

医疗文本的小样本命名实体识别

秦 健, 侯建新, 谢怡宁, 何勇军

(哈尔滨理工大学 计算机科学与技术学院, 哈尔滨 150080)

摘 要: 针对医疗文本命名实体识别缺乏足够标记数据的问题, 提出了一种新的命名实体识别神经网络和数据增强方法。该方法首先利用汉字的拼音和笔画来扩展 Bert 词向量, 使 Bert 词向量能够包含更多先验知识; 接着将命名实体识别模型与分词模型进行联合训练, 以增强模型对于实体边界的判别能力; 最后采用改进的数据增强方法处理训练数据, 能够在避免模型过拟合的同时增加模型对于命名实体的识别效果。在 CCKS-2019 提供的电子病历文本上的实验结果表明, 所提出的方法在小样本情况下能有效提高命名实体识别的准确率, 在训练数据减少一半的情况下, 识别率仍能保持没有明显下降。

关键词: 命名实体识别; 小样本; 数据增强; 联合训练; 特征融合

DOI: 10.15938/j.jhust.2021.04.013

中图分类号: TP391.1 **文献标志码:** A **文章编号:** 1007-2683(2021)04-0094-08

Few-shot Named Entity Recognition for Medical Text

QIN Jian, HOU Jian-xin, XIE Yi-ning, HE Yong-jun

(School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China)

Abstract: Aiming at the problem that medical text named entity recognition lacks sufficient labeled data, a newly named entity recognition deep neural network and data enhancement method is proposed. First of all, the Bert word vector is extended with pinyin and strokes of Chinese characters to make it contain more useful information. Then the named entity recognition model and the word segmentation model are jointly trained to enhance the model's ability to recognize entity boundaries. Finally, an improved data enhancement method is used to process the training data, which can increase the recognition effect of the model on named entities while avoiding overfitting of the model. The experimental results on the electronic medical record text provided by CCKS-2019 show that the proposed method can effectively improve the accuracy of named entity recognition in the case of small samples and the recognition rate can still be maintained without a significant decrease when the training data is reduced by half.

Keywords: named entity recognition; few-shot; data augmentation; joint training; feature fusion

收稿日期: 2020-04-29

基金项目: 国家自然科学基金面上项目(61673142); 黑龙江省自然科学基金杰出青年项目(JJ2019JQ0013); 哈尔滨市杰出青年人才基金(2017RAYXJ013); 黑龙江省自然科学基金面上项目(F2017013); 黑龙江省普通本科高等学校青年创新人才项目(UNPYSCT-2016034); 黑龙江省教育厅科学技术研究项目(12511096); 哈尔滨理工大学青年拔尖创新人才(20152); 中国博士后基金(20132303120003)。

作者简介: 秦 健(1997—), 男, 硕士研究生;

侯建新(1997—), 男, 硕士研究生。

通信作者: 何勇军(1980—), 男, 博士, 博士研究生导师, E-mail: Holywit@163.com.

0 引言

随着互联网、5G 技术和人工智能的不断深入发展,文本信息化处理已成为当今时代的必然要求。作为文本处理的关键技术之一,命名实体识别的目的在于提取和识别文本中的特定词汇。目前基于深度学习的命名实体识别方法已成为主流,虽然提高了识别性能,但模型训练需要大量的标注数据。由于缺少足够的训练数据导致一些新兴领域很难得到快速发展。因此研究小样本的命名实体识别具有重要的现实意义。

早期命名实体识别大多使用基于字典和规则的方法^[1],其构建需要专家的专业知识,识别的效果和可移植性较差。后来,机器学习逐渐成为命名实体识别的主流方法,主要有隐马尔可夫模型^[2]、最大熵^[3]与条件随机场^[4-5]模型等。这些方法与传统的规则和字典相结合取得了很好的效果。随着技术的不断深入发展,深度学习技术在命名实体识别中也取得了巨大的成功^[6],其中代表模型为 BiLSTM + CRF^[7] (bi-directional long short-term memory, BiLSTM) 模型。研究人员在深度学习网络的基础上引入汉字的字形特征^[8]或者与其他模型联合训练^[9]来改良网络也取得了很好的效果。与此同时,语言模型也取得了长足发展,能够更好的将文本材料中的蕴含信息融入到词向量中^[10-11],出现了 Word2vec^[12],ELMo 词向量^[13],通用语言模型 GPT^[14]以及 Bert^[15]等模型,有力地促进了自然语言的发展。

