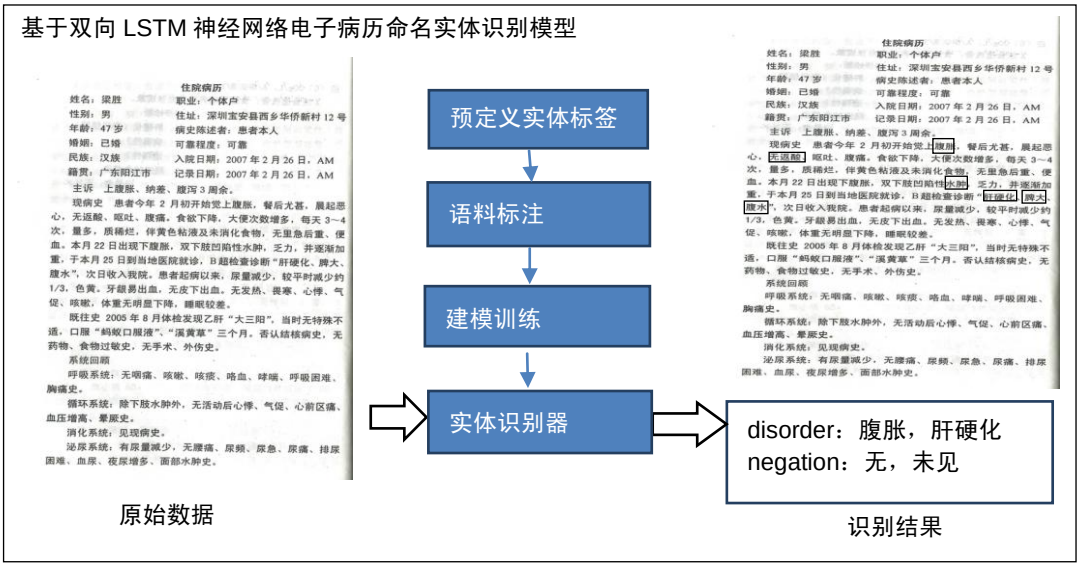
·研究原著·

基于双向LSTM神经网络电子病历命名实体的识别模型

杨红梅[1]，李　琳[2]，杨日东[1]，周　毅[1, 2] ([1]中山大学中山医学院，广东省广州市　510080；[2]新疆医科大学，新疆维吾尔自治区乌鲁木齐市　830011)

DOI:10.3969/j.issn.2095-4344.0302　　ORCID: 0000-0002-3254-0429(杨红梅)

文章快速阅读：



杨红梅，女，1993 年生，云南省楚雄市人，汉族，中山大学在读硕士，主要从事医学信息抽取，医学数据挖掘研究。

通讯作者：周毅，博士，副教授，中山大学中山医学院，广东省广州市 510080；新疆医科大学，新疆维吾尔自治区乌鲁木齐　830011

中图分类号:R318
文献标识码:B
稿件接受：2018-03-13

文题释义：

命名实体识别：又称作“专名识别”，是指识别文本中具有特定意义的实体。命名实体识别是信息提取的重要基础工具，在自然语言处理技术走向实用化的过程中占有重要地位。

语言模型：语言模型是用来计算一个句子(或者词序列)的概率的模型。一个长度为 n 的句子 W 可以用词序列 w_1，w_2，…，w_n 表示。那语言模型就是求这个词序列 W 的概率 $P(W)=P(w_1, w_2, \cdots, w_n)$。

电子病历：是在临床治疗过程中产生的，由医务人员撰写的描述患者医疗活动的记录，包括患者所患疾病、症状、检查、治疗、检查结果以及所发生的时间。这些信息相互联系，是患者的身体状况和医疗知识的体现，可以支持临床辅助决策系统、精准医学研究和疾病监控等应用。

摘要

背景：电子病历数据是医疗领域大数据的重要源头，是医学知识的体现。电子病历是患者就医过程的记录，是临床辅助决策系统、精准医学研究和疾病监控等应用的重要数据支撑。

目的：研究电子病历的信息抽取技术，提取中文电子病历中的重要医学实体，支持肝细胞癌的知识发现。

方法：数据集来自广东省某三甲医院的电子病历数据库。共收集了 240 例患有肝细胞癌的病历记录(18 542 个句子)，包括入院记录和出院小结。按照预先定义的标准进行标注。随机抽取 180 例患者病历(13 839 个句子)进行训练，并保留 60 个病例记录(4 703 个句子)作为测试集。利用双向的 LSTM 网络结合 CRF 训练命名实体识别模型。在测试数据集上评估 NER 系统的性能，并计算出严格匹配的准确率、召回率和 F1 值。

