一种中文医疗事件的联合抽取方法

纪斌，刘慧君*，李莎莎，余杰，马俊

国防科技大学 计算机学院, 湖南 长沙 410073
*lhj12uestc@163.com

摘要. 随着电子病历在医疗领域的推广应用，越来越多的研究者关注如何高效地从电子病历中抽取高价值科研信息。2020 年全国知识图谱与语义计算大会将中文电子病历临床医疗事件抽取作为评测任务，具体来说就是从中文肿瘤电子病历中抽取三种恶性肿瘤相关的属性。在本次评测任务中，结合三种属性的特点和属性间的依赖关系，我们提出了一种中文医疗事件的联合抽取方法。此外，我们提出了一种基于关键信息的全域随机替换的伪数据生成方法，提升了联合抽取方法在不同类型恶性肿瘤数据间的迁移应用能力。在测试数据集上的实验结果显示，我们提出的方法的 F1 值为 73.521%，获得了本次评测任务的第三名。

关键词: 神经网络; 中文电子病历; 医疗事件抽取; 迁移学习;

1 概述

随着电子病历的迅速普及和医疗大数据时代的到来，自然语言处理(Natural Language Processing, NLP)技术在医学领域的应用与发展，已经成为当前的研究热点。NLP 相关技术，如句子的分词，实体识别等，可以从临床医疗记录中提取有科研价值信息，以帮助科研人员进行的学术研究，从而可以支持医疗研究和辅助治疗方案决策[1]。

2020 年全国知识图谱与语义计算大会（CCKS 2020）发布中文电子病历医疗事件抽取评测任务[4]，即给定主实体为肿瘤的电子病历文本数据，定义肿瘤事件的若干属性，如肿瘤大小，肿瘤原发部位等，识别并抽取事件及属性，进行文本结构化。本次评测任务发布了 1000 份人工标注的病历作为训练数据，300 份无标注的病历作为测试数据，在本文中分别用 train 和 test 标识。针对此评测任务，我们提出了一种中文医疗事件的联合抽取方法。并且提出了一种基于关键信息的全域随机替换的伪数据生成方法，提升了联合抽取方法在不同类型恶性肿瘤数据间的迁移能力。我们提出的方法在 CCKS 2020 中文电子病历医疗事件抽取评测任务中取得了第三名的成绩。

2 相关研究

医学信息抽取指的是确定医学领域文本中的专业术语的边界，然后基于领域信息对它们进行分类[5]。目前医学信息抽取的主要方法分为浅层机器学习和深层神经网络的方法。浅层机器学习方法主要包括 HMM、ME、CRF、SVM 以及上述分类模型的改进等[6]。2015 年 Wang[7]等验证了基于 CRF 的 Gimli 方法，在 JNLPBA 2004 数据集上取得了 72.23%的 F1 值；2018 年于楠[8]等提出了多特征融合的条件随机场方法，可以准确识别中文电子病历中疾病和症状实体，同时也可准确识别未登录词。浅层机器学习方法在很大程度上依赖于人工特征的设计。为减少复杂的人工特征，2014 年 Tang[9]等采用 CRF 模型进行生物医学实体识别，在基本人工特征的基础上加入不同的词向量特征，在 JNLPBA 2004 数据集上取得了 71.39%的 F1 值。2015 年 Chang[10]等利用少量的人工特征和词向量结合的方式构建 CRF 模型并添加后处理，在 JNLPBA 2004 语料上取得了 71.77%的 F1 值。

在使用深层神经网络进行医学信息抽取的研究中，2015 年 Yao[11]等首先在无标注的生物医学文本上利用神经网络生成词向量，然后建立多层神经网络，在 JNLPBA 2004 数据集上取得了 71.01%的 F1 值。2016 年 Li[12]等采用 BiLSTM 模型在 BioCreative II GM 的数据集上取得了 88.6%的 F1 值，同时在 JNLPBA 2004 语料上取得了 72.76%的 F1 值。2018 年李丽双[13]等提出了一种基于 CNN-BLSTM-CRF 神经网络模型，在 Biocreative II GM 和 JNLPBA 2004 数据集上达到了最优的 F1 值。

