基于预训练模型和领域词典的医疗命名实体识别方法研究

温超杰[1], 陈涛*[1], 朱江[1]

[1] 五邑大学智能制造学部

klaywen15@163.com, {chentao1999, pacer0112}@gmail.com

摘要：医疗命名实体识别是要从医疗文本中识别出如疾病和诊断、影像检查、实验室检验、手术、药物、解剖部位等类型的医疗命名实体。本文针对传统医疗命名实体识别方法对无标签医疗文本数据的利用不充分的问题，提出基于预训练模型和领域词典融合的医疗命名实体识别方法。首先，从有标注的医疗文本中提取医疗实体，结合外部搜集的医疗实体构建医疗领域词典。然后，使用领域词典和未标注的医疗文本对预训练模型 WoBERT 进行同领域预训练，同时，对未标注的医疗文本使用伪标签的方法，训练得到带伪标签的医疗文本。接下来，在 RoBERTa-wwm-ext-large 预训练模型上对序列标注模型 BiLSTM-CRF 进行微调，构建基于不同预训练模型的医疗命名实体识别模型。最后，将上述多个模型的结果进行融合，得到最终的识别结果。该方法在“CCKS 2020 面向中文电子病历的医疗实体及事件抽取（一）医疗命名实体识别”任务的最终测试数据集上显示：上述方法在该任务中严格指标和松弛指标 F1 值分别达到 0.887 和 0.953，利用无标签医疗文本数据和领域词典后，严格指标 F1 值提升了 1 个百分点。

关键词: 医疗命名实体识别; 预训练模型; 领域词典

1 引言

医疗文本涵盖了患者在医院中的诊疗活动记录，是重要的医学知识资源。在数字化盛行的时代，传统的手写病历正在被数字化病历逐渐替代，各级医疗机构在电子病历信息化建设中做了大量的工作，提升了临床诊疗决策能力。其中，为实现病历数据的结构化可计算、可推理，更凸显了医疗文本的命名实体识别等自然语言处理任务的重要性 [1]。医疗命名实体识别是要从医疗文本中识别出如疾病和诊断、影像检查、实验室检验、手术、药物、解剖部位等类型的医疗命名实体。这类命名实体能够用于后续的医疗信息分析和研究，如构建临床决策系统和医疗领域的知识图谱等。

目前主流的基于深度学习的医疗命名实体识别方法采用有监督的方式从有标签数据中学习特征，对大量存在的无标签医疗文本数据的利用不充

*通信作者

分，导致一些在训练集中未出现过的实体很难被识别出来。针对上述问题，本文提出了基于预训练模型和领域词典融合的方法，利用无标签医疗文本数据对预训练模型进行再训练，同时将医疗领域知识融合到预训练模型中，提高了医疗命名实体识别的泛化能力。

该方法在 CCKS 2020 面向中文电子病历的医疗实体及事件抽取（一）医疗命名实体识别任务上取得了较好的效果。该任务是对既定的中文电子病历纯文本文档，识别、抽取出相关医学实体提及，并归类到 6 种预定义类别，即疾病和诊断、影像检查、实验室检验、手术、药物、解剖部位。

本文的贡献主要包括以下几点：

(1) 使用全词 mask 预训练模型 RoBERTa-wwm-ext-large 加上 BiLSTM-CRF 模块应用到医疗命名实体识别任务中，并使用无标注医疗文本数据对该模型进行同领域预训练，提高了通用的语言预训练模型在医疗领域的泛化能力；

(2) 将非标注数据通过使用伪标签方法，转变成带伪标签的数据文本，扩充了训练集；

(3) 构造了医疗领域的实体词典，并将词典融合到预训练模型中，提高了命名实体识别模型的识别效果。

2 相关工作

命名实体识别是中文电子病历信息结构化的基础任务，主要分为基于浅层机器学习的方法和基于深度神经网络的方法。典型的基于浅层机器学习的命名实体识别方法包括基于隐马尔科夫模型（HMM）的方法 [2]、基于条件随机场（CRF）的方法 [3] 等。Collobert 和 Weston[4] 是首批将深度神经网络技术应用于 NLP 任务的研究人员，他们的方法的目标是通过查找表将词转换为矢量表示，作为模型的输入。近年来随着深度神经网络技术的发展，基于神经网络的模型在众多序列标注任务中均取得了不错的成效。Huang 等 [5] 针对序列标注问题提出了双向 LSTM 和 CRF 的方法。Chiu 和 Nichols[6] 提出一种双向 LSTM-CNNs 架构，该架构可自动检测单词和字符级别的特征。Transformer 以增强的并行化和更好的长依赖建模的特点 [7]，为自然语言处理任务提供了新的解决方案。Devlin[8] 提出的预训练模型 BERT（Bidirectional Encoder Representation from Transformers），在当时的命名实体识别任务等多项自然语言处理基准测试中获得了较好效果。