结果与结论：对测试数据集的评估表明，入院记录中实体识别 F1 值为 0.853 5，出院小结中实体识别的 F1 值为 0.726 5，总体 F1 值为 0.805 2。研究实现了电子病历文本自动命名实体识别模型，下一步的研究重点将改进实体抽取的准确率。

关键词：

电子病历，命名实体识别，BiLSTM，CRF；组织构建

主题词：

病案系统，计算机化；神经网络(计算机)；肝肿瘤；组织工程

基金资助：

国家重点研发计划精准医学专项基金项目(2016YFC0901602)；NSFC-广东大数据科学中心联合基金项目(U1611261)；广东省前沿与关键技术创新专项基金项目(2014B010118003)；广州市 2017 年产学研协同创新重大专项(201604016136)；广州市健康医疗协同创新重大专项(201604020016)

Yang Hong-mei, Master candidate, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou 510080, Guangdong Province, China

Corresponding author: Zhou Yi, M.D., Associate professor, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou 510080, Guangdong Province, China; Xinjiang Medical University, Urumqi 830011, Xinjiang Uygur Autonomous Region, China

Named entity recognition based on bidirectional long short-term memory combined with case report form

Yang Hong-mei[1], Li Lin[2], Yang Ri-dong[1], Zhou Yi[1, 2] ([1]Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou 510080, Guangdong Province, China; [2]Xinjiang Medical University, Urumqi 830011, Xinjiang Uygur Autonomous Region, China)

Abstract

BACKGROUND: Electronic medical record (EMR) is an important source of medical source, reflecting medical knowledge. There are patient clinical features in EMR, which enables decision support system and precision medicine.

OBJECTIVE: To extract important medical entities of EMR using information extraction, and to discover hepatocellular carcinoma knowledge.

METHODS: The EMR database of a Grade-A Tertiary hospital in Guangdong Province was used. We retrieved clinical records (18 542 sentences) of 240 patients suffering from hepatocellular carcinoma, including admission notes and discharge summaries. The records were remarked according to the predetermined standards. Totally 180 patients' records (13 839 sentences) were selected randomly for training and 60 patients' records (4 703 sentences) were remained for testing. Bidirectional long short-term memory combined with case report form was used to identify the model. The performance of NER systems was evaluated on the test datasets, and precision, recall, F1 of strict matching were caculated.

RESULTS AND CONCLUSION: Evaluation on the dataset showed that an F1-measure of 0.853 5 was for admission, F1-measure of 0.726 5 was for the discharge summaries, and an overall F1-measure was 0.805 2. In this study, we have achieved the auto-name entity identification model of EMR, but the accuracy of entity extraction needs further investigation.

Subject headings: Medical Records Systems, Computerized; Neural Networks (Computer); Liver Neoplasms; Tissue Engineering

Funding: the National Key Research & Development Precision Medicine Project of China, No. 2016YFC0901602; the Project of NSFC-Guangdong Big Data Science Center, No. U1611261; the Advanced and Key Technology Innovation Project of Guangdong Province, No. 2014B010118003; the Major University-Industry Cooperation Innovation Project of Guangzhou in 2017, No. 201604016136; the Major Project of Health Medicine Cooperation Innovation Project of Guangzhou, No. 201604020016

0 引言 Introduction

电子病历(electronic medical record，EMR)是在临床治疗过程中产生的，由医务人员撰写的描述患者医疗活动的记录，病历电子化使得大规模病历的自动分析成为可能。电子病历记录患者所患疾病、症状、检查、治疗、检查结果，以及所发生的时间点，这些信息是重要的临床证据，自动抽取这些信息能够更加高效、精确地收集证据支持临床辅助决策系统[1]、精准医学研究和疾病监控等应用[2-5]。据调查结果表明，阻碍电子病历二次利用的主要原因是受数据结构化程度的制约[6-8]。因此，迫切需要探索出自动将自然语言描述的文本转化为质量较高的、可计算的知识，便于计算机理解和使用的信息抽取方法。

