中文医疗预训练语言模型MC-BERT的基础上，使用领域内的无标注中医临床记录数据集，对MC-BERT的掩码语言模型进行微调，使其可以更好地适应本文

任务领域的语义表达。在领域适应的过程中，更新的掩码语言模型$f_{LM}(\cdot;\theta_{enc},\theta_{LM})$的参数包括从MC-BERT模型上初始化的编码器参数θ_{enc}和分类头参数θ_{LM}。

3.3.3 批数据过采样

由于带标注的中医临床记录数据集存在严重的类别分布不均衡问题，如图3所示，数据集中闻诊描述的数量远少于其他三诊描述的数量，这会严重影响模型对于闻诊描述的抽取性能。为解决这个问题，本文提出在利用小批量梯度下降算法训练四诊信息序列标注模型的过程中，采用批数据过采样的方式去增加批量中闻诊信息的数量，从而在一定程度上消除类别分布不均衡问题对模型抽取性能的影响。批数据过采样的伪代码如图6所示，其中批数据过采样比例P为超参数，其取值参见4.3节。

在图6的批数据过采样算法中，闻诊信息数据集$\boldsymbol{W}$由训练数据集$\boldsymbol{D}$中所有包含闻诊信息的数据样本构成，将在4.1节中具体介绍。并且，批数据过采样在模型训练过程中的每个Epochs中都会执行一次。

4 实验

在测试数据集上，本文将所提出的方法与HMM、CRF、BiLSTM、BiLSTM+CRF等模型进行了比较。本章节后续将依次具体介绍实验中使用的数据集、评价指标、实验设置，以及实验得到的结果。

4.1 数据集

数据集	标签数量	样本数量	抽取信息的数量			
			望	闻	问	切
无标注中医临床记录数据集	-	11251	-	-	-	-
带标注中医临床记录数据集	9	10594	9388	132	12607	8381
训练数据集	9	6346	5652	82	7570	5028
验证数据集	9	2124	1881	28	2545	1661
测试数据集	9	2124	1855	22	2492	1692
闻诊信息数据集	9	79	94	82	139	65

表 1. 所有实验数据集的详细信息

本文实验使用的无标注和带标注的中医临床记录数据集均是基于真实的中医临床记录数据创建，该数据由中医专家在日常诊疗疾病的过程中收集，包含11251条中医临床记录。其中，无标注的中医临床记录数据集由此11251条无标注的中医临床记录直接构成。带标注的中医临床记录数据集则是在11251条中医临床记录的基础上，经过一系列处理得到，具体处理步骤如下：

(1) 讨论并定义中医临床记录中的四诊描述，然后制定标注指南，用于指导后续的数据标注。

(2) 中医专家按照制定好的标注指南，利用Zhang等人(2020)论文中所构建的标准化实验语料构建系统[2]，对11251条中医临床记录数据样本进行四诊信息标注。

(3) 中医专家对标注好的所有数据样本反复审查并修改，形成高质量的标注数据。

(4) 将步骤(3)中得到的高质量的标注数据，按照预定义的标签集合$\boldsymbol{L}$，转化为以字为基本标注单元的BIO标注数据。

(5) 将步骤(4)处理后的数据中包含多重标签（即数据样本中的字具有多个不同标签）的数据样本移除，并将剩余数据样本中的空格和\t移除。

经过上述处理过程，最终得到10594条带标注的中医临床记录。实验中将该数据集按6:2:2的比例随机划分为三部分，得到的训练数据、验证数据和测试数据大小分别为6346条、2124条和2124条带标注的中医临床记录。实验中还将训练数据集中所有包含闻诊信息的数据样本单独复制，组成闻诊信息数据集。各类实验数据集具体信息如表1所示。

[2]实验语料构建系统: http://hknlprel.it.sunsheen.cn/HKKSNLP/build/index.html#/LoginRelation

4.2 评价指标

本文利用F1值和准确率(Accuracy)评价各模型的中医临床记录四诊描述抽取性能，F1和Accuracy的计算公式如下：

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{7}$$

$$Accuracy = \frac{PT}{PT + PN} \tag{8}$$

在公式(7)中，Precision和Recall分别表示模型在测试集上对各类标签预测的精确率和召回率，它们的具体计算方法可参见文献(Wang et al., 2014)。在公式(8)中，PT表示模型预测标签正确的字单元数量，PN表示模型预测标签错误的字单元数量。

