基于混合式迁移学习的命名实体识别算法

余肖生 张合欢 陈 鹏*
(三峡大学计算机与信息学院 湖北 宜昌 443002)

摘 要 针对命名实体识别领域中大量标注数据难于获取而带来的问题,提出基于混合式迁移学习的命名实体识别算法——MT-NER。利用样本之间的距离作为权衡样本相似性的标准,进行样本迁移以扩充目标域样本;利用模型迁移建立带有 finetune 的新命名实体识别网络结构,用扩充后的目标域数据集来训练网络。以医疗领域为例的实验结果分析表明,MT-NER 算法在小样本数据中的实体识别效果最佳,精度达到 93.31%,召回率达到 89.5%,F1 值达到 0.931 7,与 BiLSTM-CRF 模型相比分别提升了 6.33 百分点、3.65 百分点和 0.089 1。

关键词 命名实体识别 迁移学习 双向 LSTM-CRF 分布自适应

中图分类号 TP3 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2024.08.044

NAMED ENTITY RECOGNITION ALGORITHM BASED ON MIXED TRANSFER LEARNING

Yu Xiaosheng Zhang Hehuan Chen Peng*
(*College of Computer and Information, China Three Gorges University, Yichang 443002, Hubei, China*)

Abstract In the field of named entity recognition, it is difficult to obtain a large number of labeled data. To solve this problem, this paper proposes a named entity recognition algorithm based on mixed transfer learning named MT-NER. The distance between the samples was used as the criterion to balance the similarity of the samples, and the instances-based transfer learning was carried out to expand the target domain samples. A new named entity recognition network structure with finetune was established by the models-based transfer learning, and the expanded target domain data set was used to train the network. Taking the medical field as an example, experiments show that MT-NER algorithm has the best effect in entity recognition in small sample data, with an accuracy of 93.31%, a recall rate of 89.5% and a F1 value of 0.931 7. Compared with the BiLSTM-CRF model, the accuracy, recall rate and F1 value of MT-NER are improved by 6.33, 3.65 and 8.91 percentage points.

Keywords Named entity recognition Transfer learning Bidirectional LSTM-CRF Distribution adaptation

0 引 言

命名实体识别(Named Entity Recognition,NER)技术旨在从非结构化文本信息中提取出具有特定意义的实体以及识别出实体所属的预定义类别,是自然语言处理领域中不可或缺的重要基础性任务,也是诸如机器翻译、关系提取、情感评价和知识图谱构建等众多自然语言处理任务的基础工作。

早期的 NER 方法包括基于规则的方法、基于统计机器学习的方法。基于规则的方法主要通过人为地制定规则来实现,而基于统计机器学习的方法主要通过特征选择、模型改进等方法来实现[1]。近年来,随着深度学习成为机器学习的新领域,不少学者尝试使用深度学习技术来解决 NER 问题。尽管基于深度学习的 NER 方法取得了较好的效果,但是在实际应用中,获取足够的训练数据是非常困难的,训练数据的匮乏会导致深度学习的学习效果不佳。迁移学习能够将已经学习过的知识迁移并引用到新的问题中,目的是利用已经在大量数据中学好的知识来提高目标任务的性能,其已成为解决数据集规模较小这一问题的重要方法[2]。

本文提出基于混合式迁移学习的命名实体识别算

收稿日期:2021-04-27。国家重点研发计划项目(2016YFC0802500)。**余肖生**,副教授,主研领域:自然语言处理,健康医疗大数据技术。**张合欢**,硕士生。**陈鹏**,教授。

法 MT-NER。该算法在命名实体识别模型 BiLSTM-CRF 中引入两种迁移学习方式:样本迁移和模型迁移。对于样本迁移,通过计算源域样本相对于目标域样本的相似度来权衡样本之间的权值大小,权值按降序排列后,由最佳迁移数来确定 k 个最相似的样本,即最终的迁移样本。对于模型迁移,首先利用大规模的源域样本来训练 BiLSTM-CRF 模型,得到性能较优的预训练模型,保存其参数特征,然后使用经过样本迁移扩充后的新目标域数据作为训练数据,利用预训练模型的参数初始化新的 BiLSTM-CRF 模型,采用 finetune 技术调整参数,并在损失函数中引入数据分布自适应项。MT-NER 算法混合两种迁移学习,能够更好地将源域中学习到的知识迁移至目标域中,样本迁移中最佳迁移数的设定能够动态地得到迁移效果的反馈,可以防止一定程度的负迁移现象;模型迁移中使用了 finetune 技术,能够帮助模型适应文本域的变化,并且数据分布自适应的设定能够解决数据集中训练集和测试集分布不一致的问题。实验结果显示,相比基线,本文方法具有更好的识别效果,能够在一定程度上解决因训练数据匮乏导致的命名实体识别效果不佳的问题。

1 相关工作

1.1 命名实体识别

命名实体识别方法包括基于规则的方法、基于统计机器学习的方法、基于深度学习的方法。基于规则的方法通过人为制定规则来实现,比如制定特定领域词典、句法词汇模板和正则表达式等。已有的基于规则的命名实体识别系统有:LaSIE-II、NetOwl、Facile、SAR、FASTUS 和 LTG。当词汇表规模足够大时,基于规则的方法能够取得满意的效果,但是人为地总结规则模板需要花费大量的时间与精力,并且当词汇表的规模达不到要求时,就会普遍降低实体识别的召回率。

