计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201200016

面向中文医疗事件的联合抽取方法

余 杰[1] 纪 斌[1] 刘 磊[2] 李莎莎[1] 马 俊[1] 刘慧君[1]
1 国防科技大学计算机学院 长沙 410073
2 军事科学院后勤科学与技术研究所 北京 100091
(yj@nudt.edu.cn)

摘 要 临床病历电子化的推广普及使得利用自动化的方法从病历中快速抽取高价值的信息成为可能。作为一种重要的医学信息,肿瘤医疗事件由描述恶性肿瘤的一系列属性构成。近年来,肿瘤医疗事件抽取已成为学术界的一个研究热点,众多学术会议将其发布为评测任务,并提供了一系列高质量的标注数据。针对肿瘤医疗事件属性离散的特点,文中提出了一种中文医疗事件的联合抽取方法,实现了肿瘤原发部位和原发肿瘤大小两种属性的联合抽取和肿瘤转移部位的抽取。此外,针对肿瘤医疗事件标注文本的数量和类型少的问题,提出了一种基于关键信息全域随机替换的伪数据生成算法,提升了联合抽取方法对不同类型肿瘤医疗事件抽取的迁移学习能力。所提方法获得了 CCKS2020 中文电子病历临床医疗事件抽取评测任务的第三名,在 CCKS2019 和 CCKS2020 数据集上的大量实验验证了所提方法的有效性。

关键词:中文电子病历;医疗事件抽取;迁移学习;联合抽取;肿瘤事件
中图法分类号 TP391

Joint Extraction Method for Chinese Medical Events

YU Jie[1],JI Bin[1],LIU Lei[2],LI Sha-sha[1],MA Jun[1] and LIU Hui-jun[1]
1 College of Computer,National University of Defense Technology,Changsha 410073,China
2 Institute of Logistics Science and Technology,Academy of Military Sciences,Beijing 100091,China

Abstract The popularization of electronic clinical medical records (EMRs) makes it possible to use automated ways to quickly extract high-value information from EMRs. As a kind of crucial medical information,tumor medical event is typically composed of a series of attributes describing malignant tumors. Recently,tumor medical event extraction has become a research hotspot in the academic community,and many influential academic conferences publish it as an evaluation task and provide a series of high-quality manually annotated data. Aiming at the discrete characteristic of tumor event attributes,this paper proposes a joint extraction method,which realizes the joint extraction of tumor primary site and primary tumor size and also the extraction of tumor metastasis sites. In addition,aiming to alleviate the small counts and types of annotated tumor medical texts,this paper proposes a pseudo-data generation algorithm based on the global random replacement of key information,which improves the transfer learning ability of the joint extraction method for different types of tumor events. The proposed method wins the third place in the clinical medical event extraction evaluation task of CCKS2020,and extensive experiments on CCKS2019 and CCKS2020 datasets verify the effectiveness of the proposed method.

Keywords Chinese electronic medical record,Medical event extraction,Transfer learning,Joint extraction,Tumor event

1 引言

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)技术在医学领域的落地应用已经成为当前的研究热点。NLP 相关技术,如事件抽取、关系抽取等,可以利用自动化的方法快速从临床医疗记录中提取有科研价值的信息,从而提升科研人员的工作效率以及加速药物研究进度等[1]。

事件抽取是 NLP 的一项基础任务,其目的是从非结构化信息中抽取出用户感兴趣的事件,并以结构化形式呈现给用户。近年来,肿瘤医疗事件抽取成为学术界的一个研究热点,第四届中国健康信息处理会议(CHIP2018[2]),第十三届、十四届全国知识图谱与语义计算大会(CCKS2019[3],CCKS2020[4])均将其作为重磅评测任务,吸引了大量业界人员的参与,并提供了一系列高质量的标注数据,极大地促进了医疗事件抽取的研究。

到稿日期:2020-12-02 返修日期:2021-03-11
基金项目:国家自然科学基金(61532001)
This work was supported by the National Natural Science Foundation of China(61532001).
通信作者:刘慧君(lhj12uestc@163.com)

肿瘤医疗事件抽取，即给定主实体为肿瘤的病历文本数据，定义肿瘤医疗事件的若干属性，如肿瘤大小，肿瘤原发部位等，识别并抽取事件及属性。CHIP2018，CCKS2019 和 CCKS2020 发布的数据集定义了肿瘤原发部位、原发肿瘤大小和肿瘤转移部位 3 种属性。然而，这 3 种属性是相对离散的，即它们可以相对独立地存在，而不受其他属性的影响。例如：医学上，任何身体部位均可能成为肿瘤转移部位，而不受肿瘤原发部位的约束；原发肿瘤大小和肿瘤转移部位既无医学上也无现实意义上的关系。唯一存在部分联系的是作为描述肿瘤原发部位的度量。原发肿瘤大小通常与肿瘤原发部位句子级别共存，但这种情况又不是绝对的。针对肿瘤医疗事件抽取，笔者在以前的工作中提出了 CCMNN（Collaborative Cooperation of Multiple Neural Networks）[5]，以多神经网络协作的方式实现对 3 种属性的抽取。其中，基于肿瘤原发部位和原发肿瘤大小句子级共现的结论，CCMNN 采用基于规则的方法抽取原发肿瘤大小。然而，由于自然语言的随意性和病历文本书写的不规范性，上述方法的实际性能不佳，图 1 给出了一个示例。