4.3 实验设置

在采用领域适应方法微调MC-BERT时，初始学习率被设置为5e-5，批量大小被设置为512，最大句子长度被设置为256。本文提出的模型在训练时，采用了AdamW优化器，初始学习率η被设置为3e-5，β_1被设置为0.9，β_2被设置为0.999。此外，批量大小M被设置为64，最大句子长度被设置为256，批数据过采样比例P被设置为0.4，E被设置为400，Dropout被设置为0.1。

在对比实验中，HMM模型是基于Rabiner等人(1989)的论文实现。CRF模型使用了CRF++开源工具包[3]，其特征定义为在窗口大小为2的上下文中的一元组和二元组。BiLSTM、BiLSTM+CRF等深度神经网络模型是基于Lample等人(2016)论文中的开源代码实现，它们的输入为2451（即实验数据集中包含的字表大小）维的one-hot向量，中间层字向量的维度设置为128。

4.4 实验结果及分析

方法	F1 (%)									Acc (%)
	O	B-望	I-望	B-闻	I-闻	B-问	I-问	B-切	I-切	
HMM(Rabiner, 1989)	73.07	87.78	91.41	22.99	26.15	84.13	96.02	93.07	92.45	92.80
CRF(Lafferty et al., 2001)	80.90	93.96	94.71	61.54	57.97	88.86	97.05	94.82	93.94	94.92
BiLSTM(Lample et al., 2016)	80.22	93.91	94.07	28.57	12.24	87.44	96.85	94.43	92.65	94.49
BiLSTM+CRF(Lample et al., 2016)	80.97	94.08	94.81	32.26	35.09	87.70	96.93	94.48	93.21	94.67
BERT+BiLSTM+CRF	83.46	94.00	94.92	54.55	52.63	89.64	97.45	94.22	93.81	95.37
BERT+BiLSTM+CRF+BDO[1]	83.71	94.58	95.47	60.00	56.34	89.18	97.44	94.47	94.41	95.47
本文方法-DA[2]-BDO[1]	84.49	94.39	95.26	54.05	51.28	89.70	97.41	94.82	94.39	95.49
本文方法-BDO[1]	84.32	94.25	95.37	50.00	53.73	89.11	97.45	94.67	94.57	95.52
本文方法	**85.14**	**94.62**	**95.51**	**62.22**	**61.54**	**89.93**	**97.54**	**94.91**	**94.67**	**95.70**

[1] “BDO”指**B**atch **D**ata **O**versampling，即“批数据过采样”
[2] “DA”指“**D**omain **A**daptation”，即“领域适应”

表 2. 实验结果

表2列出了本文方法和对比方法在测试数据集上获得的最佳F1值和准确率（在表2中以Acc表示）结果。从表2可以看出，本文方法在各标签的预测结果上均优于所对比的方法。本文方法的Acc达到了95.70%，相比所有对比方法有0.78%到2.9%的提升。本文方法相比最优的对比方法，在每种标签的预测F1值上，平均提升了1.37%。上述结果充分验证了本文方法在中医临床记录四诊描述抽取任务上的预测性能。

此外，通过消融实验，本文还进一步验证了所提出方法的各主要部分的有效性。从表2中可以看到，当本文方法移除领域适应和批数据过采样之后，准确率仍优于其它对比方法。具体地，除少见类别“B-闻”和“I-闻”以外的其他标签的F1值均高于其它对比方法。这证明了本文将MC-BERT+BiLSTM+CRF模型应用于中医临床记录四诊描述抽取的有效性。少见类别预测性能较差的主要原因是基于BERT的深度神经网络模型结构复杂，参数量巨大，对训练数据集

[3]CRF++开源工具包：https://taku910.github.io/crfpp/#source

中包含的少见类别学习不充分，导致其预测性能低于模型复杂度相对较低的统计机器学习模型CRF。

当本文方法只移除批数据过采样方法时，模型预测的准确率仍然优于所有对比方法。并且，在“I-望”、“I-闻”、“I-问”、“I-切”等标签上的F1值优于同时移除领域适应和批数据过采样的情况。这说明领域适应方法能够有效提升模型抽取四诊描述的整体性能，且对于非边界四诊描述的判别有强的促进作用。当本文方法不移除任何组件时，其性能在准确率以及每个标签的F1值上均优于所有的对比方法，这进一步验证了本文方法的领域适应与批数据过采样的有效性。