随着机器学习的兴起,研究者开始使用统计机器学习的方法来实现命名实体识别。这种方法将命名实体识别归类为分类方法或者序列化标注方法,为了提高命名实体的识别效果,研究人员从特征选择、模型改进等研究方面着手,其中常使用的模型包括:隐马尔可夫模型(Hidden Markov Model,HMM)、条件随机场(Conditional Random Field,CRF)、支持向量机(Support Vector Machine,SVM)、混合 HMM 和混合多个 SVM 等。

基于深度学习的方法一般分为三步:分布式表示、特征提取和标签解码,该方法的特点在于利用网络的独特结构对数据特征进行提取。常采用的深度学习方法包括 RNN(LSTM 和 GRU)、CNN 等。为了解决命名实体识别的问题,研究者提出了不同网络结构的模型,比如 Lattice 结构的 LSTM 模型[3]、混合双向 LSTM 和 CNN 的网络结构[4]、BiLSTM-CRF 模型[5]等,这些模型都从修改网络结构着手,提高了模型的识别能力,其中 BiLSTM-CRF 模型是目前主流的命名实体识别模型。

1.2 迁移学习

迁移学习的形式化定义如下:

条件:给定一个有标签的源域 D_S 和源域上的学习任务 T_S,无标签的目标域 D_T 和目标域上的学习任务 T_T。

目标:利用 D_S 和 T_S 的知识来学习在目标域上的预测函数 f。

限制条件:$D_S \neq D_T$ 或 $T_S \neq T_T$。

迁移学习即将已解决问题的方法应用到待解决问题中,将已存在的知识方法迁移解决相关领域的学习任务,属于一种新的机器学习方法[6]。传统的迁移学习方法分为基于实例的迁移学习、基于特征的迁移学习和基于模型的迁移学习等三类,可以在样本数量过少的情况下有效避免过度拟合[7]。基于实例的迁移学习是指借助丰富的源域数据来扩充目标域数据,通过对源域数据的实例加权实现。比如核均值匹配法(Kernel Mean Matching,KMM)[8]通过再生核希尔伯特空间(Reproducing Kernel Hilbert Space,RKHS)计算源域和目标域实例之间的均值,估计样本的概率分布,然后通过计算源域与目标域之间的分布比值为样本分配权重,使得加权后的源域和目标域之间的概率分布尽可能相近;混合迁移学习的 Adaboost 算法[9]通过提高可用源域实例的权重,降低不可用源域实例的权重来扩展目标域,两者都利用了基于实例的迁移学习来扩展数据。基于实例的迁移学习实现相对简单,能达到一定的正向迁移效果,但存在着对权重选择、相似度度量选择的依赖。

基于特征的迁移学习旨在采用特征变换的形式,将两个域的数据映射到同一特征空间中,从而实现特征的迁移。迁移成分分析(Transfer Component Analysis,TCA)[10]试图减小源域和目标域的距离,将两个领域的数据一起映射到一个高维的再生核希尔伯特空间,在此空间中,最小化源与目标的数据空间距离,同时最大程度地保留它们各自的内部属性。大多数方法会采用基于特征的迁移学习,因为特征的变换可以解决数据分布不同的问题,但在实际应用中容易发生过适配的现象。

基于模型的迁移学习是指构建参数共享的模型,利用源域与目标域之间的相似性将已经训练好的模型迁移到目标任务上,比如基于 TrAdaBoost 的

TaskTrAdaBoost 模型[11]、域适应机框架(Domain Adaptation Machine,DAM)[12]等,通过预训练分类器来学习目标域中的实例标签预测分类器。

2　算法设计

由于迁移学习能够很好地从相关领域中迁移学习好的知识,可以在一定程度上解决命名实体识别领域中标注数据的匮乏问题。目前已有研究者从迁移学习的角度对命名实体识别进行研究[13-17],在训练数据缺乏的情况下通过不同形式的迁移学习获得了较好的命名实体识别效果。

以上方法都在一定程度上解决了因训练语料库不足而造成的命名实体识别模型性能不佳的问题,但在实际迁移的过程中往往面临着概念漂移和负迁移问题[18]。

本文提出基于混合式迁移学习的命名实体识别算法,在训练数据匮乏情况下使用迁移学习提高命名实体识别效果,并在一定程度上解决概念漂移和负迁移问题。MT-NER 算法引入混合式迁移学习技术:样本迁移(图 1(a)部分)和深度网络模型迁移(图 1(b)部分)。混合式迁移学习能够更好地将源域中学习到的知识迁移到目标域。在样本迁移过程中,MT-NER 算法利用不同的距离来计算样本之间的相似度 D,在有序相似矩阵 $\boldsymbol{M}$ 中选取前 k 个样本作为迁移样本,通过实验调整最佳的样本迁移数 k,这样可以降低一定程度的负迁移现象;在深度网络模型迁移过程中,首先在相似的源域数据集上进行训练,得到预训练模型,保留其参数 W,用 W 初始化目标域模型参数。目标域模型在源域模型的神经网络结构上添加了一层适应层,如图 1(b)Adaptation 所示,该适应层能够衡量数据集的分布差异,并将数据分布差异值作为损失反馈项,辅助神经网络的迭代训练。

