计算机科学
COMPUTER SCIENCE

融合标签知识的中文医学命名实体识别

尹宝生, 周澎

引用本文

尹宝生, 周澎. 融合标签知识的中文医学命名实体识别[J]. 计算机科学, 2024, 51(6A): 230500203-7.

YIN Baosheng, ZHOU Peng. Chinese Medical Named Entity Recognition with Label Knowledge[J]. Computer Science, 2024, 51(6A): 230500203-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230500203

融合标签知识的中文医学命名实体识别

尹宝生　周　澎
沈阳航空航天大学人机智能研究中心　沈阳 110136
(541951941@qq.com)

摘　要　医学领域命名实体识别是信息抽取任务重要的研究内容之一,其训练数据主要来源于临床实验数据、健康档案、电子病历等非结构化文本,然而标注这些数据需要专业人员耗费大量人力、物力和时间资源。在缺乏大规模医学训练数据的情况下,医学领域命名实体识别模型很容易出现识别错误的情况。为解决这一难题,文中提出了一种融合标签知识的中文医学命名实体识别方法,即通过专业领域词典获得文本标签的释义后,分别将文本、标签及标签释义编码,基于自适应融合机制进行融合,有效平衡特征提取模块和语义增强模块的信息流,从而提高模型性能。其核心思想在于医学实体标签是通过总结归纳大量医学数据得到的,而标签释义是对标签进行科学解释和说明的结果,模型融入这些蕴含了丰富的医学领域内的先验知识,可以使其更准确地理解实体在医学领域中的语义并提升其识别效果。实验结果表明,该方法在中文医学实体抽取数据集(CMeEE-V2)3个基线模型上分别取得了0.71%,0.53%和1.17%的提升,并且为小样本场景下的实体识别提供了一个有效的解决方案。

关键词:中文医学命名实体识别;标签知识;先验知识;自适应融合机制;小样本
中图分类号　TP391

Chinese Medical Named Entity Recognition with Label Knowledge

YIN Baosheng and ZHOU Peng
Human-Machine Intelligence Research Center,Shenyang Aerospace University,Shenyang 110136,China

Abstract　Named entity recognition in the medical field is one of the important research contents of information extraction tasks. Its training data mainly comes from unstructured texts such as clinical trial data,health records,electronic medical records. However,labeling these data requires professionals to spend a lot of manpower,material resources and ime. In the absence of large-scale medical training data,named entity recognition models in the medical field are prone to recognition errors. In order to solve this problem,this paper proposes a Chinese medical named entity recognition method that integrates label knowledge,that is,after obtaining the interpretation of the text label through a professional field dictionary,the text,label and label interpretation are encoded separately,and the fusion is performed based on an adaptive fusion mechanism,to effectively balance the information flow of the feature extraction module and the semantic enhancement module,thereby improving the model performance. The core idea is that the medical entity label is obtained by summarizing a large amount of medical data,and the label interpretation is the result of scientific explanation and explanation of the label. The model incorporates these rich prior knowledge in the medical field to make it more accurate. Accurately understand the semantics of entities in the medical domain and improve their recognition. Experimental results show that the method has achieved 0.71%,0.53% and 1.17% improvement on the three baseline models of the Chinese medical entity extraction dataset(CMeEE-V2),and provides an effective method for entity recognition in small sample scenarios.

Keywords　Chinese medical named entity recognition,Label knowledge,Prior knowledge,Adaptive fusion mechanism,Few shot

1　引言

命名实体识别(Named Entity Recognition,NER)是一项重要的信息抽取技术,旨在从文本中识别出具有特定意义的命名实体,如人名、地名、组织机构名等[1]。在医学领域中,命名实体识别技术发挥着重要的作用,它可以自动从医学文献、病历、影像等数据中提取出关键实体,例如患者姓名、疾病名称、药品名称等,从而帮助医生更准确地做出诊断并制定治疗方案。此外,命名实体识别技术还被应用于医学知识图谱的构建,进一步促进了智慧医疗的研究和发展。

医学命名实体识别的目标是从非结构化的文本中提取实体并判断它们的类型[2]。然而,现有的中文医学命名实体识别面临着标注数据不足的问题。医学文献、专利、病历等数据中常常含有大量的医学术语和缩写,因此,非医学专业人员往往难以理解和做出准确的标注。虽然医学数据规模庞大,但其中大部分是非公开数据,或者仅有少量带有标注的数据可用,这给医学命名实体识别模型的训练带来了困难[3]。

在命名实体识别任务中,实体的定义通常根据具体任务

基金项目:辽宁省教育厅项目(LJKMZ20220536)
This work was supported by the Project of the Education Department of Liaoning Province,China(LJKMZ20220536).
通信作者:周澎(1132187866@qq.com)

需求来确定，不同的数据集规定了不同的实体类型。例如2020年全国知识图谱与语义计算大会评测任务三的数据集中规定了实验室检验、手术、影像检查等6种实体类型；瑞金医院MMC人工智能辅助构建知识图谱大赛数据集中标注出疾病名称、病因、药品名称、手术等15五种实体类型。研究人员通常依靠医学领域的专业知识和经验，深入分析和理解医学文本数据，提取出其中的重要实体类型，并给出相应的定义和范围[4]。不同数据集中的实体类型及其释义包含的知识各不相同，这些信息是研究人员对大量医学数据进行总结和归纳得出的，因此，它们蕴含着丰富的先验知识。如何更好地将这些信息融合到模型中，需要一个合理的解决方案。

