基于 RoBERTa – CRF 的肝癌电子病历实体识别研究*

邓嘉乐[1] 胡振生[1] 连万民[2] 华赟鹏[3] 周 毅[1]

([1] 中山大学中山医学院 广州 510080 [2] 广东省第二人民医院 广州 510317
[3] 中山大学附属第一医院 广州 510080)

〔摘要〕 **目的/意义** 肝癌电子病历中蕴涵大量医学专业知识，且大部分以非结构化数据形式存在，难以自动化提取。肝癌电子病历实体识别研究有助于构建肝癌领域医疗辅助决策系统和医学知识图谱。**方法/过程** 构建 RoBERTa 算法与 CRF 算法相结合的命名实体识别模型，利用自标注肝癌电子病历真实数据进行模型训练与测试。**结果/结论** RoBERTa – CRF 模型优于其他基线模型，具有较好实体识别效果。

〔关键词〕 肝癌电子病历；实体识别；知识提取；RoBERTa – CRF 模型；自然语言处理
〔中图分类号〕 R – 058 〔文献标识码〕 A 〔DOI〕 10. 3969/j. issn. 1673 – 6036. 2023. 06. 007

Study on Entity Recognition of Liver Cancer Electronic Medical Records Based on RoBERTa – CRF

DENG Jiale[1], *HU Zhensheng*[1], *LIAN Wanmin*[2], *HUA Yunpeng*[3], *ZHOU Yi*[1]

[1] *Zhongshan School of Medicine, Sun Yat – sen University, Guangzhou 510080, China*; [2] *Guangdong Second Provincial General Hospital, Guangzhou 510317, China*; [3] *The First Affiliated Hospital of Sun Yat – sen University, Guangzhou 510080, China*

〔**Abstract**〕 **Purpose/Significance** Electronic medical records (EMR) of liver cancer contain a large amount of medical knowledge, and most of the knowledge is in the form of unstructured data which is difficult to extract automatically. The research on entity recognition of liver cancer EMR is important in the construction of clinical decision support systems and medical knowledge graphs in the area of liver cancer. **Method/Process** A named entity recognition (NER) model combined with RoBERTa algorithm and CRF algorithm is built, and the model achieves excellent effect. The real data of self – labeled EMR of liver cancer are used for model training and testing. **Result/Conclusion** RoBERTa – CRF model is better than other baseline models and has good entity recognition effect.

〔**Keywords**〕 electronic medical records of liver cancer; entity recognition; knowledge extraction; RoBERTa – CRF model; natual language processing (NLP)

1 引言

肝癌患者在医院就诊时产生大量电子病历数据，记录患者入院检查、治疗以及出院全过程，蕴含大量医学专业数据和专业知识。然而这些数据多是非结构化文本数据。对其进行结构化处理，将相关知识自动提取出来，有助于后续医疗辅助决策系

〔修回日期〕 2023 – 03 – 18

〔作者简介〕 邓嘉乐，硕士研究生；通信作者：周毅，教授，博士生导师。

〔基金项目〕 国家重点研发计划（项目编号：2021YFC2009402）；国家重点研发计划（项目编号：2022YFC3601600）；广东省自然科学基金项目（项目编号：2021A1515011897）。

统、医疗知识图谱的构建以及精准医学研究。此外，结构化数据有助于患者和医生直观获取所需信息。如何将这些信息从无结构化文本中自动抽取出来成为研究的重点问题。而应用自然语言处理技术中的命名实体识别技术可以完成电子病历自动化知识抽取任务。

2 研究意义及背景

2.1 研究意义

肝癌又名肝脏恶性肿瘤，分为原发性肝癌和继发性肝癌。世界卫生组织国际癌症研究机构发布的 2020 年最新全球癌症负担数据显示，中国肝癌新发病例数约占全球 45.3%，死亡例数约占全球 47.1%，是目前我国第 4 位常见恶性肿瘤及第 2 位致死肿瘤。在实体识别领域，国内肝癌电子病历实体识别研究较少。运用机器学习算法对肝癌电子病历进行实体识别有利于肝癌电子病历规范化、结构化数据存储，以及相关潜在医学知识挖掘。