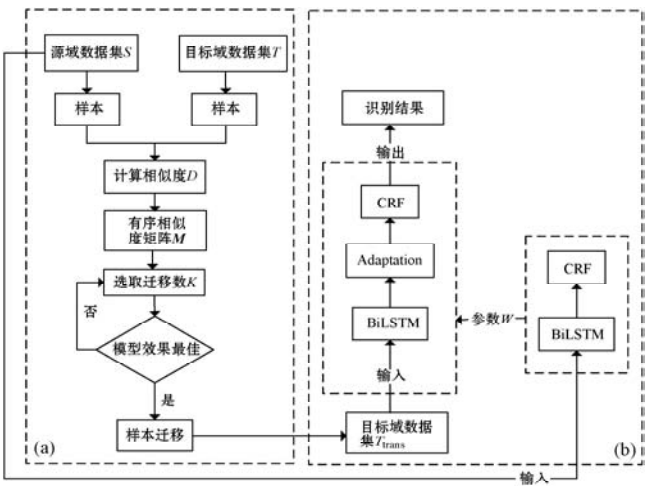

图 1　MT-NER 算法流程

2.1　BiLSTM-CRF 命名实体识别方法

本文采用的命名实体识别方法是由百度研究院提出的双向 LSTM-CRF 模型。BiLSTM-CRF 方法首先将字词转换成向量序列,然后送入 BiLSTM 网络中,进行特征提取,网络的隐藏状态序列接入线性层,输出每个单词对应的预测得分。然后 CRF 层产生约束,通过维特比解码动态计算最优解,得到模型最终输出的实体标记结果。BiLSTM-CRF 方法的框架[5]如图 2 所示。

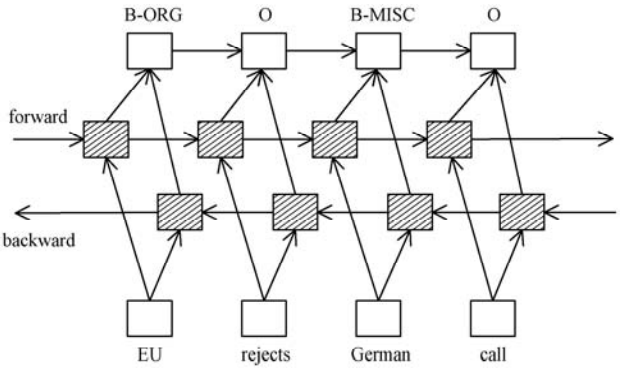

图 2　BiLSTM-CRF 框架

图 2 中 BiLSTM 是 LSTM 的变体,能够解决由于序列过长而造成的梯度爆炸或梯度消失问题,同时 BiLSTM 是双向的,能够捕捉正向和反向的序列信息。LSTM 的单元结构[19]如图 3 所示。

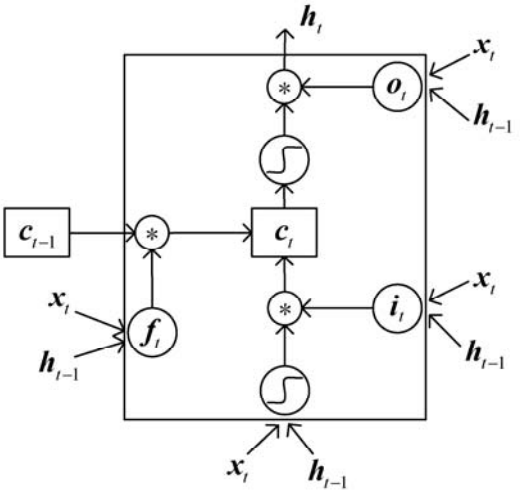

图 3　LSTM 单元结构

在图 3 中,黑色箭头代表向量的传输,即表示从一个节点的输出传送到其他节点的输入。其中:$\boldsymbol{i}_t$、$\boldsymbol{o}_t$和 $\boldsymbol{f}_t$分别表示时刻 t 的输入门、输出门和遗忘门函数;$\boldsymbol{C}_t$ 是激活向量;$\boldsymbol{x}_t$、$\boldsymbol{h}_t$是时刻 t 的输入向量和隐藏层输出向量。σ 代表 sigmoid 激活函数,tanh()代表双曲正切激活函数。

遗忘门确定前一个步长中哪些相关的信息需要被保留,用于更新记忆细胞的状态。该门读取 $\boldsymbol{h}_{t-1}$和 $\boldsymbol{x}_t$,由 sigmoid 函数输出一个在 0 到 1 之间的值给 $\boldsymbol{C}_{t-1}$(上个时刻的记忆单元)。1 表示“完全保留”,0 表示“完

全舍弃”,如式(1)所示。

$$f_t = \sigma(\boldsymbol{W}_f \times [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_f) \tag{1}$$

式中:$\boldsymbol{W}_f$表示当前单元的参数;$\boldsymbol{b}_f$表示偏置参数。

输入门确定当前输入中哪些信息是重要的,通过 sigmoid 函数决定需要更新哪些值,如式(2)所示;tanh 层读取 $\boldsymbol{h}_{t-1}$和 $\boldsymbol{x}_t$创建候选向量 $\boldsymbol{C}'_t$,如式(3)所示;输入门、遗忘门结合用来确定当前时刻的细胞状态,如式(4)所示。

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_i \times [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_i) \tag{2}$$

$$\boldsymbol{C}'_t = \tanh(\boldsymbol{W}_c \times [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_c) \tag{3}$$

$$\boldsymbol{C}_t = \boldsymbol{f}_t \times \boldsymbol{C}_{t-1} + \boldsymbol{i}_t \times \boldsymbol{C}'_t \tag{4}$$