此外，本文方法对于存在类别分布不均衡问题的“B-闻”和“I-闻”标签的抽取效果有显著提升，对应的F1值分别达到了62.22%和61.54%。该结果说明，在训练模型的过程中策略地增加批量中的闻诊信息，使模型更充分地学习闻诊描述特征，进而在一定程度上缓解了因类别分布不均衡问题给模型预测性能带来的负面影响。从表2中还可以看出，将批数据过采样应用于通用领域的BERT+BiLSTM+CRF模型时，模型在少见的四诊描述类别标签“B-闻”和“I-闻”上的抽取性能F1值也出现了显著提升，这验证了本文3.3.3节设计的批数据过采样方法的有效性。

为进一步证明批数据过采样的有效性，对模型在测试数据集上的标注结果进行了进一步的分析。发现移除批数据过采样后的模型往往会将闻诊信息错误地标注为问诊信息。例如：将“肠鸣，少腹重坠略有缓解”一起标注为问诊，而“肠鸣”实则应标注为闻诊。这是由于训练数据集中闻诊信息的数据量极少，直接利用小批量梯度下降算法对模型进行训练时，闻诊信息仅出现在少数用于计算更新梯度的批量中，在大多数批量中其出现次数甚至为0。

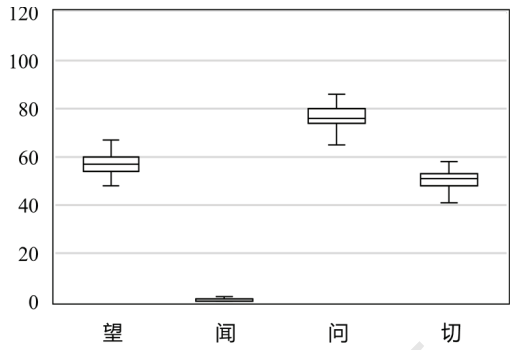

图 7. 批量中的四诊信息数量统计(P = 0)

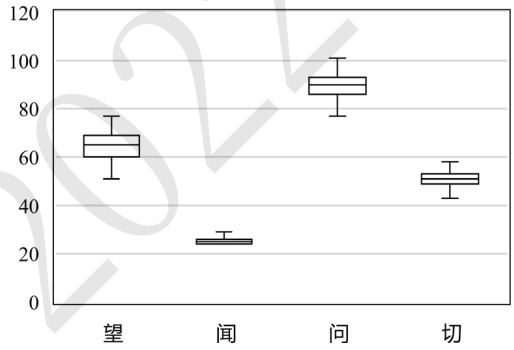

图 8. 批量中的四诊信息数量统计(P = 0.4)

图7是在移除了批数据过采样的模型训练过程中，对训练数据集一轮遍历时，以划分的批量为单位，对批量中包含的四诊描述的出现次数的统计结果。从图7中可以明显地看出，批量中闻诊描述出现的次数极少，近乎为0，与问诊描述在平均出现次数上的差值接近80。这导致模型无法充分地学习到闻诊描述特征，将闻诊描述错误地预测为其它类型的描述。

从图8中可以看出，在不移除批数据过采样且P值被设置为0.4的情况下（采用图7相同的统计方法），批量中包含的闻诊描述的数量大幅提升，这将使模型能够在训练过程中更充分地学习闻诊描述特征，同时使模型对“B-闻”和“I-闻”标签的预测性能显著提升。

4.4.1 批数据过采样比例P对模型抽取性能的影响

为验证不同批数据过采样比例P的设置，对本文所提出的中医临床记录四诊描述抽取方法的影响，本文进一步实验了在P被设置为0、0.2、0.4、0.6、0.8或1时，模型在测试集上，对$\boldsymbol{L}$中的各类标签的预测性能，实验结果如图9所示。

从图9可以看出，当$P = 0.4$时，所有标签的F1值均达到最高，并且相较于其它标签，“B-闻”和“I-闻”的F1值增幅最大。该结果说明，当$P = 0.4$时，本文模型能够最有效地从批量中学习到闻诊描述特征，能够更好地消除类别分布不均衡对模型预测性能的影响。该结果进一步说明，在批量中策略地增加包含闻诊描述的实例，间接地降低其他三诊描述在批量中出现的占比，能够有效地避免模型在训练时过度地拟合望诊、问诊和切诊类别标签，让模型更充分地学习少见的闻诊类别标签，进而增强模型的预测性能和泛化能力。

此外，当P<0.4时，本文模型在“B-闻”和“I-闻”标签上的F1值有明显降低，而在其他标签上的F1值无明显波动。该结果说明，在模型训练过程中，闻诊描述在批量中出现的次数降