命名实体识别(named entity recognition，NER)是信息抽取的关键组件，对信息检索、自动问答、机器翻译和知识库构建等研究和应用有重要的意义。近年来，深度学习在自然语言处理领域取得良好的效果，尤其是应用词嵌入技术训练词向量作为词语的特征替代人工提取特征。而电子病历命名实体和实体关系标注语料库的构建是首当其冲的。国内电子病历命名实体和实体关系标注语料库构建的基础上在标注体系完整和规模方面都存在局限。该研究调研了国内外电子病历命名实体和实体关系标注语料库构建，结合项目的需求特点，提出适合该研究的标注体系，在医生的指导和参与下，标注了内容丰富、规模较大、质量较高的命名实体和实体关系语料库。此外，文章将命名实体识别作为序列标注任务，训练词向量作为特征，基于双向LSTM(Bidirectional Long Short-Term Memory，BiLSTM)算法训练病历文本的命名实体识别模型。

中文生物医学文本命名实体识别研究方法大致分为基于规则和词典的方法、基于机器学习的方法。基于规则和词典的方法是命名实体识别中最早使用的方法，该方法大多采用语言学专家手工构造规则模板，以模式和字符串相匹配为主要手段，这类系统大多依赖于知识库和词典的建立。何林娜等[9]采用一种基于特征耦合泛化(FCG)的半监督学习方法生成药名词典，然后将药名词典和条件随机场结合进行药名实体识别。栗伟等[10]提出了一种基于CRF与规则相结合的医学病历实体识别算法，先用CRF进行病历实体的初始识别，然后基于规则进行病历实体识别结果优化。近年来，机器学习方法的应用取得比较好的效果。Lei等[11]研究SVM、ME、CRF、SSVM等机器学习方法，组合字袋、词袋、词性等不同的特征进行组合特征抽取电子病历中的医疗问题、医疗过程、用药等实体。

目前，深度学习框架应用于命名实体识别的研究越来越受到关注，CCKS 2017发布了中文电子病历命名实体识别评测任务，该任务的目标是从给定电子病历数据集中抽取疾病、症状、身体部位、检验检查等实体，吸引了国内许多从事医学健康信息处理的研究者的参与。其中，Hu等[12]，Wu等[13]，Xia等[14]构建了双向RNN网络训练命名实体识别模型，并训练词向量作为输入特征，实验结果对比同类数据集上训练的生物实体命名识别模型有较高的F1值。因此，作者尝试将BiLSTM应用于病历文本的命名实体识别。

1 材料和方法 Materials and methods

1.1 设计 电子病历的信息抽取技术研究。

1.2 时间及地点 实验于2017年7月至2018年1月在中山大学中山医学院地点完成。

1.3 材料 研究所使用的数据集来自广东省某三甲医院的电子病历数据库。共收集了240例患有肝细胞癌的病历

记录(18 542个句子)，包括入院记录和出院小结。按照预先定义的标准进行标注。随机抽取180例患者病历(13 839个句子)进行训练，并保留60个病例记录(4 703个句子)作为测试集。

1.4 基于BiLSTM的命名实体识别模型方法

1.4.1 LSTM 传统的神经网络是从输入层到隐含层再到输出层，层与层之间是全连接，每层之间的节点无连接。与传统神经网络相比，RNN网络是一种节点定向连接成环的人工神经网络，能够记忆前面的信息并应用于当前输出的计算中，即隐藏层之间的节点不再无连接而是有连接的，并且隐藏层的输入不仅包括输入层的输出还包括上一时刻隐藏层的输出，这使得它更容易处理语言模型。尽管如此，RNN在处理语言模型建模时存在梯度爆炸和梯度消失的问题，为了解决这个问题，Hochreiter和Schmidhuber提出了LSTM网络，LSTM网络是RNN的一种特殊形式，其引入了记忆门单元和门限限制，实现了对长距离信息的有效利用，并解决了梯度消失的问题。t 时刻，给定输入 x，LSTM的隐层的输出表示h_t的具体计算过程如下：

$$
\begin{aligned}
i_t &= \sigma(W_i[h_{t-1}, x_t] + b_i)\\
f_t &= \sigma(W_f[h_{t-1}, x_t] + b_f)\\
\tilde{c}_t &= \tanh(W_c[h_{t-1}, x_t] + b_c)\\
c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t\\
o_t &= \sigma(W_o[h_{t-1}, x_t] + b_o)\\
h_t &= o_t \odot \tanh(c_t)
\end{aligned}
$$