2.2 国内外研究现状

命名实体识别[1]是自然语言处理任务中的重要一环，对关系抽取[2]等任务有重要影响。其目的是从非结构化文本中抽取实体，医疗实体包括症状、疗效以及疾病名称等。在机器翻译[3]任务中，实体翻译有特殊规则或固定表达。精准找出实体并通过预先规则或字典库翻译，可提升机器翻译的效果和流畅性。Wang Q 等[4]将命名实体识别技术应用于中文实体链接方法，实现了将中文诊断和程序术语规范化为国际疾病分类（international classification of diseases，ICD）代码的应用。实体识别相关技术在生物医学[5]、专病库建设[6]和健康医疗大数据[7]等多个领域得到较好应用。

命名实体识别发展经历 3 个阶段。第 1 阶段是基于词典和规则的方法，需要不同领域专家制定词典与规则模板，在文本中进行匹配。第 2 阶段是基于统计机器学习的方法，基于序列化标注方式并最大化联合概率进行求解。第 3 阶段是深度学习方法，代表模型有长短期记忆人工神经网络（long - short term memory，LSTM）[8]、双向编码器表征（bidirectional encoder representations from transformers，BERT）[9]等。

电子病历研究在国外已开展多年，Clark C 等[10]提出基于最大熵模型和条件随机场的方法。Jiang M 等[11]采用支持向量机方法，*F*1 值达 0.848。Bruijn B 等[12]则采取统一医学语言系统（unified medical language system，UMLS）资源半监督隐马尔科夫模型，*F*1 值达 0.852 3。Gligic L 等[13]使用基于注意力的序列到序列模型。英文电子病历实体识别经过数年发展完善，已形成标准处理流程。

中文医疗实体识别[14]难度更大，边界定义模糊，存在嵌套关系，缺少统一公开的大规模语料库。用统计学习方法进行实体识别[15]时，特征选择影响结果。Lei J 等[16]根据出入院记录，分别采用条件随机场（conditional random field，CRF）、支持向量机（support vector machine，SVM）、最大熵等模型对 4 类医疗实体进行识别。Wang Y 等[17]在自标注中医文本语料上，采用分句形式，*F*1 值达 0.95。李博等[18]采用 Transformer[19]模型，使数据中长依赖关系进一步提升。杨红梅等[20]对肝癌电子病历采用 BiLSTM - CRF 模型取得较好效果；马欢欢等[21]利用 CNN - BiLSTM - CRF 模型对癫痫电子病历进行实验，达到较好效果。

近年来，研究者提出 BERT 算法，通过注意力学习字符间上下文关系。核心创新点在于两个训练任务：掩码语言模型训练和下句话预测任务。掩码策略下，模型在训练中随机掩盖单词再进行预测。下句话预测任务目的是服务问答、句主题关系等任务。针对 BERT 的缺点，研究人员提出 RoBERTa 模型[22]。该模型有更大参数量、更大批大小、更多训练数据。在掩码策略上，修改关键超参数，实现动态掩码策略，删除下句话预测任务，使模型能更好推广到下游任务。

3 构建 RoBERTa - CRF 实体识别模型

3.1 RoBERTa 层

该层输入向量由词嵌入向量、位置嵌入向量、

段嵌入向量组成。词嵌入向量词表为 50 000 比特级别的文本编码词汇。位置嵌入向量可以记录词位置，弥补自注意力模型无法感知词位置的缺点。段嵌入向量记录句子位置关系。

