基于概率分布差异的医学命名实体识别方法

刘聪[1]，吕雪峰[1]，王宏林[1]，王晓伟[2]，陆瑾[2]，孙顺[1]，胡松奇[1]
1. 中国共产党中央军事委员会后勤保障部信息中心，北京 100190；
2. 长沙军民先进技术研究有限公司，湖南 长沙 410205

摘要
医学命名实体识别是从医学文本中抽取出指代特定概念的医学实体，是医学信息抽取的基础性任务。当前主流的医学命名实体识别算法普遍基于深度学习技术，需要大量高质量的标注样本进行模型训练。然而医学领域的样本标注成本很高，严重限制了模型性能的提升。为了降低模型对标注样本的需求，一种重要的方法是基于主动学习思想，设计合理的样本采样策略，自动选取高价值样本优先标注，从而使模型提前收敛。现有算法普遍基于样本长度、样本识别的概率等特征来设计采样策略，忽视了样本类别分布这一深层次特征，导致命名实体识别召回率较低。提出了一种基于概率分布差异的主动学习算法，通过计算样本间的概率分布差异来评估样本的标注价值，并在标注样本更新时动态优化模型。在真实的医学检查文本上的实验表明，相比已有算法，达到同等的模型性能，该算法所需要的标注数据可缩减10%以上；在相同标注样本量的情况下，本算法F1值提高5%以上。

关键词
医学命名实体识别；深度学习；主动学习；概率分布

中图分类号：TP391 文献标志码：A doi: 10.11959/j.issn.2096-0271.2023008

Medical named entity recognition algorithm based on probability distribution difference

LIU Cong[1], LYU Xuefeng[1], WANG Honglin[1], WANG Xiaowei[2], LU Jin[2], SUN Shun[1], HU Songqi[1]
1. Information Center, Logistic Support Department of CMC, Beijing 100190, China
2. Changsha Civi-military Advanced Technology Research Limited Company, Changsha 410205, China

Abstract
With the improvement of data abilities and the development of emerging technologies, there are profound changes occurring in economic patterns and competitive structure of industries. In order to better respond to future opportunities and challenges, and to improve competitiveness of enterprises in new situations, it is necessary to understand and master the knowledge of digital transformation. The new competitive situation was discussed in which traditional enterprises would gradually be replaced by digital-transformed ones, digital transformation was differentiated from

digitalization. Main challenges facing traditional enterprises while undergoing digital transformation were pinpointed, which were the lack of funds, talents, data and consciousness. A digital transformation service platform oriented to new competitive situation was proposed, which provided a feasible solution to enhancing enterprise competitiveness and conducting digital transformation.

Key words
digital transformation, emerging technologies, data asset, digital economy

0 引言

医学命名实体识别旨在从医学病历、医学研究文献等文本中将检查项目、疾病诊断、解剖部位、药品耗材等指代特定概念的医学实体识别出来。医学命名实体识别是医学信息抽取的基础性任务，是医学关系抽取、医学文档分类等技术研究的重要步骤，是医学知识图谱构建、临床病历质控、医保病历核查、临床辅助诊断等下游应用的前提[1-3]。相比通用领域，医学领域的命名实体识别任务更具挑战性。首先，医学领域的文本信息通常是描述性的，实体缺少上下文信息，如“胸部高分辨率”这一检查文本实际上省略了部分信息，单纯通过常规信息抽取技术很难得到“高分辨率扫描”这一检查方法实体；其次，多个实体经常共享同一核心名词，如“1.5T MRI上、中腹部平扫”实际上包含“上腹部”和“中腹部”两个实体，而原始数据中省略了“上腹部”中的部分名词；再次，同一实体通常有多种不同的写法，如“腓骨”和“小腿长骨”属于同一类型的实体，这是医学领域的常见情形；最后，医学实体经常出现缩写形式，如“肺A”表示“肺动脉”，“冠状A”表示“冠状动脉”。

深度学习模型因其数据驱动的上下文编码能力，已经成为命名实体识别任务的主流模型。基于深度学习的医学领域命名实体识别要达到较高的准确率和召回率，通常要收集大量医学领域标注数据对模型进行充分的训练。然而，医学领域数据的标注需要具备一定程度的医学知识，相比通用领域，标注成本较高[4]。一种有效的解决途径是采用基于主动学习[5]的方法，采样少量高价值样本优先标注，在不损失准确率和召回率的前提下，减少模型的迭代次数。已有算法的采样策略普遍基于样本的最大长度[18]、样本的识别概率[19]等特征，主要关注的是已识别实体的准确率低的情况，无法解决召回率低的问题。

针对上述问题，本文提出了基于概率分布差异的主动学习算法。该算法通过比较已标注样本与待标注样本的实体类别分布差异度来量化样本的标注价值，将标注聚焦在低召回率的样本上，并通过循环迭代训练来不断增强模型的性能。在医学检查文本实际数据上的实验表明，相比已有算法，该算法能够显著减少模型收敛所需要的迭代次数以及达到相应模型性能所需要的标注样本量。同时，在相同标注样本量的情况下，该算法能够提升实体识别模型的召回率，从而有效提升模型的总体性能。

1 相关研究

1.1 医学命名体识别

早期的医学命名实体识别主要基于规则

匹配或统计学习的方法。规则匹配[6]是根据实体识别的目标，由人工设置一系列的规则进行实体提取。这类方法在具有先验知识的场景中能够发挥较好的效果，但规则的制定需要投入大量的人工，并且不同的情况需要不同的规则覆盖，自动化程度很低。

基于统计学习的方法[7]通过对序列进行标注，构建文本序列观测模型[8]，如隐马尔可夫模型[9]和条件随机场模型[10]，从大量文本数据中提取出隐藏序列，实现命名实体的识别。基于统计学习的方法依据文本的统计特征进行建模，具有一定的普适性。

