# 基于多特征融合的预训练医疗实体和事件抽取模型

晏阳天，张昕楠，吴喆，赵新宇，葛屾，吴贤

腾讯

{yangtianyan, xinnanzhang, kerrzwu, joyxyzhao, shenge, kevinxwu}@tencent.com

**摘要** 在电子医疗病历的分析任务中，命名实体识别以及事件抽取一直都是最受关注的技术手段。全国知识图谱与语义计算大会（CCKS2021）评测四提出了识别并抽取出六种医疗实体以及肿瘤事件的比赛任务。针对于这两项子任务，本文在利用BERT预训练模型获取大规模数据下的语义信息的基础上，通过BiLSTM-CRF模型增加了词语间的约束，并融合字形与字音特征，增强了对中文医学词汇的表征能力。经过测试，本文所使用的方法在两个子任务的综合F1评分上达到了0.67536，证明了方法的有效性。

**Keywords:** 医疗命名实体识别 · 医疗事件抽取 · BERT · 深度学习

## 1 引言

在大数据时代，医疗决策支持与信息服务的潜在需求日益增加，通过自然语言处理（Natural Language Processing, NLP）对医疗电子病历进行信息提取与处理已成为一个关键问题。命名实体识别（Named Entity Recognition, NER）和事件抽取（Event Extraction)的对象均为非结构化文本，NER会识别提及的命名实体并将它们分类到预定义的类别中，而事件抽取会提取事件实体及其属性，两者本质上都是文本序列的标注任务。电子病历的自动标注可以处理医院不断生成的海量电子病历，帮助医生快速掌握患者信息，节省他们阅读不同医院、不同医生以及不同书写风格病历的时间。

针对于文本序列标注的方法有很多，最早期的研究是通过制定人工归纳的规则来进行的，这种方式逻辑性较强，但对于复杂不规范的文本往往很难找出统一完整的规则。随着机器学习领域的发展，文本标注算法开始转变为以数据驱动的方式自动获取抽象规则，这些方法包括隐马尔可夫模型 [8]、支持向量机 [6]等，但此类方法在应对包含复杂语义信息的文本时依旧难以胜任。直到预训练语言模型的出现，由于模型中包含了海量数据的语义信息，

文本序列标注任务在效果上才有了质的飞跃。但是，相对于常规的语言文本，医疗电子病历包含了大量的专业词汇，数据获取难度大，且不同的医生存在不同的标注标准，这些都使得电子病历自动标注任务变得更加困难。本次的全国知识图谱与语义计算大会（CCKS 2021）评测四为了促进医疗信息领域的发展，提供了一批标注的高质量电子病历数据，需要进行两个子任务：命名实体识别以及事件抽取。NER任务需要抽取以下六种实体：疾病和诊断、影像检查、实验室检验、手术、药物、解剖部位，事件抽取需要抽取肿瘤事件以及属性，包含肿瘤的原发部位，原发部位大小以及转移部位。针对于以上任务，本文实现了一个基于多特征融合的预训练医疗实体和事件抽取模型，并在官网所提供的测试数据集得到了综合F1分数0.67536的结果。接下来，本文分别从方法与数据处理、实验细节上进行介绍。

## 2 方法与数据处理

本文使用的方法流程流如图1所示。首先，本文会对原始数据集执行一些预处理操作，包括数据清理、数据规整和数据增强等。其次，本文在训练中使用了5折交叉验证，所有训练数据被平均分为五部分，每部分采用三种不同的模型参数设置，总共生成了15个模型。最后，再将15个模型的集合输入后处理模块，以产生最终的推断结果。

