Team MSIIP at CCKS 2019 Task 1

Minglu Liu[1], Xuesi Zhou[2], Zheng Cao[2], and Ji Wu[2]

[1] Tsinghua-iFlytek Joint Laboratory
[2] Department of Electronic Engineering, Tsinghua University
liuml@mail.tsinghua.edu.cn zhouxs16@mails.tsinghua.edu.cn,
caozheng12@foxmail.com
wuji_ee@mail.tsinghua.edu.cn

摘要 我们在本文中针对医疗命名实体识别任务,构建了一个融合多种深度神经网络技术和专家知识的混合系统，旨在从自然语言病历文本中准确、全面地识别出疾病与诊断、解剖部位、影像检查、实验室检验、药物、手术等多种类型的命名实体。系统包括一个基于多种深度神经网络的融合模型，和一个基于词典、上下文模式等人工总结知识的后处理模型。与大多数命名实体识别工作类似，我们将任务建模为经典的序列标注任务，依赖神经网络融合模型从训练数据中学习大部分的命名实体基础模式，采用多个不同类型的深度神经网络模型以互补性地引入多个角度的信息，并采用投票的方式完成序列标注任务。另一方面，由于医疗文本标注数据的稀疏性问题，我们针对性地通过引入人为总结的知识来弥补数据驱动的神经网络方法的不足，主要是基于网络爬取的医疗领域内各类别实体的词典和人工总结的上下文模式，实现模型预测结果的后处理，优化框架的整体性能。我们将训练数据随机地按照 7:3 的比例分割为训练集和验证集，进行了一系列实验。实验结果表明我们的系统性能超过了基线模型，验证了我们所构建的框架的有效性。

Keywords: 命名实体识别 · 序列标注 · 深度神经网络.

1 引言

临床 EHRs 中包含着大量的自然语言文本，通常涵盖患者症状描述、实验室检验指标、用药安排等一系列重要医疗信息。医疗命名实体识别是理解医疗文本、挖掘文本中蕴含的信息、发现新的知识等工作的基础。

CCKS 2019 医疗命名实体识别子任务针对中文电子病历语义化，对于给定的电子病历纯文本文档，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 6 种预定义类别，疾病和诊断、检查、检验、手术、药物以及解剖部位。

命名实体识别，通常被建模为序列标注任务，其最常使用的模型分为两类，基于统计的模型和基于神经网络的模型。隐马尔科夫模型（HMM）、条件随机场（CRF）等基于统计的模型，试图在给定输入序列的条件下以最优联合概率为目标，去推断整个标注序列。这些模型通常会使用人工总结的特征和针对特定任务的资源 [7, 9, 10]。与基于统计的模型相比，基于神经网络的模型试图在解决序列标注问题的过程减少特征工程，甚至放弃特征工程。

Collobert 等 [1] 最先采用基于神经网络的模型解决序列标注问题，以海量无标注文本训练得到的词向量作为模型的输入特征，大大减少了特征工程的比重。随着神经网络技术的不断发展，不仅仅限于 NER 任务，基于神经网络的模型在许多序列标注任务中取得了 state-of-the-art 性能 [5, 8]。在医疗领域，诸如 CNNs 和 RNNs 等神经网络方法，也已成功的应用于命名实体识别、事件检测、概念抽取等任务 [3, 6, 11–13]。这些工作验证了神经网络方法在序列标注任务上的强大能力。

然而，医疗领域命名实体识别任务中使用基于神经网络的方法首先面临着训练数据不足的问题，由于医疗领域知识的专业性，导致缺少大量的准确标注的训练数据，影响神经网络模型的充分训练。另一方面，医疗文本有着更加特殊的文本内容、语言规范，包含大量专业命名实体的同时，并不严谨地遵循语法组织语言，这就导致了依赖海量无标注通用语料训练的词向量表示无法在医疗领域取得良好性能。