3 方法

3.1 任务分析

CCKS 2020 评测任务中的肿瘤原发部位、原发肿瘤大小以及肿瘤转移部位定义[15,16]如下所示。

1. 肿瘤原发部位：肿瘤原发的身体部位，区别于肿瘤转移部位。通常情况下，肿瘤原发部位的下文为“癌”、“恶性肿瘤”、“MT”、“CA”等。部位的选择趋向于身体部位小区，如当“左肺癌”、“左肺下叶癌”同时出现于电子病历中时，选择“左肺下叶”作为肿瘤原发部位。
2. 原发肿瘤大小：描述原发肿瘤长度、面积或体积的量度，包括，常见度量单位有 MM、CM 等。
3. 肿瘤转移部位：原发肿瘤的转移部位，理论上除肿瘤原发部位外，肿瘤可向身体任何其它部位转移。

从上述三种实体的定义中我们可以得出，作为一种描述肿瘤大小的量度，原发肿瘤大小依赖于肿瘤原发部位。一个基于统计得到的事实是原发肿瘤大小与

肿瘤原发部位在电子病历中是句子级别共存的，也就是说在绝大多数情况下原发肿瘤大小和肿瘤原发部位出现在同一个句子中。但需要注意的是，并非所有与肿瘤原发部位共现的原发肿瘤大小候选词均属于原发肿瘤大小，如图 1 所示。

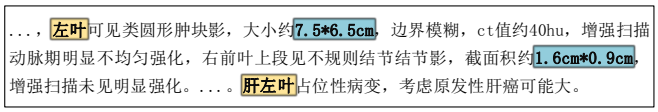

图 1 肿瘤电子病历示例，其中黄色阴影字体表示肿瘤原发部位候选词，蓝色阴影字体表示原发肿瘤大小候选词

从图 1 可以看出，1.6cm*0.9cm 是原发肿瘤大小候选词，但并不是正确的原发肿瘤大小。如何将类似的候选词正确的去除决定了原发肿瘤大小的抽取性能。Ji 等人[21]首先使用正则表达式抽取原发肿瘤大小候选词，然后通过统计肿瘤电子病历的普遍规律编写正则表达式去除错误的原发肿瘤大小候选词。由于自然语言的随意性以及电子病历书写的不规范等原因，上述方法难以在各种类型的肿瘤电子病历中推广应用。

肿瘤原发部位和肿瘤转移部位都属于身体部位或组织，在电子病历中这两种实体较为稀疏。一般情况下，一份病历中只有一个肿瘤原发部位，数个肿瘤转移部位。但电子病历中包含大量的不属于两类实体的身体部位。并且对于肿瘤转移部位来说，只有“转移”这一特征描述词可以用于辨别一个身体部位是否属于肿瘤转移部位，但这种辨别能力随着句子的长度增加而削弱。

此外，本次评测任务的训练数据集和测试数据集数据分布差异较大，主要体现在电子病历记录的肿瘤类型上的差异，如表 1 所示。

表 1 Train 和 Test 数据集记录恶性肿瘤类型信息统计

Train (%)		Test (%)	
肺	62.67	肝	28.72
乳	20.81	肠	13.18
肠	4.00	胃	12.16
肾	2.38	肺	8.11
肝	1.92	胰	7.43
食管	1.13	子宫	5.41
其它	7.09	其它	24.99

从表 1 可以看出，train 主要包含的是记录的是肺、乳两个身体部位相关的恶性肿瘤，占比 83.48%，其中肺相关的恶性肿瘤占比 62.67%。而 test 中记录的众多恶性肿瘤类型在 train 中没有出现，如胃、胰、子宫等。即使 train 和 test 中共现的恶性肿瘤类型，其在具体的恶性肿瘤描述上也存在很大的差异。

基于上述分析，我们提出了一种中文医疗事件的联合抽取方法。并且提出了一种基于关键信息的全域随机替换的伪数据生成方法，提升了联合抽取方法在不同类型恶性肿瘤数据间的迁移能力。