在医疗文本命名实体识别任务中，Yang[9] 等利用基于双向 LSTM 和 CRF 结合的实体识别模型，抽取入院记录和出院小结中的医疗命名实体。Wan[10] 等提出基于字词联合训练的 Bi-LSTM 中文电子病历命名实体识别方法，识别中文电子病历中疾病、症状等相关实体。Chowdhury[11] 等提出一种多任务的双向循环神经网络模型，从中文电子病历中抽取医疗命名实体。

3 数据

本次评测任务中的中文电子病历数据是由医渡云（北京）技术有限公司编写，标注数据由医渡云公司组织专业的医学团队进行人工标注。标注的数据集包含 1500 条文本，标注了“疾病和诊断”、“影像检查”、“实验室检验”、“手术”、“药物”、“解剖部位”6 种共 26414 个医疗命名实体。另外还包含 1000 条非标注数据文本和 6292 个实体词。标注数据集的统计信息和示例分别如表 1 和图 1 所示。

表 1. 标注数据集统计信息

	文本	疾病	检查	检验	手术	药物	部位	总数
训练集	1500	6211	1490	1885	1327	2841	12660	26414

4 方法

本文提出的基于预训练模型和领域词典的医疗命名实体识别方法的框架如图 2 所示。

首先，在对医疗文本数据进行预处理，包括数据清洗、格式转换等。然后，分别使用伪标签方法扩充训练集，从医疗文本中抽取医疗实体构造医疗领域词典和对预训练模型使用同领域预训练方法，得到新的预训练模型。接下来，对新的预训练模型和其他的预训练模型，分别使用任务数据集进行微调，在最终测试集上进行预测。最后，将分别预测得到的结果按规则进行融合。