输出门确定下一个隐藏状态的值,通过 sigmoid 函数得出的值来控制经过 tanh 层处理的当前细胞状态 $\boldsymbol{C}_t$,从而得到输出,如式(5)、式(6)所示。

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o \times [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_o) \tag{5}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \times \tanh(\boldsymbol{C}_t) \tag{6}$$

2.2 样本迁移

样本迁移即基于实例的迁移学习,旨在计算源域样本和目标域样本之间的关系权重,通过源域具有充足样本的特性来扩展目标域,并且解决深度学习中样本数较少的问题。

2.2.1 相似度计算

本文使用两种方式度量样本的相似度:基于字符的相似度度量和基于向量的相似度度量。

基于字符的相似度度量方法包括最小编辑距离和 Jaccard 相似系数。基于向量的相似度度量方法包括欧几里得距离和皮尔森相关系数。这四种度量方法分别从字符、集合、几何距离、变量的角度来衡量本文相似度。

2.2.2 数据引力

数据引力[20]模拟了万有引力来分析数据,用来计算源域样本对于目标域样本的权重。根据万有引力定律,引力的大小与质量的乘积成正比,与距离的平方成反比:

$$F = G\frac{m_1 m_2}{r^2}$$

式中:G 为万有引力常量;m_1和 m_2分别是物体 1 和物体 2 的质量;r 是两个物体之间的距离。

将万有引力模拟成源域样本和目标域样本之间的“力”。样本的相似度包括最小编辑距离 S_{Med}、Jaccard 相似系数 S_{Jacc}、欧几里得距离 S_{Euc}和皮尔森相关系数 S_{Pear}。对于第 i 个样本,具有的权值如式(7)-式(10)所示,其中 K 为常数。

$$p_i^{(0)} = G\frac{m_1 m_2}{r^2} = G\frac{Km_1 m_2}{(1+S_{\text{Med}})^2} \propto \frac{1}{(1+S_{\text{Med}})^2} \tag{7}$$

$$p_i^{(1)} = G\frac{m_1 m_2}{r^2} = G\frac{Km_1 m_2 S_{\text{Jacc}}}{(2-S_{\text{Jacc}})^2} \propto \frac{S_{\text{Jacc}}}{(2-S_{\text{Jacc}})^2} \tag{8}$$

$$p_i^{(2)} = G\frac{m_1 m_2}{r^2} = G\frac{Km_1 m_2}{(1+S_{\text{Euc}})^2} \propto \frac{1}{(1+S_{\text{Euc}})^2} \tag{9}$$

$$p_i^{(3)} = G\frac{m_1 m_2}{r^2} = G\frac{Km_1 m_2 S_{\text{Pear}}}{(2-S_{\text{Pear}})^2} \propto \frac{S_{\text{Pear}}}{(2-S_{\text{Pear}})^2} \tag{10}$$

计算样本的最终权值 p_i,如式(11)所示。

$$p_i = \text{avg}(p_i^j) \quad j = 0,1,2,3 \tag{11}$$

式中:avg 表示取平均值。

2.3 深度网络模型迁移

在源域数据集 MedDialog 中训练 BiLSTM-CRF,使用词嵌入算法 Word2vec[21]实现字词向量化,窗口大小设为 5,向量维度选取 100,保存预测性能最佳时的模型参数,得到本地的预训练模型,即迁移学习的源模型。

2.3.1 Finetune

基于模型的迁移主要通过构建参数共享的模型来实现,Finetune 就是一种基于模型的迁移方式,也是深度网络的迁移形式。即 Finetune 通过调整已经训练过的模型来适应新的任务,因此,对于新的任务,不需要从头开始对模型进行训练,减少了时间代价,除此之外,预训练模型是基于大数据集训练的,扩充了训练数据,使得模型的鲁棒性、泛化能力更好。

MT-NER 算法中,基于模型的迁移方法首先读取预训练模型的参数 W,初始化目标域模型,训练过程中神经网络层迭代更新参数,参与到目标域任务的学习之中,这样就将预训练模型学习到的知识迁移到目标领域中,并针对目标域的任务进行调整,不仅能够提高网络训练的速度,还能使模型适应文本域的变化,提高目标域任务的精准度。

2.3.2 数据分布自适应

理想情况下,训练集与测试集应该有相同的数据分布,然而在现实中,各种现实因素都会打破这个假设,产生协变量偏移,从另一个领域迁移样本也会加重这一问题。样本迁移丰富了目标域的样本,但是使得数据的来源不一,同时过度地迁移样本会使目标域整体的分布发生较大的变化,从而产生负迁移现象。另一方面,在新领域上仅仅通过 Finetune 是不可行的,在新数据上重新训练时会导致知识遗忘的现象,并且 Finetune 无法解决训练集与测试集之间数据分布不一致的问题。

为了将源域上训练的模型更好地应用在目标域中,并且保证最小的精度损失和最大的泛化能力,本文在原有的深度网络结构中加入数据分布自适应层,来

最小化目标域中训练集和测试集数据之间的分布距离。广泛应用于度量分布距离的方法包括 KL 散度(Kullback-Leibler Divergence)、JS 散度(Jensen-Shannon)、最大均值差异(Maximum Mean Discrepancy,MMD)、多核 MMD 等。由于 MMD 在样本数量不同的背景下能够计算两组数据之间的分布差异,并在迁移学习中常用来减少源领域和目标领域之间的分布不匹配[22],故选择标准分布距离度量——MMD 作为度量方法。