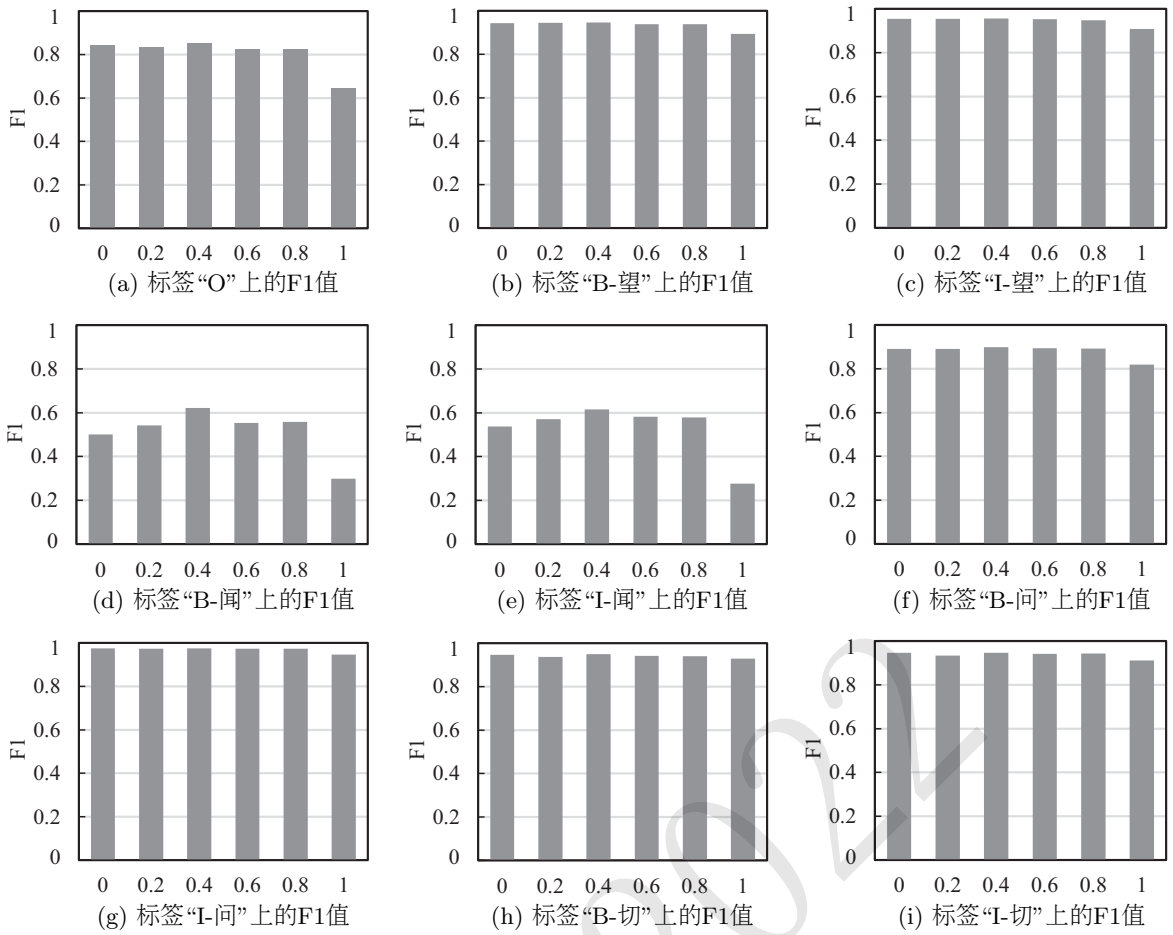

图 9. 不同批数据过采样比例下的F1值

低，导致类别分布不均衡问题加重，使得模型对闻诊类别标签的特征学习不充分，导致模型对“B-闻”和“I-闻”标签的预测能力降低。而其他三诊描述在批量中出现占比变化不大，模型仍能充分学习，因此，它们的F1值并没有明显变化。

最后，当P>0.4时，模型在各类标签上的F1值均有不同程度的降低，特别是在P被设置为1时。这是由于批数据过采样是从闻诊信息数据集中选取数据样本放入批量中导致的。P值越高，模型越近似于在闻诊信息数据集上进行模型的训练，然而，闻诊信息数据集仅包含训练数据集中所有包含闻诊描述的数据样本，数据规模小，包含信息少，这将直接影响模型的训练效果，最终导致所有标签的F1值下降。