其中，W表示连接两层的权重矩阵，b表示偏移列向量。LSTM将信息存放在循环网络正常信息流之外的门控单元中。这些单元可以通过开关判断存储的信息内容以及写入或读取信息的时机。**图1**所示是LSTM网络单元的结构。f遗忘门，表示对于当前时刻的输入 x，它决定了从上一时刻传来的信息要丢弃的部分。i表示输入门，它决定在t时刻应该更新哪些值，$\tilde{C}$是一个候选值的向量，将i和$\tilde{C}$组合起来得到C对神经元状态进行更新。o是输出层，决定神经元状态需要输出的部分。h是整个网络的输出。

1.4.2 BiLSTM-CRF LSTM模型神经元信息只能从前向后传递，也就意味着，当前时刻的输入信息仅能利用之前时刻的信息。然而对于序列标注任务来说，当前状态之前的状态和之后的状态应该是平权的。命名实体的标签之间具有强烈的依赖关系，BiLSTM则既能利用当前时刻之前的信息，又能利用之后的信息，非常适用于命名实体识别任务。

BiLSTM的结构如**图2**所示，t 时刻，给定序列$X=(X_1, X_2, X_3, \ldots, X_T)$，对于每一个词语$X_t$，用其对应的词向量$e_t$表示，以前k个词作为窗口，那么用词向量将该序列表示为$x_t=<e_{t-k}, e_{t-k+1}, e_{i-k+2}, \ldots, e_{t-1}, e_t>$。那么BiLSTM的输入就可以表示为$[x]_1^T=(x_1, x_2, x_3, \ldots, x_T)$，输出$[h]_1^T$定义为向前传递的LSTM的输出$h_f=<h_{f1}, h_{f2}, h_{f3}, \ldots, h_{fT}>$和向后传递的LSTM的输出$h_b=<h_{b1}, h_{b2}, h_{b3}, \ldots, h_{bT}>$拼接得到，$h_t$的意义为$x_t$的深层特征表示。之后通过softmax函数将$h_t$映射到标注结果的决策概率，记为

P_{ht}，y_t=softmax(h_t)

在CRF层是通过过去的输入以及输入所属的状态来预测当前输入所属的状态。通过定义概率转移矩阵$T_{i,j}$来表示从状态 i转移到状态j的评分。定义$[O]_1^n$为输出状态序列，那么整个网络和输出状态序列的转移评分为

$$S([x]_1^T, [O]_1^T, \theta, T)=\sum_{t=1}^{T}(P_{ht, yt}+T_{t-1, t})$$

为了让训练的过程不断优化模型的参数，定义损失函数£(θ，T)，用log似然在所有可能的状态序列$[j]_1^T$的评分正则化以避免过拟合。

$$£(\theta, T)=S([x]_1^T, [O]_1^T, \theta, T)-\log\sum_{\forall[j]_1^T} e^{S([x]_1^T, [j]_1^T, \theta, T)}$$

最后，利用Adam优化器使整个模型的损失达到最小。Adam是一种基于低阶矩估计的参数优化方法，对内存的需求较小，能为不同的参数计算不同的自适应学习率，适用于本研究的模型参数估计[15]。在预测阶段，利用Viterbi算法搜索转移评分最高的状态序列作为预测结果。

1.4.3 分词 分词是中文自然语言处理的基础。中文病历的语言常常不同于规范的中文句子结构，具有领域性强，句子凝练的特点。常用的分词工具在开放领域中有比较好的应用，但无法满足电子病历中特定表达的分词，例如：“肝细胞癌”作为一个独立的表达，往往被分为“肝细胞”、“癌”，这意味着命名实体识别将纳入分词的误差。在该研究中，希望抽取有意义的信息表达，而不仅限于医学术语。为了解决这个问题，作者收集了现有医学术语以建立一个专用字典，包括国际疾病分类编码第十版(international Classification of diseases，ICD-10)和医学网站的常用疾病描述，此外，还纳入了数据集中的标注实体，基于结巴分词[16]，对语料进行分词。

1.4.4 词嵌入 词嵌入也叫做词语的分布式表示，通过词嵌入技术能够将表示词语之间的语义关系。在one-hot向量表示方法中，任意2个词之间是孤立的，没有联系的，一般会出现维灾难和矩阵稀疏的问题。词嵌入通过训练神经网络语言模型得到词语的低维连续实数向量，以向量之间的距离衡量词语之间的相关性。文章利用word2vector工具基于skip-gram方法训练字向量，以500份分词的病历文本训练128维的词向量，大多数训练参数的设置参照Mikolov等[17]的研究，与之不同的是，此项研究采用128作为向量维度，窗口大小为5。