随着词的分布式表示技术的出现，深度学习模型因为其优异的性能，逐步成为各领域命名实体识别的主流方法。Lample G等人[11]提出了一种基于双向LSTM和条件随机场（BiLSTM-CRF）的模型，用于医学命名实体识别，该模型基于字符和单词级信息进行建模。Ouyang E等人[12]通过n-gram字符来获取更为丰富的上下文信息作为特征，并融合医学分词信息。Dong X S等人[13]基于迁移学习构建了双向递归神经网络（RNN）模型，该模型利用医学领域数据，在通过通用领域数据训练的模型上进行知识迁移。Zhang Z C等人[14]通过卷积神经网络（CNN）训练汉字的字符级嵌入表示，并将其与大规模医学语料库中获得的预训练字符嵌入到向量组合传，同时引入注意力机制，确定每个时间步骤将特征融合到查询向量中。Wang Q等人[15]在BiLSTM-CRF体系结构的基础上，结合5种不同特征的方案来处理中文医疗数据。Qiu J H等人[16]提出了一种基于条件随机场的残差扩张卷积神经网络，通过将汉字和医学字典特征投射到密集的矢量表示中，将它们输入的残差扩张的卷积神经网络获取上下文特征。Li X Y[17]提出了将预训练词向量作为特征输入的命名实体识别模型BERT-BiLSTM-CRF。

1.2 主动学习

构建深度学习模型通常需要大量的标注数据对模型进行训练，对标注数据的质量要求也较高。近年来，基于主动学习思想降低对标注数据的需求是一个重要的研究方向。主动学习根据筛选数据的流程不同，分为基于流（stream-based）和基于池（pool-based）两种类型[18]。其主要区别是，基于流的主动学习对未标注的样本逐个进行评估，并通过阈值决策来选择样本；基于池的主动学习对整个未标注的样本集中进行评估，然后再通过排序选择最具有价值的样本进行标注。基于流的方式的阈值需要根据不同的任务和不同的数据进行调整，因此在实际场景中应用程度不高[19]。

主动学习对于采样数据质量的优劣的影响，核心取决于采样策略，不同采样策略获得的样本空间不同。目前，主流的采样策略包括随机采样策略、基于熵的采样策略[20]、基于最低置信度的采样策略[21]和边缘采样策略[22]。

- 随机采样策略（random sampling，RS）：从未标注样本中随机选择样本进行人工标注，一般作为基准用来与其他策略进行对照。
- 基于熵的采样策略（maximal entropy，ME）：熵常用来形容一个事件发生所含有的信息量，即事件发生的不确定程度/惊奇程度，熵越大代表不确定性越高。该策略选择熵比较大的样本数据作为待定标注数据。
- 基于最低置信度的采样策略（least confidence，LC）：该策略是在每次迭代过程中选出模型识别把握最低（即概率最小）的样例。该方法通常应用于概率模型，例如在二分类问题中，选择后验概率最接

近0.5的样例进行标注，可以显著提高分类效果。而在多分类问题中，往往选择置信度最低的样例作为标注的依据：

$$x_{\mathrm{LC}}^{*}=\arg\max_{x}(1-P_{\theta}(\hat{y}\mid x)) \quad (1)$$

其中，$\hat{y}=\arg\max_{y}P_{\theta}(y\mid x)$ 表示样例x最有可能的类别 $\hat{y}$。

● 边缘采样策略（margin sampling，MS）：基于最低置信度的采样策略仅仅考虑了最有可能的类别的信息，而忽略了其他类别的信息。为了解决这种问题，出现了边缘采样[18]策略：

$$x_{M}^{*}=\arg\min_{x}(P_{\theta}(\hat{y}_{1}\mid x)-P_{\theta}(y_{2}\mid x)) \quad (2)$$

上述采样策略主要关注的是样本长度、样本的预测概率等特征，没有利用样本分类的整体分布这一细粒度特征。利用这些特征采样得到的训练样本进行模型训练，容易导致模型召回率不够理想，从而影响模型的整体性能。

2 基于概率分布差异的主动学习算法

基于概率分布差异的算法（DS算法）结合了主动学习思想和深度学习技术，利用深度学习技术构建医学命名实体识别模型，基于主动学习思想来设计合理的采样策略，通过样本采样和模型预测的循环迭代，逐步提升模型的预测推理效果。基于概率分布差异的主动学习算法流程见**表1**。

首先，医学专家人工标注少量样本，训练医学命名实体识别深度学习模型作为算法框架中的基本模型；然后，利用基本模型对未标注样本进行推理，算法框架根据预测结果使用基于概率分布差异的样本采样策略，从未标注样本集中选取对模型最有价值的样本，交由医学专家进行标注；最后，将标注后的样本加入已标注样本集中。医学专家标注完成后，将标注后的样本加入新一轮训练集中，再进行新一轮模型训练。通过不断迭代，模型的推理能力不断加强，直到满足停止条件。不同于LC、MS等采样策略，DS算法对于每轮迭代后新增的带标注的数据集L，不只用于模型M的更新，同时重新计算协方差矩阵和均值，优化采样策略函数P，使概率更加趋近真实场景的分布，在学习过程中可以动态优化采样策略函数P的效果。

算法的核心主要是前3个步骤，即样本的标注、医学命名实体识别模型M的训练和采样策略函数P的筛选。对于样本标注，需要针对概率分布差异的计算需求，对常用的标注方法做适应性修改。利用医学命名实体识别模型M对未标注样本进行预测，得到通过模型抽取的实体与各个实体的类别。采样策略函数P根据模型的预测结果，通过计算概率分布差异值来筛选未标注样本。如果样本的差异值较大，说明模型没有准确抽取出当前样本的特征信息，当前样本对增强模型特征抽取能力有更大价值，因此有必要提交标注。