该层使用编码器双向编码，使词可以无视前后远近被其他词参与编码。编码器中包含多头自注意力，以全连接前馈网络方式连接。多头自注意力将关注句子不同信息的注意力融合计算。以 3 个向量为例，多头自注意力计算方式如下：

$$MultijheadAttention(X,Y,Z) = Concat(head_1,head_2,\ldots,head_h)\ W^O \quad (1)$$

$$head_i = Attention(X\ W_i^X, Y\ W_i^Y, Z\ W_i^Z) \quad (2)$$

自注意力对 X,Y,Z 3 个向量进行向量运算，编码器输入的字向量在整个输入中点积和加权求和，得到此位置的自注意力，计算方式如下：

$$Attention(X,Y,Z) = softmax(\frac{X\ Y^T}{\sqrt{d_K}})Z \quad (3)$$

训练方法上，去掉下一句预测任务，实现动态掩码策略。每次输入序列随机选择 15% 的词，其中 80% 被［Mask］标签代替，10% 用另一个词代替，10% 不改变。数据分多序列输入过程中，不同掩码模式生成，模型在不同掩码模式中预测被掩码覆盖或者被替换的原词汇。

3.2 CRF 层

CRF 对类之间的决策边界进行建模。命名实体识别可看成多分类问题，因此该层任务是由输入序列对输出序列预测，形式为对数线性模型，学习方法是极大似然估计或正则化极大似然估计。假定句子长度为 n，句子序列为 $X = (x_1,x_2,\ldots,x_n)$，对应预测标签序列为 $Y = (y_1,y_2,\ldots,y_n)$，该预测序列总分数如下：

$$score(X,Y) = \sum_{i=0}^{n} T_{y_i,y_{i+1}} + \sum_{i=1}^{n} P_{i,y_i} \quad (4)$$

其中，T 表示标签间的转移分数，P_{i,y_i} 表示每个字到对应 y_i 标签的分数。预测序列有多种可能性，只有一种正确，应对所有可能序列进行全局归一化。学习方向是调整参数将模型正确预测的概率提升。归一化预测序列概率如下：

$$P(y \mid X) = \frac{e^{s(X,y)}}{\sum_{\tilde{y} \in Y_X} e^{s(X,y)}} \quad (5)$$

3.3 RoBERTa－CRF 实体识别模型

将 CRF 的输入变为含有更准确语言表征的词向量，组合模型预测能力会得到提升。因此将 CRF 层作为第 2 层，在 CRF 层之前加入 RoBERTa 层。用分词工具对原始语料句子分词后，RoBERTa 层可学习到更准确的字符向量表示。再将字符向量输入到 CRF 层中，可得到更好预测结果，见图 1。

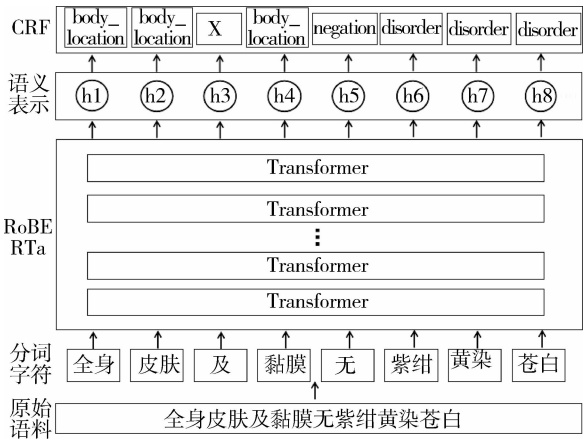

图 1 RoBERTa－CRF 命名实体识别模型

4 实验及结果分析

4.1 数据集介绍

所用数据集为 500 位肝癌患者的电子病历数据，已进行脱敏处理。电子病历中包含入院记录、出院记录和超声检查 3 部分。结合相关医学知识，本文定义 23 类肝癌数据相关实体（不包含无标签实体），并对数据进行 BIO 方法标注（实体第 1 个字符标签为“B－实体类型”，后续字符为“I－实体类型”，无标签字符为“O”）。为验证语料标注一致性，3 名标注人员分别标注同样 5 份数据，计算标注一致性达 88%。全部标注后得到有效标注 242 份，见表 1。

表 1 实体类型

序号	实体类型	含义
1	age	年龄
2	body_ location	身体部位
3	condition_ change	条件改变
4	disorder	疾病
5	drug	药物
6	exposure	暴露
7	hospital	医院
8	image_ feature	图像特征
9	marital_ history	婚姻史
10	negation	否定
11	negative_ symptom	阴性症状
12	number	数量
13	onset_ characteristic	发作特点
14	operation	手术
15	pathology_ feature	病理学特征
16	period	时期
17	person	人
18	sexuality	性别
19	size	大小
20	test	检查
21	test_ result	检查结果
22	time_ point	时间点
23	uncertainty	不确定