针对上述问题，我们构建了一个融合多种深度神经网络技术和专家知识的混合系统，旨在从自然语言病历文本中准确、全面地识别出疾病与诊断、解剖部位、影像检查、实验室检验、药物、手术等多种类型的命名实体。系统包括一个基于多种深度神经网络的融合模型，和一个基于词典、上下文模式等人工总结知识的后处理模型。我们实现了 5 种神经网络单模型，试图有不同侧重地完成医疗命名实体识别任务，然后以投票的方式融合多个模型的预测结果，所使用的模型主要分为三类：1) 引入更细粒度的标签体系；2) 引入语言分词信息；3) 引入 BERT 模型以丰富字向量的语义表示。在后处理模型部分，重点解决训练数据稀疏的问题，基于融合产生的预测结果，进行 BIO 标签校正等简单的后处理，同时还利用网络爬取的术语词典和人工总结的上下文模式实现对模型预测结果的进一步校正。

2 实验方法

2.1 引入细粒度分层标签的模型

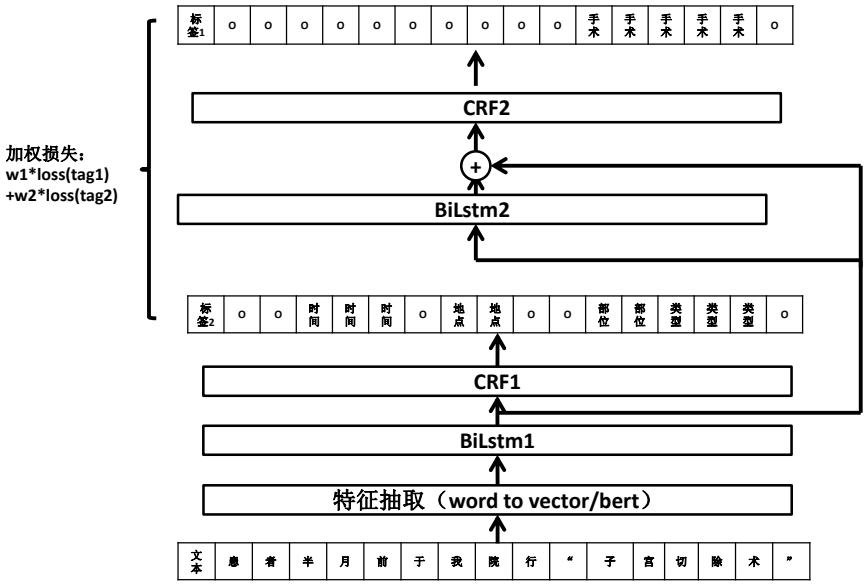

图 1. 细粒度分层标签模型

医学实体本身具有其内部结构：如疾病“胃（部位）炎”，中的“胃”本身是一个部位，手术“全麻上子宫切除术”，其中“全麻”是麻醉方式，“子宫”是一个部位，“切除术”是一个手术方法。我们采用 Nested NER[4]，通过对相同语料不同层次，不同角度的标注，增加文本标注的多样性，使少量的文本能够提供更多的信息。具体的，我们使用双层标签提取细粒度医疗文本语义表示，针对评测使用的六种实体类型，我们额外设计了“症状”，“部位”，“数值”，“单位”，“麻醉方式”，“手术类型”等 30 余种额外的标签类型，其中“症状”，“手术”，“疾病”，“检查”，“检验”作为第一层 NER 输出结果，“部位”，“单位”等其他 20 余种标签作为第二层 NER 输出结果。通过两层 NER 的混叠情况，筛选出符合规范的“部位”标签。我们使用双层