2.1 样本标注

由于DS算法中需要使用医学实体类别的数量来计算概率分布，因此在样本标注的过程中，笔者在传统的BIO形式的命名实体识别标注的基础上进行了改进，在标注结果中加入了实体所属类别的数量。

表2以医学检查项目文本为例，展示了改进后的标注输出格式。其中b、c分别为所属的医学实体类型，即检查部位（body）和检查方法（check）。随后的数值表示部位在文本中的起始位置，字符串表示对

应的实体，最后的数字表示该条样本中该类实体的数量。例如，在b:{[11,12,‘右手’],1}中，11,12表示偏移量为11和12的字符串（即“右手”）为检查部位实体，该类实体在样本中的数量为1。

2.2 医学命名实体识别模型

医学命名实体识别模型基于深度学习技术，抽取医学检查类文本中的医学实体。以医学检查项目文本“CT腓骨扫描+前臂三维”为例，将该文本输入命名实体识别模型中，模型输出为“检查部位”类型的有“腓骨”和“前臂”，输出为“检查方法”类型的有“扫描”和“三维”。

医学命名实体识别模型基于BERT-BiLSTM-CRF模型进行训练，其结构如**图1**所示。采用预训练模型（BERT）[23]作为编码器，对输入文本的每一个字符进行编码，再将编码后的向量输入双向长短时记忆网络（bi-directional LSTM RNN，BiLSTM）[24]中进行特征提取，输出的是每个字符对应的预测标签，然后将其输入条件随机场（conditional random field，CRF）[25]，CRF将会对输入进行条件约束，对输出的分数进行纠正，确保输出的是最优标签序列。

该模型的输入形式被设置为字符级别，这是因为以词级别作为输入时，中文分词可能存在分词错误的问题，而且针对医学文本的中文分词的难度更大，相对于通用领域来说其分词准确率更低，分词错误容易传播到模型后续模块中，影响最终的结果。

2.3 样本采样策略

样本采样策略的目标是选取出最有标

表1 基于概率分布差异的主动学习算法流程

算法 1:基于概率分布差异的主动学习算法流程
输入：未标注的样本集U
Step1：从U中随机抽取部分样本L，通过标注平台A，进行样本标注； Step2：构建实体识别BERT-BiLSTM-CRF模型M，使用现有标注样本训练模型M； Step3：通过采样策略P从未标注样本集合U中筛选出差异值较大的数据集合； Step4：通过标注平台A进行标注，得到标注样本集； Step5：更新标注样本集； Step6：基于更新的样本集L，更新采样策略函数P； Step7：基于更新的样本集L，更新训练模型M； Step8：将更新后的模型M在测试集中验证； if 达到收敛的条件：停止迭代； else：重复step3-step8；
输出：新增后的样本集L，最终训练的模型M

表2 样本标注样例

数据	标注结果
DR摄片（二次曝光）[右手正斜位]	[b:{[11,12,‘右手’],1},c:{[13,15,’正斜位’],1}]
磁共振3.0T平扫	[b:{[],0},c:{[7,8,’平扫’],1}]
头颅神经外科移动CT平扫+三维（神外专用）	[b:{[0,1,”头颅”],1},c:{[9,10,”平扫”],[12,13,”三维”],2}]

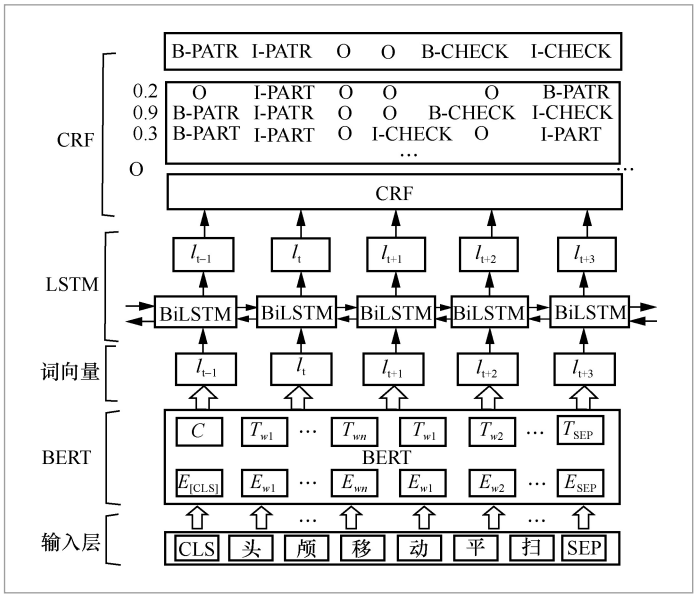

图1 医学命名实体识别模型结构

注价值的样本供人工标注。通过不断迭代训练，可以在减少训练所需数据的同时，保证模型的效果。

基于医学领域先验知识，医学文本中每个样本由不同类别的实体形成一定的分布，实体类别之间存在固有的联系。假设样本类别数据服从高斯分布，通过现有的 n 条已标注的数据集 L 的标注结果，获取每条样本的不同类别的数量 $[x_i^c, x_i^b]$，i 表示样本的标号，x_i^b 是第 i 条样本中出现检查部位类别实体的数量，x_i^c 是第 i 条样本中出现检查方法类别实体的数量，分别统计检查部位和检查方法两类医学实体在每条样本中出现的次数 X^b 和 X^c。

$$X^b = [x_1^b, \cdots, x_n^b] \tag{3}$$

$$X^c = [x_1^c, \cdots, x_n^c] \tag{4}$$

求得现有样本类别数据的高斯概率密度：

$$f(x) = \frac{1}{(\sqrt{2\pi})^n \times |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \times e^{-\frac{(x-\mu_X)^T \times \Sigma^{-1} \times (x-\mu_X)}{2}} \tag{5}$$