4.1 数据预处理

对用作训练集的 1500 条原始文本进行预处理，主要包括如下 4 步：

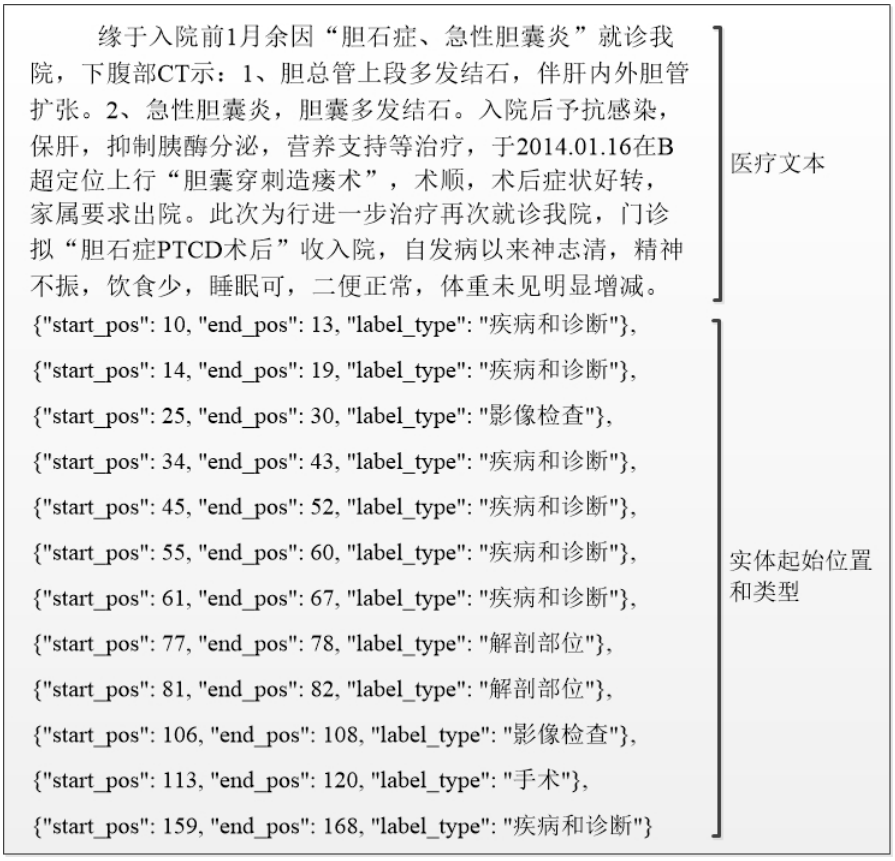

图 1. 标注数据集示例

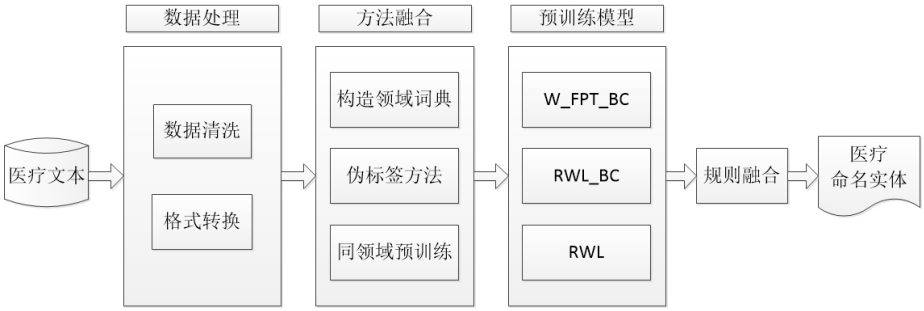

图 2. 医疗命名实体识别方法框架图

1. 数据清洗。标注的中文电子病历数据中存在部分明显的标注错误，如“直肠癌根治术（”、“胃角）高级别下皮内瘤变（重度异型增生）”等被标注成一个医疗命名实体，应修改为“直肠癌根治术”、“（胃角）高级别下皮内瘤变（重度异型增生）”。这里统一修正了有关符号问题的实体边界错误。

2. 文本规范化。完成中文电子病历数据中文本和符号的全半角统一，英文大小写转换等处理。

3. 语料格式转换。将训练集和测试集语料格式转换成“char label”形式的文本文件，并以 BIO 格式对命名实体进行标注。

4. 语料分割。由于模型输入的最大长度限制和过长序列对 LSTM 的恶劣影响，在确保病历中的句子相对完整的前提下，对输入的病历文本进行分割。使用句号等能保留语义信息的符号分割，同时考虑 [CLS]、[SEP] 符号的加入，限制序列长度在 200 个字符以内。

4.2 预训练模型

RoBERTa[12] 是一个在大量的文本语料上使用无监督学习方法训练的语言理解模型，而命名实体识别作为自然语言处理的子任务，是在这个模型上设置的特定下游任务接口去执行。哈工大开源的预训练模型 RoBERTa-wwm-ext-large 使用了 Whole Word Masking（wwm），会对组成同一个词的汉字全部 mask，即全词 mask[13]。

在训练集中，文本内容被打上不同的标签，“B_”代表实体的开始，“I_-”代表实体的中间和结尾部分，“O”标签代表与实体无关的字符。将医疗文本数据输入到模型的过程中，用标签“[CLS]”表示一句话的开始，标签“[SEP]”表示结尾。由于系统要求每句话的输入长度是固定的，用“[PAD]”标签自动填充输入语句未达到系统设定的最大长度的部分，对于超出最大长度的句子，系统自动去除超出的部分。至此，医疗文本数据输入的标签为“B_疾病和诊断”、“I_疾病和诊断”、“B_影像检查”、“I_影像检查”、“B_实验室检验”、“I_实验室检验”、“B_手术”、“I_手术”、“B_药物”、“I_药物”、“B_解剖部位”、“I_解剖部位”、“O”、“[CLS]”、“[SEP]”和“[PAD]”16 类标签。本文在预训练模型 RoBERTa-wwm-ext-large 加上 BiLSTM-CRF 模块，使用任务训练集对该模型进行微调，构建医疗文本的命名实体识别模型。本文的 RoBERTa-wwm-ext-large-BiLSTM-CRF 的模型结构图如图 3 所示。