BiLSTM+CRF 的结构来使用这种数据类型。具体的输入文本 X 首先经过特征提取层（可以使用 word2vec 词向量，或者使用 bert feature extraction 功能）转换为字的向量表达，然后经过第一层 BiLSTM1+CRF1 预测标签 2 的信息，然后使用第一层的 BiLSTM1 作为输入，输出第二层 BiLSTM2，最后将 BiLSTM1 和 BiLSTM2 进行拼接，输入到 CRF2 中，预测最终的标签序列。损失函数采用两次标签预测的多任务损失函数：

$$Loss(x) = w_1 * loss(tag1) + w_2 * loss(tag2) \tag{1}$$

，其中 w1 和 w2 是超参数。

2.2 引入分词信息的模型

汉语中词是最重要的语义单元，分词信息的引入能够帮助 NER 任务的完成。我们首先选择了在中文 NER 任务中表现较好的 Lattice LSTM 模型 [14]，Lattice LSTM 是基于字符的 LSTM-CRF 模型的改进版，其通过使用网格结构的 LSTM 和预先设置的潜在词词典，将潜在的词信息加入基于字符的 LSTM-CRF 中。该模型在利用了字信息的基础上，又很好地利用了词和词序列的信息。同时，相对于基于词的方法，Lattice 模型不会受到词语分割错误的影响，在中文 NER 有不错的表现。

此外，我们还尝试优化了经典的 CharCNN+BiLSTM+CRF 模型 [8]。通常使用的 CharCNN+BiLSTM+CRF 模型中的 CharCNN 部分是利用每个词所包含的字母的向量表示得到词的向量表示，然而汉字不同于英文中的字母，不仅汉字数远大于字母数，每个汉字也蕴含着远大于单独字母的语义信息，字粒度的上下文与词粒度的上下文有着各自独特的模式。针对这一问题，我们实现了一个基于字上下文的词表示模型，改进 CharCNN+BiLSTM+CRF 模型中的 CharCNN 部分，如图 2所示，CharCNN 部分打破分词的界限，使每个字向量都能够通过 CNN 学习到局部上下文模式，然后依据每个词所包含的字，通过动态的池化得到基于字上下文的词向量表示。

2.3 基于 BERT 模型丰富字语义信息的模型

BERT（Bidirectional Encoder Representation from Transformers）[2] 是一种基于 masked language model 的大规模语料预训练模型，再多项文

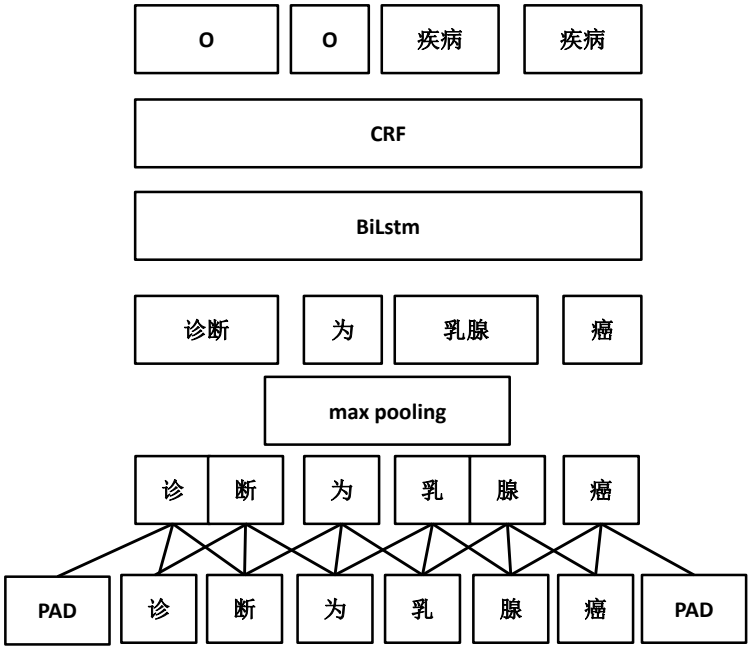

图 2. 基于字上下文的词表示模型

本任务中取了任务的最好效果。在本次评测中，我们也尝试了使用 BERT（BERT-Base，Chinese，12-layer，768-hidden，12-heads，110M-parameters）作为文本的特征表示层，后拼接 BiLSTM+CRF 作标签提取。

