基于 BERT 与模型融合的医疗命名实体识别*

乔锐[1]，杨笑然[1]，黄文亢[1]

阿里健康信息技术有限公司[1]
{qiaorui.qr,xiaoyang.yxr,wenkang.hwk}@alibaba-inc.com

Abstract. 医疗命名实体识别是将临床电子病历中的自由文本由信息转化为数据的第一步，具有极高的研究价值和应用价值。此次全国知识图谱与语义计算大会（CCKS）委员会针对医疗命名实体识别设立了一项评测任务，要求对包括疾病和诊断、影像检查、实验室检验、手术、药物以及解剖部位在内的六种类型的实体进行识别。针对这项任务，本文提出了一种基于 BERT 与模型融合的医疗命名实体识别方法。本文所提出的方法在最终的测试数据集上得到了严格指标 0.8562 的结果，排名第一。

Keywords: 电子病历，命名实体识别，BERT

1 引言

随着电子病历的应用范围越来越广，各级医院产生了海量的电子病历，如何对电子病历中的信息进行挖掘逐渐成为了近年来的研究热点。其中，医疗命名实体识别作为将临床电子病历中的自由文本由信息转化为数据的第一步，逐渐成为一个重要的研究课题。

但是目前来看，由于数据获取和标注的困难，仍缺乏统一的标准和公共数据集，医疗命名实体识别在我国的研究进展相当缓慢。针对这一问题，2019 年的 CCKS 组委会设立了医疗命名实体识别的评测任务，提供了相当数量的标记文本，在评测的同时也推动了我国医疗命名实体的标注规范和公共数据集的建立，客观上为相关研究者提供了更加丰富的研究资源。

目前，针对命名实体识别任务，最常用也最有效的方法是基于机器学习的方法，如支持向量机（SVM）[1]，条件随机场（CRF）[9]，结构化支持向量机（SSVM）[6]，递归神经网络（RNN）及其变体模型 [5] 和卷积神经网络（CNN）及其变体模型 [10]。另外，近年来，预训练模型得到了越来越多的关注，谷歌 AI 团队于 2018 年发布了 BERT 模型 [4]，在 11 种不同自然语言处理评测任务中创造了最佳成绩，被认为是自然语言处理领域里程碑式的进步。BERT 基于海量数据进行预训练并提供开源的预训练模型，已有很多研究借助 BERT 获得了不错的精度。

本文中，我们参与了 CCKS 2019 医疗命名实体评测任务并提出了基于 BERT 与模型融合的医疗命名实体识别的方法。通过使用预训练的 BERT 模型并将基于 BERT 模型扩展成的多种模型进行融合，我们基于 CCKS 2019 医疗命名实体评测任务所提供的数据集得到了严格指标 F1 分数 0.8562 的结果。接下来，本文分别对问题定义、方法、实验结果以及结论进行介绍。

* 通讯邮箱：wenkang.hwk@alibaba-inc.com

2 问题定义

对于给定的一组电子病历纯文本文档，任务的目标是识别并抽取出与医学临床相关的实体提及并将它们归类到预定义类别中。CCKS2019 组委会针对评测任务给出了 1000 份标注好的训练数据用于模型的训练和调优，共需识别包括疾病和诊断、影像检查、实验室检验、手术、药物以及解剖部位在内的六种实体。一般来讲，中文电子病历命名实体识别是一个序列标注问题。在这里，我们使用 BMESO（Begin, Medium, End, Single, 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（char level）的标记。例如“患者 3 月余前于我院诊断为“直肠癌””的标签序列如图 1 所示。

患	者	3	月	余	前	于	我	院	诊	断	为	“	直	肠	癌	”
O	O	O	O	O	O	O	O	O	O	O	O	O	B-DIS	M-DIS	E-DIS	O

图 **1.** 标签序列举例

3 方法

在这个部分中，首先对数据的预处理进行简单的介绍，接下来分别介绍条件随机场（CRF），双向 LSTM 以及 BERT 算法，最后对本文所用的模型以及模型融合策略进行介绍。

3.1 数据预处理

首先我们注意到官方提供的数据集（以下称原始数据集）有部分标注在某些词上存在一定程度上的标注标准不统一的问题，手动对数据集进行了修正（以下称修正数据集）。同时考虑到标注标准应该可以由模型从原始数据集中自动学习到，泛化能力也会更强，所以也保留了原始数据集。此时产生了原始数据集和修正数据集这两个独立的数据源。

由于医学领域的用词造句较为特殊，目前的公共分词工具在医学术语中表现不佳，所以在这里我们使用字符而不是词语作为序列标注模型的单位。本文将原始数据集和修正数据集中的每一条数据分别拆分为单个字符，按照前文所述的 BMESO（Begin, Medium, End, Single, Other）标签方案将数据集给出的标签映射到每一个字符上。至此数据预处理流程结束。

