基于 BERT-CRF 的中药说明书实体识别

刘兰华 闭水清

南宁学院人工智能学院，南宁，530200 广西壮族自治区人民医院神经外科，南宁，530200

摘 要 中药说明书实体识别是智慧医疗服务中一项重要的基础任务，当前医院诊疗过程中采用人工分析中药说明书文本的方法，容易产生关键信息遗漏且效率低下。为此，提出一种结合 BERT 与条件随机场的中文实体识别模型，通过对抗训练优化、混合精度训练优化、多模型融合优化、数据集半监督学习增强优化等操作，可有效提升实体抽取识别的效果和性能。实验结果表明，论文提出的模型能实现 77%以上的实体识别 F1 分数，显著优于传统的循环神经网络和卷积神经网络模型。

关键字 深度学习,BERT,条件随机场,命名实体识别,中药说明书

Entity Identification of Chinese Medicine Instructions Based on BERT-CRF

Liu Lanhua

School of Artificial Intelligence
Nanjing University
Nanning 530200, China;
lhliu5@iflytek.com

Bi Shuiqing

Department of Neurosurgery
People's Hospital of Guangxi Zhuang Autonomous Region
Nanning 530200, China;
clearwaterb@qq.com

Abstract—The physical identification of Chinese medicine instructions is an important basic task in smart medical services, and the manual analysis of the text of Chinese medicine instructions is adopted in the current hospital diagnosis and treatment process, which is prone to key information omissions and inefficiency. Therefore, a Chinese entity recognition model combining BERT and conditional random airport is proposed, which can effectively improve the effect and performance of entity extraction recognition through adversarial training optimization, mixed-precision training optimization, multi-model fusion optimization, and data set semi-supervised learning enhancement optimization. Experimental results show that the model proposed in the paper can achieve more than 77% of the entity recognition F1 score, which is significantly better than the traditional recurrent neural network and convolutional neural network model.

Key words—Deep learning, BERT, CRF, Named entity recognition, Instructions for chinese medicine

1 引 言

人工智能加速了中医药领域的传承创新发展,特别是在疫情催化下，人工智能正在加速助力中医药传承创新加速发展，其中中医用药知识体系沉淀挖掘是一个基础工作，对中医药文本的信息抽取是构建中医药知识图谱的核心部分，为上层应用如临床辅助诊疗系统的构建（CDSS）等奠定了基础。

中药说明书实体识别指从中药说明书的文本中识别医疗实体的边界并判断医疗实体的类别,包括药品适用疾病种类、药品适用人群、药品成分和药品服务方法等，对后续构建医疗实体关系[1]、分析中药描述句法[2]、构建疾病知识图谱[3]和搭建智能医疗问答平台等发挥重要作用。

但是中药说明书通常包含大量的医学术语，构词十分复杂,实体常常存在嵌套现象并且实体边界模糊；医学实体描述具有多样性，没有固定规则；随着医学技术的发展，新的实体不断涌现；公开的医疗领域命名实体标注数据集较少，人工标注价格昂贵，运用深度学习技术缺乏足够的训练数据[4]。这些特点进一步加大了中药说明书实体识别的难度，性能难以达到可用的程度。

针对上述问题，本文引入哈工大讯飞联合实验室发布的基于 Whole Word Masking(WWM)的中文训练 BERT 模型 chinese-roberta-wwm，并将该训练模型做为 CRF 模型的发射矩阵，获得中文 BERT-CRT 的 Baseline 模型，Baseline 模型的 F1 值只有 71.36%,本文对 Baseline 模型做了多种优化，包含使用 FGM 和 PGD 进行模型对抗训练、使用混合精度训练以提高识别速度、融合 BERT + SPAN 和 BERT + MRC 模型、采用多种数据划分、随机数种子和句子长度以及使用半监督学习方式训练以充分使用数据集，最终模型效果提升了 6.6%。