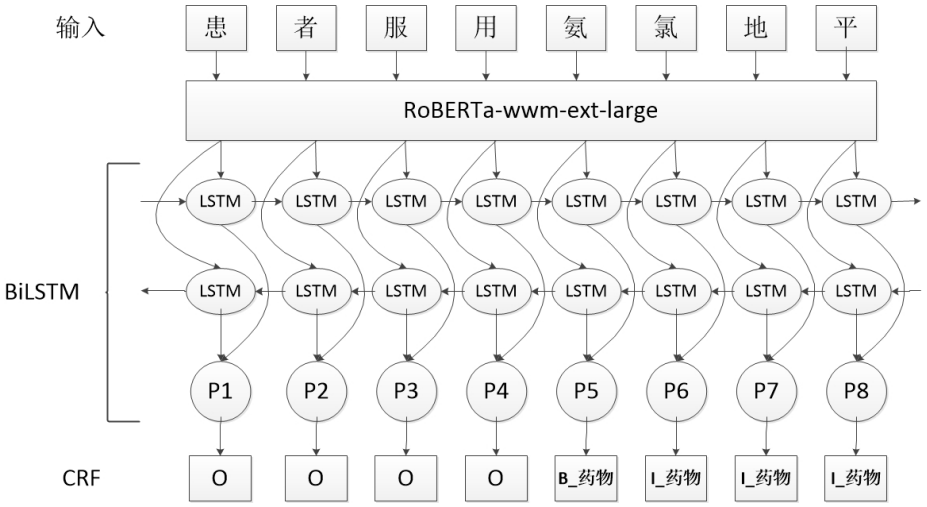

图 3. 面向医疗文本 RoBERTa-wwm-ext-large-BiLSTM-CRF 模型框架图

4.3 伪标签方法

伪标签（Pseudo Label）方法 [14] 是一种无监督学习方法，它可以通过使用非标注数据来提高预测模型的性能。方法的主旨是首先使用标注数据作为训练集训练模型，然后使用该模型预测未标注数据的标签，这类标签称作伪标签。最后，将标注数据和伪标签数据结合作为新的训练集训练模型。

本文使用 (4.2) 构建的模型“RoBERTa-wwm-ext-large-BiLSTM-CRF”，对 1000 条非标注数据文本使用伪标签的方法。即使用已有的标注数据文本作为训练集，使用非标注数据文本作为测试集，得到带伪标签的数据文本。最后筛选出高置信度的带伪标签数据文本，结合原有的标注数据文本，重新训练模型。本文使用伪标签方法的框架如图 4。

4.4 同领域预训练

若所执行任务的标注数据较少，所属的领域与初始预训练语料越不相关，而又能获取到充分的、任务相关的无标注数据时，领域预训练和任务预训练是应继续进行的，这也称为持续预训练 [15]。持续预训练可分三个方向：任务内预训练、同领域预训练、跨领域预训练。由于大多预训练模型都不是面向某一特定领域的，对预训练模型使用领域语料进行持续预训练，可以使它在该领域的表现更好。

本文对预训练模型 WoBERT 进行同领域预训练。追一科技发布的预训

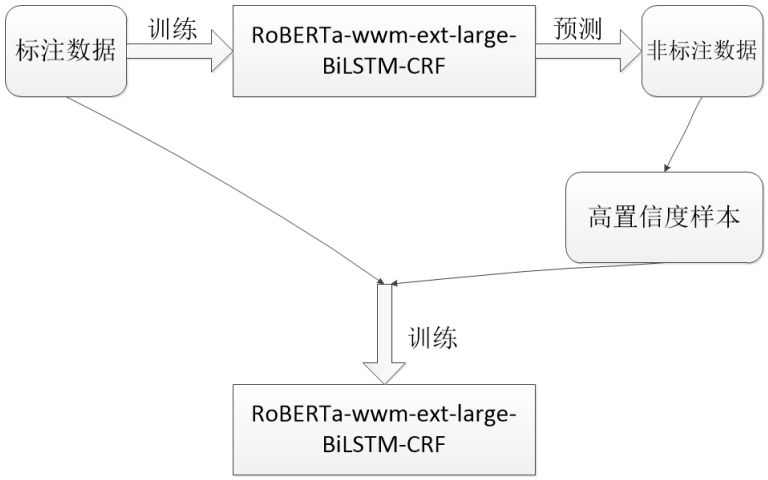

图 4. 伪标签方法框架图

练模型 WoBERT 是在 RoBERTa-wwm-ext 基础上继续预训练得到的，其不同点在于模型处理的是以中文词为单位的。这里使用构造的医疗领域字典 (4.5) 作为训练的词典文件，使用任务提供的 1000 条非标注数据作为持续预训练的数据集，训练 100000 步，得到医疗领域的预训练模型，这里称为 WoBERT_Further Pretraining（WoBERT_FPT）。