3.2 条件随机场（CRF）

条件随机场是一种无向概率图模型，是一种判别模型，长期以来广泛的应用于序列标注问题 [2]，[11]，[3]。给定字符序列 $z = \{z_1, \cdots, z_n\}$，z_i 代表第 i 个字符及其特征所组成的输入向量；再给定 z 的标签序列 $y = \{y_1, \cdots, y_n\}$，$\gamma(z)$ 代表 z 的所有可能标签。CRF 模型定义了在给定字符序列 z 时，标签序列为 y 的概率公式:

$$p(y|z;\theta) = \frac{\sum_{t=1}^{n} \exp(S(y^t, z^t, \theta)}{\sum_{t=1}^{n} \sum_{j \in \gamma(z)} \exp(S(y_j, z^t, \theta))} \tag{1}$$

其中，$S(y^t, z^t, \theta)$ 为势函数，θ 是 CRF 模型的参数。

3.3 双向 LSTM

LSTM 的全称是 Long Short-Term Memory，是 RNN 的一种。简单来说，LSTM 是通过对细胞状态中旧信息遗忘和新信息的记忆来传递对后续时刻计算有用的信息，同时丢弃无用的信息，并在每个时间步都会对隐层状态进行输出。

LSTM 的特性使其可以更好的捕捉到较长距离的依赖关系。但是利用 LSTM 对句子进行建模仍存在无法编码从后到前的信息的问题，所以 BiLSTM 应运而生。BiLSTM 的全称是 Bi-directional Long Short-Term Memory，由前向 LSTM 与后向 LSTM 组合而成。BiLSTM 在自然语言处理任务中都常被用来建模上下文信息，在命名实体识别任务中获得了极为广泛的应用 [8]，[7]。

3.4 BERT

BERT 模型由谷歌 AI 团队于 2018 年发布，其作为一个预训练模型在 11 种不同自然语言处理评测任务中创造了最佳成绩，一经发布即轰动了整个自然语言处理研究界 [4]。简单来说，BERT 是一种预训练语言表示的方法，在大量文本语料上使用无监督的方式训练了一个通用的语言理解模型，然后在这个模型上设置轻量级的下游任务接口去执行特定的自然语言处理任务。由于 BERT 采用的是无监督的训练方式，这意味着 BERT 的预训练无需繁重的语料标注过程，而海量的文本数据可以直接在网络上得到，这意味着 BERT 拥有着极大的发展空间。

BERT 模型凭借 Masked Language Model（Masked LM），双向 Transformer encoder 以及句子级别的负采样得到了一个强大的、深度双向编码的、包含着充分的描述了字符级、词级、句子级甚至句间关系的特征的预训练模型，针对特定任务，只需简单设置下游任务接口，同时使用任务数据对模型进行微调，即可完成整个模型的构建。从某种意义上来讲，如果忽略掉 BERT 模型复杂的网络结构，其实可以将其看作一个提供 char embedding 的字符嵌入生成器，不过 BERT 所生成的嵌入包含了更多以及更深层次的信息，这是传统 embedding 方式所无法比拟的。

3.5 模型

前面提到，本文所使用的的特征为纯字符特征，在这里，我们利用 BERT 模型自带的词典将数据集的单个字符映射为 ID 后，再经 BERT 模型的 Embedding 层对字符 ID 进行 char embedding 后进入网络中进行学习。

本文基于原始数据集和修正数据集分别构建以下三种模型（下面分别称模型 (1)，模型 (2)，模型 (3)）：

(1) BERT + TimeDistributed_Dense（BT/FBT）
(2) BERT + BiLSTM + TimeDistributed_Dense（BBT/FBBT）

(3) BERT + BiLSTM + CRF（BBC/FBBC）

另外，以 BERT + BiLSTM + CRF 为例，BBC 为使用原始数据集训练得到的模型的简称，FBBC 为使用修正数据集训练数据得到的简称。

3.6 模型融合

模型构建完毕后，综合验证集和测试集对以上三种模型的表现进行分析，发现基于原始训练集构建的模型 (3) 表现较为突出，对于各个种类的实体的识别性能较为均衡，不过仍存在部分如“N”, “K”等长度为 1 的实验室检验实体识别不到以及较长的疾病或诊断实体识别不全的问题。经过观察发现，基于原始训练集的模型 (2) 对于长度较短的实体的识别效果较好，而基于修正数据集构建的模型 (3) 对于较长的疾病和诊断实体识别效果较好，这也确实符合 CRF 的特性，故将以上三个模型得到的结果在相应实体类别上进行融合，得到最终的识别结果。

3.7 规则

针对由于某些实体在训练集中出现次数较少所导致的模型识别结果存在边界模糊、合并或分裂错误的情况，本文通过频繁模式挖掘等方法构建了一系列的规则约束，如“<LAB> 数”这一规则表示识别出的实验室检验后面如有“数”则需合并进实体之中。