中文作为一种象形文字，文字的结构具有特殊含义，如“胸”“腹”“脑”等，均具有“月”作为结构的一部分，因此字形结构是提示 NER 发现实体的一种重要的信息。很多药品名称来源于英文的音译如:“奥沙利柏”和“奥沙利铂”，由于翻译和书写习惯的不同，有时会选择同音字来表示相同的实体，因此我们也希望能够利用中文的拼音作为实体发现的一项信息。借鉴[8]，我们实现了一种基于字，字形，拼音的三层表示提取结构，其中每个字视为一个 word，字的拼音视为一个单词的字母序列，网络结构如图 3所示。

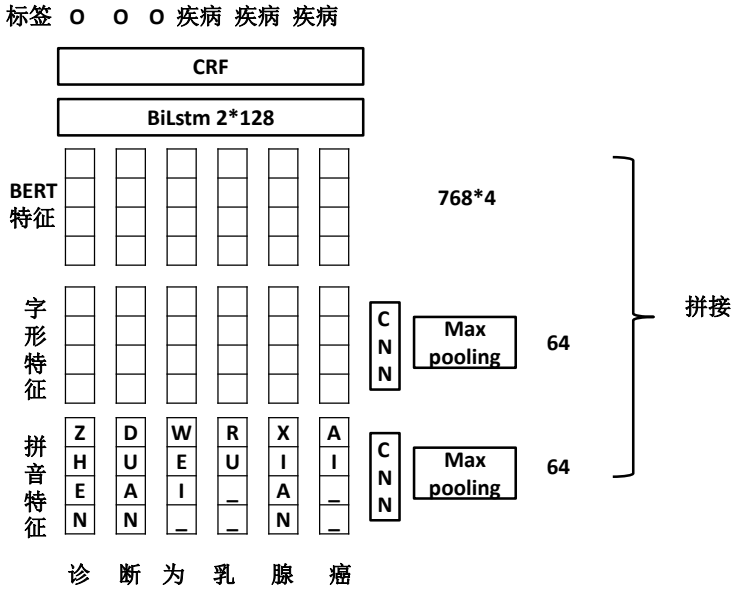

图 3. 引入拼音字形的 BERT+CRF 模型

2.4 基于词典和上下文的后处理

通过对公开的医疗相关网站在线抓取的词典信息，我们构建了一个规模在 3 万个词的医学术语词典，词典内容涉及手术，药物，疾病，部位等多种实体类型。同时基于自举策略，使用标注语料挖掘了一部分高频的实体上下文。另外，我们也使用了人工编辑的实体上下文规则。通过添加词表的白名单，以及上下文的正则匹配，构成了我们的规则后处理模块。

3 实验结果分析

3.1 实验数据介绍

数据集介绍: 数据集由医渡云（北京）技术有限公司编写，并由医渡云公司组织专业的医学团队进行人工标注。数据集共标注了“疾病和诊断”，“解剖部位”，“手术”，“药物”，“检查”，“检验”六种实体类型，数据集共标注了 5363 个实体。

3.2 预处理

首先对于输入的原始文本进行预处理，主要包括如下两步：

1. 切句：由于过长的序列会导致 LSTM 等模型的性能恶化，需要在保证句内语义信息相对完整的前提下，对输入的病历文本进行切分，实现中使用句号切割，同时限制最长 200 个字符，共产生 6355 个句子，按照 0.7:0.3 的比例随机拆分训练、测试集合；
2. 文本规范化：这一部分主要实现输入病历中的文本与符号的全半角统一，英文大小写转换，不可见字符的处理等功能。

3.3 模型实现

细粒度分层标签模型：在实现中取了 Bert 最后四个前向层，共 768*4 作为 biLSTM 的输入，biLSTM 隐层单元取 128 个，最后通过一个 128 个隐藏单元的全连接层连接 CRF 计算最后输出 tag。batch size 取 256，学习率 0.001，训练 10 个 epoch，并使用最后 5 个 epoch 的参数平均作为最后模型，w1=w2=0.5，使用 Adam 作为优化器。