3.2 方法设计

图 2 给出了我们提出的中文医疗事件联合抽取方法架构图。我们通过肿瘤原发部位和原发肿瘤大小的联合抽取以及肿瘤转移部位的抽取，实现了中文医疗事件的联合抽取。

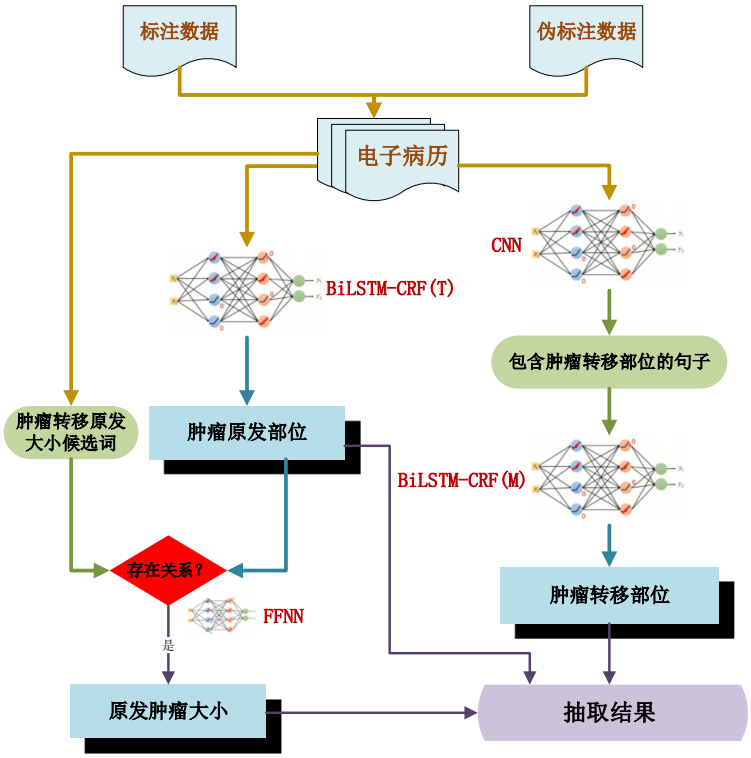

图 2 临床医疗命名实体抽取方法架构图

3.2.1 肿瘤原发部位和原发肿瘤大小联合抽取

肿瘤原发部位候选词的抽取是一个典型的命名实体识别过程。如 3.1 节所述，在电子病历中肿瘤原发部位有明显的特征描述词，因此我们采用经典的 BiLSTM-CRF 模型抽取肿瘤原发部位，其框架结构图如图 3 所示。

BiLSTM-CRF 模型实现句子级别的命名实体识别。模型的第一层是 embedding 层，在将句子输入到模型之前，需要将句子转换为向量表达。从图中可以看出，本文中的 BiLSTM-CRF 模型基于字符 embedding。具体来说就是将句子中的每个字符用字符 embedding 表示，最后得到关于句子的向量表示序列。假设一个句子 X 含有 n 个字，则该句的向量表达可表示为 X =

$(x_1, x_2, x_3, \ldots, x_n)$，其中$x_i \in R^d$，d 是字符 embedding 的维度。在输入下一层之前，设置 dropout 以缓解过拟合。

模型的第二层是双向 LSTM 层，用于自动提取句子的特征。将一个句子的各个字的字符 embedding(用 x_i表示)作为双向 LSTM 各个时间步的输入，再将正向 LSTM 输出的隐状态序列$(\overrightarrow{h_1}, \overrightarrow{h_2} \ldots\ldots, \overrightarrow{h_n})$与反向 LSTM 的$(\overleftarrow{h_1'}, \overleftarrow{h_2'}, \ldots, \overleftarrow{h_n'})$在各个位置输出的隐状态按位置拼接$H_t = [\overrightarrow{h_t}; \overleftarrow{h_t'}] \in R^m$，得到句子完整的隐状态序列$(H_1, H_2, H_3, \ldots, H_n) \in R^{n \times m}$，其中 m/2 为 LSTM 的隐藏维度。