2 结果 Results

2.1 语料及标注结果 研究所用的数据是广东省某大型三甲医院的自2008至2016年的肝肿瘤患者的住院病历中

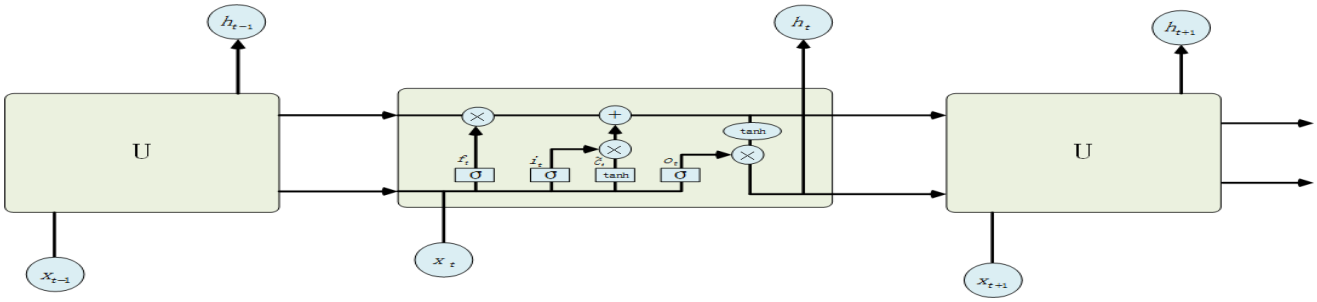


图 1　LSTM 网络单元结构

Figure 1　LSTM unit structure

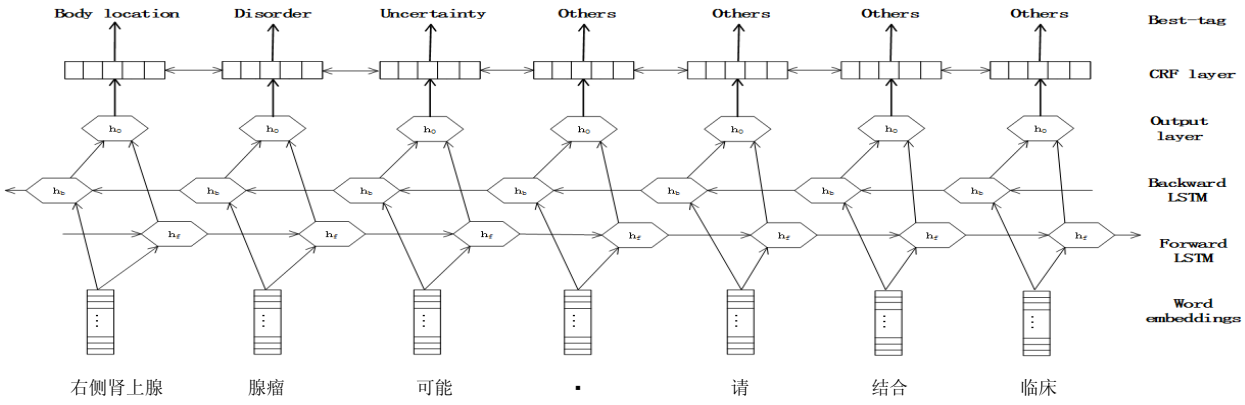


图 2　BiLSTM-CRF 网络结构图

Figure 2　Structure of BiLSTM-CRF

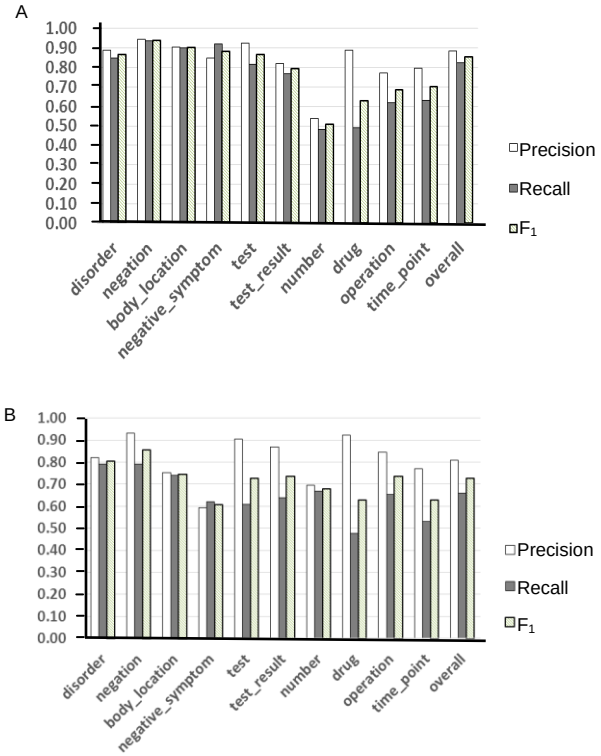


图 3　入院记录和出院小结中的各命名实体识别结果

Figure 3　Entity identification results in admission notes and discharge summaries

图注：图 A 为入院记录，B 为出院小结。在数据集中分布较多的实体，模型在该类实体是那个的表现较优。

表 1　实体识别结果

Table 1　Entity identification results

实体类型	实体定义	入院记录		出院小结	
		训练集	测试集	训练集	测试集
disorder	疾病和症状	2 289	1 567	1 772	676
negation	否定	266	231	134	79
uncertainty	表示不确定	112	83	32	22
body_location	身体部位	1 140	683	1 215	400
negative_symptom	阴性特征	351	414	285	83
test	检查项目	612	349	110	376
test_result	检查结果	229	41	1 112	161
image_feature	影像特征	309	148	173	83
pathology_feature	病理特征	58	5	291	46
condition_change	病情变化	58	25	66	20
onset_characteristic	发病特征	72	40	32	12
size	肿瘤大小	65	46	133	27
number	肿瘤数量	68	24	66	19
drug	用药	109	117	261	114
operation	手术	146	75	402	77
exposure	外部暴露因素	41	193	9	1
time_point	时间点	205	187	370	80
period	时间段	216	124	142	53
person	人物	23	47	10	2