…，左叶可见类圆形肿块影，大小约7.5*6.5 cm，边界模糊，ct值约40hu，增强扫描动脉期明显不均匀强化，右前叶上段见不规则结节结节影，截面积约1.6 cm*0.9 cm，增强扫描未见明显强化。….肝左叶占位性病变，考虑原发性肝癌可能大。

注：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

图 1 肿瘤病历文本示例（电子版为彩色）

Fig. 1 Tumor clinical text example

在图 1 的示例中，首先，“左叶”是“肝左叶”的简写形式，这并不影响人类的理解，但对于机器识别来说却是一个巨大的挑战；其次，“7.5 cm * 6.5 cm”和“1.6 cm * 0.9 cm”均与“左叶”句子级别共现，但后者不是原发肿瘤大小。

针对 CCMNN 中存在的问题，本文改进了 CCMNN 并提出一种中文医疗事件的联合抽取方法。本文方法实现了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。本文方法在 CCKS2020 中文电子病历临床医疗事件抽取评测任务中取得了 73.52 的 $F1$ 值，获得了本次评测任务的第三名。为了验证本文方法的有效性，在 CCKS2019 和 CCKS2020 中文医疗事件抽取数据集上针对本文方法和 CCMNN 进行了大量的对比实验。实验结果表明，本文方法相比 CCMNN 绝对 $F1$ 值有显著提升。进一步的实验分析表明，本文方法在原发肿瘤大小的抽取上性能获得了极大的提升，达到了本文的研究目的。

此外，针对肿瘤医疗事件标注病历文本数量和种类少的问题，本文提出了一种基于关键信息全域随机替换的伪数据生成算法。在 CCKS2020 中文医疗事件抽取数据集上的实验结果表明，该算法可以有效扩充标注病历文本的数量和类型，提升本文方法在不同类型肿瘤医疗事件间的迁移学习能力。

2 相关研究

类似于通用领域的信息抽取[6-8]，医学信息抽取指确定医学文本中专业术语的边界，然后基于领域信息对它们进行分类[9-10]。目前医学信息抽取的方法主要有浅层机器学习方法和深层神经网络两类方法。浅层机器学习方法主要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）等[11]。Wang 等[12]验证了基于 CRF 的 Gimli 方法，在 JNLPBA 2004 数据集上的 $F1$ 值达到 72.23；Yu 等[13]提出了多特征融合的 CRF 方法，可以准确识别病历文本中的疾病和症状实体，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。为解决上述问题，Tang 等[14]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 数据集上的 $F1$ 值达到 71.39。Chang 等[15]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 语料上的 $F1$ 值为 71.77。

在使用深层神经网络进行医学信息抽取的研究中，Yao 等[16]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。Li 等[17]采用 BiLSTM（Bidirectional Long Short-Term Memory）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。Li 等[18]提出了一种基于 CNN-BLSTM-CRF 的神经网络模型，在 Biocreative II GM 和 JNLPBA 2004 数据集上达到了最优的 $F1$ 值。

在肿瘤医疗事件抽取研究中，Liang 等[19]提出了一种基于模式匹配的抽取方法，该方法在 CHIP2018 数据集上取得了 69.7 的 $F1$ 值。Ji 等[5]提出了一种多神经网络联合协作的抽取方法，在 CCKS2019 数据集上取得了 76.35 的 $F1$ 值。Zhao 等[20]提出了一种基于多序列标注模型的方法，在 CCKS2019 数据集上取得了 76.17 的 $F1$ 值。Song 等[21]提出了一种基于 ELMo 的序列标注方法，结合规则在 CCKS2019 数据集上取得了 70.69 的 $F1$ 值。最新的相关研究中，Dai 等[22]提出了一种基于 RoBERTa 的抽取方法，结合大量外部资源对 RoBERTa 进行微调，并使用规则对数据进行预处理，该方法在 CCKS2020 数据集上取得了 76.23 的 $F1$ 值。Zhang 等[23]使用 BiLSTM-CRF 模型抽取肿瘤医疗事件，该模型基于 RoBERTa 并使用数据增强和规则对数据进行处理，在 CCKS2020 数据集上取得了 74.58 的 $F1$ 值。

医学信息抽取的研究紧跟通用信息抽取研究的步伐，但研究进展相对滞后，主要原因在于缺乏大规模高质量的医学标注数据。当前的中文电子病历肿瘤医疗事件抽取方法或采用大量的规则，大幅降低了抽取方法的泛化能力[5,19]；或高度依赖预训练语言模型[21-23]和外部资源[21-22]，提高了抽取方法对计算资源和领域知识的需求，阻碍了抽取方法的实际落地应用。