深度学习和机器学习方法依赖于大量的标注语料,因此针对小样本的无监督学习、半监督学习和数据增强越来越受到关注。无监督学习大多使用聚类算法将同类实体进行聚类,之后再对其类别进行判断。无监督学习在缺少标注数据的垂直领域具有一定的实用性,但在特定领域内存在领域扩展性差和实体歧义的问题^[16]。半监督学习借助于少量标注语料和大量未标注语料来训练模型,首先借助于标记语料对于模型进行训练,并用此模型去预测未标注语料。若预测结果概率大于设定阈值,则更新这个语料的标签并加入到标记语料中。之后重复这个过程用于扩大训练语料,从而获得一个效果较好的模型^[17]。但随着半监督学习训练的不断进行,自动标记的数据中噪声容易不断积累导致模型效果的下降。因此研究小样本的监督学习显得尤为重要。

医疗文本中实体相较于其他领域有其自身的特点。首先医疗实体有比较明显的形态特征,疾病名和身体部位等实体常与人的身体器官相关,如“胸痛”、“腹积水”、“脑”等实体,这些实体都表现出比较明显的偏旁特征;医疗实体的很多名称来源于英文的音译,如“奥沙利铂”和“奥沙利铂”。由于翻译和书写习惯的不同,有时会选择同音字来表示相同的实体,这样情况下拼音对于命名实体识别起到重要的帮助。医疗文本中实体的多样性复杂性也增加了对于实体边界的判别能力,因此可以借助于分词模型来增加命名实体识别模型对于实体边界的判别能力。

为了能够在较小训练数据的情况下进行命名实体识别,本文提出了基于特征融合的联合训练模型。网络共包含 4 层,第一层为字特征卷积层 char_CNN,使用 CNN (convolutional neural networks, CNN) 网络提取汉字的笔画和拼音特征;第二层为 Bert 词向量层,并且使用 char_CNN 的输出来扩展 Bert 词向量;之后使用双向长短时记忆网络 BiLSTM 层从长距离上下文中学习特征表示,最后采用 CRF 层对于标签进行解码。此外,我们提出一个统一的框架来联合训练命名实体识别和中文分词模型。在此框架中,命名实体识别模型和中文分词模型共享 char_CNN + Bert 层作为网络的词向量,但分词模型拥有一个独立的 CRF 用于分词标签解码。最后还对训练数据进行了合理的增强,首先把每一个训练语句的实体部分随机替换为同类实体,非实体的部分进行随机删除和插入^[18]。这样可以在保持句子结构性的前提下扩展出更多有效的句子,并通过随机删除和插入来减少句子之间的相似性。实验结果表明,我们提出的模型效果优于如今主流模型,而且通过我们的数据增强,模型在小样本的情况下依然可以取得较好的识别效果。

1 网络模型和数据增强

1.1 模型框架

本文设计的特征融合的联合训练模型如图 1 所示。网络主要由命名实体识别模型和分词模型两部分构成,它们共享 char_CNN + Bert 层作为词向量特征提取层,但各自拥有不同的独立 CRF 解码层。在实体识别模型中,输入的语句首先进入到 char_CNN 层和 Bert 层,char_CNN 层对于汉字的拼音笔画特征进行特征提取,并且与 Bert 层输出的字特征

向量进行拼接; 拼接后的字特征向量输入到 Bi-LSTM 进行进一步的特征提取,并学习语句包含的上下文信息,最后将 Bi-LSTM 输出结果输入到 CRF 中进行最终的实体预测。而分词模型使用的词向量同样是由经过 char_CNN 扩展后的 Bert 提供,并将扩展后的词向量直接输入到 CRF 模块中进行分词的预测。下面将对网络的各层信息进行详细介绍。