MMD 将数据映射到再生希尔伯特空间(Reproducing Kernel Hilbert Spaces),两组数据在 RKHS 中的均值的距离常用于度量两个不同但相关的分布之间的距离。设$\{X_S^i\}_{i=1,2,\cdots,m}$和$\{X_T^j\}_{j=1,2,\cdots,n}$分别是从数据空间 $\boldsymbol{X}$ 上的分布 D_S和 D_T中提取的数据向量,并且存在一个再生希尔伯特空间 $\boldsymbol{H}$,并有特征空间映射函数 $\phi(\cdot):\boldsymbol{X}\rightarrow\boldsymbol{H}$,MMD 表达式如式(12)所示。

$$MMD^2(X_s,X_t)=\left\|\frac{1}{m}\sum_{i=1}^{m}\phi(X_S^i)-\frac{1}{n}\sum_{j=1}^{n}\phi(X_T^j)\right\|_H^2 \tag{12}$$

本文使用搭建好的 BiLSTM-CRF 作为模型框架,将 MMD 度量嵌入神经网络的学习之中,本文算法是通过最小化扩展后的目标域中存在的分布差异、最小化实体识别损失值来优化模型,为了避免过拟合,加入正则化项,得到本算法的损失函数如式(13)所示。

$$L=L_N(x,y)+\alpha MMD^2(X_s,X_t)+\beta L_W \tag{13}$$

式中:L_N表示 BiLSTM-CRF 模型的命名实体识别损失值;$\alpha>0$ 是惩罚系数,控制 MMD 对损失函数的贡献程度;L_W是权值正则项;β 是正则化参数。

2.4　算法描述

输入:源域数据集 S,目标域数据集 T。

2.4.1.　样本迁移

1)读取源域数据集和目标域数据集中的样例文本 $W_S=\{W_{S_1},W_{S_2},\cdots,W_{Sm}\}$ 和 $W_T=\{W_{T_1},W_{T_2},\cdots,W_{Tn}\}$($m$、$n$ 分别表示源域数据集和目标域数据集的样例数,$m>n$)。

2)针对每一个目标域数据集样本 W_{T_i},计算 W_{T_i}与源域数据集中所有样本在不同度量下的相似度距离 $\boldsymbol{D}\in\mathbf{R}^{n\times m\times j}$($j$ 为选取的相似度距离度量方法种类),通过数据引力计算式(7)-式(10)得到相对于源域样本的权值 $\boldsymbol{P}\in\mathbf{R}^{n\times m\times j}$,计算其平均值得到最终的权重矩阵 $\boldsymbol{M}\in\mathbf{R}^{n\times m}$。

3)将权重矩阵 $\boldsymbol{M}$ 按从大到小排序得到有序矩阵 $\boldsymbol{M}_{\text{order}}\in\mathbf{R}^{n\times m}$,选取正整数 k 值($0<k<m$),依次从源域数据集中选择权重最大的 k 个样本组成样本集 T_k,将其原始文件迁移至目标域原始文件所在文件夹中(不重复迁移),得到样本迁移后的目标域数据集:$T_{\text{trans}}=T\cup T_k$。

2.4.2.　深度网络模型迁移

1)使用源域数据集 S 训练 BiLSTM-CRF 模型,保留参数 W。

2)构建新的 BiLSTM-CRF 模型,用 W 做参数初始化。

3)以扩充后的目标域数据集 T_{trans} 训练新的 BiLSTM-CRF 模型,通过损失函数计算式(13)迭代优化模型,W 参与网络模型的更新迭代。

2.4.3.　超参数调整

独立更新样本迁移数 k、MMD 惩罚系数 α,重复执行 2.4.1 节和 2.4.2 节的步骤,训练模型 MT-NER,返回最终结果。根据反馈的结果,选择最优 k 值和最优 α 系数。

3　实　验

3.1　实验数据集

本文采用 MedDialog[23] 数据集作为源域数据集,该数据集包含 1 145 231 个患者和医生之间的中文咨询,共计 110 万个对话和 400 万个话语,每次会诊包括以下部分:1)描述病人的医疗状况和病史;2)病人与医生的对话;3)医生给出的诊断和治疗建议(可选)。本实验选取 2019 年和 2020 年的数据,共包括 145 600 个对话记录,文本的命名实体标注由实体词典 + jieba 词性自动标注和人工辅助共同完成。

目标域数据集选取 CCKS 2017 任务二(http://www.sigkg.cn/ccks2017/)的电子病历数据集,包含病史特点、出院情况、一般项目和诊疗经过四个目录,总共有 1 596 篇中文电子病历记录,每一份电子病历对应有一份实体标注文件,标注文件包含此文本包含的实体内容、实体类型和实体内容的位置索引值。

实验中的数据标注采用{B,I,O}三元标注体系,实体类别如表 1 所示。

表 1　实体类型

实体类型	定义	例子
DIS	疾病	糖尿病、肝炎
SUR	手术	子宫全切术、冠脉造影术
PRE	措施(非手术)	穴位贴敷
ORG	部位词	肱二头肌、跟腱
TES	检查	头颅 CT、血管彩超
SYM	症状	言语清晰、双肺呼吸音

续表 1

实体类型	定义	例子
DRU	药品	抗生素、硝苯地平片
PSB	可能性词	可能、不排除
FW	频率词	偶有、长期
PT	既往信息词	传染病病史、外伤史
O	非实体部分	； ， 。 （）