3 肿瘤医疗事件联合抽取方法

3.1 任务分析

肿瘤原发部位、原发肿瘤大小和肿瘤转移部位的定义如下。

肿瘤原发部位：最先出现某种恶性肿瘤的组织或者器官。

通常情况下，肿瘤原发部位的上下文中有明显的特征词，如“癌”“恶性肿瘤”“MT”“CA”等。

原发肿瘤大小：描述原发肿瘤大小的量度，通常的形式有长度、面积、体积。

肿瘤转移部位：恶性肿瘤从肿瘤原发部位转移到的其他组织或器官。

图 2 给出了 3 种属性示例。

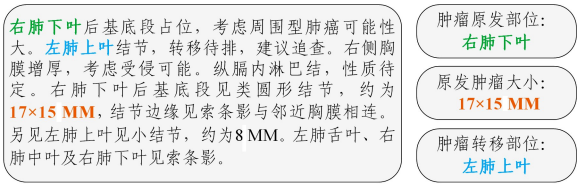

图 2　肿瘤医疗事件 3 种属性的定义

Fig. 2　Definitions of three tumor medical event attributes

CCKS2020 中文电子病历临床医疗事件抽取评测任务将研究方法跨不同类型肿瘤医疗事件抽取的迁移学习能力作为一项重要指标，因此提供的训练集和测试集数据分布差异较大，主要体现在肿瘤类型上。本文统计了 CCKS2020 医疗事件抽取数据集的训练集(train)和测试集(test)中的肿瘤类型信息，如表 1 所列。

表 1　CCKS2020 的 train 和 test 包含的肿瘤医疗事件类型统计

Table 1　Statistics on types of tumor medical events included in train and test of CCKS2020

（单位：%）

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
其他	7.09	其他	24.99

从表 1 可以看出，train 中主要包含肺、乳两种肿瘤医疗事件，占比 83.48%，其中肺相关的肿瘤医疗事件占比 62.67%。而 test 中包含的众多肿瘤医疗事件并没有在 train 中出现，如胃、胰、子宫等相关的肿瘤医疗事件。此外，train 和 test 中共现的肿瘤医疗事件，在具体的描述上也存在很大的差异。

3.2　方法设计

图 3 给出了本文提出的中文医疗事件联合抽取方法的架构图。本文方法分为两部分：1)肿瘤原发部位和原发肿瘤大小的联合抽取；2)肿瘤转移部位的抽取。

本文方法首先抽取肿瘤原发部位候选词，将抽取过程形式化为命名实体识别，并使用 BiLSTM-CRF 模型抽取，其模型结构如图 4 所示。

BiLSTM-CRF 模型的第一层是 embedding 层，将病历文本包含的每个 token 映射到 token 嵌入表示(一个 token 指病历文本中的一个汉字，或标点符号，或英文字母，或其他符号)，最后得到关于文本的嵌入表示序列。若一个病历文本 X 含有 n 个 token，则 X 的嵌入表示序列可表示为 $\boldsymbol{X}'=(\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3,\cdots,\boldsymbol{x}_n)$，其中 $\boldsymbol{x}_i\in\boldsymbol{R}^d$，$d$ 是 token 嵌入表示的维度。在输入下一层之前，设置 dropout 以缓解过拟合。