另外对于本次药物专业性强，出现次数较少的特点，我们在百度文库上获取了常见化疗药物及其英文缩写并将这些知识以词表的形式融入到模型中。

4 实验

4.1 数据

CCKS 2019 医疗命名实体识别评测任务共提供了 1000 例语料作为训练数据集，语料中共有包括疾病和诊断、影像检查、实验室检验、手术、药物以及解剖部位在内的六种类型实体的标注。另外，CCKS 2019 组委会还提供了 379 例未标记的语料作为测试数据集对评估模型进行评估。在训练过程中，出于模型调优以及超参数选择的需求，我们从 1000 例训练预料随机抽取 800 例作为训练语料，200 例作为测试语料。

4.2 实验设置

前面提到的模型 (1)，模型 (2) 和模型 (3) 的参数是统一的，其中 BERT 模型的最长序列长度选择为 512，优化算法为 Adam 算法，学习率设为 1e-5，batch_size 为 8，epoch 为 10，另外，双向 LSTM 层的隐层节点数均为 32（这里指单个方向的隐层节点数）。

表 **1.** 模型性能对比（严格）

模型	评估方式	实验室检验	手术	药物	影像检查	解剖部位	疾病和诊断	综合
BBC	严格	0.7047	0.8167	0.9248	0.8322	0.8587	0.8249	0.8397
BBC+ 规则	严格	0.7538	0.8208	0.9548	0.8580	0.8595	0.8292	0.8495
BBC+BBT+ 规则	严格	0.7694	0.8333	0.9602	0.8616	0.8610	0.8395	0.8549
BBC+BBT+FBBC+ 规则	严格	**0.7694**	**0.8333**	**0.9602**	**0.8629**	**0.8618**	**0.8429**	**0.8562**

表 **2.** 模型性能对比（松弛）

模型	评估方式	实验室检验	手术	药物	影像检查	解剖部位	疾病和诊断	综合
BBC	松弛	0.8885	0.9068	0.9573	0.8999	0.9310	0.9234	0.9251
BBC+ 规则	松弛	0.9003	0.9186	0.9774	0.9217	0.9318	0.9250	0.9301
BBC+BBT+ 规则	松弛	0.9102	0.9231	0.9826	0.9272	0.9328	0.9283	0.9331
BBC+BBT+FBBC+ 规则	松弛	**0.9102**	**0.9231**	**0.9826**	**0.9293**	**0.9344**	**0.9308**	**0.9346**

4.3 实验结果

表 1 和表 2 分别列举了不同模型组合在六种实体上的严格和松弛指标，性能指标的度量方式为 F1 分数，其中，性能最佳的模型的识别效果由粗体标出。

从表 1 和表 2 中可以看出，使用 BBC、BBT、FBBC 以及规则的融合模型取得了最高的精度。从表格中可以看到，在加入模型融合策略后，模型的精度有着明显的提高，表明我们所采取的模型融合策略是有效的。另外观察模型识别结果可以看出，加入 BBT 后，对于长度较短的实体的识别效果有所提升，而加入 FBBC 后，对于较长的疾病和诊断实体识别效果的提升也有所帮助。

还有，通过对松弛指标与严格指标的观察可以看出，模型已经找到了绝大多数实体，表示我们所提出的模型发现实体的能力较强，但是由于边界以及是否合并等因素的影响，相比于松弛指标，严格指标的精度较低。

5 结论

本文中提出了一种基于 BERT 与模型融合的医疗命名实体识别方法。相比于单纯的 BERT + BiLSTM + CRF 算法，本问所提出的模型取得了更高的精度，在 CCKS 2019 医疗命名实体评测中以严格指标 0.8562 排名第一。通过对识别结果的分析，我们发现由于语料规模等限制，很多实体的边界以及合并情况并不能被很好的识别出来，我们未来的工作将侧重如何更精确的地提取实体的边界以及确定实体的合并情况，同时也将注重于提升未出现在训练集中的新实体的识别精度。

References

1. Asahara, M., Matsumoto, Y.: Japanese named entity extraction with redundant morphological analysis. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. pp. 8–15. Association for Computational Linguistics (2003)

2. 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)
3. 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 p. 103252 (2019)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
5. Huang, Z., Xu, W., Yu, K.: Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991 (2015)
6. Lee, Y.J., Mangasarian, O.L.: Ssvm: A smooth support vector machine for classification. Computational optimization and Applications **20**(1), 5–22 (2001)
7. Luo, L., Yang, Z., Yang, P., Zhang, Y., Wang, L., Lin, H., Wang, J.: An attention-based bilstm-crf approach to document-level chemical named entity recognition. Bioinformatics **34**(8), 1381–1388 (2017)
8. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional lstm-cnns-crf. arXiv preprint arXiv:1603.01354 (2016)
9. McCallum, A., Li, W.: Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. pp. 188–191. Association for Computational Linguistics (2003)
10. Strubell, E., Verga, P., Belanger, D., McCallum, A.: Fast and accurate entity recognition with iterated dilated convolutions. arXiv preprint arXiv:1702.02098 (2017)
11. Yang, X., Huang, W.: A conditional random fields approach to clinical name entity recognition. In: CCKS Tasks. pp. 1–6 (2018)