由于标注数据十分有限，模型往往面临过拟合和缺乏鲁棒性等问题，难以取得令人满意的结果，因此，本文提出了一种通过融合标签及标签释义的方法，使用自适应融合机制将医学领域的先验知识融入到模型中，从而提升模型在全数据场景下的性能，并在小样本数据中取得了显著的性能提升。

2 相关工作

以往的工作主要集中于全数据场景下的命名实体识别，下面主要针对小样本学习及小样本场景下的命名实体识别进行介绍。

小样本学习[5](Few-shot Learning)的目标是让模型具备区分事物的能力，但面临着训练数据有限以及不能很好地反映真实数据分布的问题。此外，小样本场景下模型容易出现过拟合的问题。目前小样本学习领域主要采用模型微调[6]、数据增强[7]、迁移学习[8]等方法进行研究。为了提升模型的泛化能力，可以通过数据增强的方法来提高数据多样性[9]。Dixit等[10]提出的属性引导增强学习了一种允许合成数据的映射，使得合成样本的属性处于期望值或强度。Schwartz等[11]提出了一种基于自编码器的样本生成方法，通过分辨训练类别的实例对之间的差异，生成应用于新类型的样本数据。Chen等[12]提出小样本命名实体识别任务需要从额外的资源中利用和迁移有用的知识，且通过使用概念集合描述标签集合可以解决实体类型限制的问题。因为小样本数据中的知识是有限的，并且知识存在可能与新任务不直接匹配的情况。例如，“America”在Wikipedia，OntoNotes和WNUT17数据集中分别被分类为地理实体、GPE和位置。

Lai等[13]提出了一种用于小样本中文命名实体识别的基于提示的遗传Bert，通过在高资源数据集上训练模型，然后在低资源数据集上发现更多的隐式标签信息，并进一步设计了标签扩展策略，实现了从高资源数据集学习标签知识。Ma等[14]提出了一种分解的元学习方法，通过使用元学习一次处理小样本跨度检测和小样本实体类型来解决小样本命名实体识别的问题，将小样本跨度检测作为序列标注问题，并通过引入模型不可知元学习算法来训练跨度检测器，从而找到可以快速适应新实体类的良好模型参数。Wang等[15]提出了开创性的基于跨度的原型网络，它通过两阶段方法处理小样本命名实体识别，包括跨度提取和实体分类，在跨度提取阶段，将顺序标签转换为全局边界矩阵，使模型能够关注显示的边界信息，对于实体分类，利用原型学习来捕获每个标记范围的语义表示，并使模型更好地适应新实体。Li等[16]定义了一个小样本命名实体识别的N-way K-shot设定，同时提出FewNER模型，将整个网络分为任务无关部分和任务特定部分，从而使模型不易过拟合并提高计算效率。Ma等[17]提出使用两个编码器对文本和标签进行独立编码，然后进行融合，该方法在小样本实体识别任务中取得了性能提升。而Fritzler等[18]则将小样本命名实体识别任务作为半监督学习任务进行学习，通过修改原型网络解决实体抽取任务，结果表明其在资源匮乏的情况下优于最新模型。

3 融入标签知识的中文医学命名实体识别

本文提出了融入标签语义信息及其释义信息的中文医学实体抽取模型LK4CMeNER(Label Knowledge For Chinese Medical NER)，图1展示了该模型的体系结构，其中包含4个模块：向量表示模块、文本编码模块、特征融合模块、输出模块。