3.2 实验设置及评估

为保证实验的公平性，除一些特别设置外，都选取一样的参数值，具体的参数设置如表 2 所示。数据集按照 7∶2∶1的比例分为训练集、验证集和测试集，模型均采用 Adam 优化器，用早停法防止过拟合。

表 2 模型参数设置

参数	值
向量维度	100
批处理大小	32
学习率	3E－4
L2 正则化	1E－5
Dropout	0.5
最佳样本迁移数 k	400

本实验使用精度（Precision）、召回率（Recall）、F1 值（F1-score）来评价模型的性能，其计算式分别为：

$$P_{\text{recision}} = \frac{T_{\text{P}}}{T_{\text{P}} + F_{\text{P}}} \tag{14}$$

$$R_{\text{ecall}} = \frac{T_{\text{P}}}{T_{\text{P}} + F_{\text{N}}} \tag{15}$$

$$F_1 = 2 \times \frac{p_{\text{recision}} \times r_{\text{ecall}}}{p_{\text{recision}} + r_{\text{ecall}}} \tag{16}$$

式中：T_{P} 表示正确识别的命名实体数；F_{P} 表示错误识别的命名实体数；F_{N} 表示未被识别的命名实体数。

3.3 实验结果与分析

为了验证 MT-NER 算法的有效性，本节进行不同的对比实验，使用的模型包括 IDCNN-CRF[24]、BiLSTM-CRF、BiLSTM-CRF-Ts（BiLSTM-CRF-样本迁移）和 BiLSTM-CRF-Tm（BiLSTM-CRF-深度网络模型迁移）。

3.3.1 样本迁移对比实验

为探讨 MT-NER 算法中样本迁移数量对最终性能的影响，在其余条件不变的情况下，进行样本迁移数量的对比实验。图 4 显示了样本迁移数量对 BiLSTM-CRF-Ts 模型性能的影响。实验结果表明，样本迁移数量越多，目标域中迁入的其他领域的数据也就越多，一定程度上能够扩充数据集，达到迁移学习的效能，但由于数据领域不同，引入的数据过多，会出现负迁移现象。如图 4 所示，最优样本迁移数为 400，此时 F1 值为 0.896 3、准确度为 91.36%、召回率为 87.96%。当迁移数小于 400 时，模型的 F1 值随着迁移数量的增加而呈现明显的上升趋势；当迁移数超过最佳迁移样本数之后，迁移学习为模型带来的负迁移现象也随之越来越明显，模型的 F1 值和精度呈下降趋势，并逐渐收敛。虽然召回率的变化波动比较大，但仍存在降低的趋势。因此选取合适的样本迁移数是决定迁移效果的一个重要因素。根据图 4 中展示出来的对比实验性能，本文将最佳样本迁移数 k 设定为 400。

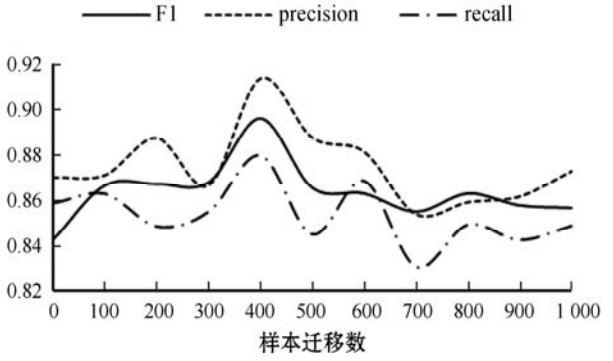

图 4 不同样本迁移数量下 BiLSTM-CRF-Ts 模型的性能

3.3.2 不同 MMD 惩罚系数的对比实验

本文采用的数据分布自适应方法能够有效解决数据分布不一致的问题，为了衡量数据分布自适应在 MT-NER 算法中的重要性，针对 MMD 惩罚系数做一组对比实验，选取不同的惩罚系数 α 进行训练，实验结果如图 5 和图 6 所示。

