Adam 优化算法, 学习率设置为 0.001, batch size 的大小为 64, epoch 的大小为 80.

5.3 实验结果和分析

实验结果的评价指标有 3 个, 分别为精确率, 召回率和 F 值. 计算公式如式 (10), (11), (12).

$$Precision(P)=\frac{\text{系统正确识别的实体个数}}{\text{系统识别的实体个数}} \tag{10}$$

$$Recall(R)=\frac{\text{系统正确识别的实体个数}}{\text{文档中的实体个数}} \tag{11}$$

$$F-measure=\frac{2\times P\times R}{P+R} \tag{12}$$

不同模型的实验结果分别见表 2, 3 和 4.

表 2 BiLSTM-CRF 的实验结果

实体类别	精确率	召回率	F1 值
SYMPTOM	0.8243	0.8145	0.8194
DISEASE	0.9458	0.9172	0.9313
TREATMENT	0.7727	0.7961	0.7843
CHECK	0.7171	0.6267	0.6689
平均值	0.8150	0.7886	0.8009

对比实验结果可以看出, IndRNN-CRF 模型在精确率上比基准模型 BiLSTM-CRF 高, 召回率的值为 0.6848, 相比于模型 BiLSTM-CRF 的召回率比较低. IDCNN-BiLSTM-CRF 模型在精确率, 召回率和 F1 值上均超过了基准模型 BiLSTM-CRF. 图 5, 图 6 和图 7 分别是模型 BiLSTM-CRF, IndRNN-CRF 和 IDCNN-BiLSTM-CRF 的 *Loss* 曲线图, 纵坐标代表 *Loss* 值, 横坐标代表的是迭代次数. 从图中可以看出在经过了 24 000 次的迭代后模型 BiLSTM-CRF 的 *Loss* 值大于 2.0, 模型 IndRNN-CRF 和 IDCNN-BiLSTM-CRF 的 loss 值小于 2.0, 其中模型 IndRNN-CRF 的 loss 值最低.

表 3 IndRNN-CRF 的实验结果

实体类别	精确率	召回率	F1 值
SYMPTOM	0.8640	0.7171	0.7837
DISEASE	0.9492	0.9066	0.9274
TREATMENT	0.7935	0.6226	0.6977
CHECK	0.7643	0.4931	0.5994
平均值	0.8427	0.6848	0.7521

表 4 IDCNN-BILSTM-CRF 的实验结果

实体类别	精确率	召回率	F1 值
SYMPTOM	0.8669	0.8023	0.8334
DISEASE	0.9469	0.9172	0.9318
TREATMENT	0.8076	0.8044	0.8060
CHECK	0.7727	0.6313	0.6949
平均值	0.8485	0.7888	0.8165

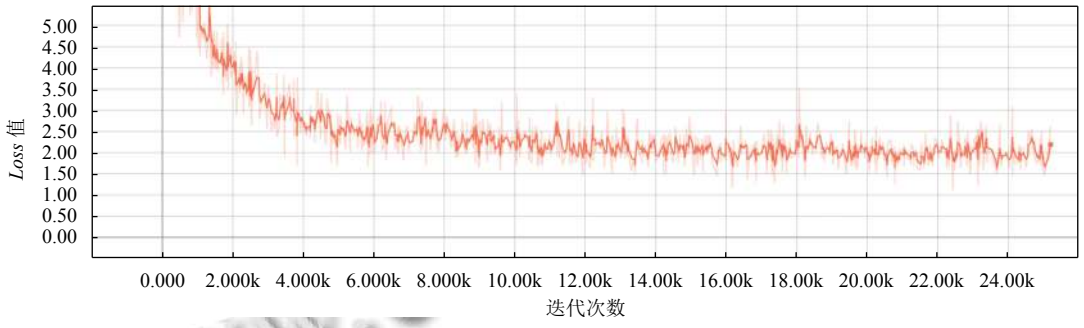

图 5 BiLSTM-CRF 的 loss-step 曲线图

由于 IDCNN-BiLSTM-CRF 模型的总体性能最好, 可以用在互联网在线问诊医疗文本的实体识别上, 该模型也可用在医学文献, 电子病历等文本的命名实体识别上. 模型 IndRNN 可以用在对精确率要求较高, 但对召回率要求不高的任务中.