5 总结与展望

本文初探了中医临床记录四诊描述抽取任务，以万余条中医临床记录自建了标准实验语料，针对四诊描述类别分布不均衡带来的挑战，提出了一种基于批数据过采样的中医临床记录四诊描述抽取方法。在标准实验语料上，与对比方法相比，本文提出方法取得了最优结果。与对比方法的最优结果相比，本文提出的方法将少见类别的抽取性能F1值平均提升了2.13%。此外，通过多个角度的细致分析，进一步验证了本文所提出方法的有效性。

目前，中医临床记录四诊描述抽取模型的预测性能还有待进一步提升，未来将深入探究中医临床记录四诊描述抽取任务的特点及存在的问题，设计并实践更优的抽取方法，进一步提升中医临床记录四诊描述抽取方法的性能，并将方法应用于实践，达到为中医临床辨证论治提质增效的目标。

参考文献

李红岩, 李灿, 郎许锋, 杨涛, 周作建, 战丽彬. 2022. 中医四诊智能化研究现状及热点分析. 南京中医药大

学学报, 38(02):180-186.

刘树栋, 张可. 2019. 类别不均衡学习中的抽样策略研究. 计算机工程与应用, 55(21):1-17.

屈丹丹, 杨涛, 胡孔法. 2021. 基于字向量的BiGRU-CRF肺癌医案四诊信息实体抽取研究. 世界科学技术-中医药现代化, 23(09):3118-3125.

肖瑞, 胡冯菊, 裴卫. 2020. 基于BiLSTM-CRF的中医文本命名实体识别. 世界科学技术-中医药现代化, 22(07):2504-2510.

印会河. 2005. 中医基础理论. 上海科学技术出版社, 上海, 中国.

Zhenjin Dai, Xutao Wang, Pin Ni, Yuming Li, Gangmin Li, and Xuming Bai. 2019. Named entity recognition using BERT BiLSTM CRF for Chinese electronic health records. *12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 1-5.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Ozan Irsoy and Claire Cardie. 2014. Opinion mining with deep recurrent neural networks. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 720-728.

John Lafferty, Andrew Mccallum, and Fernando C.N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001)*, 282-289.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234-1240.

Chin Teng Lin, Tsung Yu Hsieh, Yu Ting Liu, Yang Yin Lin, Chieh Ning Fang, Yu Kai Wang, Gary Yen, and Nikhil R. Pal. 2017. Minority oversampling in kernel adaptive subspaces for class imbalanced datasets. *IEEE Transactions on Knowledge and Data Engineering*, 30(5): 950-962.

Zara Nasar, Syed Waqar Jaffry, and Muhammad Kamran Malik,. 2021. Named entity recognition and relation extraction: State-of-the-art. *ACM Computing Surveys (CSUR)*, 54(1): 1-39.

Lawrence R. Rabiner. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2): 257-286.

Sebastian Ruder. 2016. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Shaukat Ali Shahee and Usha Ananthakumar. 2018. An adaptive oversampling technique for imbalanced datasets. *Industrial Conference on Data Mining*, 1-16.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 30.

Yaqiang Wang, Yiguang Liu, Zhonghua Yu, Li Chen, and Yongguang Jiang. 2012. A preliminary work on symptom name recognition from free-text clinical records of traditional Chinese medicine using conditional random fields and reasonable features. *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, 223-230.

Yaqiang Wang, Zhonghua Yu, Li Chen, Yunhui Chen, Yiguang Liu, Xiaoguang Hu, and Yongguang Jiang. 2014. Supervised methods for symptom name recognition in free-text clinical records of traditional Chinese medicine: an empirical study. *Journal of Biomedical Informatics*, 47: 91-104.

Ningyu Zhang, Qianghuai Jia, Kangping Yin, Liang Dong, Feng Gao, and Nengwei Hua. 2020. Conceptualized representation learning for chinese biomedical text mining. *arXiv preprint arXiv:2008.10813*.

Tingting Zhang, Zonghai Huang, Yaqiang Wang, Chuanbiao Wen, Yangzhi Peng, and Ying Ye. 2022. Information Extraction from the Text Data on Traditional Chinese Medicine: A Review on Tasks, Challenges, and Methods from 2010 to 2021. *Evidence-Based Complementary and Alternative Medicine*.

Tingting Zhang, Yaqiang Wang, Xiaofeng Wang, Yafei Yang, and Ying Ye. 2020. Constructing fine-grained entity recognition corpora based on clinical records of traditional Chinese medicine. *BMC Medical Informatics and Decision Making*, 20(1): 1-17.