基于 LSTM-CRF 模型的电子病历临床命名实体识别

叶振鹏 蔡凯文 张家豪

韶关学院，广东韶关，512005；

摘要：本文提出一种基于 LSTM-CRF 模型的中文电子病历命名实体识别方法。该模型通过结合 LSTM 的序列建模能力和 CRF 的标签转移优化，有效解决了传统 CRF 在长距离依赖建模上的局限性。实验表明，该方法在不增加数据开销的情况下显著提升了实体识别性能，F1 值达到 93.4%，为临床文本分析提供了高效解决方案。

关键词：LSTM-CRF 模型；临床命名体识别；电子病历

DOI：10.69979/3029-2808.25.05.045

引言

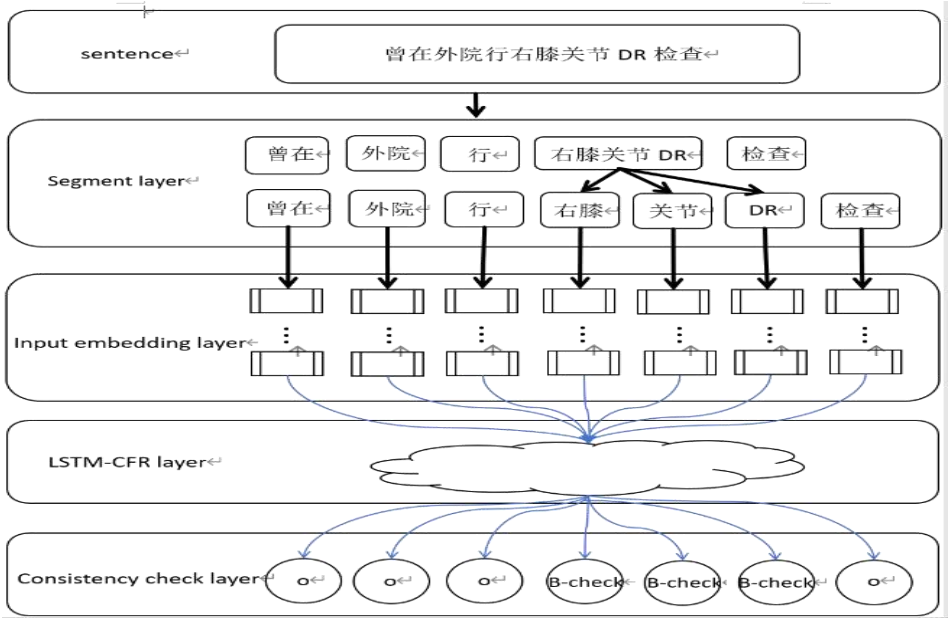
电子病历（EMR）以数字化形式取代纸质病历，确保医疗信息的完整性与实时共享，显著提升临床效率。中文电子病历命名实体识别（Chinese-CNER）通过从非结构化文本中提取医学实体（如疾病、症状），为智能诊疗系统提供结构化数据支持，助力医疗信息化建设。

随着深度学习的发展，基于长短期记忆网络（LSTM）和条件随机场（CRF）的神经网络模型在电子病历命名实体识别中表现出良好效果。相比传统方法如 HMM 和标准 CRF，LSTM-CRF 模型无需复杂特征工程，能更有效地捕捉上下文信息，从而提升识别准确率。有研究者提出结合自学习模型与医学知识库 UMLS，通过语义概念预测进行语义分类，在该任务中取得 F 值高达 85.23%，为本研究提供了参考。

当前中文医学实体识别仍较薄弱，语料资源匮乏。LSTM 与 CRF 各具优势：前者擅长建模长距离依赖，后者适合处理序列标注。本文结合两者特性，提升中文电子病历命名实体识别效果。