hospital	医院名称	100	76	30	6

的入院记录和出院小结。依据该研究的数据抽取需求，由3名临床医学专业人员标注共计20种实体，**表1**列举出了相应的实体类型及示例。为了验证语料标注的一致性，让3名标注人员分别标注同样的5份数据(共377个句子)，计算标注一致性达88%。

2.2 研究结果及分析 实验中将240份病历的数据集按照3∶1的比例随机划分为训练集和测试集。训练集和测试集中各个实体的的数量分布如表1所示。基于tensorflow编程在训练集上拟合模型，大多数参数参照以往的研究进行设置和调整[18-20]。评价指标采用准确率Precision、召回率Recall和F1值[21]，具体计算公式如下：

$$\text{Precision}=\frac{\text{识别正确实体数}}{\text{识别实体数}}$$

$$\text{Recall}=\frac{\text{识别正确的实体数}}{\text{总的实体数}}$$

$$\text{F1}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$$

分别在入院记录和出院小结和全部训练集上训练NER模型，模型识别结果如**表2**所示。**图3**显示在数据集中分布较多的实体，模型在该类实体是那个的表现较优。例如，入院记录中“ “exposure”、“negative_symptom”和“body_location”实体的F1值较高，而出院记录中“exposure”实体极少出现，而“negative_symptom”实体也分布较少，因此，F1值相对较低。

表 2 实体识别结果
Table 2 Entity identification results

数据集	Precision	Recall	F1
入院记录	0.885 3	0.823 8	0.853 5
出院小结	0.809 3	0.659 0	0.726 5
总体	0.830 9	0.781 0	0.805 2

3 讨论 Discussion

随着世界范围内电子病历的广泛使用，由自由文本所描述的表型信息语料库不断增长，电子病历的信息抽取成为国内外医学信息研究者广泛关注的问题。电子病历是在临床治疗过程中产生的，由医务人员撰写的描述患者医疗活动的记录，包括患者所患疾病、症状、检查、治疗、检查结果，以及所发生的时间。这些信息相互联系，是患者的身体状况和医疗知识的体现，可以支持临床辅助决策系统、精准医学研究和疾病监控等应用。据调查结果表明，阻碍电子病历二次利用的主要原因是受数据结构化程度的制约[22]。电子病历的信息抽取是有价值的研究问题。

