文章编号:1672－7010(2020)01－0021－06

# 基于双层 Bi-LSTM-CRF 模型的糖尿病领域命名实体识别

何春辉[1],王梦贤[2],何小波[3]

(1.湘潭大学 数学与计算科学学院,湖南 湘潭,411105;
2.湖南城市学院 管理学院,湖南 益阳,413000;
3.75841 部队,湖南 长沙,410000)

**摘　要**:随着信息技术的发展,电子文档在糖尿病领域的信息记录中得到了大量应用,通过自动化技术对这些电子文档进行分析具有重大的意义。由于现有的命名实体识别方法在糖尿病领域中识别准确率偏低。为了改变这种现状,提出了双层的双向长短时记忆神经网络条件随机场模型(Bi-LSTM-CRF),并将其应用到糖尿病领域命名实体识别任务中。实验结果表明该模型在包含 15 种实体类别的数据集上准确率达到了 89.14%,且在外部测试集上平均 $F_1$ 值为 72.89%,充分揭示了双层 Bi-LSTM-CRF 模型的有效性。

**关键词**:糖尿病;命名实体识别;字符嵌入;Bi-LSTM-CRF
**中图分类号**:TP29;TP391.1;R331　　**文献标志码**:A

# Named entity recognition in the field of diabetes based on double-layer Bi-LSTM-CRF model

HE Chunhui[1], WANG Mengxian[2], HE Xiaobo[3]

(1.School of Mathematics and Computational Sciences, Xiangtan University, Xiangtan 411105, China;
2.College of Management, Hunan City University, Yiyang 413000, China;
3.Troop of 75841, Changsha 410000, China)

**Abstract**: With the development of information technology, electronic documents have been widely used in the information record of diabetes. Analysis of these electronic documents through automation technology has a great significance. Due to the low accuracy of existing named entity recognition methods in the field of diabetes, a double-layer bidirectional long-short-term memory neural network conditional random field model (Bi-LSTM-CRF) was proposed and applied to the task of named entity recognition in the field of diabetes. Experimental results show that the accuracy of the model is 89.14% on a dataset containing 15 entity categories, and the average $F_1$ score on the external test dataset is 72.

**收稿日期**:2019－03－20
**基金项目**:湖南省教育厅科研项目(17C0293)
**作者简介**:何春辉,男,数据挖掘工程师,硕士,主要从事数据挖掘与信息处理方面的研究;E-mail:xtuhch@163.com

89%, which fully reveals the effectiveness of the double-layer Bi-LSTM-CRF model.

**Key words**: diabetes; NER; character-embedding; Bi-LSTM-CRF

从医学角度来看,糖尿病主要有 1 型糖尿病和 2 型糖尿病。据相关统计数据显示,在所有糖尿病患者中有超过 90%的患者属于 2 型糖尿病[1]。无论是哪种类型,都会对患者的正常生活带来极大的影响。近年来,公开的研究数据表明中国糖尿病患者人数位居全球第一,且仍在快速增长。国际糖尿病联合会公布的调查数据显示,2040 年中国糖尿病患者数预计会超过 1.5 亿[2]。对于这些糖尿病患者而言,大多数医院都会通过电子文档的形式对他们的病情进行记录和存档。针对这些电子文档,目前缺乏高效的自动识别方法对命名实体进行快速识别。为了改变这种现状,提升糖尿病领域的命名实体识别准确率,为人们做出辅助决策提供数据支撑,已成为糖尿病领域的一个重要的研究课题[3]。在其他领域中,命名实体识别已经取得了许多高质量研究成果[4]。但是国内外在糖尿病领域的命名实体识别方法研究还处于起步阶段。因此,提出高效的糖尿病领域命名实体识别方法,这对糖尿病领域电子病历的自动分析而言具有深远的意义。有研究成果表明,基于传统机器学习和规则[5-7]相结合的条件随机场 CRF[8-9]和结合深度学习的条件随机场 CRF[10-13]是中文命名实体识别领域性能非常高的两类方法[14]。下面引入某一段真实的糖尿病临床指南样例原文:"自从 2005 年国际上第一个肠促胰素药物上市以来,此类药物的研究和临床应用有了飞速的发展。目前在我国已经上市 2 种胰高血糖素样肽 1(GLP-1)受体激动剂和 5 种二肽基肽酶Ⅳ(DPP-4)抑制剂。"上述样例包含了糖尿病领域中部分药物类实体名称。如何应用智能算法从类似的文本中自动识别糖尿病领域命名实体是本文的研究重点。为了将智能方法应用到糖尿病领域从而解决糖尿病领域的命名实体识别任务,分别对经典的 CRF 和基于深度学习的 CRF 这两类方法在公开的糖尿病领域标注数据集上开展实证研究。

## 1 双层 Bi-LSTM-CRF 模型

由于深度学习的框架可以对文本特征进行自动抽取,从而无需使用传统的特征工程来做特征筛选和建模。因此,在文本分类和序列标注问题中,深度学习模型通常会具有更好的性能。图 1 展示了基于字符嵌入的双层 Bi-LSTM-CRF 模型的体系结构。