1 系统实现原理

研究使用 LSTM-CRF 模型来实现电子病例的 NER。如图所示，模型主要包含四个部分



分段层：对句子进行分段，并使用 BIO 来标记模型所需的格式。

输入嵌入层：将每个词映射成一个低纬向量。

LSTM-CRF 层：利用双向 LSTM-CRF 来标记每个单词，如图 2 所示。

一致性检查层：检查 LSTM-CRF 的结果一致性。

1.1 分段层

在中文命名实体识别中，词语边界难以界定，分词

是 NER 系统生成特征的关键第一步，处理不当会引入初始错误。例如“上腹部”作为医学术语常被误切分。为解决此问题，本项目提出先从训练集中提取所有临床命名实体，并保留其在原句中的形式，再在此基础上进行分词，获取最终结果。采用 BIO（Beginning,Inside,Outside）标注方案：词组起始标为 B，内部为 I，其他为 O。训练文本中，每个词占一行，每句后留空行，一行必须至少包含两列，第一列是单词本身，最后一列是命名的实体。

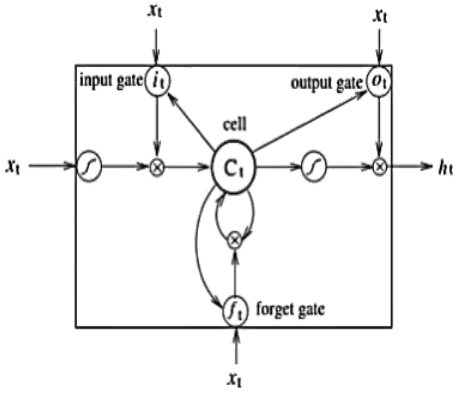
1.2 输入嵌入层

本项目的输入嵌入包含两部分，字符嵌入和单词嵌入。使用有限训练的词嵌入来初始化本项目的查找表，其中分别使用 Word2vec[8]和 Glove[9]进行预训练。词嵌入是基于训练数据集。嵌入的维度为 50，窗口大小为 4，迭代次数为 100。

1.3 LSTM-CRF 层

循环神经网络(RNNs)是一个在连续数据上运行的神经网络。虽然理论上来讲，RNNs 能够捕捉到长距离的依赖信息，但在实践中，RNNs 由于梯度消失或梯度爆炸的问题，基本无法完成。长短期记忆网络（LSTM）是 RNN 的变种，意在解决这些梯度消失和梯度爆炸的问题[11]。总的来说，一个 LSTM 单元包含三个乘法门，它们控制着下一个时间步骤中遗忘和传递信息的比例。从形式上看，在时刻 t 更新一个 LSTM 单元的公式为：

$$it= \partial (Wxixt+Whiht-1+Wcict-1+bi)$$
$$ct=(1-it)\odot ct-1+it\odot \tanh(Wxcxt+Whcht-1+bc)$$
$$Ot= \partial (Wxoxt+Whoht-1+Wcoct+bo)$$



其中，sigma 表示的是元素对应的 sigmoid 函数。o dot 是元素对应的积。xt 表示当前输入，it 表示一个输入门，其对应的权重矩阵为 Wxi,Whi,Wci,bi。Ot 表示一个输出门，对应的权重矩阵是 Wxo,Who,Wco，bo。ht-1 代表上一步生成的状态。ct 代表当前状态，结构如图所示：

使用该模型提倡的词语表示方法是将词语的左边和右边的上下文表示。这种表示方法有效的包括了一个词在上下文中的代表，对许多标签的应用都很有启发。本项目没有对标签决策进行独立的建模。而是使用条件随机场对他们进行联合建模[10]。

1.4 一致性检查层

一致性检查是一种用于检查预测的标签类别是否与训练的数据集中的标签命名类别一致的方法。对训练数据中 1344 个不同的命名标签对进行了统计评估。如果预测的标签类别与训练数据集中的原始标签不一致，试验人员将手动修改标签为原始标签。

LSTM-CRF 方法来解决中文电子病历的命名实体识别问题的特点：

1.4.1 传统的统计模型与神经网络的结合

该模型将传统统计方法与神经网络结合，充分发挥了 LSTM 在序列建模中的优势，有效弥补了 CRF 在上下文建模方面的不足。该结构具备良好的稳健性，无需额外数据即可实现更优性能。测试表明，该模型能够准确识别中文命名实体的词语边界，并建模标签之间的转移关系，弥补了单独使用 LSTM 无法处理标签转移的局限。