带注释的语料库使监督机器学习系统的训练能够自动执行相同类型的注释[7, 23]。但是，这些系统的性能仍然与他们的训练数据密切相关。测试数据与训练数据的差异性越显著(例如新闻与临床文本数据之间的差距)系统的性能表现越低。为了捕捉临床文本中复杂的语义类型，此项研究面向中文电子病历信息抽取的需求[24-25]，拟抽取肝癌患者住院病历中20种实体，制定一套标注规范，并标注了240份肝癌病历语料库，包括240份入院记录，240份出院小结和240份放射报告，是目前规模最大的中文肝癌专科语料库[26]。基于该语料库，文章利用BiLSTM网络结合CRF训练命名实体识别模型，抽取医学实体信息。

研究入院记录的实体识别结果较优，入院记录中体格检查和过往史的句子描述部分相对规范，因此，身体部位、检查和暴露因素的识别率偏高。而出院小结内容高度概括，覆盖实体复杂，实例不足造成识别率偏低。例如时间和检查结果，分词的粒度与实体组词不匹配的问题仍然大量存在，尤其是在症状描述、时间和检查结果等实体中[27-28]。显然，分词错误将直接传递到实体识别，影响识别准确率。

目前也存在大量的基于词位的命名实体识别方案，基于词位的命名实体识别以字为单位，将实体类别标签和词位标签组合进行识别，避免分词的错误传递[29-32]。基于词位的命名实体识别对于预料的规模提出了更高的要求，如果没有足够训练语料，实例覆盖不全造成模型的泛化性能差。因此就当前的语料规模来说，基于分词的命名实体识别有显著的优势。下一步，作者将不断补充语料，并尝试基于词位训练命名实体识别模型[33]。

作者贡献：第一作者杨红梅负责研究设计、实现及论文撰写，第二作者李琳负责研究设计、论文修改，第三作者杨日东负责编码实现、论文修改，通讯作者周毅负责研究选题、论文修改。

经费支持：该文章接受了“国家重点研发计划精准医学专项基金项目(2016YFC0901602)；NSFC-广东大数据科学中心联合基金项目(U1611261)；广东省前沿与关键技术创新专项基金项目(2014B010118003)；广州市2017年产学研协同创新重大专项(201604016136)；广州市健康医疗协同创新重大专项(201604020016)”的资助。所有作者声明，经费支持没有影响文章观点和对研究数据客观结果的统计分析及其报道。

利益冲突：文章的全部作者声明，在课题研究和文章撰写过程，没有因其岗位角色影响文章观点和对数据结果的报道，不存在利益冲突。

伦理问题：实验研究的实施符合《赫尔辛基宣言》和医院对人体研究的相关伦理要求。文章的撰写与编辑修改后文章遵守了《观察性临床研究报告指南》(STROBE 指南)。

文章查重：文章出版前已经过专业反剽窃文献检测系统进行3次查重。

文章外审：文章经小同行外审专家双盲外审，同行评议认为文章符合本刊发稿宗旨。

作者声明：第一作者和通讯作者对研究和撰写的论文中出现的不端行为承担责任。论文中涉及的原始图片、数据(包括计算机数据库)记录及样本已按照有关规定保存、分享和销毁，可接受核查。

4 参考文献 References

[1] Lossio-Ventura JA, Hogan W, Modave F, et al.Towards an obesity-cancer knowledge base: Biomedical entity identification and relation detection. Proceedings (IEEE Int Conf Bioinformatics Biomed). 2016;2016:1081-1088.

[2] Kohane IS.Using electronic health records to drive discovery in disease genomics.Nat Rev Genet. 2011;12(6): 417-428.

[3] Razavian N, Blecker S, Schmidt AM, et al.Population-Level Prediction of Type 2 Diabetes From Claims Data and Analysis of Risk Factors.Big Data. 2015;3(4): 277-287.

[4] Rotmensch M, Halpern Y, Tlimat A, et al.Learning a Health Knowledge Graph from Electronic Medical Records.Sci Rep. 2017;7(1):5994.

[5] Small AM, Kiss DH, Zlatsin Y, et al.Text mining applied to electronic cardiovascular procedure reports to identify patients with trileaflet aortic stenosis and coronary artery disease.J Biomed Inform.2017;72:77-84.

[6] Payne TH, Corley S, Cullen TA,et al.Report of the AMIA EHR-2020 Task Force on the status and future direction of EHRs.J Am Med Inform Assoc.2015;22(5):1102-1110.

[7] Wang Y, Wang L, Rastegar-Mojarad M, et al.Clinical Information Extraction Applications: A Literature Review.J Biomed Inform.2017.