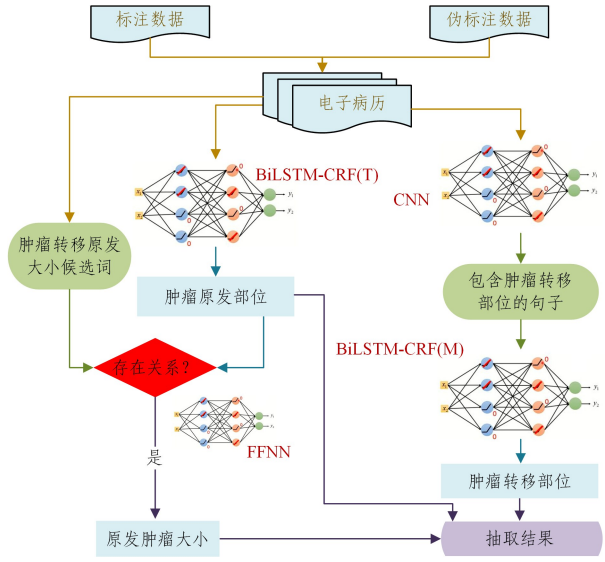

图 3　中文医疗事件联合抽取方法的架构图

Fig. 3　Architecture of joint Chinese medical event extraction method

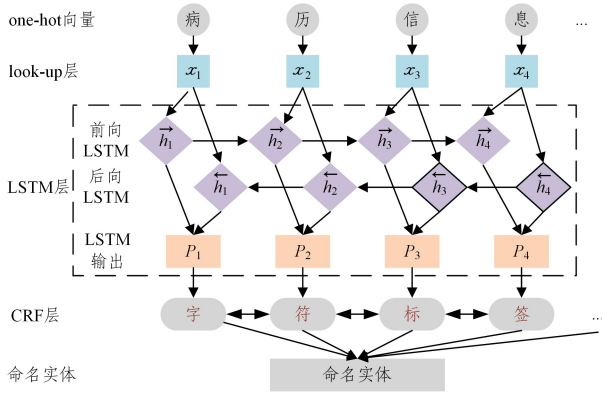

图 4　BiLSTM-CRF 模型的结构图

Fig. 4　Model architecture of BiLSTM-CRF

3.2.1　肿瘤原发部位与原发肿瘤大小的联合抽取

BiLSTM-CRF 模型的第二层是双向 LSTM 层，用于自动提取文本特征。将 $\boldsymbol{X}'$ 作为双向 LSTM 的输入，再将正向 LSTM 隐状态输出($\overrightarrow{\boldsymbol{h}_1},\overrightarrow{\boldsymbol{h}_2},\cdots,\overrightarrow{\boldsymbol{h}_n}$)与反向 LSTM 隐状态输出($\overleftarrow{\boldsymbol{h}_{1'}},\overleftarrow{\boldsymbol{h}_{2'}},\cdots,\overleftarrow{\boldsymbol{h}_{n'}}$)按位置拼接，得到文本隐状态序列($\boldsymbol{H}_1,\boldsymbol{H}_2,\boldsymbol{H}_3,\cdots,\boldsymbol{H}_n$)$\boldsymbol{R}^{n\times m}$，其中 $m/2$ 为 LSTM 的隐藏维度。

BiLSTM-CRF 模型的第三层是 CRF 层，该层训练一个状态转移矩阵 $\boldsymbol{A}$，其中 $\boldsymbol{A}_{ij}$ 表示从第 i 个标签到第 j 个标签的转移得分。$\boldsymbol{A}$ 使得模型在标注一个新 token 时可以使用此前已经标注过的标签信息。假设 $y=(y_1,y_2,y_3,\cdots,y_n)$ 为文本 X 一个的标签序列，那么模型将 X 标注为 y 的得分的计算公式如下。

$$score(X,y)=\sum_{i=1}^{n+1}\boldsymbol{A}_{y_{i-1},y_i}+\sum_{i=1}^{n}\boldsymbol{P}_{i,y_i}$$

整个得分等于各标签得分之和，各标签得分由两部分组成：1)由 CRF 的状态转移矩阵 $\boldsymbol{A}$ 决定；2)由 BiLSTM 隐状态输出计算得到。

BiLSTM-CRF 模型在解码过程时使用动态规划的 Viterbi 算法来求解最优路径，如下式所示。

$$y^* = \arg\max_{y'} score(X, y')$$

BiLSTM-CRF 模型在训练过程中使用交叉熵损失函数，模型超参数(使用“T”标识)设置如表 2 所列。

表 2 模型超参数设置

Table 2 Model hyperparameter settings

BiLSTM-CRF (T)		BiLSTM-CRF (M)		CNN	
parameter	value	parameter	value	parameter	value
bs	4	*bs*	100	*ed*	150
epoch	80	*epoch*	150	*seq_length*	600
hd	150	*hd*	300	*num_filters*	128
optimizer	*adam*	*optimizer*	*adam*	*hd*	150
learning	0.001	*learning*	0.001	*dropout*	0.5
clip	4.0	*clip*	5.0	*learning*	0.001
dropout	0.5	*dropout*	0.5	*batch_size*	100
ed	300	*ed*	300	—	—

注：“*bs*”表示 *batch_size*；“*hd*”表示 *hidden_dim*；“*ed*”表示 *embedding_dim*

BiLSTM-CRF 模型的训练数据采用 BIO[24] 的标注模式，依据数据的人工标注信息将数据处理成适合模型训练的格式。

肿瘤原发部位的候选词可能包含多个属于同一身体部位但粒度不同的候选词，因此需要对其进行筛选。筛选过程遵循精细化、小区化的原则，即保留部位描述最精确的。例如：若“肺”和“左肺上叶”同为候选词，则选择“左肺上叶”为肿瘤原发部位。

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)等按照一定的规则构成的。本文首先将病历文本中所有符合定义形式的词语抽取出来，作为原发肿瘤大小候选词。例如：图 2 中的“17X15 MM”和“8 MM”均为候选词。