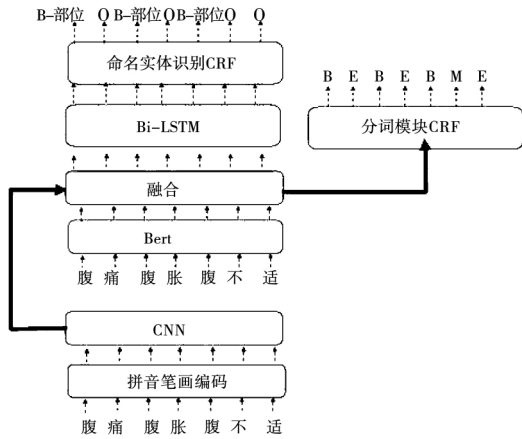


图 1 模型总体结构

Fig. 1 Model overall structure

1) 字特征卷积层 char_CNN

字特征卷积层主要包括拼音笔画编码层和 CNN 卷积神经网络两个部分,我们分别对这两个部分进行介绍。

拼音笔画编码层需要对汉字进行拆分编码,为网络引入汉字的拼音笔画特征。英文中可以根据单词的词根词缀来猜测其意义和性质,汉字的笔画以及偏旁部首中也蕴含着大量的信息。而且在医疗领域内的实体往往较为明显的形态特征,比如在检查部位和手术等试题中经常出现“腰”、“肝”、“脑”等身体部位,他们都使用“月”作为偏旁。因此引入汉字的笔画可以使网络更好的识别医疗实体。许多医疗领域的药名都来自英语,虽然它们的书写不同,但读音却很相似,所以在命名实体识别中引入拼音特征具有积极意义。使用一个“明”为例进行拆分,其结果见图 2。每个字拆分后依次按照声母、韵母、声调、笔画进行排列,其中汉字的笔画我们选用四角码来表示。四角码一共有 5 位数,前四位用 0-9 分别表示汉字四个角的单笔或复笔的笔形;第五位由汉字右下角上方的一个笔形决定,用来区分前四位同码的汉字。

目前,CNN 已广泛应用于图像局部信息的提取。因为汉字拆分后长度较短且固定,信息比较集

中,因此通过 CNN 网络来提取汉字的拼音和笔画特征。这里的 CNN 网络由一个一维卷积网络和一个一维最大池化层构成,卷积核大小设置为 3,激活函数选择 ReLU(Rectified Linear Unit, ReLU) 激活函数,最后将最大池化层的输出展开,方便对 Bert 进行扩展。

普通 CNN 由卷积层和池化层组成。对于特征的处理,CNN 是通过卷积核来完成的。卷积核是一个由用户定义大小的权重矩阵,其权值由后续的学习得到。卷积核将每一个局部的特征提取出来作为输出。池化层则是通过特定的池化操作对特征进行处理。对特征的卷积操作可以用式(1)来表示,而池化操作可由式(2)表示。

$$H_i = f(H_{i-1} \otimes w_i + b_i) \quad (1)$$

$$H_i = f(\text{pooling}(H_{i-1}) + b_i) \quad (2)$$

其中 H_i 表示第 i 层特征, $f(x)$ 是一个非线性激活函数, $\otimes$ 表示卷积核与特征的卷积操作。 $\text{pooling}(x)$ 为池化操作, b_i 表示第 i 层的偏置。

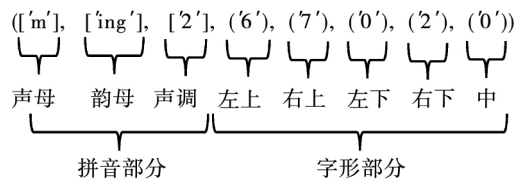


图 2 汉字拆分示意图

Fig. 2 Chinese character splitting diagram

2) Bert 字编码层

谷歌团队在 2018 年发布了 Bert 模型,作为一种预训练模型。Bert 在自然语言处理的 11 项测试任务中取得了最好的成绩,极大地推动了自然语言处理的发展^[3]。Bert 模型是由无监督学习方法在通用数据文本中训练得到,它的训练不需要标记数据,所以可以在大规模文本上训练以取得更好的先验知识。Bert 主要使用 Transformer 作为核心结构,其具体框架如图 3 所示。Transformer 完全基于自注意力机制来训练词向量,通过自注意力计算出每一个词和所有词之间的关系,根据词与词之间的相互关系得出词之间的联系和该词在句子中的权重。通过这种方式得到的词向量能够有效利用上下文的信息,因此 Bert 在不同任务中都能取得很好的效果。

3) BiLSTM 层

作为循环神经网络的改进模型,LSTM 能很好地抑制梯度下降和梯度爆炸,并且能捕获长距离序列信息,具有很强的序列建模能力。在命名实体识

别中,LSTM 可以感知到距离实体较远的信息,可以提高命名实体的识别率。例如“小明患有青光眼,疾病让他看不清周围的物体”,尽管“青光眼”和“看不清”有较长的距离,但是可以从“看不清”推断出“青光眼”是一种眼类疾病。实体的上下文一般都有较为重要的信息,因此使用双向 LSTM 从全局上学习上下文信息。每一个 LSTM 单元主要包含:输入门、输出门、忘记门和细胞状态。具体如图 4 所示。