4.5 领域词典

在 CCKS 2020 医疗命名实体识别评测任务中，医渡云（北京）技术有限公司提供了 1500 条中文电子病历标注数据集，我们从里面抽取出所有被标注的医疗实体。CCKS 2020 医疗命名实体识别评测任务还提供了包含 6292 个医疗实体的实体词表，此外，我们对公开的医疗网站爬取相关的医疗实体。最后，我们对所有的医疗实体统一进行精简处理，即删除重复或相似度较高的医疗实体。至此，我们构建了一个涉及药物、疾病、身体部位等多种医疗实体类型，共 11000 个医疗实体的医疗领域词典。

子词级粒度文本表示最早出现在英文文本翻译任务中，通过词切片 [16]（Word Piece）将英文单词切分，如对“The highest mountain”这句话进行子词级表示后，将变成“The high ##est moun ##tain”。对中文文本使用子词级表示时，会先对词汇中词首和非词首部分进行区分，在所有非词首部分的字符前面添加“##”，添加“##”前后形成两个不同的矢量，这样在词表仅增加一倍的情况下，又能保留中文词汇的语义信息 [17]。首先对此前

收集到的 11000 个医疗实体进行分词，如医疗实体“左侧颈动脉中段”经过分词后将产生“左侧”、“颈”、“动脉”、“中段”四个词汇，分别对其使用子词级表示，即“## 左侧”、“## 颈”、“## 动脉”、“## 中段”。

4.6 方法融合

根据前面构造的医疗领域词典和子词级表示方法结合，最终形成 3 万个词条。同时对预训练模型 RoBERTa-wwm-ext-large 自带的词典文件 vocab.txt 的内容进行精简，删除词典中对本次医疗命名实体识别评测任务无明显作用的词条，如“ください”这类日文或特殊符号等。最后融合得到一个包含 5 万个词条的医疗实体词典文件。

分别使用预训练模型“WoBERT_FPT-BiLSTM-CRF”（W_FPT_BC）、“RoBERTa-wwm-ext-large”（RWL）和“RoBERTa-wwm-ext-large-BiLSTM-CRF”（RWL_BC），融合标注数据和伪标签数据作为各个预训练模型的训练集，使用构造的医疗领域词典作为各个预训练模型的词典文件。最后将三个模型的预测结果进行融合。首先将模型 W_FPT_BC 和 RWL_BC 的预测结果进行比较，抽取出两者中实体起始位置相同而类型不同的实体；然后与模型 RWL 的预测结果进行比较，取与 RWL 预测结果相同的进行融合。

5 实验

本次医疗命名实体识别任务采用精确率（Precision）、召回率（Recall）以及 F1 值作为评测指标，分别从严格指标和松弛指标两个层面进行评价，主要考量识别实体的边界情况和实体类型。有关两类指标的具体定义可参考评测网站（https://www.biendata.xyz/competition/ccks_2020_2_1/evaluation/）。

设置实验参数，最大序列长度为 200，学习率为 5e-5，batch_size 为 16，epoch 为 5，优化算法为 Adam，dropout_rate 为 0.5，BiLSTM 隐层单元个数为 128。各个模型的结果如表 2 和表 3 所示。

表 2 和表 3 分别列举了不同方法组合在 6 种实体上松弛和严格指标，性能指标的度量方式为 F1 值。从表 2 和表 3 可以看出，结合预训练模型使用伪标签方法和同领域预训练方法，识别的效果得到了提升，严格指标 F1 值分别提升了 0.5 个百分点。在此基础上，与领域词典的融合后的模型，识

表 2. 各模型实验结果（松弛）

	疾病	检查	检验	手术	药物	部位	综合
W_BC	0.959	0.901	0.828	0.976	0.956	0.951	0.946
W_FPT_BC	0.980	0.893	0.810	0.970	0.961	0.961	0.955
W_FPT_BC+ 伪标签	0.973	0.886	0.819	0.971	0.963	0.962	0.954
W_FPT_BC+ 领域词典 + 伪标签	0.982	0.893	0.829	0.973	0.965	0.965	0.959
RWL_BC+ 领域词典 + 伪标签	0.976	0.889	0.820	0.967	0.977	0.960	0.956
RWL+ 领域词典 + 伪标签	0.987	0.891	0.821	0.962	0.972	0.968	0.961