接下来，组合肿瘤原发部位和原发肿瘤大小候选词，得到肿瘤大小关系候选元组，组合原则是在病历文本中肿瘤原发部位应出现在原发肿瘤大小候选词之前。由于自然语言的随意性，病例文本在书写过程中存在大量缩写、简写的情况，如图 1 中的黄色阴影标注的“左叶”和“肝左叶”所示。为解决上述问题，本文为每个肿瘤原发部位创建了一个缩写和简写列表，并将列表中的词语与对应的肿瘤原发部位同等对待。以图 1 给出的病历文本为例，关系候选元组如下：

a)〈左叶，7.5 cm * 6.5 cm〉

b)〈左叶，1.6 cm * 0.9 cm〉

以下两个元组因相对位置的原因被排除在外：

c)〈肝左叶，7.5 cm * 6.5 cm〉

d)〈肝左叶，1.6 cm * 0.9 cm〉

本文拼接肿瘤原发部位语义表示和候选元组的上下文的语义表示作为肿瘤大小关系的语义表示。以上述 a)为例，肿瘤原发部位为“左叶”，候选元组上下文为“可见类圆形肿块影，大小约”。肿瘤原发部位的语义表示由 BiLSTM-CRF 模型输出的该部位首尾字符隐状态的拼接组成，即 $\boldsymbol{H}_{\text{head}}$ 和 $\boldsymbol{H}_{\text{tail}}$；将 max pooling 函数作用于候选元组上下文的隐状态输出序列得到候选元组的上下文的语义表示，在本文中用 $\boldsymbol{H}_c$ 表示。因此肿瘤大小关系的语义表示 $\boldsymbol{H}_{\text{candidate}}$ 如下：

$$\boldsymbol{H}_{\text{candidate}} = [\boldsymbol{H}_{\text{head}}, \boldsymbol{H}_{\text{tail}}, \boldsymbol{H}_c]$$

接下来，本文首先将 $\boldsymbol{H}_{\text{candidate}}$ 输入到一个多层前馈神经网络(Feed Forward Neural Network，FFNN)；然后将输出结果输入到 sigmoid 激活函数，得到关系候选元组的概率分布，如下式所示：

$$\boldsymbol{P}_{\text{candidate}} = \text{sigmoid}(\boldsymbol{W} \cdot \boldsymbol{H}_{\text{candidate}} + \boldsymbol{b})$$

候选元组分类是一个二元分类，因此在模型训练过程中使用二元交叉熵损失函数。通过搜索 $\boldsymbol{P}_{\text{candidate}}$ 中得分最高的概率值判断候选元组是否存在肿瘤大小关系，得到预测存在肿瘤大小关系的元组，进而得到原发肿瘤大小。

通过上述方法实现了肿瘤原发部位和原发肿瘤大小的联合抽取。

3.2.2 肿瘤转移部位抽取

肿瘤转移部位独立于肿瘤原发部位和原发肿瘤大小，因此无法在三者之间建立关联关系。本文采用 CCMNN 所提出的肿瘤转移部位抽取方法，具体分为以下 3 步。

(1)筛选可能包含肿瘤转移部位的句子

一个基于统计得到的事实是若病历文本的一个句子中包含“转移”等特征词，则该句子包含肿瘤转移部位(假设一)；或该句子及其前一句包含肿瘤转移部位(假设二)；或该句子及其前二句包含肿瘤转移部位(假设三)；或该句子及前面的句子都不包含肿瘤转移部位(假设四)。依据上述 4 个假设，切割病历文本得到可能包含肿瘤转移部位的句子。

(2)句子分类

使用卷积神经网络(Convolutional Neural Network，CNN)对可能包含肿瘤转移部位的句子进行分类，得到包含肿瘤转移部位的句子集合。本文使用的 CNN 只包含一个卷积层和一个 max pooling 层，模型结构如图 5 所示。

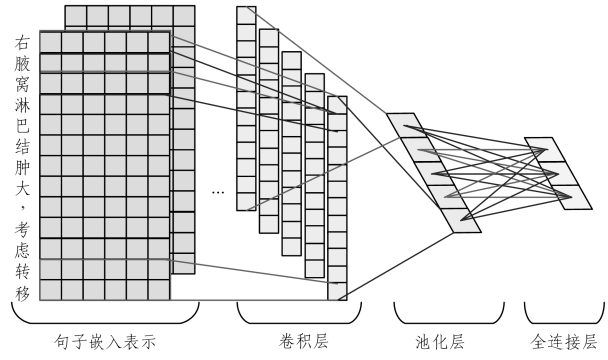

图 5 CNN 模型结构图

Fig. 5 Model architecture of CNN model

本文研究中句子的分类属于二分类，因此使用二元交叉熵损失函数作为 CNN 模型的损失函数。此外，本文研究使用的 CNN 模型超参数设置如表 2 所列，使用“CNN”标识。

(3)肿瘤转移部位实体识别

肿瘤转移部位实体识别使用 BiLSTM-CRF 模型，模型架构与图 4 一致。BiLSTM-CRF 采用交叉熵损失函数，其超参数设置如表 2 所列(使用“M”标识)。模型训练完成后用于识别步骤(2)获取的句子集合中包含的身体部位，作为最终的肿瘤转移部位。

3.2.3 基于关键信息全域随机替换的伪数据生成算法

本文提出了一种基于关键信息全域随机替换的伪数据生

成算法，对已标注的病历文本进行伪标注，获取伪标注数据，实现标注病历文本数量和类型的扩充。

具体来说，本文算法的主要思想是：在标注的病历文本中，将肿瘤原发部位用其他肿瘤原发部位随机替代，并随机替换其中的肿瘤特征词，如“癌”“CA”“cancer”“carcinoma”“bi-rads”“恶性占位病变”等，如算法 1 所示。