其中 μ_X 是均值矩阵，Σ 为协方差矩阵，现有样本类别数据概率密度如**图2**所示。

基于上述统计分析观察，提出了基于概率分布差异的采样策略。具体的样本采样策略描述如下。

首先，根据式（5）获得现有样本类别数据的高斯概率密度 $f(x)$，为更好地描述标签分布的异常程度，将其做进一步的归一化处理，得到样本采样函数 $P(x)$：

$$P(x) = \frac{f(\mu_X) - f(x)}{f(\mu_X)} = 1 - e^{-\frac{(x-\mu_X)^T \times \Sigma^{-1} \times (x-\mu_X)}{2}} \tag{6}$$

$P(x)$ 将差异值压缩在 $[0,1]$ 范围内，值越大说明未标注的数据类别分布与已有的数据集分布差异越大，越应值得关注。

下面以医学检查文本的实际数据为例，解释样本采样策略的具体计算过程。标注1 000条样本组成的数据集 L，再根据样本采样策略中描述的计算方法，分别得到检查部位和检查方法两类医学实体的向量 $\boldsymbol{X}^b$、$\boldsymbol{X}^c$。这两个1 000维向量分别表示对应样本中该医学实体出现的次数：

$$\boldsymbol{X}^b = [1,2,3\cdots,2,1]$$

$$\boldsymbol{X}^c = [1,1,3,\cdots,1,1]$$

然后计算出均值 $\boldsymbol{\mu}_X$ 和协方差矩阵 $\boldsymbol{\Sigma}$ 作为采样函数 $P(x)$ 的关键参数：

$$\boldsymbol{\mu}_X = [1.55, 1.52]$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1.023 & 0.961 \\ 0.961 & 1.025 \end{bmatrix}$$

从而得到了具体的采样函数 $P(x)$：

$$P(x) = \frac{f(\mu_X) - f(x)}{f(\mu_X)} = 1 - e^{-\frac{(x-\mu_X)^T \times \Sigma^{-1} \times (x-\mu_X)}{2}} \tag{7}$$

图3展示了基于上述1 000条样本的数据集 L 得到的概率分布函数。X、Y 轴分别对应了样本预测结果中检查部位（body）和检查方法（check）的类别个数，Z 表示 $P(x)$ 的值。Z 值描述了样本的概率分布差异，越与现有的已标注样本结果分布相似，Z 值越小。越偏离已标注的样本结果分布情况，Z 值越大，说明该样本与已标注的数据集差异大，更值得人工标注后提供给模型学习，提升模型的性能。

假设当前需要比较样本 s_1 和 s_2 的标注价值：

$$s_1 = \text{“胸部冠状面成像加收”}$$

$$s_2 = \text{“加+胸骨侧位”}$$

通过已训练的模型 M，输出的结果是：

$$p_1 = [b:\{[0,1,\text{“胸部”}],1\}, c\{[\ ],\ 0\}]$$

$$p_2 = [b:\{[2,3,\text{“胸骨”}],1\}, c\{[4,5,\text{“侧位”}],1\}]$$

通过采样策略，将s_1和s_2对应的向量$\boldsymbol{x}_1$和$\boldsymbol{x}_2$（$\boldsymbol{x}_1$=[1,0]，$\boldsymbol{x}_2$=[1,1]）输入函数$P(x)$中得到：

$$P(x_1)=0.988$$

$$P(x_2)=0.166$$

由于s_1的概率分布差异值$P(x_1)$显著大于s_2的概率分布差异值$P(x_2)$，因此标注s_1给模型带来的提升更大，DS算法将优先采样s_1，并对其进行标注。

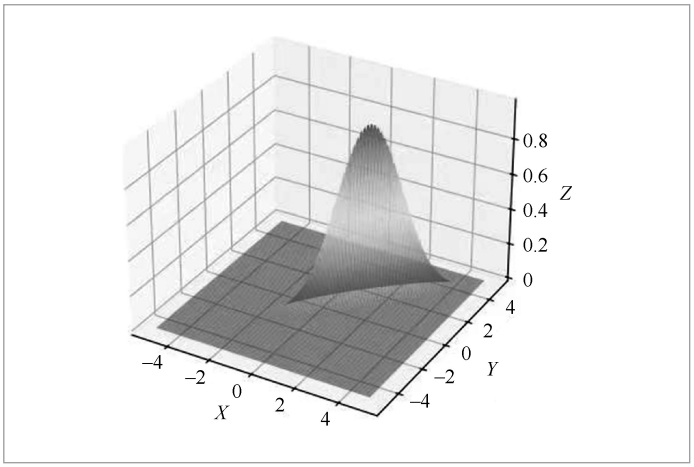

图 2 现有样本类别数据概率密度

3 实验与分析

3.1 实验数据

数据来源于某地级市30家二级以上医院2个月产生的影像科检查项目开单清单文本数据，经过数据清洗、脱敏、标注后共获取检查类文本数据28 000条，其中来源于DR、CT、MR科室的数据分布是均等的。为了验证决策算法的有效性，在全量的数据中通过科室类型等比抽取15 000条数据作为训练语料库，抽取5 000条作为验证集，剩余8 000条作为测试集。**图4**展示了部分实验所用的真实数据。其中一行表示一个独立的样本，item_name列是需要抽取实体的文本，实体类型分为两类：检查部位和检查方法。