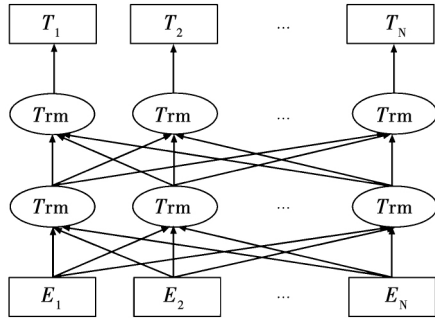


图3 Bert 模型框架

Fig. 3 Bert model framework

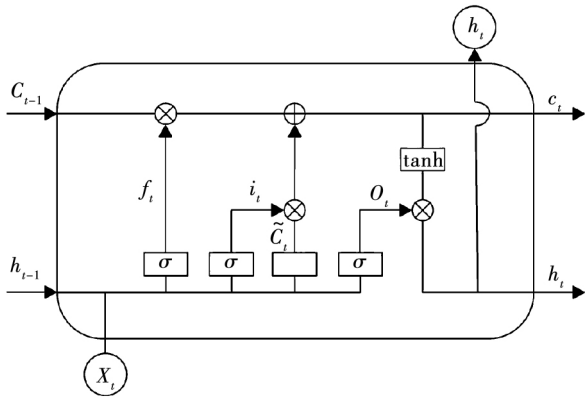


图4 LSTM 单元结构

Fig. 4 LSTM unit structure

它们的更新方式如公式(3)到公式(8)所示:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (3)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (4)$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (5)$$

$$C_t = C_{t-1} \cdot f_t + i_t \cdot \tilde{C}_t \quad (6)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (7)$$

$$h_t = o_t \cdot \tanh(C_t) \quad (8)$$

其中: f_t, i_t, C_t, o_t 分别代表忘记门、输入门、细胞状态和输出门; W_f, W_i, W_c, W_o 为不同状态的权值; b_f, b_i, b_c, b_o 是不同状态的偏置; h_t 和 W_t 分别为输入和输出; $\tilde{C}_t$ 是计算过程中的临时细胞状态; σ 为 sig-

moid 激活函数。

4) CRF 输出层

在命名实体识别任务中,相邻的标签通常有较强的相关性^[9],因此我们在命名实体识别任务中使用 CRF,而非 Softmax 这样的独立解码层。

假设给定输入序列 $X = (x_1, x_2, x_3, \dots, x_n)$, 其对应的标签序列为 $Y = (y_1, y_2, y_3, \dots, y_n)$ 。使用式(9)可以机算出标签序列的得分。

$$\text{score}(X, Y) = \sum_{i=1}^n p_{i, y_i} + \sum_{j=0}^{n+1} A_{y_{j-1} y_j} \quad (9)$$

其中: A 表示转移得分矩阵,矩阵元素 $A_{i, j}$ 表示标签 i 转移到标签 j 的转移得分。设 y_0 和 y_{n+1} 为句子的起始标签和终止标签,标签种类为 k ,则 $A \in R^{(k+2) \times (k+2)}$ 。 $P \in R^{n \times k}$ 是输出层的得分矩阵,矩阵元素 $p_{i, j}$ 表示第 i 个词在第 j 个标签下输出的得分。

5) 联合分词模型

命名实体识别可以看作是两个子任务的组合:从文本中识别出实体名称的边界并对其进行分类。识别实体的边界是一项困难的任务,很多时候实体的边界存在歧义性。比如“长江大桥”,可以被看作是“长江”和“大桥”两个实体也可以被看成一个实体。而中文分词模型的任务便是将中文文本分成单个词,也就是对于文本中词的边界进行判断。因此分词模型和实体识别模型高度相关,联合分词模型的训练能够有效的提高实体识别的准确性^[8]。引入分词模型后,网络拥有两个 CRF 输出,模型的总体损失可以使用式(10)进行表示。

$$L = (1 - \lambda) L_{NER} + \lambda L_{CWS} \quad (10)$$

其中: L_{NER} 和 L_{CWS} 分别表示命名实体识别和分词任务的损失; $\lambda \in [0, 1]$ 是控制分词任务损失在总损失中相对重要性的系数,在本文中 λ 的值为 0.3。模型优化的目标是损失函数最小化,损失函数包括分词和命名实体识别两个部分,在模型优化的过程中,分词和命名实体识别两个任务可以相互促进,从而获取更好的表现效果。

1.2 数据增强

合理的数据增强能够在保存原始数据分布和标签的情况下,使模型在小样本情况下也能够取得较好的训练效果^[13]。因此需要对训练集数据进行合理的扩充,具体实现步骤如下:

Step1: 提取训练数据中所有的实体,并且按照类别进行保存;

Step2: 通过网络收集相关实体,对现有实体集进行适当扩充;