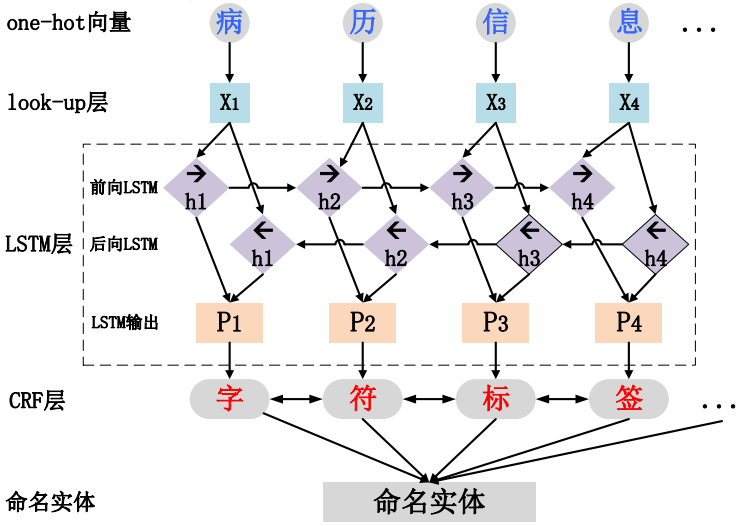

图 3 BiLSTM-CRF 模型框架结构图

模型的第三层是 CRF 层，进行句子级的序列标注。CRF 层的参数矩阵是一个维度为$(k+2) \times (k+2)$的状态转移矩阵 A，其中A_{ij}表示的是从第 i 个标签到第 j 个标签的转移得分,因此在为句子的一个字符进行标注的时候可以利用此前已经标注过的标签信息。假设$y = (y_1, y_2, y_3, \ldots, y_n)$为一个长度等于句子长度的标签序列，那么模型对于句子 X 的标签序列等于 y 的计算公式如下所示。

$$score(X, y) = \sum_{i=1}^{n+1} A_{y_{i-1}, y_i} + \sum_{i=1}^{n} P_{i, y_i}$$

可以看出整个序列的得分等于各个位置的得分之和，而每个位置的打分由两部分得到，一部分是由 LSTM 输出的p_i计算得到决定，另一部分则由 CRF 的状态转移矩阵 A 决定。模型在预测过程时使用动态规划的 Viterbi 算法来求解最优路径 [8]。

BiLSTM-CRF 模型的训练数据采用 BIO[17]的标注模式，依据人工标注信息将 train 数据集处理成适合模型训练的格式。其中，用 B-TU、I-TU 代表肿瘤原发部位首字和非首字，O 用于标注不属于命名实体的字符，一个数据标注示例如图 4 所示。

结	合	临	床	，	右	乳	腺	癌	并	右	腋	窝	淋	巴	结	肿. .
0	0	0	0	0	B-TU	I-TU	I-TU	0	0	0	0	0	0	0	0	0

图 4 语料标注示例

如前所述，原发肿瘤大小是由数字、长度单位（MM 或 CM）、表示乘法的二元符号（*、x、X 等）组成按照一定的规则构成的描述原发肿瘤的量度。一个基于统计的事实是现实世界的电子病历中并没有明显的可用于将上述度量辨别为原发肿瘤大小的特征，但可以基于其与肿瘤原发部位的依赖关系实现实体的抽取。由于自然语言的随意性，电子病历在书写过程中大量存在缩写、简写的情况，如表 1 中的黄色阴影标注的“左叶”和“肝左叶”所示，两个词语均表示“肝左叶”的含义，虽然不影响人员阅读，但给机器的识别带来了不小的挑战。

为解决上述问题，在本次评测任务中，我们为每个肿瘤原发部位候选词创建了一个缩写、简写列表，并在本次评测任务中将其与肿瘤原发部位候选词同等对待。

在本文中，肿瘤原发部位和原发肿瘤大小的联合抽取流程如下所示：

1） 使用正则表达式抽取出与肿瘤原发部位句子级别共现的所有原发肿瘤大小候选词。
2） 组合肿瘤原发部位和原发肿瘤大小候选词，得到**肿瘤大小候选关系元祖**。
3） 使用多层前馈神经网络预测肿瘤大小候选关系元祖是否存在肿瘤大小关系，获得原发肿瘤大小属性。

3.2.2 肿瘤转移部位抽取

肿瘤转移部位与肿瘤原发部位和原发肿瘤大小无明显的内在关系，因此肿瘤转移部位的抽取可以作为一个单独的任务实现。但 “转移”作为极少数的特征，难以用来定位长句中的多个肿瘤转移部位。一种启发式的方法是首先筛选出包含肿瘤转移部位的句子，然后抽取出上述句子中所有的身体部位作为肿瘤转移部位。