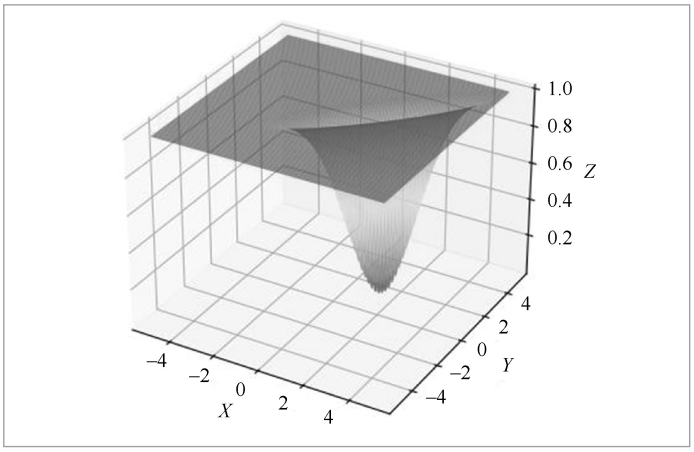

图 3 采样函数 P(x)

3.2 评价指标

对于实体级别的指标，针对预测抽取的实体和实际标注的实体进行计算。分别计算准确率（precision）、召回率（recall）和F1值，其中F1值是对准确率和召回率进行综合评价的指标。

$$\text{precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{8}$$

$$\text{recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{9}$$

$$\text{F1}=\frac{2\times\text{precision}\times\text{recall}}{\text{precision}+\text{recall}} \tag{10}$$

3.3 实验结果

3.3.1 不同采样策略对比

本实验训练模型采用BERT-BiLSTM-CRF模型，主要针对不同的采样策略进行对比，包括随机采样策略（RS）、

医院		item_name	检查代码（院内吖	dep
省中医	76	(右侧)DR肩关节岗上肌出口位片	2011231141565050	DR
省中医	65	（右侧）钼靶引导下的乳腺穿刺	2011231113456194	DR
省中医	67	(右侧)乳腺穿刺标本钼靶摄影	2011231114190414	DR
省中医	63	(右侧)乳腺钼靶片(轴位＋斜位)	2011231112582666	DR
省中医	265	(右侧)上肢动脉CTA	2011231727239092	CT
省中医	73	(左侧)DR肩关节岗上肌出口位片	2011231141211336	DR
省中医	64	（左侧）钼靶引导下的乳腺穿刺	2011231113308598	DR
省中医	66	(左侧)乳腺穿刺标本钼靶摄影	2011231113591167	DR
省中医	62	(左侧)乳腺钼靶片(轴位＋斜位)	2011231112304865	DR
省中医	266	(左侧)上肢动脉CTA	2011231727391034	CT
省中医	452	1.5T(右侧)肩胛骨磁共振平扫	2011250958597024	MR
省中医	460	1.5T（右侧）肩胛骨磁共振平扫+增强	2101061223330205	MR
省中医	606	1.5T(右侧)肩胛骨磁共振增强扫描	2011251002541792	MR
省中医	595	1.5T（右侧）上肢动脉血管造影（CEMRA）	2012221155450014	MR
省中医	451	1.5T(左侧)肩胛骨磁共振平扫	2011250958436505	MR

图 4　部分实验所用的真实数据

基于熵的采样策略（ME）的采样策略、基于最低置信度的采样策略（LC）、边缘采样策略（MS）和本文提出的基于概率分布的采样策略（DS），其中RS策略用于表示同等数据集下基础模型性能。

模型训练批大小为32，LSTM的隐藏层维度为300，采用Adam[26]作为优化器，初始学习率设为0.0001；从训练语料库中选取相同ID的1 000条医学检查类样本作为不同策略的初始训练集。模型训练实验经历10轮迭代，每一轮迭代训练集中增加使用各自的采样策略获取的1 000条样本，不同样本数量的训练集迭代所产生的模型在测试集上进行测试得到相应的指标：准确率、召回率、F1值。

DS算法与其他4种算法的准确率对比结果如**图5**所示，当达到75%的准确率时，RS算法的随机采样基础模型需要标注8 000余条样本，而DS算法的采样策略基于模型对样本的预测结果来选择样本，只需要标注5 000余条样本即可。

DS算法与其他4种算法召回率对比结果如**图6**所示，当达到80%的召回率时，其他算法的采样策略都需要标注超过9 000条样本，而DS算法的采样策略更加关注模型预测的低召回率样本，只需要标注7 000余条样本即可。

DS算法与其他4种算法F1值对比结果如**图7**所示，DS算法的采样策略筛选出7 000条样本时，其F1值达到了85%，而常用的LC、MS、ME算法的采样策略在筛选出近9 000条样本时才达到同样的效果。

从实验结果的对比图可以看出，采用不同的采样策略，3个指标呈现出了很明显的差异。从数据上来看，DS算法在召回率和F1值收敛的效果上，明显优于其他的策略。相比于随机采样，DS算法的数据量为4 000条时，模型的F1值就达到了随机采样策略9 000条数据的效果。在数据量为7 000条的情况下，DS算法的F1值达到了其他常用算法使用9 000条的数据训练得到的效果。

3.3.2 动态更新算法效果对比

为了进一步验证DS算法的采样策略动态更新的效果，将算法1中的第6步去除，作为不使用动态更新的DS算法版本（DS_），用来对比动态更新算法的效果。

动态更新算法效果对比结果如**图8**所示，采用同样的采样策略，通过样本的更新来动态更新采样策略的参数值，能够有效地提升采样策略的效果。

3.3.3 公共数据集上效果对比

为了进一步验证本文提出方法的普适性，本文采用中文医学语言理解测评[27]（CBLUE）的数据集进行测试，在实体识别任务上，一共分为9类实体，包括：疾病(dis)、临床表现(sym)、药物(dru)、医疗设备(equ)、医疗程序(pro)、身体(bod)、医学检验项目(ite)、微生物类(mic)、科室(dep)。其中训练语料库有15 000条样本，测试集有5 000条样本，验证集有3 000条样本。