Lattice LSTM：模型使用了在中文 Giga-Word 上预训练的字向量和词向量，其中词典大小为 70w 词左右。词向量和字向量的维数为 50 维。在实

验中，将给定比赛数据集切分，训练模型直到收敛为止。根据预处理中的分句情况，模型接受的最大句子长度设定为 250 字。模型使用了 dropout，设置为 0.5。优化器为 SGD，初始学习率为 0.015，decay rate 为 0.05。

基于字上下文的词表示模型: 实现中使用公开分词工具 jieba 作为分词工具，使用随机初始化方法初始化 200 维字 embedding 和 200 维词 embedding，卷积部分采用宽度为 2、3、4 的一维卷积核各 100 个，这样基于字粒度上下文获取的词粒度表示为 300 维，与原有的 200 维词 embedding 拼接，送入 BiLSTM 层，BiLSTM 隐层单元取 200 个，共两层，dropout 设为 0.6，最后通过全连接层连接 CRF 计算最后输出 tag。训练中 batch size 设为 24，训练 12 个 epoch，选取最后 10 个 checkpoint 进行模型参数平均作为最终模型，使用 Adam 作为优化器。

BERT+CRF 基础模型: 在实现中取了 BERT 最后四个前向层, 共 768*4 作为 biLSTM 的输入，biLSTM 隐层单元取 128 个，最后通过一个 128 个隐藏单元的全连接层连接 CRF 计算最后输出 tag。batch size 取 256，学习率 0.001，训练 10 个 epoch，并使用最后 5 个 epoch 的参数平均作为最后模型，使用 Adam 作为优化器。

引入拼音字形信息的 BERT+CRF 模型：在实现中取了 BERT 最后四个前向层, 共 768*4 作为字特征。拼音使用 26+1=27 个字符表示, 其中 1 用于表示英文和数字的拼音, 使用随机初始化方法初始化为 64 维 embedding。字形使用笔画的方式转为类似于拼音的序列表示，共使用 26 个常见中文笔画 +1 个用于表示英文和数字的特殊笔画。字形和拼音的卷积核尺寸均为 3，通道数选为 64。字符 + 字形 + 拼音特征选择拼接的方式融合，作为 BiLSTM 的输入。BiLSTM 隐层单元取 128 个，最后通过一个 128 个隐藏单元的全连接层连接 CRF 计算最后输出 tag。batch size 取 256，学习率 0.001，训练 10 个 epoch，并使用最后 5 个 epoch 的参数平均作为最后模型，使用 Adam 作为优化器。

3.4 模型实验结果

对于多个模型的结果，我们采用多数表决的方法确定为融合模型的输出结果。融合结果经过规则后处理模块进行校正后，作为最终的输出结果。各个模型及后处理规则融合后最终实验结果如表所示（评测指标使用 relax f1）：

表 1. 各模型及融合后实验结果

	疾病	手术	药物	检查	检验	部位	总体
Lattice	0.91	0.932	0.965	0.934	0.923	0.924	0.926
字上下文	0.912	0.942	0.961	0.921	0.911	0.889	0.909
BERT base	0.915	0.929	0.97	0.938	0.924	0.926	0.929
BERT+ 拼音字形	0.92	0.93	0.968	0.94	0.93	0.927	0.931
BERT+ 分层	0.915	0.935	0.97	0.94	0.919	0.84	0.885
融合	0.92	0.937	0.979	0.943	0.945	0.93	0.935
融合 + 规则	0.914	0.943	0.987	0.952	0.949	0.933	0.937