算法 1　基于关键信息全域随机替换的伪数据生成算法

```
已知：
  已标注病历：emr1
  已标注病历中的肿瘤原发部位：tumer
  已标注病历中的肿瘤特征词：feature
  伪标注病历：emr2
  肿瘤原发部位实体库：corpus
  肿瘤特征词库：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
```

病历文本通过关键信息的全域随机替换能够生成与原文本有一定区别的病历文本，因为全域随机替换算法具有随机性，所以生成的病历文本不一定符合真实场景，但可以极大地丰富标注的病历文本的数量和类型，减少人工标注数据的时间和人力成本。

4　实验

4.1　实验环境

本文研究所使用的软硬件环境及实验配置如表 3 所列。

表 3　软硬件环境及实验配置

Table 3　Software and hardware environments and experimental configurations

	名称	参数/版本
硬件	CPU	Intel i9 10900 k * 1
	GPU	GeForce GTX TITAN * 2
	Memory	DDR4 3200 16 GB * 2
软件	Ubuntu	16.04
	Python	3.5.2
	CUDA	8.0
	cuDNN	7.1.4
	Tensorflow	1.4.0

4.2　数据集和评价指标

针对肿瘤医疗事件抽取评测任务，CCKS2019 发布了 1 000 份标注的肿瘤病历文本作为训练集，400 份标注的肿瘤病历文本作为测试集；CCKS2020 发布了 1 000 份标注的肿瘤病历文本作为训练集，300 份标注的肿瘤病历文本作为测试集。我们基于以上两个数据集验证本文方法的有效性。

本文使用标准的准确率（P）、召回率（R）和 micro-average $F1$ 作为模型评估指标，其公式如下：

$$P=\frac{TP}{TP+FP}$$

$$R=\frac{TP}{TP+FN}$$

$$F1=\frac{2\times P\times R}{P+R}$$

其中，TP 表示真正例，即模型预测为正例、实际也为正例的属性数量；FP 表示假正例，即模型预测为正例、实际为负例的属性数量；FN 表示假负例，即模型预测为负例、实际为正例的属性数量。

4.3　实验结果

本文方法取得了 CCKS2020 中文电子病历临床医疗事件抽取评测任务的第三名。本节首先给出本次评测任务的前五名成绩，如表 4 所列。

表 4　CCKS2020 医疗事件抽取评测任务前五名成绩

Table 4　Top five results of CCKS2020 medical event extraction evaluation task

参赛队	$F1$
dst[22]	76.23
TMAIL[23]	74.58
LHJB[1)]	73.52
araloak	72.73
zhjohnchan	71.25

参赛队 dst 和 TMAIL 均使用中文 RoBERTa 预训练语言模型。dst 参赛队从网络上抓取了 96 万份医学文本用于微调中文 RoBERTa 模型，同时采用规则对数据进行清洗、字符替换等预处理操作[22]。此外，TMAIL 参赛队也采用规则对数据进行清洗、字符替换等预处理操作[23]。

dst 和 TMAIL 提出的方法和本文方法均使用 BiLSTM-CRF 模型作为方法的主要组成部分。但通过分析肿瘤原发部位和原发肿瘤大小两种属性的性质，本文方法使用两个 BiLSTM-CRF 模型实现了两种属性的抽取，因此从经验的角度来看，同等条件下本文方法对这两种属性的针对性更强，抽取效果更好。此外，同 dst 和 TMAIL 提出的方法相比，本文方法的优势有：1）未使用 RoBERTa 预训练语言模型，而是使用随机初始化的 token 嵌入表示；2）未使用任何外部资源；3）未使用任何规则对数据集进行清洗、字符替换等预处理操作。因未采用规则预处理数据，本文方法具有更好的泛化能力；因未使用外部资源和预训练语言模型，本文方法对计算资源的需求更低。

为验证本文提出的联合抽取方法的优越性，在 CCKS2019 和 CCKS2020 数据集上分别运行了本文方法和 CCMNN，即使用训练集对模型进行训练，并在对应的测试集上测试模型，实验结果如表 5 所列。为公平起见，在两种方法中，我们均未使用本文提出的伪数据生成算法。

表 5　本文方法和 CCMNN 在 CCKS2019 和 CCKS2020 数据集上的实验结果

Table 5　Experimental results of the proposed method and CCMNN on CCKS2019 and CCKS2020 datasets

方法	CCKS2019			CCKS2020		
	P	R	$F1$	P	R	$F1$
CCMNN	73.61	79.30	76.35	62.38	59.45	60.88
本文方法	76.54	82.65	79.48	69.44	61.13	65.02

1）本文方法取得第三名，模型训练集中加入了 1 000 份伪数据生成算法生成的伪数据

从表 5 可以看出，在两个数据集上，本文方法的性能一致超过了 CCMNN，证明了本文提出的联合抽取方法的有效性。具体而言，相比 CCMNN，本文方法在 CCKS2019 数据集上的绝对 $F1$ 值提升了 3.13，在 CCKS2020 数据集上的绝对 $F1$ 值提升了 4.14。