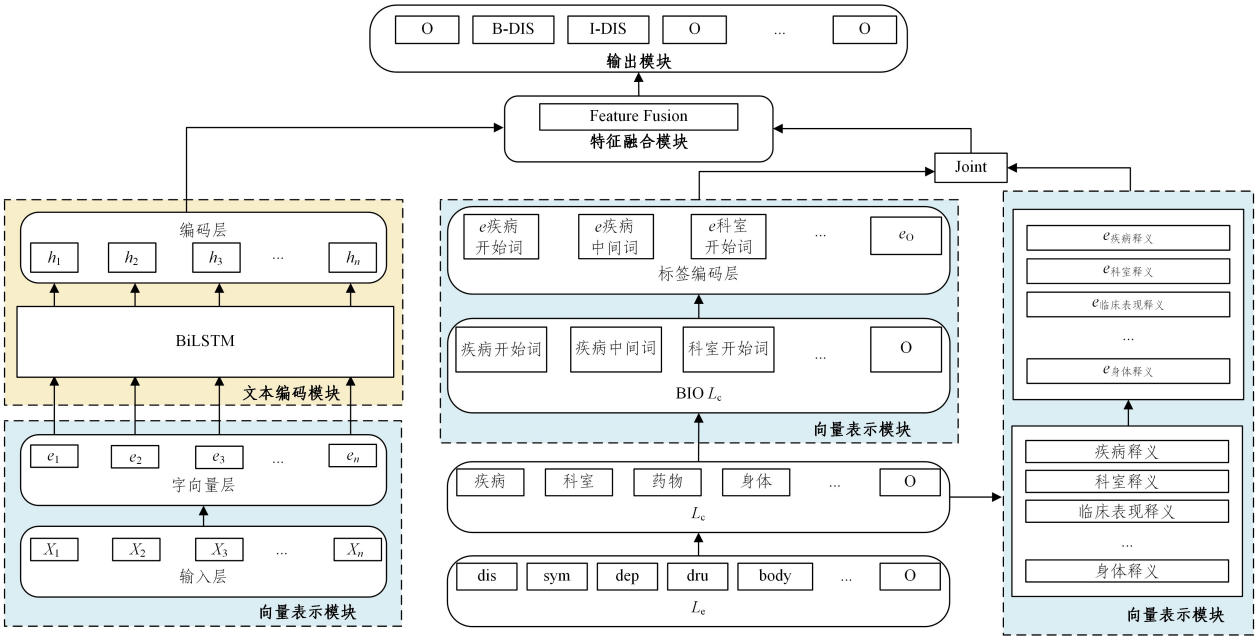

图1 LK4CMeNER模型图

Fig. 1 LK4CMeNER model diagram

3.1 向量表示模块

向量表示[19]（Vector Representation），也称为词嵌入或词向量化，是一种将对象（如词汇、图像、音频等）映射到向量空间中的技术，其主要目的是将对象转化为实数组成的向量，其中每个维度代表对象的某个特定属性。通过向量表示，能够将对象的复杂信息编码成简洁且连续的向量形式，从而方便进行计算和处理。

在自然语言处理领域中，向量表示通常用于将文本中的词汇或句子转换为向量形式。例如，词向量是一种将单词映射到向量空间中的一个向量的方法，通过向量的位置和距离来编码单词的语义和语法信息。同样地，句子向量用于将整个句子表示为向量，以便在语义匹配、文本分类、文本聚类等任务中进行操作。词向量和句子向量的生成过程通常依赖于深度学习模型，如 Word2Vec[20]，Glove[21] 和 Bert[22] 等，这些模型通过大规模语料库的训练来学习词汇和句子的表示。

3.1.1 字符向量

在处理中文命名实体识别任务时，模型接受的输入通常是一个句子 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_i,\cdots,x_n\}$，其中 n 表示输入文本的长度。中文句子与英文句子不同，没有明显的单词边界，因此需要将输入文本切分成字的形式。每个字 x_i 经过编码后得到字符向量 $\boldsymbol{e}_{\text{word}}$，如式(1)所示：

$$\boldsymbol{e}_{xi}=Bert(x_i) \tag{1}$$

其中，$\boldsymbol{e}_{xi}$ 表示字符向量，我们使用 Devli 等发布的 Bert 预训练模型对输入文本进行向量化处理。Bert 预训练模型是一种强大的自然语言处理模型，其通过大规模的预训练和微调可以提供丰富的语义表示能力。我们选择使用该模型对中文文本进行编码，旨在充分利用其在各种语言任务上取得的优异性能，来提高中文命名实体识别的准确性和效果。

3.1.2 标签语义向量

为了在模型中使用标签的语义信息，我们同样使用 Bert 对其进行向量化处理。

对于给定的英文标签集合 $L_e=\{l_1,l_2,\cdots,l_m\}$，长度为 m，我们需要将英文标签名称转换成中文标签名称，例如将“dis”转换为“疾病”，将“body”转换为“身体”，得到中文标签集合 L_c。

接下来，对 L_c 中的每个标签名称添加 BIO 标记，比如将“疾病”转换成“疾病开始词”和“疾病中间词”，这样得到的标签数量将变为 $2*m+1$。然后，使用 Bert 对每个标签进行编码，得到每个标签的向量表示，即标签语义向量 $\boldsymbol{e}_{\text{label}}$。具体而言，第 i 个标签的向量表示为：

$$\boldsymbol{e}_{li}=Bert(l_i) \tag{2}$$

3.1.3 标签释义向量

本文利用更多的医学领域知识增强标签语义信息。我们获取了每个标签的释义，并得到标签释义集合 $D=\{d_1,d_2,\cdots,d_i,\cdots,d_m\}$，其中 d_i 表示第 i 个标签的详细解释。例如，“药物”一词对应的释义是“药物是指用于预防、治疗和诊断疾病的化学物质或生物制剂。”，其他标签的释义如表 1 所列。

表 1 标签释义信息

Table 1 Label interpretation information

标签名称	释义
疾病	疾病是指导致人们陷入非健康状态的因素或医生对患者做出的诊断，通常可以通过接受治疗来缓解或治愈
临床表现	临床表现是指患者主观感受和医生通过检查获得的异常体征或生理指标
医疗程序	医疗程序指为了确诊和治疗而采取的一系列措施、方法和过程
医疗设备	医疗设备是指用于诊断、治疗或预防疾病的各种工具、器具、仪器等医疗用具
药物	药物是指用于预防、治疗和诊断疾病的化学物质或生物制剂
医学检验项目	医学检验项目是指在医学检验过程中需要采集体液、测量生理指标以及进行其他相关检查的项目
身体	身体是指由细胞、组织、器官、系统和肢体等组成的人体结构，同时还包括身体内部产生或由身体组织分泌的物质
科室	科室是指医院或其他医疗机构设立的，专门针对某一类疾病或医疗领域进行诊疗、治疗和护理的部门
微生物类	微生物类包括细菌、病毒、真菌、原生生物、藻类等微小生物，以及它们分泌的毒素、酶和其他生物活性分子