1.4.2 分段层及 BIO 标记

词语分割是 NER 系统生成特征的关键步骤。中文中词语边界模糊，例：“以上/腹部/疼痛/为/诱因”，“上腹部”容易被误分割。为解决此问题，本研究提出一种新分割方法：首先识别训练集中出现的所有临床命名实体，保留其在原句中的完整形式，再在此基础上进行分词。最终使用 BIO 标注法：词组起始位置标记为 B-label，内部为 I-label，或者为 O。

1.4.3 LSTM-CRF 模型

循环神经网络（RNNs）理论上能处理长距离依赖，但实际中常因梯度消失或爆炸而失效。长短期记忆网络（LSTM）是 RNN 的变种，通过三个乘法门控机制调节信息流，解决上述问题。模型使用词语左右上下文信息进行表示，有效捕捉上下文语义，对多个标签识别任务具有启发意义。

1.4.4 负反馈

模型还将序列建模的输出标签信息负反馈给模型，有效地减少梯度爆炸和消失引起的模型不收敛，提高中文病例命名实体识别的准确率。

2 电子病历识别

2.1 识别方法

本文所使用的项目代码均已上传 Github 开源，为了方便读者使用可克隆链接使用：

https://github.com/ZephyrYe9/Clinical-Named-Entity-Recognition-in-Electronic-Medical-Records-Based-on-the-LSTM-CRF-Model.git

本项目训练模型所用的数据来源于某医院的电子病历数据，因为其中涉及到病人的个人信息，所以在本次作品中不予公开。获得的电子病历数据中，有 1100 个标记的文件和 10427 个未标记的文件。将这些数据打乱后，随即得到训练集（80%）和验证集（10%）以及测试集（10%）。本文通过这些数据集来展示如何识别电子病历，大致流程如下：

通过该基于 LSTM-CRF 模型识别的方法所设计的系统（包含代码和算法均会共享），下面将详细讲述使用方法：

依据上述采集的数据集，通过 BIO 三元标注法进行高效且高精度标注后，进行人工核对以确保数据的准确性，部分标注如下：

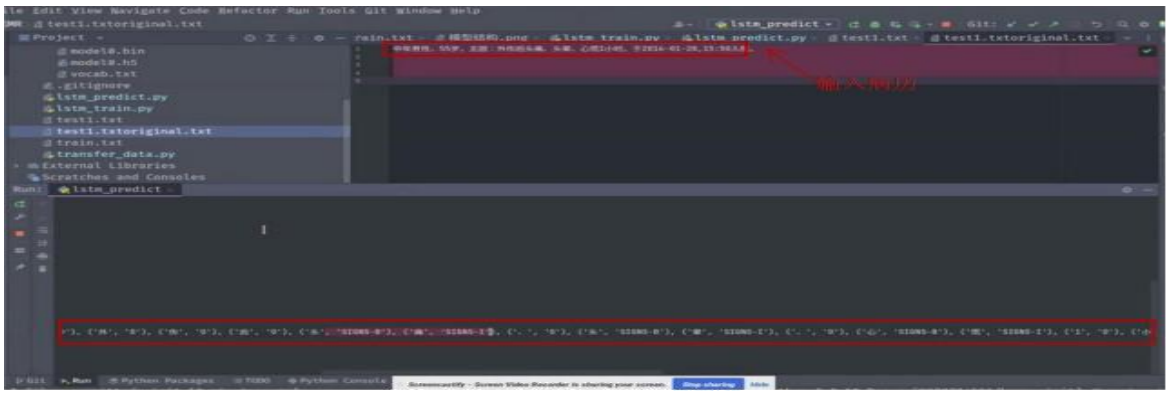
```
self.classdict={
100,TREATMENT-F:1,TREATMENT-B':2,'BODY-B':3,BOD
Y.:4.'SIGNS.I:5.'SIGNS.B':6.CHECK-B:7.CHECK-F:8,
"DISEASE-I':9,"DISEASE-B':10
```