同样地，本实验先随机从训练语料库中选取1 000条数据作为初始的训练集，之后不同的采样策略通过各自的算法从训练语料库中选取1 000条数据加入训练集中，均采用BERT-BiLSTM-CRF模型迭代训练，模型训练批大小为64，LSTM的隐藏层维度为300，采用Adam[26]作为优化器，初始学习率设为0.001；实验总共进行10次迭代。因考虑到多分类以及不同实体类别数据分布不均衡的情况，测试集的评价指标采用Micro-F1来衡量。

（1）先计算出所有类别总的准确率和召回率：

$$\text{Precision}_{\text{micro}}=\frac{\sum_{i=1}^{n}\text{TP}_i}{\sum_{i=1}^{n}\text{TP}_i+\sum_{i=1}^{n}\text{FP}_i} \quad (11)$$

$$\text{Recall}_{\text{micro}}=\frac{\sum_{i=1}^{n}\text{TP}_i}{\sum_{i=1}^{n}\text{TP}_i+\sum_{i=1}^{n}\text{FN}_i} \quad (12)$$

（2）然后利用F1公式计算出来的F1值即为Micro-F1：

$$\text{Micro-F1}=2\cdot\frac{\text{Precision}_{\text{micro}}\cdot\text{Recall}_{\text{micro}}}{\text{Precision}_{\text{micro}}+\text{Recall}_{\text{micro}}} \quad (13)$$

其中，式（11）、（12）中i代表不同类别，n代表类别总数。

DS算法与其他4种算法Micro-F1值对比结果如图9所示，采用RS算法的模型作为基础性能模型，在Micro-F1值达到55%时，RS算法需要标注8 000条样本，而DS算法只需要标注6 000条样本，其他的决策方案效果也明显低于DS算法。因此，在不同的数据集和任务上，基于概率分布差异的主动学习算法在医学命名实体识别上具有一定的普适性。

为进一步验证本文方法的有效性，选取当前几种常用的模型方法进行对比讨论。ALBERT模型[28]提出两种减少模型参数的方法：一是对嵌入层的参数进行分解，二是层间参数共享。该模型使用了一种自监