将标注后的数据按 1∶1∶3 划分为测试集、验证集和训练集。训练集实体类型频数分布，见图 2。

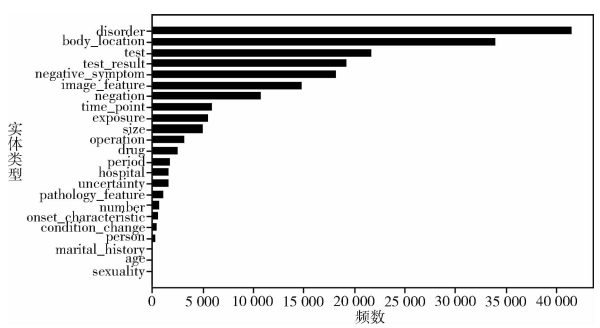

图 2 训练集实体类型频数分布

训练集各实体类别数量存在较大差异，分布极不均衡，如疾病、体位等实体数量大约在 35 000 个左右，性别、年龄等实体数目只有几百个。对每份电子病历数据的文本长度进行分析统计，主要集中在 2 000 个字符，同时也存在包含 8 000 ~ 10 000 个字符的超长文本，文本长度过长会导致模型难以捕获到文本间的长依赖关系，因此在模型训练前需要对数据进行截断切分。

4.2 数据平移

数据平移是将原始电子病历数据从不同切分点开始切分，然后将切分后的数据分别进行训练，最终对多个模型进行投票。本实验将每份电子病历数据拆分成最大长度为 256 个字符的样本，第 1 种切分方式为从起始位置累计达到 256 个字符进行切分；第 2 种切分方式是首先将数据中前两个句子作为一个样本，然后再依次切分。

4.3 交叉验证

采用 5 折交叉验证，将训练集划分为 5 部分，每次以其中 4 部分作为训练集剩余部分组成验证集，最终训练得到 5 个模型参数。对于实体识别任务，可以对 5 个模型结果进行投票，也可以将 5 个模型参数平均后再进行解码预测。本文采用投票方式。

4.4 参数设置

本实验中 RoBERTa 预训练模型使用 RoBERTa - zh - base 版本，模型层数为 24 层，使用 30G 原始文本，近 3 亿个句子，100 亿个中文字，产生 2.5 亿个训练数据。覆盖新闻、社区问答、多个百科数据等。RoBERTa 训练时使用批大小为 12，学习率为 1e - 4，优化器为 Adam 优化器，CRF 层学习率为 1e - 2。

4.5 评价指标

实验使用精确率 P = TP/（TP + FP），召回率 R = TP/（TP + FN）以及 $F1 = 2 \times P \times R/(P + R)$ 值对模型效果进行评价。其中，真阳性（true positive，TP）是指预测的实体类型与原本实体类型一致的观测样本数量，假阳性（false positive，FP）是指模型将本不属于该实体类型的观测预测为该实体类型的样本数量，假阴性（false negative，FN）是

指模型将原本属于该实体类型的观测预测为非该实体类型的数量。

4.6 结果分析

4.6.1 多模型性能对比 实验使用 RoBERTa – CRF 命名实体算法模型，并将命名实体识别中传统的隐马尔科夫（hidden markov model，HMM）模型、条件随机场、BiLSTM – CRF、BERT – CRF 算法模型作为基线模型进行对比，见表 2。

表 2 各模型测试集性能对比

模型	精确率	召回率	*F*1
HMM	0.504 4	0.620 2	0.556 3
CRF	0.666 1	0.691 9	0.678 8
BiLSTM – CRF	0.616 1	0.688 4	0.650 3
BERT – CRF	0.869 6	0.783 9	0.824 5
RoBERTa – CRF	0.901 2	0.812 6	0.854 6