此外，由表 5 还可以得到，无论是本文方法还是 CCMNN，在两个数据集上的性能差别均较大，主要有以下两个原因：1)为评估方法的迁移学习能力，CCKS2020 数据集的训练集和测试集数据分布差异较大；2)CCKS2019 和 CCKS2020 两个数据集的数据分布整体上存在较大的差异。

为进一步探究本文方法相比 CCMNN 的优越之处，我们进一步统计了本文方法和 CCMNN 在 3 种属性上的抽取性能，统计结果如表 6 所列。由于本文方法和 CCMNN 采用相同的方法抽取肿瘤转移部位，抽取结果也相同，因此我们未将其展示在表 6 中。

表 6 本文方法和 CCMNN 在 CCKS2019 和 CCKS2020 数据集上的事件属性抽取结果($F1$ 评估标准)

Table 6 Medical event attribute extraction results of the proposed method and CCMNN on CCKS2019 and CCKS2020 datasets ($F1$ measure)

方法	CCKS2019		CCKS2020	
	肿瘤原发	原发肿瘤	肿瘤原发	原发肿瘤
CCMNN	78.12	72.63	61.03	52.66
本文方法	79.44	81.56	61.44	60.17

从表 6 可以看出，在两个数据集上，本文方法和 CCMNN 方法取得近乎一致的肿瘤原发部位抽取性能，然而在原发肿瘤大小抽取上，本文方法较 CCMNN 绝对 $F1$ 值有大幅的提升，分别为+8.93(CCKS2019)和+7.51(CCKS2020)。相比 CCMNN 采用基于规则的方法抽取原发肿瘤大小，本文提出的联合抽取方法可以有效地提升原发肿瘤大小的抽取性能，达到了本文的研究目的。

为提升本文方法的迁移学习能力，本文提出了一种基于关键信息全域随机替换的伪数据生成算法。dst[22] 和 TMAIL[23] 采用了类似的数据伪标注算法。如 TMAIL 通过对病历文本中的句子进行全局重排序的方式，获取了 2800 份伪标注数据。

为验证本文提出的伪数据生成算法的有效性，我们在 CCKS2020 数据集上进行了一系列实验。首先，使用该算法生成了 2000 份伪标注数据；然后，根据不同的训练数据组合，结合本文方法训练神经网络模型，并在 CCKS2020 测试集上进行测试，共进行 5 组实验。实验结果如表 7 所列，其中 train 指代 CCKS2020 的 1000 份训练数据；test 指代 CCKS2020 的 300 份测试数据；train+500，train+1000，train+1500，train+2000 分别指代在 train 中加入对应数量的伪标注数据。此外，由于伪标注数据具有随机性，本文对上述实验过程进行了 10 次迭代，取这 10 次迭代 $F1$ 值的平均值作为最终的 $F1$ 值。由表 7 可以看出，在 train 中加入 1000 份伪标注数据时，本文方法取得了 74.68 的 $F1$ 值，超越了本文方法在 CCKS2020 医疗事件抽取评测任务中的 $F1$ 值(73.52)。此外，我们还可以总结出：随着在训练集中加入的伪标注数据量的增加，本文方法性能先上升，到达一个峰值后下降，但伪标注数据始终有益于本文方法。原因在于：1)该算法可以极大地扩充标注的病历文本数量和类型，这对模型性能的提升至关重要；2)该算法生成的伪标注样本具有随机性且不一定符合真实场景，加入过多的伪标注数据会对模型的正确性产生一定的干扰，从而影响模型性能的提升。

表 7 本文方法在 CCKS2020 数据集上针对伪标注数据的实验结果

Table 7 Experimental results of the proposed method on CCKS2020 dataset for pseudo-labeled data

训练集	测试集	$F1$
train	test	65.02
train+500	test	69.76
train+1000	test	74.68
train+1500	test	71.44
train+2000	test	68.01

未使用伪数据增强时，本文方法在 CCKS2020 测试集上取得了 65.02 的 $F1$ 值；使用伪数据增强时，本文方法取得的最优 $F1$ 值为 74.68，绝对 $F1$ 值获得了 9.66 的提升。如表 1 所列，CCKS2020 的训练集和测试集包含的肿瘤医疗事件类型存在巨大的差异，上述实验结果则验证了本文提出的伪数据生成算法可以有效提高本文方法在不同类型肿瘤医疗事件抽取间的迁移学习能力。

结束语 本文提出了一种中文医疗事件的联合抽取方法，实现了两种肿瘤事件属性的联合抽取，并且提出了一种基于关键信息全域随机替换的伪数据生成算法，提高了模型的迁移学习能力。本文方法在 CCKS2020 中文电子病历临床医疗事件抽取评测任务中取得了第三名，在 CCKS2019 和 CCKS2020 数据集上的大量实验表明，本文方法相比 CCMNN 方法性能有较大的提升，尤其在原发肿瘤大小抽取上性能取得了大幅提升，达到了本文的研究目的。但本文方法提出的伪数据生成算法具有较大的随机性，导致生成的伪数据不一定符合真实语义，在一定程度上有损模型性能。接下来，我们将研究基于语义相似度替换的伪数据生成算法，提高伪数据的生成质量，进一步提升模型性能。

参 考 文 献

[1] TANG B Z, WANG X L, YAN J, et al. Entity recognition in Chinese clinical text using attention-based CNN-LSTM-CRF [J]. BMC Medical Informatics and Decision Making, 2019, 19(S3): 74.