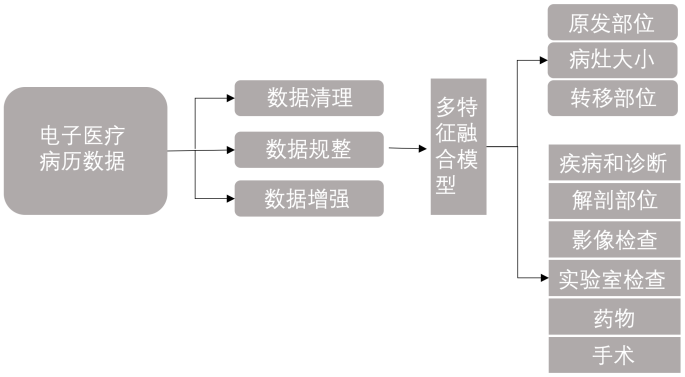


**图 1.** 方法流程图

### 2.1 数据预处理

由于命名实体识别与事件抽取均可以表示为序列标注任务，本文使用BIO(Begin, Inside, Other)标签方案将数据集给出的标签映射到每一个字符上，进行字符级别的标记。例如“进食流质及稀饭后出现中下腹阵发性隐痛”的标签序列如图2所示。


**图 2.** 标签序列举例

**数据清理** 在原始的数据中，会存在一些缺乏实际意义的符号，如“\T”、”\\U0004”以及“\\\”等。为了降低数据噪声以及数据的整理长度，本文在训练以及预测的过程中会将其去除。

**数据裁剪与规整** 考虑到BERT模型输入序列长度的限制,最多只能接受512个token的长度，本文在保证不损害本任务中病历信息完整性的情况下，手动删除了一些文本部分，这样的数据大约占到了1/3。如果存在包含关键信息且长度过长的数据，本文会采用滑动窗口的方法将数据分成几个512长度的段落。针对于数据中存在文本格式不统一的问题，本文会进行文本的规整，包括字母的大小写统一，以及全角符号、半角符号的统一。此外，在事件抽取的任务中，数据注释仅提供了目标实体属性的名称，没有提及到它们在原始病历中的具体位置，本文需要通过字符串匹配将真实标签与病历数据对齐。例如，病历描述的原发部位肿瘤大小为“2.0*3.0cm”，而数据给到的标签为“2.0cm×3.0cm”，为了确保数据注释与原始文本对齐，需要进行特殊处理来规范这些尺寸描述。

**数据增强** NER任务和事件抽取任务的数据量均在1000条左右，要让模型能够真正提取到病历数据的语义信息还是不够的。为了增强模型的鲁棒性，本文在每个实例中随机重新安排句子顺序，并生成相应的新训练样本，使整个

训练集加倍。同时，本文使用经过训练的模型对几百条官方提供的无标签数据进行了预测，生成伪标签，在一定程度上也扩充了数据集。

### 2.2 BERT与BiLSTM-CRF

**BERT** 大规模语料下的预训练语言模型的出现,给整个NLP领域带来了新的转机。2018年谷歌提出的多层双向Transformer模型——Bidirectional Encoder Representation from Transformers（BERT），通过Masked Language Mode（MLM）以及Next Sentence Prediction（NSP）任务在数十亿数量级的无标签数据上，学习到了文本信息的深层双向表征。此项研究开启了NLP的新纪元，利用预训练的BERT模型只要对下游任务进行微调，就可以在很多NLP任务上获得最优的结果 [1,2,10]，其中便包括命名实体识别任务。然而，许多文献指出，NSP任务可能会降低模型性能，直接用整个长句训练BERT模型会更为有效。在此之后，Facebook提出了一种更为鲁棒的模型：Robustly Optimized BERT（RoBERTa），该模型比原始的BERT模型具有更好的性能，基于更大的语料库，在训练过程去除NSP任务，使用全文档训练以及动态掩码机制。