Step3: 对句子进行扩充,其具体步骤是将一个句值中的每个实体随机替换为同类实体,并以该句的派生句数作为增益系数;

Step4: 对于句中实体外的字词,进行随机的删除,删除的概率为 p_1 ,此处选择 10% 的概率;

Step5: 对于句中非实体的每个位置进行随机插入,插入的词语选用哈工大停用词表,插入的概率为 p_2 ,这里选择 10% 的概率。

首先通过分析标记文本来建立实体集合。接着替换训练数据集中的实体为实体集中同类实体,可以有效增加标记数据的数量。但这样的操作会让语句有明显的人工特点,同时句子的重复性也比较高,这会造成模型过拟合和鲁棒性差。于是我们学习 Xie 等^[2]的操作,在句子中引入噪声,为了不破坏原本实体,在非实体的每个位置进行随机插入哈工大停用词。

2 实验效果及分析

2.1 数据准备

实验数据来源于中国知识图谱与语义计算大会 CCKS2019 (China conference on knowledge graph and semantic computing) 的电子病历命名实体识别任务,共 1 000 份真实的临床语料,训练集数据包含 600 条语料,而验证集和测试集分别有 200 条语料。数据集共有 6 类实体,分别为疾病与诊断、解剖部位、影像检查、实验室检验、药物、手术。

2.2 数据格式化

这里使用 BIOES (B-begin, I-inside, O-outside) 标注法,将每个元素标注为“B-X”、“I-X”或“O”。其中,B 和 I 分别表示实体的开头和中间,X 表示这个实体所属的类别,“O”表示不属于任何类型。同时我们通过北大分词工具 pkuseg^[3]来引入分词信息,其中“B”、“M”、“E”、“W”分别表示一个词语的开始、中间、结尾和单个字。标注效果如表 1 所示。

表 1 文本标注示例

Tab. 1 Examples of text annotation

数据	腹	胀	腹	不	适
实体信息	B-部位	O	B-部位	O	O
分词信息	B	E	B	M	E

2.3 评价指标

为了衡量网络模型对实体的发现能力,以及命名实体识别的准确性两项指标。选择了严格评价指

标和松弛评价指标两种评价方式。对同一文本而言,严格的评价指标是指当且仅当满足公式 (11)、(12)、(13) 的情况下,才判定识别正确。

$$P_i \bullet pos_b = Y_i \bullet pos_b \quad (11)$$

$$P_i \bullet pos_e = Y_i \bullet pos_e \quad (12)$$

$$P_i \bullet category = Y_i \bullet category \quad (13)$$

松弛标准需要同一条语句满足公式 (14) 和 (15) 才认定为识别正确。

$$\max(P_i \bullet pos_b, Y_i \bullet pos_b) \leq \min(P_i \bullet pos_e, Y_i \bullet pos_e) \quad (14)$$

$$P_i \bullet category = Y_i \bullet category \quad (15)$$

其中: P_i 和 Y_i 分别表示预测和实际的实体, $\bullet$ 符号用于连接属性和实体, pos_b 和 pos_e 表示标签的开始和结束位置,而 $category$ 表示实体的类别, $P_i \bullet pos_b$ 表示预测的第 i 个实体的开始位置。

本文通过 F1 值来评价模型的表现, F1 值的计算如公式 (16) 所示:

$$F1 = 2 \frac{precision \cdot recall}{precision + recall} \quad (16)$$

其中: $precision$ 和 $recall$ 分别表示模型的准确率和召回率。

2.4 数据增益系数

随着数据数量的增长,模型的训练时间将变得越来越长,但对数据进行扩充而获得的效果提升不会随着数据量的增长一直提高,所以我们要选择一个合适的扩充数量。训练数据集中有 600 份标记文本。随机选择其中 20% 数据,即 120 份数据进行扩充,根据前人的推荐参数^[4],扩大系数应在 8 ~ 16 之间。本文在经典的命名实体识别网络 Word2Vec + BiLSTM + CRF 网络中进行了相应的实验。实验结果如图 5 所示。