2 数据集标注与处理

2．1 数据集介绍

本文标注数据源来自中药药品说明书，共包含1997份去重后的药品说明书，说明书中的关键信息包括症状、证候、药品、疾病分组、食物、药物性味、中药功效、食物分组、人群、药品分组、药物剂型、药物成分、中药功效等13类实体。

2．2 数据集处理

加载中药说明书数据集之后，先做初步探索，发现说明书大部分都是长文本句子构成，如图1所示，而且说明书标签分布不均匀，如图2所示。

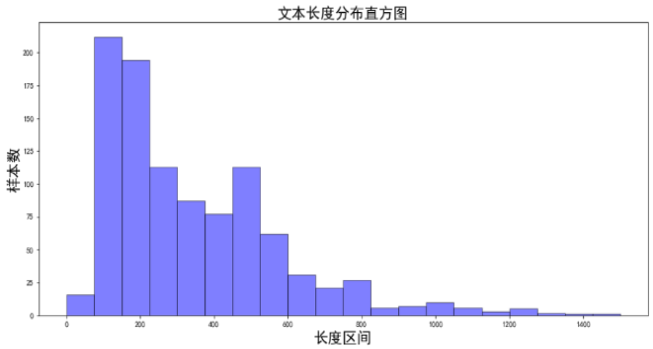

图 1 中药说明书数据集文本句子长度分布图

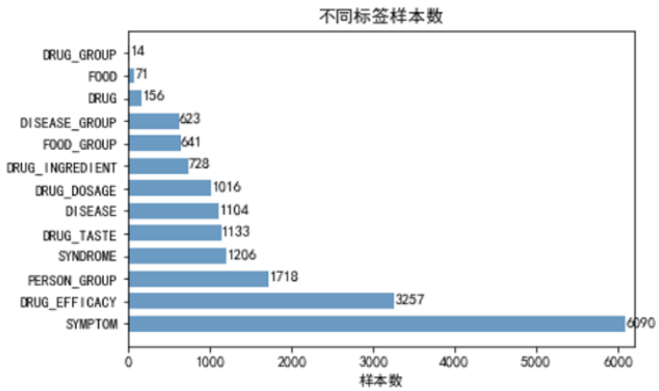

图 2 中药说明书数据集标签数目图

接下来对说明书文本进行预清洗与长文本切分。预清洗部分对无效字符进行过滤。针对长文本问题，采用两级文本切分的策略。切分后的句子可能过短，将短文本归并，使得归并后的文本长度不超过设置的最大长度。此外，利用全部标注数据构造实体知识库，作为领域先验词典。最后融合 RoBERTa+CRF+FGM 对训练集进行 5 折交叉验证，修正数据集中的漏标、错标、摸棱两可和手滑等产生误差数据，修正流程和效果如图 3 所示。

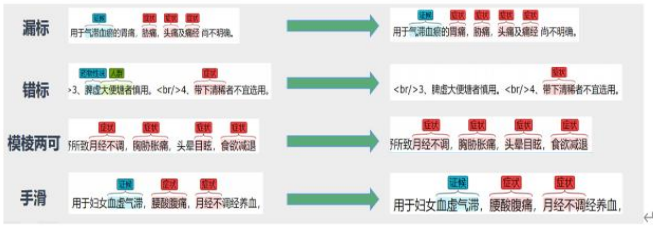

图 3 数据预处理效果

3 模型构建与优化

3．1 构建 Baseline 模型

本文构建的基础参照模型是采用腾讯开源的24层USER框架，加载哈工大讯飞联合实验室的chinese-roberta-wwm模型，模型结构如图4所示。然后采用大规模优质中文语料继续训练，CLUE 任务中单模第一。Baseline模型的BERT层学习率2e-5；其他层学习率2e-3，在加权滑动参数平均最后几个epoch模型的权重之后，得到更加平滑和表现更优的模型。