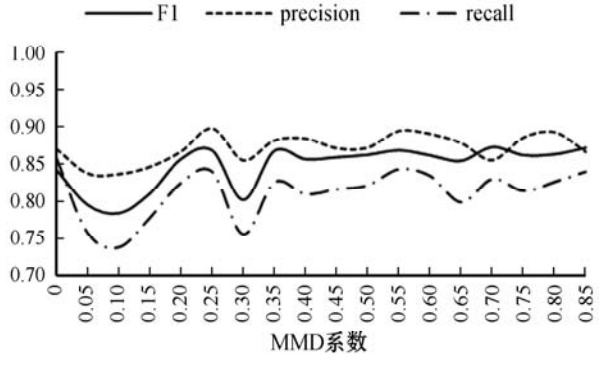

图 5 不同惩罚系数下的 BiLSTM-CRF-Tm 模型的性能

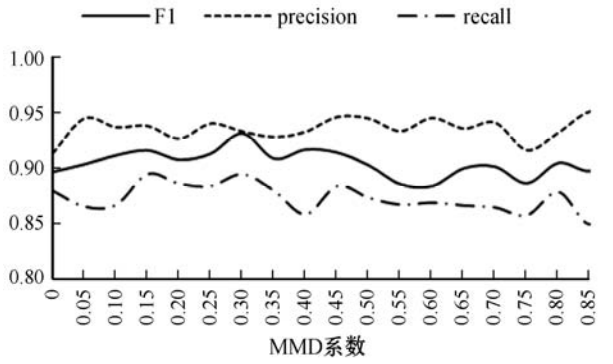

图 6 不同惩罚系数下的 MT-NER 模型的性能

从图 5 的走向可以看出,对于 BiLSTM-CRF-Tm 模型,测试集上的 F1 值、精度和召回率并不与 MMD 惩罚系数成正相关或负相关关系,过低或过高地赋予 MMD 项的贡献比重会使模型达不到最佳的性能,这是因为过低的 MMD 惩罚系数会低估数据分布对模型的影响力度,过高的 MMD 惩罚系数会过多关注数据分析不一致带来的影响,相比之下,降低了对实体识别任务的学习能力。经过对比实验发现,当 MMD 惩罚系数取 0.25 时,MMD 产生的损失值能够最有效地作为模型迭代的反馈项,在此设置下,BiLSTM-CRF-Tm 模型能够达到的 F1 值为 0.868 1,精度为 89.76%,召回率为 84.05%,此时为 MMD 惩罚系数的最佳值,当 MMD 惩罚系数大于 0.25,模型的性能趋于下降。

同样,为了探讨不同 MMD 惩罚系数对 MT-NER 算法的影响,图 6 展示了对比实验的性能曲线。与图 5 显示的结果类似,过低或过高的 MMD 惩罚系数下 MT-NER 算法都不能达到最优性能,惩罚系数的大小对精度的影响波动较为平缓,相比之下,对 MT-NER 算法的 F1 值和召回率的影响较为明显,MMD 惩罚系数为 0.3 时,MT-NER 算法能够达到的 F1 值为 0.931 7,精度为 93.31%,召回率为 89.5%。

3.3.3　迁移效果对比实验

为验证迁移学习对小样本实体识别效果的提升,基于目标域数据集进行不使用迁移学习技术、单独使用迁移学习技术、混合使用迁移学习技术的实验,结果如表 3 所示。IDCNN-CRF 使用了迭代扩张 CNN 架构重复地将相同的扩张卷积块应用于符号表示,能够达到共享参数的目的以防止过拟合。本文实验中,IDCNN-CRF 在小样本数据集上的命名实体识别效果与 BiLSTM-CRF 相比召回率略高,精度和 F1 值略低。对比不使用迁移学习技术的 BiLSTM-CRF 模型,使用样本迁移的 BiLSTM-CRF-Ts 模型精度提升了 4.38 百分点,召回率提升了 2.11 百分点,F1 值提升了 5.37 百分点,使用深度网络模型迁移的 BiLSTM-CRF-Tm 模型精度提升了 2.78 百分点,F1 值提升了 2.55 百分点,召回率降低了 1.8 百分点,两组实验证明了迁移学习对小样本的命名实体识别效果的积极作用。本文提出的 MT-NER 算法能够达到最好的实体识别效果:精度达到 93.31%,召回率达到 89.5%,F1 值达到 0.931 7,相比于未使用迁移学习的命名实体识别模型,精度提升了 6.33 百分点,召回率提升了 3.65 百分点,F1 值提升了 0.089 1。结果表明,混合式迁移学习可以更好地实现迁移效果,能够在一定程度上解决中文命名实体识别领域样本不足的问题。

表 3　不同模型对比实验结果

模型	精度/%	召回率/%	F1 值
IDCNN-CRF	86.00	87.15	0.836 3
BiLSTM-CRF	86.98	85.85	0.842 6
BiLSTM-CRF-Ts	91.36	87.96	0.896 3
BiLSTM-CRF-Tm	89.76	84.05	0.868 1
MT-NER	**93.31**	**89.50**	**0.931 7**

4　结　语

本文为解决命名实体识别领域中大量的标注数据难以获取的挑战,提出基于混合式迁移学习的命名实体识别算法,该算法混合样本迁移和模型迁移,以解决中文命名实体识别领域样本不足的问题,同时使用最大均值差异构建适应层以解决数据分布不一致的问题,实验过程中通过动态地选择超参数能够在一定程度上降低算法的负迁移现象。实验结果表明,该算法只需要利用少量的目标域数据,便可以有效地获取命名实体识别结果,能够有效地识别中文电子病历中的疾病、检查和症状等实体,在自然语言处理领域具有重要意义。不过本文研究仍有一些不足之处,例如:文本的嵌入信息没有语言预测模型中的丰富,文本原始数据中存在因人为操作不当而产生的噪声,在今后的研究中,应考虑相关方面的改进。

参　考　文　献

[1] 刘浏,王东波. 命名实体识别研究综述[J]. 情报学报,2018,37(3):329 - 340.

[2] 张驰名,王庆凤,刘志勤,等. 基于深度迁移学习的肺结节辅助诊断方法[J]. 计算机工程,2020,46(1):271 - 278.

[3] Zhang Y, Yang J. Chinese NER using lattice LSTM[C]// 56th Annual Meeting of the Association for Computational Linguistics,2018:1554 - 1564.

[4] Chiu J P C, Nichols E. Named entity recognition with bidirectional LSTM-CNNs[J]. Transactions of the Association for Computational Linguistics,2016,4:357 - 370.

[5] Huang Z H, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[EB]. arXiv:1508.01991,2015.

[6] 曹晓杰,么娆,严雨灵. 应用迁移学习的卷积神经网络花卉图像识别[J]. 计算机应用与软件,2020,37(8):142 - 148.

[7] 金祝新,秦飞巍,方美娥. 深度迁移学习辅助的阿尔兹海默氏症早期诊断[J]. 计算机应用与软件,2019,36(5):171 - 177.