接着对原始数据进行转换，例：拳打伤左面部，双上臂，用改锥划伤前胸部后出血、疼痛伴头晕 5.5 小时于 2016-02-11:22:45 入院。标注文件数据样式如下：

```
左面部 35 身体部位
双上臂 79 身体部位
出血 2021 症状和体征
疼痛 2324 症状和体征
```

接下来，通过转换脚本函数，输出模型，部分模型样式：

```
-O:O
拳 O 小 O
双 BODY-B 部 BODY-I
臂 BODY-I 疼 SIGNS-B
痛 SIGNS-I 划 O
```



随后进行模型搭建，本模型使用预训练字向量，座位嵌入层输入，然后经过两个双向 LSTM 层进行编码，编码后进入 dense 层，最后送入 CRF 层进行序列标注。

最后系统部署，本模型的训练和测试运行的平台为 Ubuntu20.04(LTS),其中依赖环境，为了方便读者使用，这里提供了 docker 镜像，镜像拉取命名为：

dockerpush2878724536/keras:init

模型训练每层的配置方式：

```
model.add(Dropout(0.5))
model.add(Bidirectional(LSTM(64,return_sequences
=True)))
model.add(Dropout(0.5))
model.add(TimeDistributed(Dense(self.NUM_CLASS
ES)))
```

```
crf_layer=CRF(self.NUM_CLASSES,sparse_target=True)
model.add(crf_layer)
model.compile('adam',loss=crf_layer.loss_function,metrics=[crf_layer.accuracy])
model.summary()
Return
```

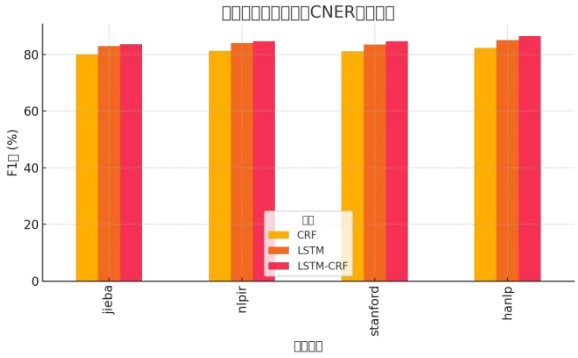
```
model.add(embedding_layer)
model.add(Bidirectional(LSTM(128,return_sequences=True)))
```

得到三个用于推理的文件，model0.bin,model0.h5,vocab.txt。（用户使用时文件后缀对上即可）推理完成后，可在“enteransent：”后面输入中文病历信息。

3 实验测试及结果

3.1 基准

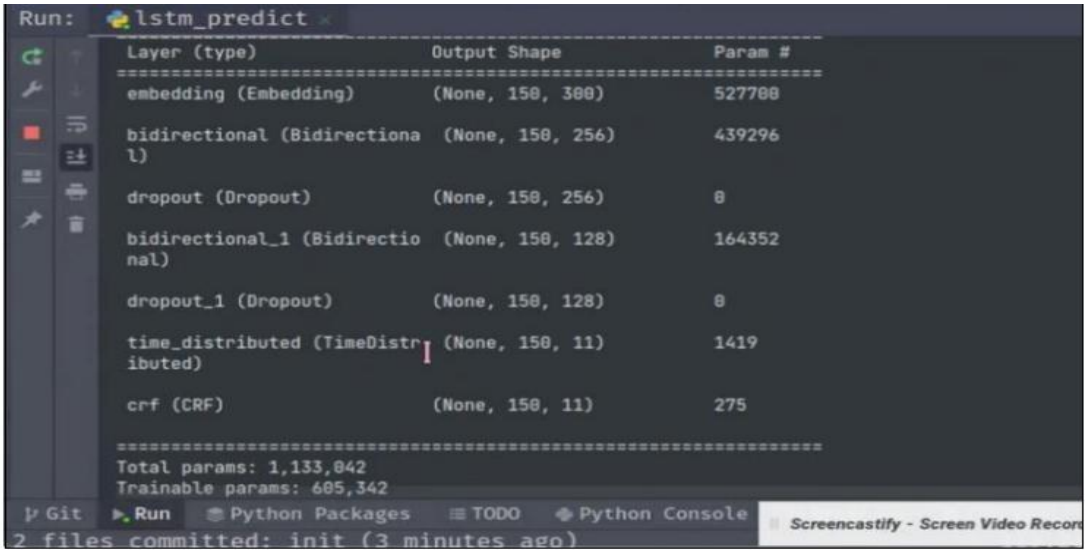
模型测试中，本项目在相同数据集下使用不同分词工具，比较三种常见模型性能，无需预处理和特征工程。三种模型分别为传统统计方法 CRF、单一神经网络结构 LSTM，以及 LSTM 与 CRF 的结合模型，其中 LSTM 的词嵌入采用随机初始化。