6 结论与展望

本文针对在线问诊医疗文本, 利用深度学习技术设计了两种不同的神经网络模型, 进行医疗文本命名实体识别的研究, 共识别 4 类医疗实体: 疾病, 症状, 治疗和检查. 对基于字级别的命名实体识别任务, 在模型 IDCNN-BiLSTM-CRF 中使用卷积神经网络和循环神经网络提取特征向量, 并将两个特征向量拼接, 形成既包含全局特征又包含局部特征的向量, 该向量经过映射层后输入到 CRF 层中, 实验结果表明该模型的整体性能最好. 但是由于医疗领域的特殊性, 仍然需要继续提高医疗实体的识别率, 获取更精确的挖掘结果. 在接下来的工作中, 可以考虑先对医疗文本分词, 然后加入

词性或者拼音等特征训练模型, 提高识别率. 此外, 对于医疗文本还要考虑文本中是否含有修饰性实体, 比如表示时间和否定的词汇等, 如“无头痛”, 症状“头痛”前的“无”就是修饰实体. 模型最终结果与参数的调试也有较大的关系, 设置不同的参数, 模型的输出值可能会不同.

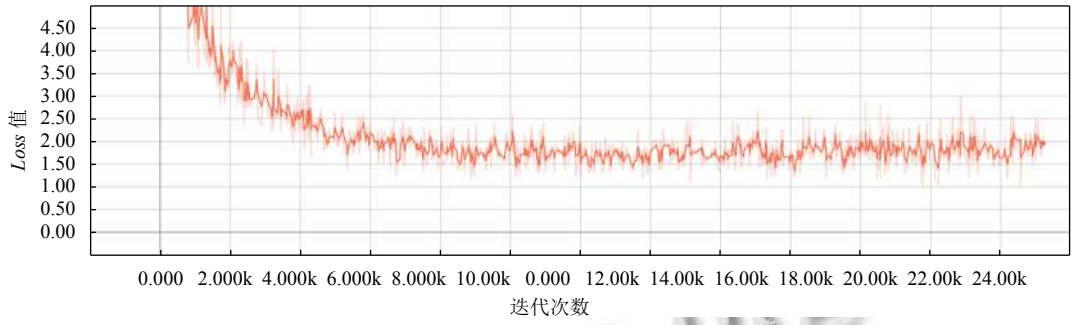

图 6 IndRNN-CRF 的 loss-step 曲线图

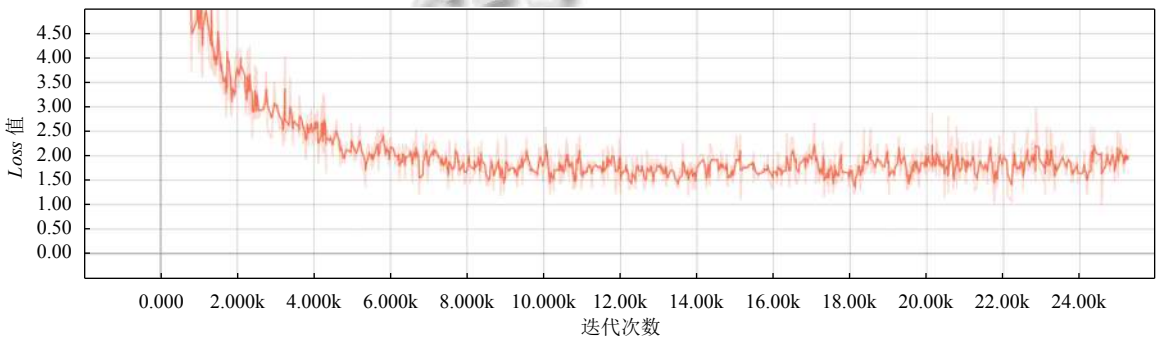

图 7 IDCNN-BILSTM-CRF 的 loss-step 曲线

参考文献

1 Sundheim BM. Named entity task definition, version 2.1. Proceedings of the Sixth Message Understanding Conference. Columbia, MA, USA. 1995. 319–332.

2 Huang ZH, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv: 1508.01991, 2015.

3 苏娅, 刘杰, 黄亚楼. 在线医疗文本中的实体识别研究. 北京大学学报 (自然科学版), 2016, 52(1): 1–9.

4 张帆, 王敏. 基于深度学习的医疗命名实体识别. 计算技术与自动化, 2017, 36(1): 123–127. [doi: 10.3969/j.issn.1003-6199.2017.01.025]

5 Li S, Li WQ, Cook C, *et al*. Independently recurrent neural network (IndRNN): Building a longer and deeper RNN. arXiv preprint arXiv: 1803.04831, 2018.

6 Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv: 1511.07122, 2016.

7 Strubell E, Verga P, Belanger D, *et al*. Fast and accurate entity recognition with iterated dilated convolutions. Proceedings of 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark. 2017. 2670–2680.

8 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. 自动化学报, 2014, 40(8): 1537–1562.

9 Mikolov T, Chen K, Corrado G, *et al*. Efficient estimation of word representations in vector space. ICLR: Proceeding of the International Conference on Learning Representations Workshop Track. AZ, USA. 2013. 1301–3781.

10 Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, NV, USA. 2013. 3111–3119.

11 Kenter T, Borisov A, de Rijke M. Siamese CBOW: Optimizing word embeddings for sentence representations. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany. 2016. 941–951.