表 3. 各模型实验结果（严格）

	疾病	检查	检验	手术	药物	部位	综合
W_BC	0.846	0.844	0.749	0.865	0.87	0.872	0.858
W_FPT_BC	0.849	0.843	0.738	0.892	0.899	0.878	0.866
W_FPT_BC+ 伪标签	0.860	0.843	0.758	0.910	0.895	0.881	0.871
W_FPT_BC+ 领域词典 + 伪标签	0.850	0.844	0.762	0.929	0.913	0.881	0.873
RWL_BC+ 领域词典 + 伪标签	0.847	0.849	0.758	0.918	0.914	0.884	0.873
RWL+ 领域词典 + 伪标签	0.854	0.855	0.765	0.931	0.922	0.893	0.882

别精度也略有提升，严格指标 F1 值提升了 0.2 个百分点。

对于多个模型的结果，我们设定一种规则确定融合后的输出结果，最后取“W_FPT_BC+ 领域词典 + 伪标签”、“RWL_BC+ 领域词典 + 伪标签”和“RWL+ 领域词典 + 伪标签”三个模型在最终测试集的结果进行融合。融合后的结果如表 4 所示。

表 4. 最终提交的评测结果

	疾病	检查	检验	手术	药物	部位	综合
松弛	0.970	0.886	0.847	0.975	0.966	0.957	0.953
严格	0.872	0.859	0.792	0.940	0.923	0.891	0.887

从表 4 可以观察到，取表现最好的三个模型的识别结果进行规则融合，识别的精度有了明显的提高，表明我们采取的融合方法是有效的。另外观察各个模型的识别结果发现，加入领域词典、伪标签方法和同领域预训练方法后，对实体识别效果的提升也有所帮助。

6 结论

本文提出一种基于医疗领域词典与预训练模型融合的医疗命名实体识别的方法。与传统的通用预训练模型直接识别的效果相比，本文提出的方法使用无标注医疗文本数据对该模型进行同领域预训练，同时使用伪标签方法利用大量无标注医疗文本数据扩充了训练集，在医疗领域的命名实体识别任务上取得更好的效果。

通过对结果分析，发现由于语料规模不大、实体标注存在问题等情况的限制，很多实体的精确边界并没有很好的识别出来。我们未来的工作将注重如何设定更有效的融合规则，同时也关注如何提升在训练集中未出现的新实体的识别效果。

参考文献

1. Lei, J., Tang, B., Lu, X., Gao, K., Jiang, M., Xu, H.: A comprehensive study named entity recognition in Chinese clinical text (2014)
2. Das, A., Garain, U.: CRF-based Named Entity Recognition @ICON 2013 (2014)
3. Gayden, V., Sarkar, K.: An HMM Based Named Entity Recognition System for Indian Languages: The JU System at ICON 2013 (2014)
4. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch, Journal of Machine Learning Research, 12(Aug), 2493–2537 (2011)
5. Huang, Z., Xu, W., Yu, K.: Bidirectional lstm-crf models for sequence tagging (2015)
6. Chiu, J.P., Nichols, E.: Named entity recognition with bidi-rectional lstm-cnns, Trans. Assoc. Comput. Linguist., 357–370 (2016)
7. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need (2017)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018)
9. Yang, H., Li, L., Yang, R., Zhou, Y.: Named Entity Recognition Based

on Bidirectional Long Short-Term Memory Combined with Case Report Form, Chinese Journal of Tissue Engineering Research, 22(20), 3237-3242 (2018)

10. Wan, L., Luo, Y., Li, Z., Qi, X.: The Recognition of Naming Entity of Bi-LSTM Chinese Electronic Medical Records Based on the Joint Training of Chinese Characters and Words, China Digital Medicine, 14(2), 54-56 (2019)
11. Chowdhury S., Dong, X., Qian, L., Li, X., Guan, Y., Yang, J., Yu, Q.: A multitask bi-directional RNN model for named entity recognition on Chinese electronic medical records, BMC Bioinfomatics, 19(17), 499 (2018)
12. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach (2019)
13. Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z., Wang, S., Hu, G.: Pre-Training with Whole Word Masking for Chinese BERT (2019)
14. Lee, D.: Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks (2013)
15. Gururangan, S., Marasovi´c, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., Smith, N.: Don' t Stop Pretraining: Adapt Language Models to Domains and Tasks (2020)
16. Wu, Y., Schuster, M., Chen, Z., et al.: Google' s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation (2016)
17. Li, S.: Subword-Level Chinese Text Classification Method Based on BERT, Computer Science and Application, 10(Jun), 1075-1086 (2020)