RoBERTa – CRF 算法模型在肝癌电子病例数据上各指标均达最高，表现最好。BiLSTM – CRF 模型 *F*1 值相对 CRF 模型略低，主要是因为肝癌电子病历数据中存在很多短句和缩略词，其中缩略词记录患者病情或者身体状况，短句间没有很强语义关系，模型未学得合理的上下文语义表示。BERT – CRF 在 3 个指标上都低于 RoBERTa – CRF，说明 RoBERTa 层学习到的词向量有更加准确的语义表示。

4.6.2 RoBERTa – CRF 模型实验结果分析 本实验中 RoBERTa – CRF 模型在各标签上的 *F*1 值、精确率、召回率结果，见图 3。

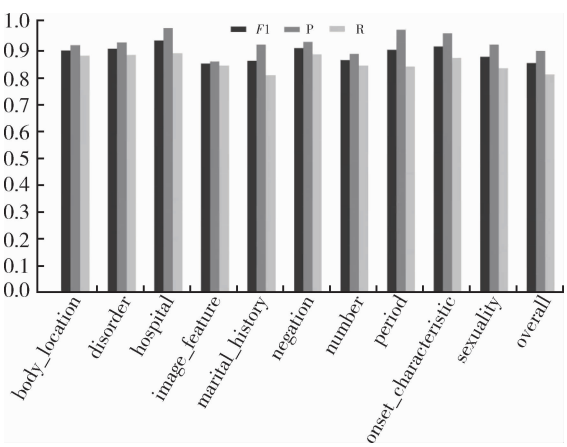

图 3 RoBERTa – CRF 模型识别结果

RoBERTa – CRF 模型在全部实体上的精确率达到 0.901 2，召回率达到 0.812 6。精确率较高代表模型实体识别正确率高，部分未正确识别可能是由于模型对标签之间的边界没有正确界定，例如 number 标签原意是病变区域个数，但是其数字表达形式可能会使模型误识别为 period；召回率相比之下较低，有部分实体被模型判断为非实体。通过学习更多高质量肝癌电子病历语料集可以提高模型性能。具体到各实体标签的识别结果，可以看到大部分实体类别有较高的精确率、召回率、*F*1 值。RoBERTa – CRF 模型在这类标签中达到较好学习效果。然而在某些标签中并没有达到理想效果。例如 pathology_feature、condition_change 等，模型由于该类标签定义比较模糊并且没有充足的样本训练，没有达到很好效果。针对该问题，可以扩大该类训练样本容量来达到更好效果。而在 test、test_result 标签上，在没有关系定义的前提下，模型很难将 test 和对应的 test_result 相联系。针对该问题，将关系抽取与实体识别相结合，会产生更加全面准确的判别能力。在 drug 标签上，模型没有在专业医药库中训练过，因此专业药物名称识别能力较差。针对该问题，可以将模型在专业医药语料中进行训练，以增强模型医药识别能力。

5 结语

预训练模型在通用语料库各项任务上都表现出较好性能，但是针对肝癌电子病历数据某些语句较短语义联系不强的特点，RoBERTa – CRF 模型、BERT – CRF 模型表现较好，BiLSTM – CRF 表现相对于 CRF 模型 *F*1 值略差。其原因是肝癌电子病历数据中存在大量短句和缩略词，导致基于 LSTM 的深度学习模型未能学到合理的上下文表示，反而通过 CRF 定义的局部特征效果较好。相较于 BERT – CRF 模型，RoBERTa – CRF 模型中，RoBERTa 有更高鲁棒性、更优训练策略以及更大规模预训练数据，因此模型性能表现更好。

针对这一问题，可以从肝癌电子病历数据本身着手，规范记录形式或者提高数据标签质量。此

外，本实验中使用的 RoBERTa 预训练模型是应用通用领域数据训练的预模型，如果能够使用大量医学文本数据对模型参数进行训练调整，训练出医学领域专用 RoBERTa 预训练模型，将会在该任务上有更好表现。而针对具体标签识别能力差的问题，应分析具体原因，采取相应措施，弥补模型短板。

参考文献

1 CHENG J, LIU J, XU X, et al. A review of Chinese named entity recognition [J]. KSII transactions on internet & information systems, 2021, 15 (6): 2012 - 2030.