通过对标签释义编码，将 Bert 模型输出的[CLS]的最后一层隐藏状态作为标签释义向量。具体的，它通过线性层和激活函数对该隐藏状态进行处理，得到最终的标签释义向量。这种编码方式能够捕捉标签释义的语义信息，并为后续的模型处理提供丰富的上下文表示。

3.2 文本编码模块

为了更全面地利用文本信息，本文采用 BiLSTM[23] 对文本序列进行建模。通过引入双向信息，模型能够捕捉前向和后向的上下文关系，从而更好地理解文本的语义，有助于提高模型在文本处理任务中的性能。

BiLSTM（Bidirectional Long Short-Term Memory）是一种由前向 LSTM 和后向 LSTM 组合而成的双向循环神经网络。其前向 LSTM 计算公式表示如下：

$$i_t^f=\sigma(\boldsymbol{W}_{xi}x_t+\boldsymbol{W}_{hi}h_{t-1}+\boldsymbol{W}_{ci}c_{t-1}+\boldsymbol{b}_i) \tag{3}$$

$$f_t^f=\sigma(\boldsymbol{W}_{xf}x_t+\boldsymbol{W}_{hf}h_{t-1}+\boldsymbol{W}_{cf}c_{t-1}+\boldsymbol{b}_f) \tag{4}$$

$$C_t^f=f_t^f\odot c_{t-1}+i_t^f\odot\tanh(\boldsymbol{W}_{xc}x_t+\boldsymbol{W}_{hc}h_{t-1}+\boldsymbol{b}_c) \tag{5}$$

$$o_t^f=\sigma(\boldsymbol{W}_{xo}x_t+\boldsymbol{W}_{ho}h_{t-1}+\boldsymbol{W}_{co}c_t^f+\boldsymbol{b}_o) \tag{6}$$

$$h_t^f=o_t^f\odot\tanh(c_t^f) \tag{7}$$

其中，i_t^f，f_t^f，c_t^f，o_t^f，h_t^f 分别表示输入门、遗忘门、细胞状态更新、输出门和隐状态；x_t 为输入序列在时间步 t 的输入，h_{t-1} 为前一个时间步的隐状态；c_{t-1} 为前一个时间步的细胞状态；σ 表示 sigmoid 函数；$\odot$ 表示逐元素乘法；tanh 表示双曲正切函数；$\boldsymbol{W}_{xi}$，$\boldsymbol{W}_{hi}$，$\boldsymbol{W}_{ci}$，$\boldsymbol{b}_i$ 分别为输入门的权重矩阵和偏置向量；$\boldsymbol{W}_{xf}$，$\boldsymbol{W}_{hf}$，$\boldsymbol{W}_{cf}$，$\boldsymbol{b}_f$ 分别为遗忘门的权重矩阵和偏置向量；$\boldsymbol{W}_{xc}$，$\boldsymbol{W}_{hc}$，$\boldsymbol{b}_c$ 分别为细胞状态更新的权重矩阵和偏置向量；$\boldsymbol{W}_{xo}$，$\boldsymbol{W}_{ho}$，$\boldsymbol{W}_{co}$，$\boldsymbol{b}_o$ 分别为输出门的权重矩阵和偏置向量。

后向 LSTM 的计算与前向 LSTM 类似，只是将输入序列从后向前进行计算。前向 LSTM 和后向 LSTM 的输出可以进行拼接，得到 BiLSTM 的输出：

$$h_t^b = \text{concat}(h_t^f, h_t^b) \tag{8}$$

其中，h_t^f 为前向 LSTM 在时间步 t 的隐状态，h_t^b 为后向 LSTM 在时间步 t 的隐状态，concat 表示将两个向量在某个维度上进行拼接。BiLSTM 可以同时利用前向和后向的信息，从而更好地捕捉输入序列的上下文关系，提高模型在序列数据处理任务中的性能。

3.3 特征融合模块

在获得字符向量、标签语义向量和标签释义向量后，我们首先使用拼接标签语义向量和标签释义向量，得到标签知识向量，并采用以下几种方式将字符向量与标签所含的先验知识进行结合：

(1)点积：采用逐元素相乘的方式将字符向量和标签知识向量进行融合。此方法能够有效地将两者的信息交互起来，使模型对于字符和标签之间的关联性有更好的理解。

$$\boldsymbol{u}_i = \boldsymbol{e}_{\text{word}} \odot \boldsymbol{e}_{\text{label}} \tag{9}$$