句子筛选

一个基于统计得到的事实是若中文电子病历的一个句子中包含“转移”、“骨质破坏”等，则该句子包含肿瘤转移部位（情况一），或该句子及其前一句包含肿瘤转移部位（情况二），或该句子及其前二句包含肿瘤转移部位（情况三）；还有一种情况为句子虽然包含“转移”，但该句本身及前面的句子都不包含转移部位（情况四）。在本次评测任务中，我们用上述关键字分类情况一、情况二。在本次评测任务中，我们使用 CNN 实现句子的筛选

CNN 模型

图 5 给出了 CNN 模型的结构。对于一个包含 n 个字的句子 X，对于其每个字x_i，用$v_i \in R^m$表示，其中v_i是 m 维的列向量。由于句子是变长序列，为了能够使用 CNN 实现句子分类，我们设置了一个句子的最大长度值 length，which is larger than length of any given sentence.当句子长度小于 length 时，我们用 0 向量补全句子，实现句子长度对齐。因此一个句子的向量表达可以表示为：

$$X_{1,length} = x_1 \oplus x_2 \oplus \ldots \oplus x_n \oplus \ldots \oplus x_{length}$$

其中，⊕符号表示向量的拼接。因此，$X_{1,length}$是一个 m*length 维的向量。若卷积窗口大小为 j，则用$X_{i,i+j-1}$表示一个卷积函数处理的字符序列$x_i, x_{i+1}, \ldots, x_{i+j-1}$的向量表达。假设一个卷积算子$w \in R^{mj}$，则卷积函数从每个卷积窗口计算得到的特征$f_i$由以下公式计算得到。

$$f_i = f(w \cdot X_{i,i+j-1} + b)$$

其中，$b \in R$是一个偏置项，f 是一个非线性激活函数，如 softmax、tanh 等。将上述卷积算子应用于完整的句子序列，将得到以下的特征表达。

$$F = [f_1, f_2, \ldots, f_{i+j-2}, f_{length-j+1}]$$

由上述公式可以得出$F \in R^{length-j+1}$，上述例子是一个卷积函数对字符序列的处理过程，卷积层有多个卷积函数，如图中所示。在本次评测任务中，不同的卷积函数的卷积窗口相同，但卷积算子不同。所有卷积函数关于输入字符序列特征映射即为卷积层的输出。在卷积层之后是一个 max-pooling 层，用于获取每个卷积函数输出的最大特征映射值，即$\tilde{F} = \max\{F\}$。按照上述方法获取每个卷积函数输出的最大特征映射值，并将其组合为一个特征向量，该向量即为 pooling 层的输出，如图中所示。

将 pooling 层的输出经过一个全连接层后得到字符序列经过卷积神经网络后的向量表达，在此向量表达上使用 softmax 函数得到一个概率分布，该概率分布即为各文本类别在输入字符序列上的概率分布值。

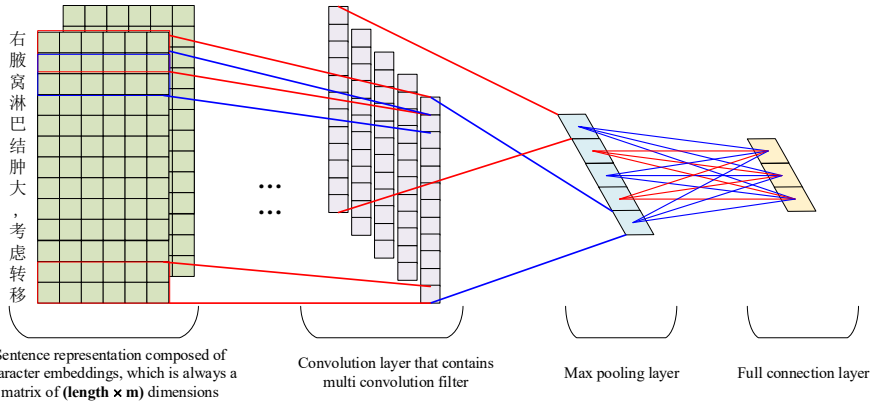

图 5 CNN 网络结构

在本次评测任务中，我们使用上述 CNN 模型判断一个句子是否包含肿瘤转移部位。类似于语言模型中的 bigram，在本次评测任务中，我们定义两种句子组合方式，uni-sentence 和 bi-sentence，其定义如下所示。