[8] Jensen PB, Jensen LJ, Brunak S.Brunak,Mining electronic health records: towards better research applications and clinical care.Nat Rev Genet.2012;13(6): 395-405.

[9] 何林娜,杨志豪,林鸿飞,等.基于特征耦合泛化的药名实体识别[J].中文信息学报, 2014,28(2):72-77.

[10] 栗伟,赵大哲,李博,等.CRF与规则相结合的医学病历实体识别[J].计算机应用研究, 2015,32(4):1082-1086.

[11] Lei J, Tang B, Lu X,et al.A comprehensive study of named entity recognition in Chinese clinical text.J Am Med Inform Assoc.2014;21(5):808-814.

[12] Hu JL,Shi X, Liu ZJ,et al.HITSZ_CNER: A hybrid system for entity recognition from Chinese clinical text.

[13] Wu J,Hu X, Zhao R,et al.Clinical Named Entity Recognition via Bi-directional LSTM-CRF Model.

[14] Xia Y, Wang Q.Clinical Named Entity Recognition: ECUST in the CCKS-2017 Shared Task 2.

[15] Kingma DP,Ba J.Adam: A Method for XStochastic Optimization. Computer Science.2014.

[16] fssjy. “Jieba” (Chinese for “to stutter”) Chinese text segmentation: built to be the best Python Chinese word segmentation module.https://pypi.python.org/pypi/jieba/.

[17] Mikolov T,Chen K,Corrado G,et al.Efficient Estimation of Word XRepresentations in Vector Space.Computer Science. 2013.

[18] Huang Z, Xu W, Yu K.Bidirectional LSTM-CRF Models for Sequence Tagging.Computer Science.2015.

[19] Cogswell M,AhXmed F,Girshick R,et al.Reducing Overfitting in Deep Networks by Decorrelating Representations. Computer Science.2015.

[20] Srivastava N,Hinton G,Krizhevsky A,et al.Dropout:a simple way to prevent neural networks from overfitting.Journal of Machine Learning Research.2014;15(1): 1929-1958.

[21] Uzuner Ö,South BR,Shen S,et al.2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc. 2011;18(5):552-556.

[22] 杨红梅,田翔华,周毅.电子病历对基于知识网络的精准医学的支撑及模式研究[J].中国数字医学,2017,12(8):29-31+75.

[23] Jonnalagadda SR,Adupa AK,Garg RP,et al.Text Mining of the Electronic Health Record: An Information Extraction Approach for Automated Identification and Subphenotyping of HFpEF Patients for Clinical Trials.J Cardiovasc Transl Res. 2017;10(3): 313-321.

[24] Kumar V,Stubbs A,Shaw S,et al.Creation of a new longitudinal corpus of clinical narratives. J Biomed Inform. 2015;58 Suppl:S6-S10.

[25] Kübler S,Zinsmeister H.Corpus linguistics and linguistically annotated corpora. Bloomsbury Academic.2015.

[26] 杨锦锋,关毅,何彬,等.中文电子病历命名实体和实体关系语料库构建[J].软件学报,2016,27(11):2725-2746.

[27] Zhang S, Kang T, Zhang X, et al.Speculation detection for Chinese clinical notes: Impacts of word segmentation and embedding models.J Biomed Inform.2016;60:334-341.

[28] Lishuang Li, Liuke Jin, Yuxin Jiang, et al.Recognizing Biomedical Named Entities Based on the Sentence Vector/Twin Word Embeddings Conditioned Bidirectional LSTM. 2016: Springer International Publishing.

[29] Wang Y, Wang L, Rastegar-Mojarad M, et al. Clinical information extraction applications: A literature review. J Biomed Inform. 2018;77:34-49.

[30] Varone M, Varone M, Varone M, et al. Conditional random fields with semantic enhancement for named-entity recognition. in International Conference on Web Intelligence, Mining and Semantics. 2017.

[31] Shao Y, Hardmeier C, Tiedemann J, et al. Character-based Joint Segmentation and POS Tagging for Chinese using Bidirectional RNN-CRF.2017.

[32] Pham TH, Lehong P.End-to-end Recurrent Neural Network Models for Vietnamese Named Entity Recognition: Word-level vs. Character-level.2017.

[33] Gridach M.Character-Level Neural Network for Biomedical Named Entity Recognition. J Biomed Inform. 2017;70:85-91.