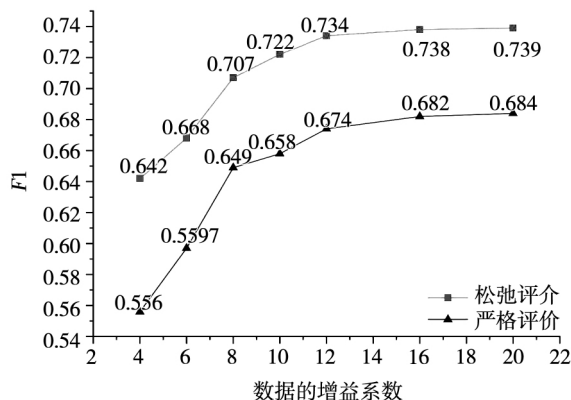


图 5 不同增益系数对于模型的影响

Fig. 5 Impact of different expansion sizes on the model

实验选取 20% 数据进行数据增强操作,并设置不同的增益系数作为对比。随着增益系数的不断增加,模型的表现也越来越好。增益系数从 4 增长到 8 时模型效果提升明显,但是增益系数在 12 以上时模型的提升明显减小,增益系数达到 16 以上时模型基本没有任何提升。所以合理的增益系数应在 10 ~ 12 之间,本文选择 10 为增益系数。

2.5 数据增强实验

在 Word2Vec + BiLSTM + CRF 模型中分别对 20%、40%、50%、60%、80% 和 100% 数据进行增强,增益系数选择 10。具体效果如图 6 和表 2 所示。从实验结果可以看出,训练数据在 50% 及以下时模型训练的效果下降明显,在 20% 的数据时松弛评价标准下 $F1$ 值仅有 0.203,准确性仅为原始数据的 $\frac{1}{4}$ 。

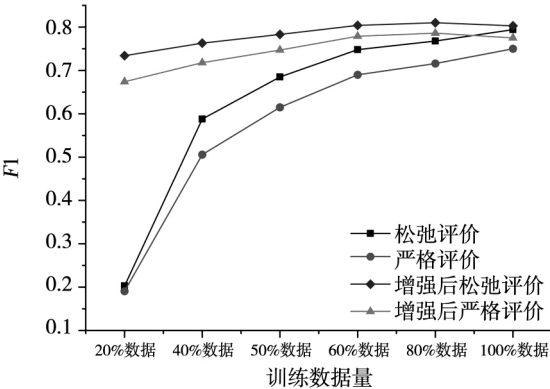


图 6 数据增强后表现

Fig. 6 Performance after data expansion

表 2 数据增强后表现($F1$ 值)

Tab. 2 Performance after data enhancement ($F1$ -score)

	20% 数据	40% 数据	50% 数据	60% 数据	80% 数据	100% 数据
松弛评价	0.203	0.588	0.685	0.748	0.768	0.794
严格评价	0.191	0.506	0.615	0.690	0.716	0.750
增强后松弛评价	0.734	0.763	0.783	0.804	0.810	0.803
增强后严格评价	0.674	0.718	0.747	0.779	0.786	0.775

本文提出的数据增强方案能有效提高小样本下模型的训练效果,在不同数量的训练数据下,都取得了较为明显的提升。20% 的原始数据在增强之后,严格和松弛评价标准较没有经过数据增强的模型分别提升了 252% 和 261%。60% 的训练数据经过数据增强后效果超过同一模型在原数据上的训练效果。利用本文提出的数据增强方法能够有效提高模型在小样本下的表现。

2.6 模型评价

对于本文提出的神经网络,分别选取在命名实体识别领域常见的神经网络模型作为对比。这里选择 Word2Vec + BiLSTM + CRF 模型、Word2Vec + BiLSTM + Softmax 模型、Bert + CRF 模型、Bert + Softmax 模型、Bert + BiLSTM + CRF 模型、特征融合模型和联合训练模型作为对比模型。其中特征融合模型采用 char_CNN + BERT + CRF 结构。联合训练模型则是 Bert + BiLSTM + CRF 命名实体识别模型和 BERT + CRF 分词模型进行联合训练,他们共享同一个 Bert 的输出。具体结果如表 3 和表 4 所示。

表 3 松弛评价标准下不同模型 $F1$ 值

Tab. 3 $F1$ -score of different models under relaxation evaluation criteria

模型	疾病与 诊断	解剖 部位	影像 检查	实验室 检查	药物	手术	总体
Word2Vec + BiLSTM + CRF	0.718	0.780	0.746	0.719	0.946	0.871	0.794
Word2Vec + BiLSTM + Softmax	0.633	0.724	0.745	0.555	0.870	0.810	0.728
Bert + CRF	0.764	0.799	0.780	0.793	0.953	0.887	0.819
Bert + Softmax	0.611	0.777	0.696	0.468	0.670	0.548	0.689
Bert + BiLSTM + CRF	0.761	0.810	0.749	0.756	0.952	0.882	0.822
特征融合模型	0.782	0.822	0.743	0.722	0.942	0.856	0.829
联合训练模型	0.783	0.806	0.771	0.709	0.933	0.886	0.823
文章中融合模型	0.778	0.829	0.762	0.669	0.929	0.889	0.832