(2)拼接：将字符向量和标签知识向量按维度进行拼接，形成一个更丰富的特征表示。

$$u_i = \boldsymbol{e}_{\text{word}} \oplus \boldsymbol{e}_{\text{label}} \tag{10}$$

(3)自适应融合点积

本文使用自适应融合机制，通过动态分配权重来有效融合自注意力机制和标签注意力机制所提取的文本信息。这种方法可以让模型更好地学习文本中的关键信息和标签信息，并在实体识别任务中获得更好的性能。

$$\alpha_{\text{word}} = sigmoid(\boldsymbol{e}_x \cdot \boldsymbol{W}_1) \tag{11}$$

$$\beta_{\text{label}} = sigmoid(\boldsymbol{e}_l \cdot \boldsymbol{W}_2) \tag{12}$$

$$\alpha_{\text{word}} + \beta_{\text{label}} = 1 \tag{13}$$

$$o_i = \alpha_{\text{word}} * \boldsymbol{e}_{\text{word}} \odot \beta_{\text{label}} * \boldsymbol{e}_{\text{label}} \tag{14}$$

其中，$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 是可学习的参数矩阵，α_{word} 和 β_{label} 用于控制标签知识向量对字符向量的影响程度，o_i 表示自适应融合后的最终输出。

(4)自适应融合拼接

$$\alpha_{\text{word}} = sigmoid(\boldsymbol{e}_x \cdot \boldsymbol{W}_1) \tag{15}$$

$$\beta_{\text{label}} = sigmoid(\boldsymbol{e}_l \cdot \boldsymbol{W}_2) \tag{16}$$

$$\alpha_{\text{word}} + \beta_{\text{label}} = 1 \tag{17}$$

$$o_i = \alpha_{\text{word}} * \boldsymbol{e}_{\text{word}} \oplus \beta_{\text{label}} * \boldsymbol{e}_{\text{label}} \tag{18}$$

3.4 输出模块

条件随机场(CRF)[24]是一种用于建模序列标注问题的概率图模型，其原理如下：给定输入序列 $X=\{x_1, x_2, \cdots, x_n\}$ 和对应的标签序列 $Y=\{y_1, y_2, \cdots, y_n\}$，其中 n 是序列的长度，CRF 模型的概率定义为：

$$P(Y|X) = \frac{1}{Z(X)} \exp\left(\sum_{i=1}^{n} \sum_{j=1}^{k} w_j \cdot f_j(x_i, y_{i-1}, y_i)\right) \tag{19}$$

其中，$P(Y|X)$ 是给定输入序列 X 条件下，标签序列 Y 出现的概率。$Z(X)$ 是归一化因子，用于保证概率的和为 1，定义为：

$$Z(X) = \sum_Y \exp\left(\sum_{i=1}^{n} \sum_{j=1}^{k} w_j \cdot f_j(x_i, y_{i-1}, y_i)\right) \tag{20}$$

其中，Y 表示所有可能的标签序列；w_j 是模型参数，用于表示特征函数 $f_j(x_i, y_{i-1}, y_i)$ 的权重，这些特征函数用于描述输入序列和标签序列之间的关系；$f_j(x_i, y_{i-1}, y_i)$ 是特征函数，用于捕捉输入序列、前一个标签和当前标签之间的局部信息；k 是特征函数的总数。

CRF 模型通过学习参数 w_j 来最大化给定训练数据的对数似然函数，从而得到最优的参数值，用于预测新的输入序列对应的标签序列。

4 实验

4.1 数据集

本文的实验数据集源自于 CHIP2020 数据集测评任务一———中文医学文本命名实体识别任务[25]。该数据集的语料来源于临床儿科学领域，涵盖了身体(BOD)、疾病(DIS)、临床表现(SYM)、医疗程序(PRO)、药物(DRU)、医学检验项目(ITE)、微生物类(MIC)、医疗设备(EQU)、科室(DEP)共 9 种实体类型。CMeEE 数据集详情如表 2 所列，实体类型分布如表 3 所列。

表 2 CMeEE 数据集详细介绍

Table 2 Detailed introduction of CMeEE dataset

数据集	句子数量/个	平均句子长度/字	平均每句含实体数量/个
训练集	11032	44.24	3.58
验证集	3677	43.99	3.55
测试集	3678	44.53	3.62

表 3 CMeEE 数据集实体类型分布

Table 3 CMeEE dataset entity type distribution

实体类型	训练集	验证集	测试集
身体	11259	3642	3802
疾病	9915	3316	3455
临床表现	8122	2687	2761
医疗程序	4129	1435	1372
药物	2639	841	782
医学检验项目	1571	538	520
微生物类	1224	378	394
医疗设备	555	184	177
科室	182	63	68

实验数据采用 BIO 方式进行标记，并在此数据集上进行了 1-shot，5-shot，20-shot 和 50-shot 的对比实验设置。通过使用这一医学命名实体识别数据集模拟实际应用中的不同数据情景，旨在评估模型在不同样本数量下的性能表现，帮助我们了解在小样本场景下模型的鲁棒性和泛化能力。