[8] Huang J Y, Smola A, Gretton A, et al. Correcting sample selection bias by unlabeled data[C]//19th International

Conference on Neural Information Processing Systems,2006:601 - 608.
[9] Dai W Y, Yang Q, Xue G R, et al. Boosting for transfer learning[C]//24th International Conference on Machine Learning,2007:193 - 200.
[10] Pan S J, Tsang I W, Kwok J, et al. Domain adaptation via transfer component analysis[J]. IEEE Transactions on Neural Networks,2011,22(2):199 - 210.
[11] Yao Y, Doretto G. Boosting for transfer learning with multiple sources[C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010: 1855 - 1862.
[12] Duan L X, Xu D, Tsang I W. Domain adaptation from multiple sources: A domain-dependent regularization approach [J]. IEEE Transactions on Neural Networks and Learning Systems,2012,23(3):504 - 518.
[13] 王红斌,沈强,线岩团. 融合迁移学习的中文命名实体识别[J]. 小型微型计算机系统,2017,38(2):346 - 351.
[14] 朱艳辉,李飞,翼相冰,等. 反馈式 K 近邻语义迁移学习的领域命名实体识别[J]. 智能系统学报,2019,14(4):820 - 830.
[15] Yang H Y, Huang S J, Dai X Y, et al. Fine-grained knowledge fusion for sequence labeling domain adaptation[C]//Conference on Empirical Methods in Natural Language Processing,2019:4195 - 4204.
[16] Chen L Z, Moschitti A. Transfer learning for sequence labeling using source model and target data[EB]. arXiv:1902.05309,2019.
[17] Zhou J T, Zhang H, Jin D, et al. Dual adversarial neural transfer for low-resource named entity recognition[C]//57th Conference of the Association for Computational Linguistics, 2019:3461 - 3471.
[18] 陈美杉,夏晨曦. 肝癌患者在线提问的命名实体识别研究:一种基于迁移学习的方法[J]. 数据分析与知识发现,2019,3(12):61 - 69.
[19] Pham T, Tran T, Phung D, et al. DeepCare: A deep dynamic memory model for predictive medicine[C]//20th Pacific-Asia Conference on Advances in Knowledge Discovery and Data,2016:30 - 41.
[20] Ma Y, Luo G C, Zeng X, et al. Transfer learning for cross-company software defect prediction[J]. Information & Software Technology,2012,54(3):248 - 256.
[21] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[EB]. arXiv:1301.3781,2013.
[22] Ghifary M, Kleijn W B, Zhang M J. Domain adaptive neural networks for object recognition[C]//Pacific Rim International Conference on Artificial Intelligence,2014:898 - 904.
[23] He X H, Chen S, Ju Z Q, et al. MedDialog: Two large-scale medical dialogue datasets[C]//Conference on Empirical Methods in Natural Language Processing,2020:9241 - 9250.
[24] Strubell E, Verga P, Belanger D, et al. Fast and accurate entity recognition with iterated dilated convolutions[C]//Conference on Empirical Methods in Natural Language Processing,2017:2670 - 2680.

(上接第 258 页)

[5] 赵玉文,敖玉龙,杨超,等. 申威 26010 众核处理器上一维 FFT 实现与优化[J]. 软件学报,2020,31(10):3184 - 3196.
[6] Ozkan I, Yilmaz A, Celebi G. Improved segmentation with dynamic threshold adjustment for phonocardiography recordings[C]//41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society,2019:6681 - 6684.
[7] Lei B, Fan J L. Image thresholding segmentation method based on minimum square rough entropy[J]. Applied Soft Computing,2019,84:79 - 83.
[8] Xie D H, Lu M, Xie Y F, et al. A fast threshold segmentation method for froth image base on the pixel distribution characteristic[J]. PLoS One,2019,14(1):2300 - 2307.
[9] Abdel-Basset M, Chang V, Mohamed R. A novel equilibrium optimization algorithm for multi-thresholding image segmentation problems[J]. Neural Computing and Applications, 2021,33(17):10685 - 10718.
[10] 宋森森,贾振红,杨杰,等. 结合 Ostu 阈值法的最小生成树图像分割算法[J]. 计算机工程与应用,2019,55(9):178 - 183.
[11] 周力凯,江雨洋,冯亚春,等. 基于多尺度区域与类不确定性理论的局部阈值分割方法[J]. 计算机应用,2020,40(S2):66 - 72.
[12] 付云凤. 基于阈值的图像分割研究[D]. 重庆:重庆大学,2013.
[13] Liu W X, Hu J M, Li Z Y, et al. Tongue image segmentation via thresholding and gray projection[J]. KSII Transactions on Internet & Information Systems,2019,13(2):945 - 961.
[14] Chen R, Xu Y A. Threshold optimization selection of fast multimedia image segmentation processing based on Labview [J]. Multimedia Tools & Applications,2020,79(13/14):9451 - 9467.
[15] 杨琳,吴家铸,扈啸,等. 互相关运算在银河飞腾 DSP 上的实现及优化[J]. 计算机科学,2015,42(11):53 - 55.
[16] 张军阳,郭阳,扈啸. 二维矩阵卷积的并行计算方法[J]. 浙江大学学报(工学版),2018,52(3):515 - 523.
[17] 李勇,陈书明,陈胜刚. 一种基于 YHFT-Matrix DSP 的去块效应滤波算法的向量化实现[C]//第十五届计算机工程与工艺年会暨第一届微处理器技术论坛,2011.