– 由于标注数据的标注规范存在一定的不一致性，因此我们选择使用 relax f1 作为我们性能的评估指标。
– 由于基于字上下文的 NER 模型中，需要输出以词为单位的 NER 标签，而在本次评测数据中，解剖部位受分词影响较大，如“头痛”，标注为头//部位痛，而一般分词工具无法将头痛拆分为两个词，因此基于字上下文的模型中，“部位”实体性能受到较大影像，但在一些长实体识别上，如“手术”等取得了与其他模型相近，或更优的效果。
– 基于精细标注的分层模型，“部位”标签与其他模型的“部位”标签在标注上存在一定的分歧，因此部位性能影像较大，导致总体性能降低。但是在其他实体的识别中，“药物”，“检查”等要素，均取得了单模型下最优的效果。
– 引入拼音和字形后，可以对原始基于文本语料进行预训练的 BERT 性能取得一定的提升。
– 从实验结果上看，不同模型的输出结果具有一定的互补性，多模型融合后具有一定的性能提升。使用规则模型进行后处理校正后，最终性能有一定的提升，最终取得 relax 模式下 f1 值 0.937。最终评测提交结果性能如表 2 所示:

表 2. 最终提交的评测结果

	疾病	手术	药物	检查	检验	部位	总体
strict	0.943	0.820	0.95	0.878	0.781	0.855	0.86
relax	0.943	0.942	0.973	0.944	0.924	0.934	0.939

4 Conclusion

为了克服医疗文本数据欠缺，标注复杂的问题，在本次评测中，我们选择了引入外部分词信息，细粒度标签信息，拼音、字形信息等额外信息提高实体识别性能，具体实现了利用这些外部信息的 5 种模型：细粒度分层标签模型（引入额外的标签信息），Lattice LSTM（引入外部分词信息），基于字上下文的词表示模型（引入外部分词信息），引入拼音字形信息的 BERT 模型（引入拼音，字形信息），最终取得了严格模式下 f1 值 0.86，宽松模式下 f1 值 0.939 的最终性能。

参考文献

1. 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)
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018)
3. Jagannatha, A.N., Yu, H.: Bidirectional rnn for medical event detection in electronic health records. In: Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting. p. 473. NIH Public Access (2016)
4. Ju, M., Miwa, M., Ananiadou, S.: A neural layered model for nested named entity recognition. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 1446–1459 (2018)
5. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural architectures for named entity recognition. In: Proceedings of NAACL-HLT. pp. 260–270 (2016)
6. Li, P., Huang, H.: Clinical information extraction via convolutional neural network. arXiv preprint arXiv1603.09381 (2016)
7. Luo, G., Huang, X., Lin, C.Y., Nie, Z.: Joint entity recognition and disambiguation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 879–888 (2015)
8. Ma, X., Hovy, E.: End-to-end sequence labeling via bi-directional lstm-cnns-crf. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1064–1074 (2016)

9. Passos, A., Kumar, V., McCallum, A.: Lexicon infused phrase embeddings for named entity resolution. In: Proceedings of the Eighteenth Conference on Computational Natural Language Learning. pp. 78–86 (2014)
10. Ratinov, L., Roth, D.: Design challenges and misconceptions in named entity recognition. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning. pp. 147–155. Association for Computational Linguistics (2009)
11. Tutubalina, E., Nikolenko, S.: Combination of deep recurrent neural networks and conditional random fields for extracting adverse drug reactions from user reviews. Journal of Healthcare Engineering **2017** (2017)
12. Viani, N., Miller, T.A., Dligach, D., Bethard, S., Napolitano, C., Priori, S.G., Bellazzi, R., Sacchi, L., Savova, G.K.: Recurrent neural network architectures for event extraction from italian medical reports. In: Conference on Artificial Intelligence in Medicine in Europe. pp. 198–202 (2017)
13. Wu, Y., Jiang, M., Lei, J., Xu, H.: Named entity recognition in chinese clinical text using deep neural network. Studies in health technology and informatics **216**, 624 (2015)
14. Zhang, Y., Yang, J.: Chinese ner using lattice lstm (2018)