4.2 参数设置

本文使用 Pytorch1.10 框架和 Bert 预训练词向量进行实验。同时，实验对模型参数进行了微调，最后使用验证集上性能表现最好的模型对测试集进行评估。

在超参数设置时，Batch size 的取值范围集合为{8，16，32，64}，学习率取值范围集合为{1×10^{-5}，1×10^{-6}，2×10^{-5}}，条件随机场学习率取值范围集合为{1×10^{-3}，1×10^{-2}，2×10^{-3}，2×10^{-2}}，其他参数详情如表 4 所列。

表 4 超参数设置

Table 4 Hyperparameter settings

超参数	取值范围集合
训练批量大小(Batch size)	8，16，32，64
学习率(Learning rate)	1×10^{-5}，1×10^{-6}，2×10^{-5}
条件随机场学习率(CRF Learning rate)	1×10^{-3}，1×10^{-2}，2×10^{-3}，2×10^{-2}
丢弃率(Dropout rate)	0.1，0.2，0.3
双向 LSTM 层数(Number of BiLSTMlayers)	1，2，3

4.3 评估方法

本文使用精确率（Precision）、召回率（Recall）和 F1 值（F1-score）评估实体抽取的质量。

精确率指的是分类器正确识别正例（True Positive）的比例，即真正例（True Positive）除以真正例（True Positive）和假正例（False Positive）之和。其公式为：

$$P=TP/(TP+FP) \tag{21}$$

召回率指的是所有正例中分类器正确识别正例的比例，即真正例（True Positive）除以真正例（True Positive）和假反例（False Negative）之和。其公式为：

$$R=TP/(TP+FN) \tag{22}$$

F1 值是精确率和召回率的加权平均值，其公式为：

$$F1=(2*P*R)/(P+R) \tag{23}$$

4.4 结果与分析

4.4.1 CMeEE 全数据实验结果

在 CMeEE 数据集的全数据集场景下使用 UIE[26]、NegSample[27]以及本文提出的方法在 3 个基线模型上进行对比，详细实验结果如表 5 所列。

表 5 CMeEE 数据集全数据实体识别实验结果

Table 5 Experiment results of full data entity recognition in CMeEE dataset

（%）

模型	P	R	F1
UIE	60.14	65.26	62.60
NegSample	60.25	61.74	60.99
Bert+BiLSTM+CRF	68.42	64.77	66.54
Bert+BiLSTM+CRF w/LK4CMeNER	69.15	65.46	67.25
Bert+Softmax	60.26	64.92	62.50
Bert+Softmax w/LK4CMeNER	60.74	65.51	63.03
Bert-MRC	73.48	74.20	73.84
Bert-MRC w/LK4CMeNER	**74.86**	**75.16**	**75.01**

其中 UIE 是由百度开源的信息抽取大一统模型，在多个数据集上取得了 SOTA 的效果；NegSample 由腾讯 AI 发布，通过使用负采样缓解弱监督 NER 中的未标注实体问题；Bert+BiLSTM+CRF[24]是一种用于自然语言处理的深度学习模型组合，其结合了预训练的 BERT 模型、双向 LSTM 网络和条件随机场，有效提高了对文本的语义理解和序列标注能力；Bert+Softmax 模型[28]是一种基于预训练的 Bert 模型，在顶部添加 Softmax 分类层，用于解决实体识别任务，它通过将 Bert 的输出传入 Softmax 层，将文本映射到不同的类别，实现了实体类型分类；Bert-MRC 模型[29]是一种基于预训练的 BERT 模型，并结合机器阅读理解（MRC）任务的深度学习模型，用于回答自然语言问题并从文本中提取准确的答案。在精确率、召回率和 F1 值等指标上，Bert-MRC w/LK4CMeNER 明显优于其他模型。CMeEE 数据集是专门针对中文医学领域的数据集，其特点在于文本格式具有不规范性，包含大量的专有名词、缩写和简写名词，同时词汇边界相对于其他领域更加模糊。实验结果表明本文提出的方法在全数据集上取得了明显的性能提升。

实验结果表明，在使用本文提出的 LK4CMeNER 方法时，与 3 个基线模型相比，F1 值分别提高了了 0.71%，0.53%，1.17%，这证明了本文方法具有良好的泛化性能。其中，Bert-MRC w/LK4CMeNER 能够取得出色的性能，主要归因于其能够将 LK4CMeNER 方法与 Bert-MRC 模型相结合，通过融合标签和标签释义信息，模型能够学习字符编码蕴含的丰富的语义信息，从而提高命名实体识别的性能水平。

4.4.2 小样本实验结果

表 6 列出了各模型在小样本实验中的性能表现。值得注意的是，当采用本文提出的 LK4CMeNER 方法时，3 个基线模型性能均有明显提升。

表 6 CMeEE 数据集小样本实体识别实验结果

Table 6 Experimental results of few-shot entity recognition in CMeEE dataset

（%）

模型	1-shot F1	5-shot F1	20-shot F1	50-shot F1
UIE	23.14	35.73	39.25	49.36
NegSample	21.67	30.42	36.63	45.62
Bert+BiLSTM+CRF	32.35	45.25	51.72	57.41
Bert+BiLSTM+CRF w/ LK4CMeNER	35.16	**48.34**	53.67	58.95
Bert+Softmax	16.63	21.35	37.23	53.46
Bert+Softmax w/ LK4CMeNER	18.94	24.72	38.34	54.82
Bert-MRC	36.51	42.63	55.43	64.25
Bert-MRC w/ LK4CMeNER	**40.83**	46.36	**58.26**	**65.74**