2 NASAR Z, JAFFRY S W, MALIK M K. Named entity recognition and relation extraction: state - of - the - art [J]. ACM computing surveys (CSUR), 2021, 54 (1): 1 - 39.

3 ZHANG J, ZONG C. Neural machine translation: challenges, progress and future [J]. Science China technological sciences, 2020, 63 (10): 2028 - 2050.

4 WANG Q, JI Z, WANG J, et al. A study of entity - linking methods for normalizing Chinese diagnosis and procedure terms to ICD codes [J]. Journal of biomedical informatics, 2020, 105 (5): 103418.

5 金小桃，周毅，赵霞. 健康医疗大数据技术与应用 [M]. 北京：人民卫生出版社，2019.

6 LIU M, LUO J, LI L, et al. Design and development of a disease - specific clinical database system to increase the availability of hospital data in China [J]. Health information science and systems, 2023, 11 (1): 11.

7 张振，周毅，杜守洪，等. 医疗大数据及其面临的机遇与挑战 [J]. 医学信息学杂志，2014，35 (6)：2 - 8.

8 HOCHREITER S, SCHMIDHUBER J. Long short - term memory [J]. Neural computation, 1997, 9 (8): 1735 - 1780.

9 KENTON J D M W C, TOUTANOVA L K. BERT: pre - training of deep bidirectional transformers for language understanding [C]. Minneapolis: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL - HLT), 2019.

10 CLARK C, ABERDEEN J, COARR M, et al. MITRE system for clinical assertion status classification [J]. Journal of the American medical informatics association, 2011 (5): 563 - 567.

11 JIANG M, CHEN Y, LIU M, et al. A study of machine - learning - based approaches to extract clinical entities and their assertions from discharge summaries [J]. Journal of the American medical informatics association, 2011 (5): 601.

12 DE BRUIJN B, CHERRY C, KIRITCHENKO S, et al. Machine - learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010 [J]. Journal of the American medical informatics association, 2011, 18 (5): 557 - 562.

13 GLIGIC L, KORMILITZIN A, GOLDBERG P, et al. Named entity recognition in electronic health records using transfer learning bootstrapped neural networks [J]. Neural networks, 2020, 121 (1): 132 - 139.

14 咎红英，关同峰，张坤丽，等. 面向医学文本的实体关系抽取研究综述 [J]. 郑州大学学报（理学版），2020，52 (4)：1 - 15.

15 邱莎. 基于统计的生物命名实体识别研究 [D]. 成都：四川大学，2006.

16 LEI J, TANG B, LU X, et al. A comprehensive study of named entity recognition in Chinese clinical text [J]. Journal of the American medical informatics association, 2014, 21 (5): 808 - 814.

17 WANG Y, YU Z, CHEN L, et al. Supervised methods for symptom name recognition in free - text clinical records of traditional Chinese medicine: an empirical study [J]. Journal of biomedical informatics, 2014, 47 (2): 91 - 104.

18 李博，康晓东，张华丽，等. 采用 Transformer - CRF 的中文电子病历命名实体识别 [J]. 计算机工程与应用，2020，56 (5)：153 - 159.

19 VASWANI A, SHAZEER N, PARMAR N, et al. Attention is All You Need [C]. Long Beach: The 31st Annual Conference on Neural Information Processing Systems (NIPS), 2017.

20 杨红梅，李琳，杨日东，等. 基于双向 LSTM 神经网络电子病历命名实体的识别模型 [J]. 中国组织工程研究，2018，22 (20)：3237 - 3242.

21 马欢欢，孔繁之，高建强. 中文电子病历命名实体识别方法研究 [J]. 医学信息学杂志，2020，41 (4)：24 - 29.

22 LIU Y, OTT M, GOYAL N, et al. RoBERTa: a robustly optimized BERT pretraining approach [J]. Journal of computer science and technology, 2020, 35 (1): 197 - 221.