**BiLSTM-CRF** BiLSTM-CRF是处理序列标注任务最常用的模型，尤其是NER任务。LSTM（Long Short-Term Memory）是循环卷积神经网络的一种特殊形式 [7]，通过遗忘门、记忆门以及输出门来实现对信息的遗忘及更新，这种方式的循环神经网络可以捕捉较长距离的依赖，解决了梯度消失问题 [5]。而BiLSTM是将双向信息添加到传统的LSTM中，使其能够有效地使用前向和后向的语义信息。同时，作为一种传统的判别式模型，条件随机场（Conditional Random Field，CRF）也通常被用来处理序列标注任务，通过增加字符间的约束，可以减少一些无效的预测 [3,4,9]，其损失函数可以简单地表示为如下形式，其中N表示所有可能的路径：

$$\text{Loss} = -\log \frac{P_{\text{realpath}}}{\sum_i^N P_i} \tag{1}$$

### 2.3 多特征融合模型

本次评测使用的模型包含了以下几种特征：字音特征、字形特征以及经过BERT的预训练网络特征。字音本就是中文的重点特征之一，在医疗领域，更是有很多音译的专业名词，所以对于不同医生书写的电子病历就有可能存在不同的音译版本，通过加入这一特征可以大大提升模型对同音同意不同形

词语的理解。此外，中文本身作为象形文字，偏旁部首蕴藏着极多的语义信息，比如一些疾病，大多含有“疒”，身体部位大多含有“月”。在实际的处理过程中，字音会通过Python包pypinyin进行转化获取，字形则通过抓取相关网站获取笔画信息，本次评测使用的BERT为Chinese-BERT-wwm[1] 具体为以下配置：24-layer, 1024-hidden, 16-heads, 330M parameters。模型的基础结构如图3所示，字音字形通过3*3*64的卷积层之后，加上maxpooling，得到64维的特征向量，再与BERT模型输出的1024*3维向量进行拼接，便将这融合的特征向量输入到BiLSTM-CRF中。

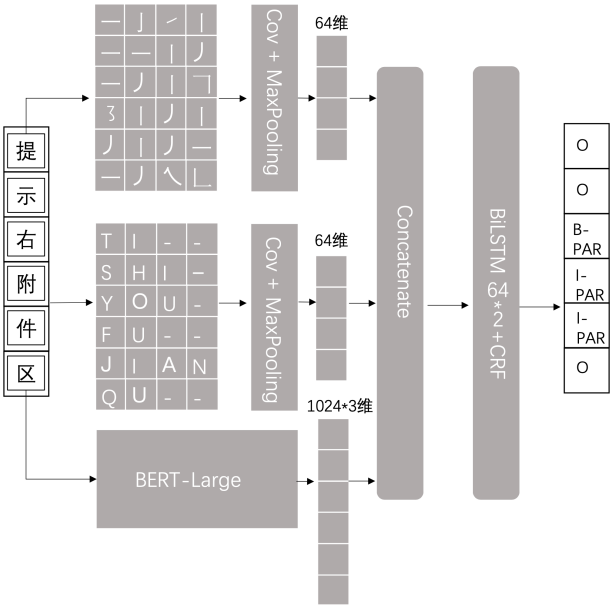


**图 3.** 多特征融合模型

## 2.4 模型融合与后处理

**模型融合** 本文使用了5折交叉验证，获取了三种不同参数下的15个模型。为了达到更好的预测效果，本文使用了两种策略来集成所有15个模型的预测结果。(策略A)：直接对15个模型进行投票，如果一个实体被8个以上的模型预测出来，占到了模型总数的1/2以上，则选择其为最终预测结果，否则将其丢

[1] https://github.com/ymcui/Chinese-BERT-wwm

弃。(策略B)：先从固定fold中选取F1分数最优的某个参数下的模型，再将得到的5个模型进行投票，投票阈值选为3。值得一提的是，本文在进行NER任务的模型融合时，特别地使用了双阈值投票。由于NER任务中的某些实体具有较高的专业性，不容易产生歧义，较容易通过词表的方式提取出来，所以在模型的融合过程，首先会使用高阈值的投票选取出置信度高的词语来充当词表，在进行较低阈值的投票时，保证不丢失这些词表中的词语，本文通过这种方式确保了模型的召回率。