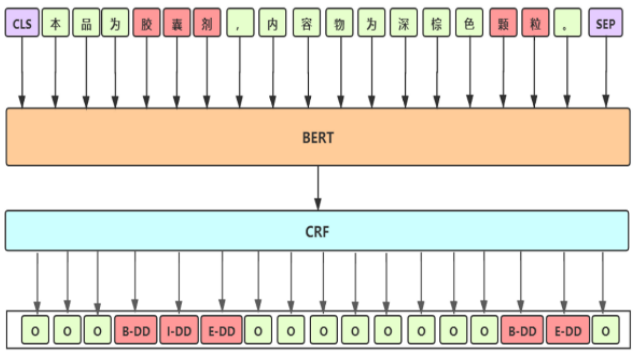

图 4 BERT-CRF Baseline模型结构图

在文本实体识别领域中，通常使用精确度 (P)、召回率 (R) 和 F 值（F-score）来评估模型的性能[5]，虽然做了一些优化设置，但是Baseline模型的三个性能指标相对较低，关键的F值只有71.36%，具体指标如表1所示。

表 1 Baseline模型性能指标

精确度P	召回率R	F-score
63.82%	70.98%	71.36%

而且Baseline模型在实体标签的类别预测时出现歧义性的错误，如“中医属热毒伤阴淤血”文本中的“淤血”应该属于症候实体标签，但是Baseline模型识别成症状实体标签，“用于气滞血瘀所致的乳腺增生病”文本中的“乳腺增生病”应该属于疾病实体标签，但是Baseline模型识别成疾病分组实体标签，也会出现嵌套性错误，“软坚散结，化淤定通”文本中的“化淤定通”应该属于疾病实体标签，但是Baseline模型识别成“化

淤”属于症状实体标签，如图5所示。

图 5 Baseline模型实体识别误差

3．2 模型对抗训练优化

对抗训练是一种防御对抗攻击的方法，是将混合一些微小的扰动而生成的对抗样本，加入到训练集中实现数据增强，让模型在训练的时候就先学习一遍对抗样本，以实现增强模型的对抗稳健性和提高模型的鲁棒性[6]。

本文使用FGM(Fast Gradient Method)对embedding层在梯度方向添加扰动和PGD(Projected Gradient Descent)对算法迭代扰动，每次扰动被投影到规定范围内，这两种引入噪声的训练方式，可以对参数进行正则化，缓解模型鲁棒性差的问题，提升模型泛化能力。

3．3 模型混合精度训练优化

FGM和PGD对抗训练增加了训练集数据，导致模型的训练计算效率下降，混合精度训练可以优化训练耗时。混合精度训练是在尽可能减少精度损失的情况下利用半精度浮点数加速训练，它使用FP16即半精度浮点数存储权重和梯度，在减少占用内存的同时起到了加速训练的效果。

本文在内存中用FP16做存储和乘法来加速，用FP32做累加避免舍入误差，同时反向传播前扩大2^k倍loss，防止loss下溢出，在反向传播后将权重梯度还原实现混合精度训练优化。

3．4 多模型融合优化

在实际的生活应用场景种，不存在一种万能的算法模型，在所有情况下都胜过其他的算法，而多模型融合优化的思想就自然而然出现了，就是充分运用不同算法模型各种的优势，取长补短，组合形成一个强大的模型。由图 2-2 可以看出 BERT-CRF 的 Baseline 模型在中医药说明书实体识别存在着比较明显的歧义误差，可以采用多级医学命名实体识别系统通过特定的方式组合的方法以消除歧义性。

本文通过更换随机种子、更换句子切分长度实现训练数据差异化，把 BERT-CRF、BERT-SPAN 和 BERT-MRC 这 3 个差异化模型框架同时做 5 折交叉验证得到的模型进行第一级概率融合，将 logits 平均后解码实体，最后把概率融合后的模型进行第二级投票融合，获取最终结果。具体的多模型融合优化流程如图 6 所示。