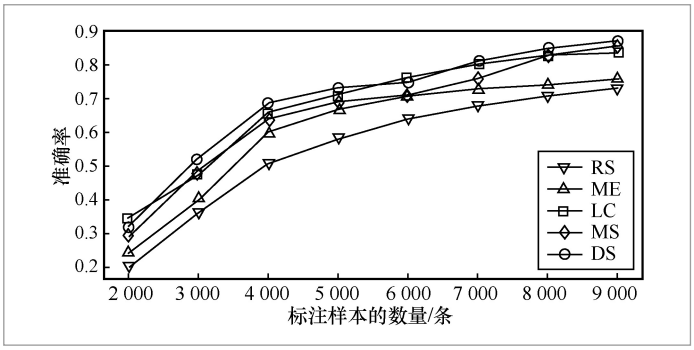

图 5 DS 算法与其他 4 种算法的准确率对比

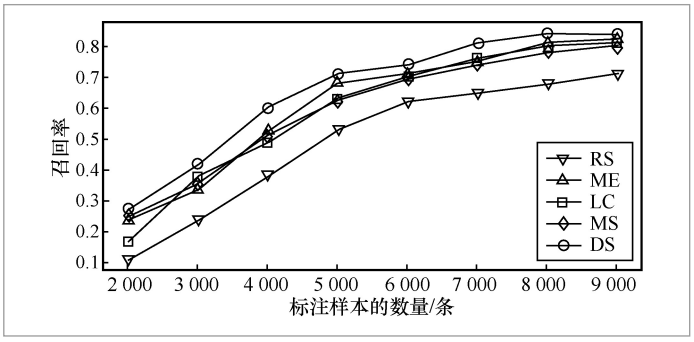

图 6 DS 算法与其他 4 种算法召回率对比

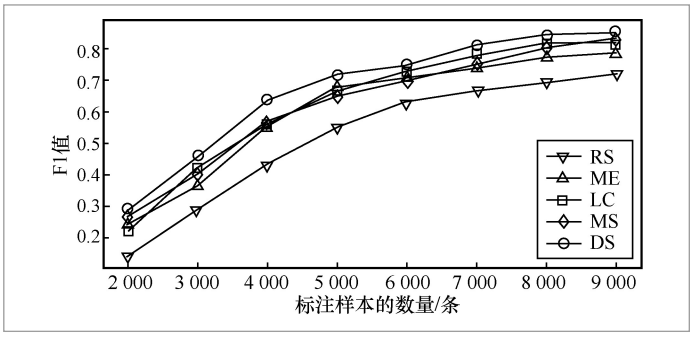

图 7 DS 算法与其他 4 种算法 F1 值对比

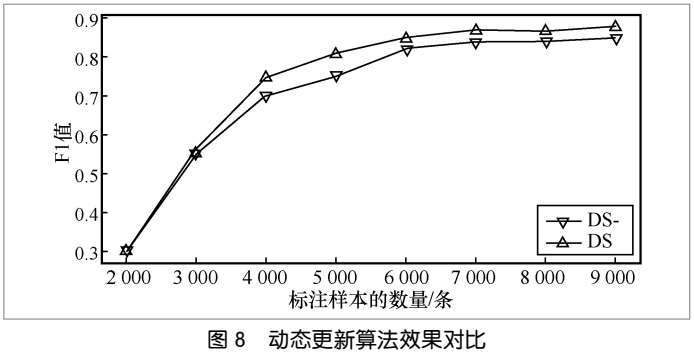

图 8 动态更新算法效果对比

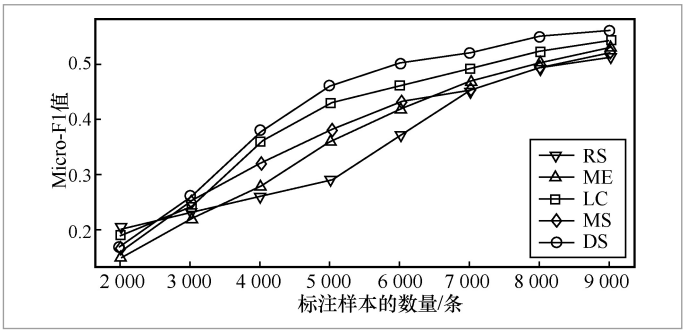

图 9 DS 算法与其他 4 种算法 Micro-F1 值对比

督的损失方法，其侧重于句子间连贯性的建模，并表明它对多句子输入的下游任务始终有帮助，本文选择其base模型及xxlarge模型进行对比；ZEN模型[29]通过在编码层融合N-gram表征，增强模型对N-gram或词语边界的感知能力，从而提升模型的文本表征能力。不同算法的训练参数设置见**表3**。不同算法模型在CBLUE数据集上训练的结果见**表4**，本文提出的基于概率分布差异的医学命名实体识别方法，在训练样本数量更少的条件下，仍然可以达到与其他算法相近的效果，表明DS算法能有效提高模型效果。

上述实验结果表明，相比于基于主动学习的算法，DS算法在样本的采样策略上更加关注模型漏检、少检的样本，通过加入这类数据，能够明显提升召回率和F1值。通过新增标注样本对采样策略函数P进行动态更新，能够持续提升模型的医学实体识别性能。此外，在公开数据集上的测试结果也显示出了决策算法的普适性。

4 结束语

本文针对医学命名实体识别任务，提出了一种基于概率分布差异的主动学习算法。该算法的核心基于概率分布差异的采样策略。该策略更关注模型低召回的样本，通过计算待选样本与已有样本集在概率分布上的差异，动态评估样本对模型的价值，从而利用较少标注语料取得较好的识别效果。此外，该算法还通过样本对模型的动态更新，进一步提升了模型的性能。该算法可有效减少医学命名实体识别任务对标注语料的依赖，挑选出更具有学习价值的样本，从而降低标注成本，提高医学命名实体识别算法的综合性能。

相比通用领域的深度学习任务，在医学领域中通常存在数据量大、数据杂乱和标注成本高的问题。因此通过基于概率分布的采样策略，可以在未标注的数据中定位到最具有价值的样本进行标注，相比于随机挑选数据进行标注的策略，大大减少了人工标注的成本和时间。

医学命名实体识别的复杂性体现在医学文本背后的深层次医学含义更丰富，其难点在于如何在模型中融入医学领域知识。未来可以考虑数据驱动和知识驱动融合的技术路线，把医学领域知识图谱集成到医学命名实体识别的算法流程中，一方面通过基于主动学习的数据标注方法将医学领域知识沉淀到知识图谱中，另一方面利用知识图谱针对医学文本的缩写、省略、同义词等现象进行替换和补全，从而提高医学命名实体识别的准确率和召回率。

参考文献:

[1] 杨威，刘艳如，孟颖，等. 浅谈临床医学术语的标准化管理[J]. 中国卫生标准管理，2021，12(12)：1-4.
YANG W, LIU Y R, MENG Y, et al. Discussion on standardization management of clinical medical terminology[J]. China Health Standard Management, 2021, 12(12): 1-4.
[2] 赵嘉莹，高鹏，朱勇俊，等. 人工智能的应用将改进中国基层医疗卫生服务效能[J]. 中国全科医学，2017，20(34)：4219-4223.

ZHAO J Y, GAO P, ZHU Y J, et al. The application of artificial intelligence could improve primary health care provision in China[J]. Chinese General Practice, 2017, 20(24): 4219-4223.

[3] 曾晓天, 徐春园, 张勇, 等. 人工智能在医学大数据标准化体系建设中的研究进展[J]. 北京生物医学工程, 2019, 38(6): 639-643.
ZENG X T, XU C Y, ZHANG Y, et al. Research progress on artificial intelligence in the standardization system construction of medical big data[J]. Beijing Biomedical Engineering, 2019, 38(6): 640-644.

[4] 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, 832.
ZHENG Q, LIU Q J, WANG Z H, et al. Research and development on biomedical named entity recognition[J]. Application Research of Computers, 2010, 27(3): 811-815, 832.

[5] SETTLES B. Active learning literature survey[J]. Machine Learning, 2010, 15(2): 201-221.

[6] HANISCH D, FUNDEL K, MEVISSEN H T, et al. ProMiner: rule-based protein and gene entity recognition[J]. BMC Bioinformatics, 2005, 6(Suppl 1): S14.

[7] 刘一佳, 车万翔, 刘挺, 等. 基于序列标注的中文分词, 词性标注模型比较分析[C]//第六届全国青年计算语言学会议论文集. [出版者不详:出版地不详], 2012: 26-34.
LIU Y J, CHE W X, LIU T, et al. A comparison study of sequence labeling methods for Chinese word segmentation, POS tagging models[C]//The 6th Youth Conference of Computational Linguistics. [S.l.:s.n.], 2012: 26-34.

[8] 王浩畅, 赵铁军. 基于SVM的生物医学命名实体的识别[J]. 哈尔滨工程大学学报, 2006, 27(S1): 570-574.
WANG H C, ZHAO T J. SVM-based biomedical Name entity recognition[J]. Journal of Harbin Engineering University, 2006, 27(S1): 570-574.