**后处理规则** 本文在模型预测的最后会通过一些启发式规则对最终结果进行校准。比如在NER任务中，某些包含“+”的实体会将其表示为同一实体：“胸内食管胃吻合术+腹腔淋巴结清扫术+纵膈淋巴结清扫术”，某些解剖部位需要加上具体的位置修饰词：“食管”会被补充为“食管下段”，“肺”会被补充为“左肺”。针对于事件抽取任务，某些解剖部位“X”会通过判断是否出现“X”癌、“XCA”、“X术后”进行校验，涉及到原发病灶大小的预测，也会“肿块影”，“结节影”或者“密度影”等词进行辅助预测，此外在考虑转移部位时，“转移”，“侵犯”，“多发某某病灶”等词也帮助约束转移部位的预测。

## 3 实验

### 3.1 数据集介绍

本次评测的数据由医渡云（北京）技术有限公司提供，数据的标注也是由医渡云公司组织的专业医生团队进行人工标注的。在NER任务中，训练数据包含了1500条标注数据，1000条非标注数据，标注的六类实体：“影像检查”，“实验室检验”，“手术”，“药物”，“疾病与诊断”以及“解剖部位”总共达到了6292个实体数。在事件抽取任务中，训练数据达到了1400条中文标注数据，1300条非标注数据，标注的三类事件属性：肿瘤原发部位，原发病灶大小以及转移部位，实体总数达到了863个。

### 3.2 模型实现

在实际的训练过程中，所有模型均以Adam作为优化器，BERT学习率设置为2e-5，BiLSTM-CRF学习率设置为2e-4。为了获取更丰富的语义信息，本文将BERT最后三层的隐藏层拼接起来，维度为1024*3，加上字形特征的64维

与字形特征的64维，输入到BiLSTM-CRF中的维度达到了3200维。此外，整个训练过程在Tesla P40上完成，数据的输入不会超过512token，训练的批次大小为8。

### 3.3 模型结果

本次评测采用提交手工评测的方式，本文所使用的方法经过评测后在两个子任务的平均F1评分上达到了0.67536，取得了第三名的好成绩。

## 4 结论

针对于电子医疗病历的命名实体识别以及事件抽取的两项任务，本文提出了一种基于多特征融合的预训练医疗实体和事件抽取模型。相较于简单的BERT模型，本文加入了字音字形的特征，加强了模型对中文医学词汇的表征能力，在最终的测试集上，本文的模型达到了第三名的成绩。

## 参考文献

1. Adhikari, A., Ram, A., Tang, R., Lin, J.: Docbert: Bert for document classification. arXiv preprint arXiv:1904.08398 (2019)
2. Alberti, C., Lee, K., Collins, M.: A bert baseline for the natural questions. arXiv preprint arXiv:1901.08634 (2019)
3. Chen, A., Peng, F., Shan, R., Sun, G.: Chinese named entity recognition with conditional probabilistic models. In: Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. pp. 173–176 (2006)
4. Chen, Y., Zhou, C., Li, T., Wu, H., Zhao, X., Ye, K., Liao, J.: Named entity recognition from chinese adverse drug event reports with lexical feature based bilstm-crf and tri-training. Journal of biomedical informatics **96**, 103252 (2019)
5. Chiu, J.P., Nichols, E.: Named entity recognition with bidirectional lstm-cnns. Transactions of the Association for Computational Linguistics **4**, 357–370 (2016)
6. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**(3), 273–297 (1995)
7. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
8. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE **77**(2), 257–286 (1989)

9. Wallach, H.M.: Conditional random fields: An introduction. Technical Reports (CIS) p. 22 (2004)
10. Yang, W., Zhang, H., Lin, J.: Simple applications of bert for ad hoc document retrieval. arXiv preprint arXiv:1903.10972 (2019)