在上图中，带有 hanlp 工具的 LSTM-CRF 模型显示出更好的性能，因此选择它作为基准模型。

3.2 对训练好的中文电子病历命名实体识别模型的测试

模型训练参数如下：



在命名行中输入要测试的病历，输出结果符合预期，测试效果良好：

4 讨论与结论

中文 NER 面临分词错误传播的显著挑战，尤以医学术语边界模糊性为甚。为了减少由单词分割造成的错误，本研究提出新的预处理方法（见下表）。其中，预 pretrain 指的是在输入嵌入层完成了预训练的模型。char 指的是基于字符的词语建模的模型，dropout 是一种通过简化模型的复杂度来降低过拟合的处理方法。

模型结构	预处理策略	F1 值	增益
CRF	无	85.7%	-
LSTM	无	89.2%	+3.5%
LSTM-CRF	无	91.4%	+2.2%
LSTM-CRF + BIO 优化	实体保留+一致性	93.4%	+2.0%

模型结构	变体	F1 值
LSTM-CRF + 预处理	pretrain	91.6%
LSTM-CRF + 预处理	pretrain + dropout	92.3%
LSTM-CRF + 预处理	pretrain + dropout + char	93.0%
LSTM-CRF + 预处理	pretrain + dropout + char + CC	93.5%

为了分析本项目预处理的影响，本项目对各模型采用相同预处理方法，表 2 显示预处理后模型性能提升，

LSTM - CRF 模型改善近 2%。这表明预处理方法有效且适用于常见 NER 模型。表 3 中，预训练输入嵌入层使整体性能在 F1 上+3.37，dropout 带来+1.43 增长，学习 char 嵌入+0.56，一致性检验层+2.26。

参考文献

[1]黄志恒,徐伟,余凯.(2015).用于序列标注的双向 LSTM-CRF 模型.arXiv 预印本 arXiv:1508.01991.

[2]于勇,司新宇,胡成,张健.(2019).循环神经网络综述：LSTM 单元与网络结构.神经计算,31(7),1235 - 1270.

[3]Pennington,J.,Socher,R.,Manning,C.D.(2014).GloVe：基于全局统计的词向量表示.EMNLP 会议论文.

[4]Mikolov,T.,Chen,K.,Corrado,G.,Dean,J.(2013).向量空间中词表示的高效估计.arXiv 预印本 arXiv:1301.3781.

[5]DeBruijn,B.,Cherry,C.,Kiritchenko,S.,Martin,J.,Zhu,X.(2011).临床信息抽取三阶段的机器学习解决方案：2010 年 i2b2 竞赛技术综述.美国医学信息学会杂志,18(5),557 - 562.

[6]李彦霖,宋煜,张佳豪,王栋.(2020).FLAT：基于扁平-网格 Transformer 的中文命名实体识别.ACL2020 会议论文集.

[7]邵一鸣,裘旭,黄萱菁.(2019).结合词典增强 BERT 与两阶段标注的中文命名实体识别.ACL2019 会议论文集.

作者简介：叶振鹏（2004—5.12），男，汉族，广东东莞人，韶关学院，本科，研究方向：人工智能、集成电路与微电子。