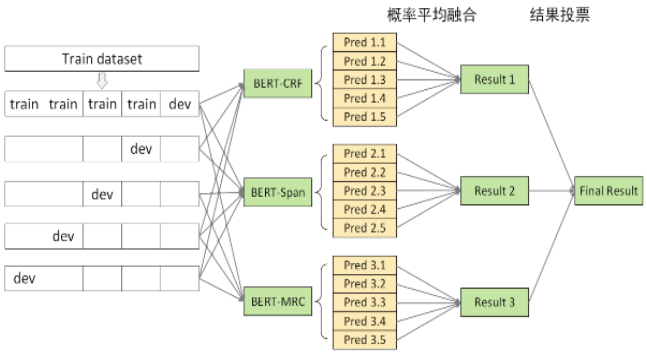

图 6 多模型融合优化流程

3．5 半监督学习优化

去重的数据集仅有 1997 份，其中有部分数据是预测数据未带标记。为了缓解医疗场景下的标注语料稀缺的问题，本文使用多模融合的基准模型在保证伪标签的准确率通过半监督学习（伪标签）充分对预测数据进行标记，从而提升数据集的利用率。具体做法是：首先使用原始标注数据训练一个基准模型 M，然后使用基准模型 M 对初赛测试集进行预测得到伪标签，最后将伪标签加入训练集，赋予伪标签一个动态可学习权重，加入真实标签数据中共同训练得到模型 M’。具体流程如图 7 所示，alpha 为动态可学习权重。

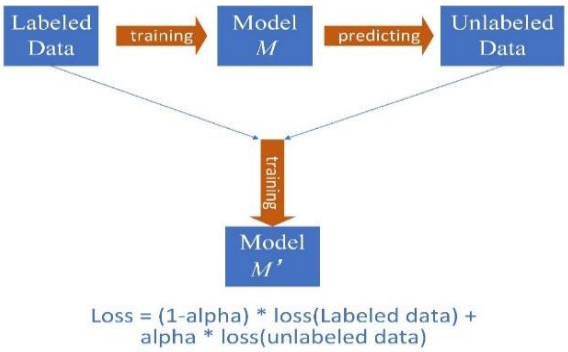

图 7 半监督学优化流程

经过四步优化之后，BERT-CRF模型在中药说明书实体识别的F-score值达到了77.95%，详细性能指标见表2。

表 2 BERT-CRF模型优化之后的最终性能指标

精确度P	召回率R	F-score
74.51%	80.85%	77.92%

模型优化之后的精确度 (P)、召回率 (R) 和 F 值（F-score）都有提升，说明优化操作十分可行。

4 结束语

本文提出了基于BERT-CRF模型对中药说明书进行实体识别的方法，验证对Baseline模型的对抗训练优化、混合精度训练优化、多模型融合优化、数据集半监督学习优化等操作的可行性和有效性，最终的模型的精度提升了6.6%，证明本文设计模型在中药说明书识别有更好的识别性能。

参 考 文 献

[1] Xue K，Zhou Y，Ma Z，et al. Fine-tuning BERT for Joint Entity and Relation Extraction in Chinese Medical Text[J]. ArXiv Preprint ArXiv：1908.07721，2019.

[2] 甘丽新，万常选，刘德喜，等. 基于句法语义特征的中文实体关系抽取[J]. 计算机研究与发展，2016,53(2):284-302

[3] 王鑫，邹磊，王朝坤等. 知识图谱数据管理研究综述[J]. 软件学报，2019,30(7):2139-2174

[4] 杨锦锋, 于秋滨, 关毅等. 电子病历命名实体识别和实体关系抽取研究综述[J]. 自动化学报，2014, 40(8): 1537-1562.

[5] 刘浏, 王东波. 命名实体识别研究综述[J]. 情报学报, 2018, 37(3): 329-340.

[6] Seyed-Mohsen Moosavi-Dezfooli S. M., Fawzi A., Frossard P.. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks.2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) ,2016, Las Vegas, NV, USA, IEEE Press, 2016:2547-2582