表4 严格评价标准下不同模型 $F1$ 值Tab. 4 $F1$ -score of different models under strict evaluation criteria

模型	疾病与 诊断	解剖 部位	影像 检查	实验室 检查	药物	手术	总体
Word2Vec + BiLSTM + CRF	0.669	0.729	0.737	0.697	0.936	0.826	0.750
Word2Vec + BiLSTM + Softmax	0.582	0.659	0.745	0.469	0.876	0.781	0.678
Bert + CRF	0.716	0.768	0.789	0.853	0.965	0.850	0.791
Bert + Softmax	0.509	0.721	0.705	0.241	0.656	0.500	0.624
Bert + BiLSTM + CRF	0.718	0.778	0.796	0.804	0.968	0.833	0.795
特征融合模型	0.762	0.782	0.743	0.689	0.942	0.87	0.798
联合训练模型	0.775	0.779	0.789	0.694	0.914	0.842	0.801
文章中融合模型	0.769	0.794	0.799	0.739	0.930	0.842	0.809

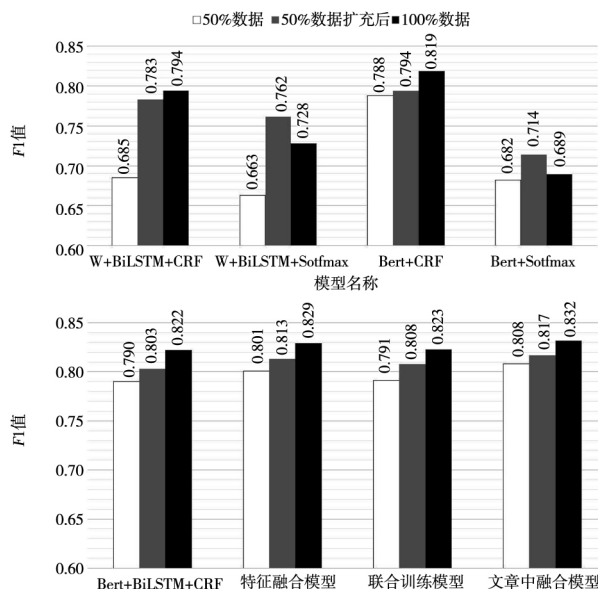
通过对比试验可以发现,随着词向量中包含的先验知识数量的提升,命名实体识别网络的总体效果也在不断提升。以 BiLSTM + CRF 为例,采用 Bert 词向量比 word2Vec 词向量的松弛评价标准 $F1$ 值提高了 0.028,当使用汉字的拼音笔画特征扩展后的 Bert 作为词向量,比 Bert 词向量本身松弛评价标准提高了 0.07。这说明我们引入汉字的拼音笔画到词向量中能够有效的优化命名实体识别的效果。CRF 层对于命名实体识别的效果也远好于使用 Softmax,因为 CRF 能有效利用上下文信息提高命名实体识别的效果。通过与分词模型联合训练,命名实体识别在松弛评价标准下并没有多少提升,但是在严格评价标准下 $F1$ 值却有 0.06 的提升,说明我们使用分词模型的联合训练能帮助我们提高实体边界的识别能力。本文提出的融合模型在松弛评价标准和严格评价标准下都取得了最好的成绩,充分的体现了改进后神经网络效果的优越性。

2.7 不同模型下的数据增强

将文中提出的数据增强技术应用到主流框架之中,也取得很好的表现,这里使用 50% 的数据进行增强。具体结果如图 7 所示,图中的 W 为 Word2Vec 的缩写。

经过数据增强操作后,模型性能都有很大的提升,与此同时文本中提出的融合模型在不同大小的数据集下都显示出了最好识别的效果。Word2Vec + BiLSTM + Softmax 和 Bert + Softmax 模型采用本文提出的数据增强操作,仅凭借一半的训练数据集就取得远超完整数据集训练的效果。使用 50% 训练数据获得的模型效果更是优于 Word2Vec + BiLSTM + CRF 模型在原数据集上的表现,与 Bert + BiLSTM

+ CRF 模型的效果也相差较小,经过数据增强后效果又有了新的提升。这说明本文提出方法能够有效提高在小样本情况下的实体识别效果。

图7 50%数据增强后模型松弛评价标准 $F1$ 值Fig. 7 $F1$ value of model relaxation evaluation standard after 50% data enhancement