表 3 不同算法训练参数设置

模型算法	轮次	批大小	初始学习率
ALBERT-base[28]	10	32	5×10^{-5}
ALBERT-xxlarge[28]	10	12	1×10^{-5}
ZEN[29]	10	20	4×10^{-5}
BiLSTM-CRF[15]	10	32	3×10^{-4}
本文方法	10	32	1×10^{-4}

表 4 不同算法模型在 CBLUE 数据集上训练的结果对比

模型算法	性能指标(F1值)	训练集样本数/条
ALBERT-base[28]	58.5	15 000
ALBERT-xxlarge[28]	61.7	15 000
ZEN[29]	60.9	15 000
BiLSTM-CRF[15]	57.6	15 000
本文方法	61.6	10 000

[9] MORWAL S, CHOPRA D. NERHMM: a tool for named entity recognition based on hidden Markov model[J]. International Journal on Natural Language Computing, 2013, 2(2): 43-49.

[10] PATIL N, PATIL A, PAWAR B V. Named entity recognition using conditional random fields[J]. Procedia Computer Science, 2020, 167: 1181-1188.

[11] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: Association for Computational Linguistics, 2016.

[12] OUYANG E, LI Y X, JIN L, et al. Exploring N-gram character presentation in bidirectional RNN-CRF for Chinese clinical named entity recognition[C]//

Proceedings of China Conference on Knowledge Graph and Semantic Computing 2017. [S.l.:s.n.], 2017.

[13] DONG X S, CHOWDHURY S, QIAN L J, et al. Transfer bi-directional LSTM RNN for named entity recognition in Chinese electronic medical records[C]// Proceedings of 2017 IEEE 19th International Conference on e-Health Networking, Applications and Services. Piscataway: IEEE Press, 2017: 1-4.

[14] ZHANG Z C, ZHANG Y, ZHOU T. Medical knowledge attention enhanced neural model for named entity recognition in Chinese EMR[C]//Proceedings of China National Conference on Chinese Computational Linguistics, International Symposium on Natural Language Processing Based on Naturally Annotated Big Data. Cham: Springer, 2018: 376-385.

[15] WANG Q, XIA Y H, ZHOU Y M, et al. Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition[J]. Journal of Biomedical Informatics, 2019, 92: 103133.

[16] QIU J H, WANG Q, ZHOU Y M, et al. Fast and accurate recognition of Chinese clinical named entities with residual dilated convolutions[C]//Proceedings of 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Piscataway: IEEE Press, 2019: 935-942.

[17] LI X Y, ZHANG H, ZHOU X H. Chinese clinical named entity recognition with variant neural structures based on BERT methods[J]. Journal of Biomedical Informatics, 2020, 107: 103422.

[18] 张岑芳. 基于主动学习的命名实体识别算法[J]. 计算机与现代化, 2021(7): 18-22.
ZHANG C F. Named entity recognition algorithm based on active learning[J]. Computer and Modernization, 2021(7): 18-22.

[19] 卢宁杰. 结合主动学习的中文医疗命名实体识别研究[D]. 上海: 华东师范大学, 2020.
LU N J. Research on Chinese medical named entity recognition combined with active learning[D]. Shanghai: East China Normal University, 2020.

[20] SHANNON C E. A mathematical theory of communication[J]. Bell System Technical Journal, 1948, 27(4): 623-656.

[21] LEWIS D D, CATLETT J. Heterogeneous uncertainty sampling for supervised learning[M]//Machine learning proceedings 1994. Amsterdam: Elsevier, 1994: 148-156.

[22] SCHEFFER T, DECOMAIN C, WROBEL S. Active hidden Markov models for information extraction[M]//Advances in intelligent data analysis. Heidelberg: Springer, 2001: 309-318.

[23] DEVLIN J, CHANG M, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint, 2018: arXiv: 1810.04805.

[24] GRAVES A, SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, 18(5/6): 602-610.

[25] SUTTON C. An introduction to conditional random fields[J]. Foundations and Trends® in Machine Learning, 2012, 4(4): 267-373.

[26] KINGMA D P, BA J. Adam: a method for stochastic optimization[J]. arXiv preprint, 2014, arXiv: 1412.6980.

[27] ZAN H Y, LI W X, ZHANG K L, et al. Building a pediatric medical corpus: word segmentation and named entity annotation[M]//Lecture notes in computer science. Cham: Springer, 2021: 652-664.

[28] LAN Z, CHEN M, GOODMAN S, et al. ALBERT: a lite BERT for self-supervised learning of language representations[J]. arXiv preprint, 2019, arXiv: 1909.11942.

[29] DIAO S Z, BAI J X, SONG Y, et al. ZEN: pre-training Chinese text encoder enhanced by N-gram representations[C]//

Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg: Association for Computational Linguistics, 2020.

作者简介

刘聪(1985-),男,博士,中国共产党中央军事委员会后勤保障部信息中心工程师,主要研究方向为医疗卫生大数据、医疗卫生信息化。

吕雪峰(1979-),男,中国共产党中央军事委员会后勤保障部信息中心高级工程师,主要研究方向为医疗卫生大数据、医疗卫生信息化。

王宏林(1988-),男,中国共产党中央军事委员会后勤保障部信息中心工程师,主要研究方向为后勤信息化。

王晓伟(1980-),男,博士,长沙军民先进技术研究有限公司高级工程师,主要研究方向为自然语言处理、大数据。

陆瑾(1993-),男,长沙军民先进技术研究有限公司工程师,主要研究方向为自然语言处理、人工智能。

孙顺(1980-),男,中国共产党中央军事委员会后勤保障部信息中心工程师,主要研究方向为卫生信息化。

胡松奇(1988-),男,中国共产党中央军事委员会后勤保障部信息中心工程师,主要研究方向为卫生信息化。

收稿日期: 2022-09-07
通信作者: 吕雪峰, 49374582@qq.com
基金项目: 军队后勤科研重点项目(No.BS220R007)
Foundation Item: Key Program of Scientific Research of Army Logistics (No.BS220R007)