综合来看，采用 LK4CMeNER 方法时，模型在小样本实体识别中有不同程度的提升。特别是在数据集较少的 1-shot 和 5-shot 情况下，模型提升效果更加显著，而在 20-shot 和 50-shot 情况下，模型仍然能够取得良好的表现。

这一结果验证了将标签和标签释义融入到模型中能够使得模型更好地利用医学领域的先验知识，在医学实体识别任务中获得更好的效果。随着数据数量不断增加，模型识别效果提升幅度随之变缓，这是因为在小样本场景下使用此方法有一定提升效果，随着数据数量的增加，提升效果逐渐减弱。

4.4.3 不同融合方式对实体识别的影响

为了深入探究标签语义信息融合方式对模型性能的影响，我们以 Bert+BiLSTM+CRF 作为基准线，对 4 种融合方式进行了实验：拼接（Cat）、点积（Dot）、自适应融合拼接（Adaptive fusion cat）、自适应融合点积（Adaptive fusion dot）。

我们对模型在不同融合方式下的表现进行了详细的统计分析，结果如表 7 所列。通过比较这些不同方式下的模型性能，我们可以进一步了解融合方式对模型的影响。

表 7 模型在不同融合方式下的结果

Table7 Results of models under different fusion methods

（%）

融合方式	P	R	F1
拼接（Cat）	62.51	63.63	63.07
点积（Dot）	65.28	63.79	64.53
自适应融合拼接（Adaptive fusion cat）	66.04	64.52	65.27
自适应融合点积（Adaptive fusion dot）	**69.15**	**65.46**	**67.25**

根据实验结果可知，相对于拼接方法，使用点积的方法可以显著提高实体识别的性能，表明模型更准确地理解了点积后的向量表示。具体而言，使用拼接方法可能会引入一些噪声，因为它将两个向量的所有维度都进行了简单拼接。

相比之下，点积方法可以避免这种情况，其只保留两个向量对应维度的乘积作为新的向量维度，因此能够更好地消除

噪声影响。此外，点积方法还能更好地捕捉两个向量之间的关系，因为点积可以表示其在相应维度上的相似度或相关性。因此，模型对于点积后的向量表示的理解更准确、合理。

此外，自适应融合点积方法在实验中取得了最好的性能结果，该方法通过调整点积后向量每个维度的权重，使得模型更关注对实体识别有用的信息，并根据不同信息的贡献来调整权重，有效平衡了特征提取模块和语义增强模块的信息。

4.4.4 训练标签语义向量对实体识别的影响

为了验证训练标签向量对模型性能的影响，本研究采用了冻结标签语义向量的方法，即对标签语义向量不进行训练。实验结果如表 8 所列。

表 8 训练标签语义向量的影响

Table 8 Impacts of training label semantic vector

(%)

	P	R	F1
Original	**69.15**	**65.46**	**67.25**
w/o train	68.22	64.57	66.34

通过对标签语义向量进行训练，可以使字符向量和标签向量在表征空间上更加一致，从而使模型能够更好地捕捉实体之间的语义关系，提高实体识别的性能水平。

4.4.5 添加标签释义信息对实体识别的影响

通过实验比较，我们探究了添加标签释义信息对模型性能是否有影响，实验结果如表 9 所列。

表 9 添加标签释义信息的影响

Table 9 Impacts of adding label paraphrase information

(%)

	P	R	F1
Original	**69.15**	**65.46**	**67.25**
w/o label definition	68.07	64.38	66.17

标签释义包含了丰富的医学领域信息，涵盖了实体的定义、属性和上下文等关键信息。在实体识别任务中，LK4CMeNER 方法通过将这些标签释义信息显式地融入到端到端的神经网络中，有效提升了模型的识别能力。

该方法不仅使模型更加注重医学领域专业知识，还从语义层面增强了对实体的识别能力。通过将标签释义信息融入到神经网络，模型能够更准确地理解实体在医学领域中的语义含义，并更好地判断实体的边界和类型，从而使得模型在医学实体识别任务中具备更高的精确性和准确性。

结束语 本研究提出了一种融合标签知识的中文医学命名实体识别方法，旨在应对中文医学领域实体识别任务中的挑战。实验结果表明，相较于传统方法，该方法在实体识别任务中表现出色，具有更高的准确率和召回率。其核心思想是将标签知识显式融入到神经网络中，以更准确地捕捉实体之间的语义关系，从而提升模型的性能。此外，通过自适应融合机制，可以有效平衡特征提取模块和语义增强模块的信息流，进一步提高模型性能。该研究在中文医学领域的命名实体识别任务中取得了明显的效果，并为小样本场景下的实体识别问题提供了一个有效的解决方案。在未来，我们将在其他数据集上验证所提方法的性能并考虑引入其他外部知识。

参 考 文 献

[1] LIN H, LU Y, TANG J, et al. A Rigorous Study on Named Entity Recognition: Can Fine-tuning Pretrained Model Lead to the Promised Land? [C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020: 7291-7300.