3 结 语

本文针对小样本下的医疗命名实体识别提出了一种新的解决方法,包括新的神经网络和改进后的数据增强操作。新的神经模型采用拼音笔画特征扩展后的 Bert 作为词向量,联合分词模型进行训练,模型在不同大小的数据集下都取得了性能提升。改进后的数据增强操作能够在保持原有句式不变的情

况下扩展出大量数据,提高了模型在小样本情况下的性能。实验表明通过我们的方法能够有效提高小样本下的命名实体识别效果。

参考文献:

- [1] SHAALAN K, RAZA H. NER: Named Entity Recognition for Arabic [J]. Journal of the Association for Information Science and Technology, 2009, 60(8): 1652.
- [2] 俞鸿魁,张华平,刘群,等. 基于层叠隐马尔可夫模型的中文命名实体识别 [J]. 通信学报, 2006, 4(2): 87.
YU Hongkui, ZHANG Huaping, LIU Qun, et al. Chinese Named Entity Identification Using Cascaded Hidden Markov model [J]. Journal on Communications, 2006, 4(2): 87.
- [3] BO C, YI-DAN S, QI D. Automatic Recognition of Chinese Name Based on Maximum Entropy [J]. Computer Engineering and Applications, 2009, 45(4): 227.
- [4] LI L, ZHOU R, HUANG D, et al. Brief Communication: Two-phase Biomedical Named Entity Recognition Using CRFs [J]. Computational Biology and Chemistry, 2009, 33(4): 334.
- [5] DANG T H, LE H Q, NGUYEN T M, et al. D3NER: Biomedical Named Entity Recognition Using CRF-biLSTM Improved with Fine-tuned Embeddings of Various Linguistic Information [J]. Bioinformatics, 2018, 34(20): 3539.
- [6] LECUN Y, BENGIO Y, HINTON G. Deep Learning [J]. Nature, 2015, 521(7553): 436.
- [7] LUO L, YANG Z, YANG P, et al. An Attention-based BiLSTM-CRF Approach to Document-level Chemical Named Entity Recognition [J]. Bioinformatics, 2018, 34(8): 1381.
- [8] LIU Z, YANG M, WANG X, et al. Entity Recognition from Clinical Texts Via Recurrent Neural Network [J]. BMC Medical Informatics and Decision Making, 2017, 17(2): 67.
- [9] WU F, LIU J, WU C, et al. Neural Chinese Named Entity Recognition Via CNN-LSTM-CRF and Joint Training with Word Segmentation [C] //The World Wide Web Conference. 2019: 3342.
- [10] BOJANOWSKI P, GRAVE E, JOULIN A, et al. Enriching Word Vectors with Subword Information [J]. Transactions of the Association for Computational Linguistics, 2017, 5: 135.
- [11] YU J, JIAN X, XIN H, et al. Joint Embeddings of Chinese Words, Characters, and Fine-grained Subcharacter Components [C] //Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 286.
- [12] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed Representations of Words and Phrases and Their Compositionality [C] //Advances in Neural Information Processing Systems. 2013: 3111.
- [13] PETERS M E, NEUMANN M, IYYER M, et al. Deep Contextualized Word Representations [C] //Proceedings of NAACL-HLT. 2018: 2227.
- [14] RADFORD A, WU J, CHILD R, et al. Language Models are Unsupervised Multitask Learners [J]. OpenAI Blog, 2019, 1(8): 9.
- [15] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding [J]. ArXiv Preprint ArXiv: 1810. 04805, 2018.
- [16] REN X, EL-KISHKY A, WANG C, et al. Clustype: Effective Entity Recognition and Typing by Relation Phrase-based Clustering [C] //Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2015: 995.
- [17] PÉREZ, ALICIA, WEEGAR R, et al. Semi-supervised Medical Entity Recognition: A Study on Spanish and Swedish Clinical Corpora [J]. Journal of Biomedical Informatics, 2017, 71: 16.
- [18] WEI J, ZOU K. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks [C] //Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019: 6382.
- [19] MA X, HOVY E. End-to-end Sequence Labeling Via Bi-directional Lstm-cnns-crf [J]. ArXiv Preprint ArXiv: 1603. 01354, 2016.
- [20] XIE Z, WANG S I, LI J, et al. Data Noising as Smoothing in Neural Network Language Models [J]. ArXiv Preprint ArXiv: 1703. 02573, 2017.
- [21] LUO R, XU J, ZHANG Y, et al. PKUSEG: A Toolkit for Multi-domain Chinese Word Segmentation [J]. ArXiv Preprint ArXiv: 1906. 11455, 2019.

(编辑: 温泽宇)