Uni-sentence：若句子中 s 中包含“转移”、“骨质破坏”等，则该句子即为 uni-sentence

bi-sentence：若句子 s 中包含“转移”、“骨质破坏”等，则该句子及该句子的前一句组合成为一个新句子。得到的句子称为 bi-sentence。

CNN 模型的目标是将 bi-sentence 类型的句子分类，筛选出包含肿瘤转移部位的句子。对训练集进行预处理并人工标注后得到训练数据，句子类别只有两类，分别用 “anatomy” 和“other”。

训练数据处理完成后，将其输入到 CNN 模型中进行训练，其中 CNN 模型采用 early stopping 的训练策略。CNN 模型训练完成后，对于测试集中的每一份中文电子病历，对其按照 bi-sentence 的方式组合后使用训练好的模型进行预测，获取分类为“anatomy”的句子。

对于测试集中的每一份中文电子病历，最终的包含肿瘤转移部位的句子有两部分构成，{uni-sentences，bi-sentences}，对上述句子集合做去重处理，去除已经包含在 bi-sentence 中的 uni-sentence。最后将得到的句子集合组合成一个长句。对上述长句进行命名实体识别得到的解剖部位即为肿瘤转移部位。

肿瘤转移部位实体识别

肿瘤转移部位实体识别使用一个 BiLSTM-CRF 模型，模型架构上述的 BiLSTM-CRF 模型架构相同，其训练数据来自于 CCKS 评测任务训练数据集。

用训练好的 BiLSTM-CRF 模型识别上一部分获取的长句中的肿瘤转移部位，获取粗粒度的肿瘤转移部位，接下来对肿瘤转移部位进行后处理，获取细粒度的肿瘤转移部位。

3.2.3 基于关键信息的全域随机替换的伪数据生成方法

在本评测任务中，我们提出了基于关键信息的全域随机替换的伪数据生成方法，实现肿瘤电子病历类型的扩充，得到伪标注数据集。

具体来说，基于关键信息的全域随机替换 (RandomReplace) 的算法的主要思想是在原始病历文本中，将肿瘤原发部位实体用同其它肿瘤原发部位实体替代，并随机替换其中的恶性肿瘤特征词，如“癌”、“CA”、“cancer”、“carcinoma”、“bi-rads”、“恶性占位病变”等，如算法 1 所示。

算法1 基于关键信息的全域随机替换算法
已知: 原始病历：emr1 原始病历肿瘤原发部位: tumer 原始病历中的肿瘤原发部位特征词：feature 伪标注病历：emr2 肿瘤原发部位实体库：corpus 肿瘤原发部位特征词库：base

在corpus中随机选取一个 entity
在base中随机选取一个item
for tumer in emr1 do:
replace(tumer, entity)
for feature in emr1 do:
replace(feature, item)
emr2←emr1
return emr2

电子病历文本通过关键信息的全域随机替换能生成与原文本有一定区别的新电子病历，因为全域随机替换算法具有随机性，所以生成的新电子病历不一定符合真实的语义， 故而称为伪标注数据。

4 实验

本文提出的方法在 test 上的获得了 73.521%的权重 F1 值（评测结果由 CCKS2020 评测平台提供），在此次评测任务中排名第三。

CCKS2019 的发布了一项与本次评测任务形式相同的评测任务，我们提出的方法在 CCKS2019 评测任务中取得了第一名的成绩。

但应用于 CCKS2019 评测任务中方法采用规则的方法抽取原发肿瘤大小，如文献[21]中所述。此种方法在方法的泛化和迁移上表现不佳。为解决上述问题，我们提出了中文医疗事件的联合抽取方法，摒弃了基于规则的方法，联合肿瘤原发部位和原发肿瘤大小，提高了方法的泛化能力和迁移能力。改进后的方法在 CCKS2019 数据集上取得了 79.48%的 F1 值，相对于原方法提高了 3.13%的绝对 F1 值，进一步验证了我们方法的有效性和泛化能力。

5 结束语

在本文中，我们提出了一种中文医疗事件的联合抽取方法，并且提出了一种基于关键信息的全域随机替换的伪数据生成方法，解决联合抽取方法在不同类型恶性肿瘤数据间的迁移问题。我们的方法在 CCKS 2020 评测任务中取得了第三名的成绩。

但仍然需要许多的工作来完善我们的方法。我们的方法中使用的所有神经网络模型均是基于随机初始化的字符 embeddings，已经可以证明的是领域相关的预训练的字符 embeddings 可以有效提高相关任务性能[19,20]，因此我们未来的工作是将预训练语言模型应用到我们的方法中，以进一步提高方法性能。

参考文献

[1] Buzhou Tang, Xiaoling Wang, Jun Yan and Qingcai Chen. Entity recognition in Chinese clinical text using attention-based CNN-LSTM-CRF. BMC Medical Informatics and Decision Making [J]. 19(supply 3):74.