[2] Extraction of clinical medical entities and attributes from Chinese electronic medical records [EB/OL]. [2020-11-28]. http://icrc.hitsz.edu.cn/chip2018/task.html.

[3] Named entity recognition for Chinese electronic medical records [EB/OL]. [2020-11-28]. http://www.ccks2019.cn/?page_id=62.

[4] Medical entity and event extraction for Chinese electronic medical records [EB/OL]. [2020-11-28]. http://sigkg.cn/ccks2020/?page_id=69.

[5] JI B, LI S S, YU J, et al. Research on Chinese medical named entity recognition based on collaborative cooperation of multiple

neural network models[J]. Journal of Biomedical Informatics, 2020,104:103395.

[6] LYU J N, XING C Y, LI L. Video Character Relation Extraction Based on Multi-feature Fusion and Fine-granularity Analysis[J]. Computer Science, 2021, 48(4): 117-122.

[7] DING L, XIANG Y. Chinese Event Detection with Hierarchical and Multi-granularity Semantic Fusion[J]. Computer Science, 2021, 48(5): 202-208.

[8] ZHANG D, CHEN W L. Chinese Named Entity Recognition Based on Contextualized Char Embeddings [J]. Computer Science, 2021, 48(3): 233-238.

[9] ZHOU X J, XU C M, RUAN T. Multi-granularity Medical Entity Recognition for Chinese Electronic Medical Records[J]. Computer Science, 2021, 48(4): 237-242.

[10] SUN X, SUN C Y, REN F J. Biomedical named entity recognition based on deep conditional random fields[J]. Pattern Recognition and Artificial Intelligence, 2016, 29(11): 997-1008.

[11] 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.

[12] WANG X, YANG C, GUAN R. A comparative study for biomedical named entity recognition[J]. International Journal of Machine Learning & Cybernetics, 2018, 9(3): 373-382.

[13] YU N, WANG P, WENG Z, et al. Named entity recognition in Chinese electronic medical records based on multi-feature integration[J]. Beijing Biomedical Engineering, 2018, 37(3): 279-284.

[14] TANG B, CAO H, WANG X. Evaluating word representation features in biomedical named entity recognition tasks [J]. BioMed Research International, 2014: 240403.

[15] 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.

[16] 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.

[17] 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.

[18] LI L S, GUO Y K. Biomedical named entity recognition with CNN-BLSTM-CRF[J]. Journal of Chinese Information Processing, 2018, 32(1): 116-122.

[19] LIANG Z, CHEN J, XU Z, et al. A Pattern-Based Method for Medical Entity Recognition From Chinese Diagnostic Imaging Text[J]. Frontiers in Artificial Intelligence, 2019, 2: 1-8.

[20] ZHAO G, ZHANG T, WANG C Y, et al. Team MSIIP at CCKS 2019 Task 2 [EB/OL]. [2020-11-11]. https://conference.bj.bcebos.com/ccks2019/eval/webpage/pdfs/eval_paper_1_2_2.pdf.

[21] SONG Y W, LUO L, LI N, et al. NER-PS-MS: Medical Attribute Extraction based on Medical Named Entity Recognition [EB/OL]. [2020-11-09]. https://conference.bj.bcebos.com/ccks2019/eval/webpage/pdfs/eval_paper_1_2_3.pdf.

[22] DAI S T, WANG Q, HUANG P P, et al. Small sample medical event extraction based on pre-trained language model. [EB/OL]. [2020-11-28]. CCKS2020 evaluation paper, https://bj.bcebos.com/v1/conference/ccks2020/eval_paper/ccks2020_eval_paper_3_2_1.pdf.

[23] ZHANG X N, ZHAO X Y, GE S, et al. ccks2020 medical event extraction based on named entity recognition [EB/OL]. [2020-11-28]. CCKS2020 evaluation paper, https://bj.bcebos.com/v1/conference/ccks2020/eval_paper/ccks2020_eval_paper_3_2_2.pdf.

[24] JI B, LIU R, LI S S, et al. A hybrid approach for named entity recognition in Chinese electronic medical record[J]. BMC Medical Informatics and Decision Making, 2019, 19(S2): 64.

YU Jie, born in 1982, Ph.D, research fellow, master supervisor, is a member of China Computer Federation. His main research interests include operating system, artificial intelligence and natural language processing.

LIU Hui-jun, born in 1993, Ph.D. Her main research interests include natural language processing and text against attack and defense.