[2] LEE J, YOON W, KIM S, et al. BioBERT: a pre-trained biomedical language representation model for biome-dical text mining [J]. Bioinformatics, 2019, 36(4): 1234-1240.

[3] MI F, ZHOU W, CAI F, et al. Self-training improves pre-training for few-shot learning in task-oriented dialog systems[C]// Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2021: 1887-1898.

[4] HA S, KERSNER M, KIM B, et al. Marionette: Few-shot face reenactment preserving identity of unseen targets[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 10893-10900.

[5] WANG Y, YAO Q, KWOK J T, et al. Generalizing from a Few Examples: A Survey on Few-shot Learning[J]. ACM computing surveys (csur), 2020, 53(3): 1-34.

[6] LIUH, TAM D, MUQEETH M, et al. Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning[J]. Advances in Neural Information Processing Systems, 2022, 35: 1950-1965.

[7] OSAHOR U, NASRABADI N M. Ortho-shot: low displacement rank regularization with data augmentation for few-shot learning [C]// Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022: 2200-2209.

[8] SUN Q, LIU Y, CHUA T S, et al. Meta-Transfer Learning for Few-Shot Learning[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019: 403-412.

[9] LUO X, XU J, XU Z. Channel importance matters in few-shot image classification[J]. In International Conference on Machine Learning (PMLR), 2022, 162: 14542-14559.

[10] DIXIT M, KWITT R, NIETHAMMER M, et al. AGA: Attribute-Guided Augmentation[J]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, 35: 7455-7463.

[11] SCHWARTZ E, KARLINSKY L, SHTOK J, et al. Delta-encoder: an effective sample synthesis method for few-shot object recognition[J]. Advances in neural information processing systems, 2018, 31: 2850-2860.

[12] CHEN J, LIU Q, LIN H, et al. Few-shot named entity recognition with self-describing networks[J]. arXiv: 2203.12252, 2022.

[13] LAI P, YE F, ZHANG L, et al. PCBERT: Parent and Child BERT for Chinese Few-shot NER[C]// Proceedings of the 29th International Conference on Computational Linguistics. 2022: 2199-2209.

[14] MA T, JIANG H, WU Q, et al. Decomposed Meta-Learning for Few-Shot Named Entity Recognition [J]. arXiv: 2204.05751, 2022.

[15] WANG J, WANG C, TAN C, et al. SpanProto: A Two-stage Span-based Prototypical Network for Few-shot Named Entity Recognition[C]// Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. 2022: 3466-3476.

[16] LI J, CHIU B, FENG S, et al. Few-shot named entity recognition

via meta-learning[J]. IEEE Transactions on Knowledge and Data Engineering，2020，34(9)：4245-4256.

[17] MA J，BALLESTEROS M，DOSS S，et al. Label semantics for few shot named entity recognition[J]. arXiv：2203.08985，2022.

[18] FRITZLER A，LOGACHEVA V，KRETOV M. Few-shot classification in named entity recognition task[C]// Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing. 2019：993-1000.

[19] LI J，SUN A，HAN J，LI C. A survey on deep learning for named entity recognition[J]. IEEE Transactions on Knowledge and Data Engineering，2020，34(1)：50-70.

[20] CHURCH K W. Word2Vec[J]. Natural Language Engineering，2017，23(1)：155-162.

[21] PENNINGTON J，SOCHER R，MANNING C D. Glove：Global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP)，2014：1532-1543.

[22] DEVLIN J，CHANG M W，LEE K，et al. Bert：Pre-training of deep bidirectional transformers for language understanding[J]. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)，2019，21：4171-4186.

[23] LU W，LI J，WANG J，et al. A CNN-BiLSTM-AM method for stock price prediction[J]. Neural Computing and Applications，2021，33：4741-4753.

[24] LAFFERTY J，MCCALLUM A，PEREIRA F. Conditional random fields：probabillstic models for segmenting and labeling sequence data[C]// Proceedings of the 18th International Conference on Machine Learning. 2001：282-289.

[25] ZAN H Y，LI W X，ZHANG K L，et al. Building a Pediatric Medical Corpus：Word Segmentation and Named Entity Annotation[C]// The 21st Chinese Lexical Semantics Workshop. 2021：652-664.

[26] LU Y，LIU Q，DAI D，et al. Unified structure generation for universal information extraction[C]// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics(ACL). 2022：5755-5772.

[27] LI Y，LIU L，SHI S. Empirical analysis of unlabeled entity problem in named entity recognition[J]. arXiv：2012.05426，2020.

[28] FU Y，LIN N，YANG Z，et al. Towards Malay named entity recognition：an open-source dataset and a multi-task framework [J]. Connection Science，2023：35(1)：2159014.

[29] LI X，FENG J，MENG Y，et al. A unified MRC framework for named entity recognition[J]. arXiv：1910.11476，2019.

YIN Baosheng，born in 1975，professor. His main research interests include deep learning and natural language processing.

ZHOU Peng，born in 1999，postgraduate. His main research interests include natural language processing and named entity recognition.