[2] Zong Q. Statistical Natural Language Processing [M] .Beijing：Tsinghua University Press, 2008.

[3] CCKS 2018 named entity recognition of Chinese electronic medical record [EB/OL], https:// www.biendata.com/competition/CCKS2018_1/.

[4] 面向中文电子病历的医疗实体及事件抽取 [EB/OL]. http://sigkg.cn/ccks2020/?page_id=69

[5] 孙晓, 孙重远, 任福继. 基于深层条件碎即成的生物医学命名实体识别[J]. 模式识别

与人工智能. 2016, 29(11):997-1008.
[6] Dong X S, Qian L J, Guan Y. A multiclass classification method based on deep learning for named entity recognition in electronic medical record [C]. //Proceedings of the International 2016 New York Scientific Data Summit (NYSDS), 2016:1-10.
[7] Wang X, Yang C, Guan R. A comparative study for biomedical named entity recognition [J]. International Journal of Machine Learning & Cybernetics, 2015:1-10.
[8] 于楠, 王普, 翁壮, 方丽英. 基于多特征融合的中文电子病历命名实体识别[J]. 北京生物医学工程. 2018, 37(3):279-284.
[9] Tang B, Cao H, Wang X. Evaluating word representation features in biomedical named entity recognition tasks [J]. BioMed Research International, 2014:1-6.
[10] Chang F, Guo J, Xu W. Application of word embeddings in biomedical named entity recognition tasks [J]. Digital Inf. Manage. 2015, 13(5):321-327.
[11] Yao L, Liu H, Liu Y. Biomedical named entity recognition based on deep natural network [J]. International Journal of Hybrid Information Technology. 2015, 8(8):279-288.
[12] Li L, Jin L, Jiang Y. Recognizing biomedical named entities based on sentence vector/twin word embeddings conditioned bidirectional LSTM [C]. //Proceedings of China National Conference on Chinese Computational Linguistics. Springer International Publishing, 2016:165-176.
[13] 李丽双, 郭元凯. 基于 CNN-BLSTM-CRF 模型的生物医学命名实体识别[J]. 中文信息学报. 2018,32(1):116-122.
[14] Vikas Yadav, Steven Bethard. A survey on recent advances in named entity recognition from deep learning models [C]. //Proceedings of the 27th international conference on computational linguistics. 2018: 2145-2158.
[15] Buzhou Tang, Qingcai Chen, J. Z. I. W. (2018a). Brief for chip shared task [EB/OL].
[16] Buzhou Tang, Qingcai Chen, J. Z. I. W. (2018b). Manual for structuralizing medical imaging examination results [EB/OL].
[17] Bin Ji, Rui Liu, Shasha Li, Jie Yu, Qingbo Wu, Yusong Tan and Jiaju Wu. A hybrid approach for named entity recognition in Chinese electronic medical record [J]. BMC Medical Informatics and Decision Making. 19(supply 2):64.
[18] Zihong Liang, Junjie Chen, Zhaoping Xu, Yuyang Chen, and Tianyong Hao. A pattern-based method for medical entity recognition from Chinese diagnostic imaging text [J]. Frontiers in Artifical Intelligence. 2019,2: 1-8.
[19] Yanran Li, Wenjie Li, Fei Sun, and Sujian Li. Component-enhanced Chinese character embeddings [C]. //Proceedings of the 2015 Conference on empirical methods in natural language processing. 2015: 829-834.
[20] Shao Yan, Christian Hardmeier, and Joakim Nivre. Multilingual named entity recognition using hybrid neural networks [C]. //The sixth Swedish language technology conference. 2016.
[21] Bin Ji, Shasha Li, Jie Yu, et al. Research on Chinese medical named entity recognition based on collaborative cooperation of multiple neural network models [J]. Journal of Biomedical Informatics, 104, 2020..