图 1 中的体系结构从下至上总共分为 7 个层次。首先是输入层,图中以[胰岛素治疗]这个中文字符串当作模型的输入序列;然后达到字分割层(按照字符进行切分),得到[胰][岛][素][治][疗]这 5 个不同的字符,这些字符分别调用预先训练好的字符嵌入向量映射成对应的向量形式;将相应的向量依次输入到第一层深度学习网络的前向传播和反向传播网络中进行参数训练;将计算结果继续依次输入到第二层深度学习网络的前向传播和反向传播网络中进行参数训练;将计算结果输入到条件随机场 CRF 层,完成相应的序列标注;最后,在输出层输出最终的序列标注结果。需注意的是,在深度学习网络中,采用双层 Bi-LSTM网络结构。实验中,模型的主要超参数设置如表 1 所示。

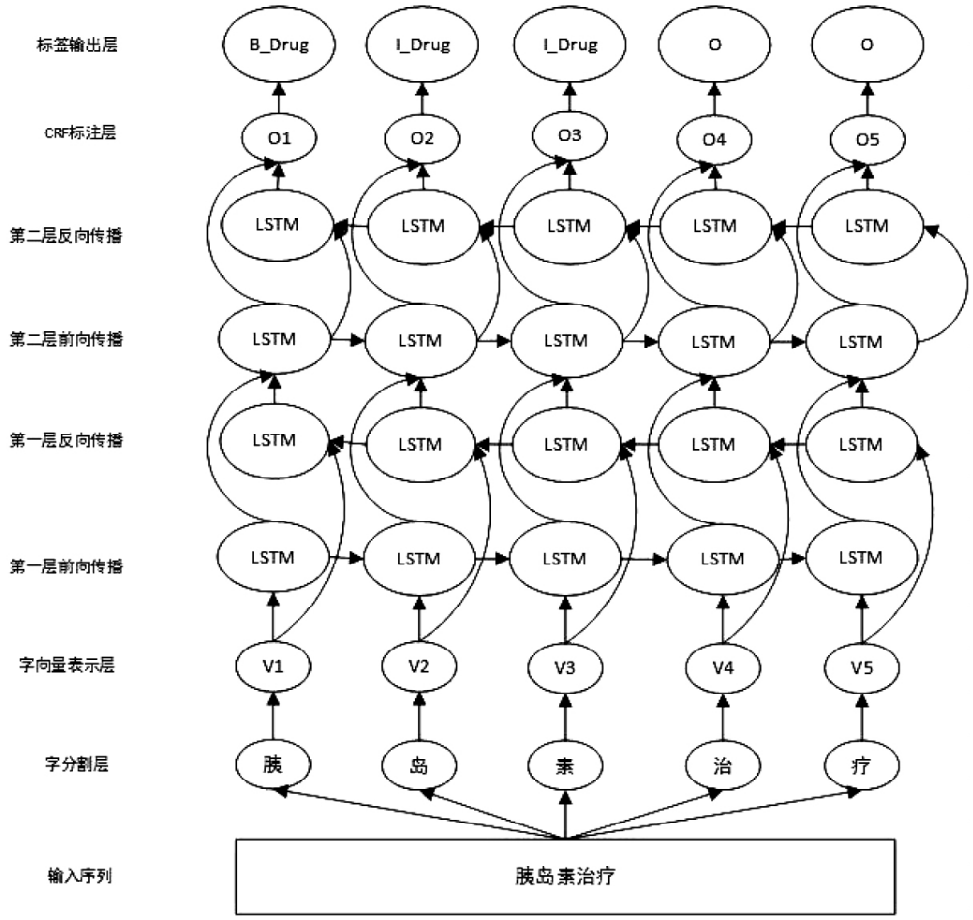


**图 1　基于字符嵌入的双层 Bi-LSTM-CRF 模型体系结构**

**Fig.1**　Double-layer Bi-LSTM-CRF model architecture based on character embedding

**表 1　双层的 Bi-LSTM-CRF 模型核心超参数设置**

**Table 1**　Core hyperparameters setting for the double-layer Bi-LSTM-CRF model

| 参数名称 | 参数取值 | 参数名称 | 参数取值 |
|---|---|---|---|
| EMBED_DIM(向量维度) | 300 | 第一层 BiLSTM_UNITS | 500 |
| DROPOUT_RATE | 0.25 | 第二层 BiLSTM_UNITS | 250 |
| Batch_Size | 128 | Epoch | 15 |
| Loss_function | crf_layer.loss_function | Optimizer function | Adam |

注：表中所列的超参数及取值情况，是指实验中性能最优的模型对应的参数取值。

## 2　数据集及预处理

为验证模型的性能，实验数据集由阿里云天池大赛平台提供的瑞金医院 MMC 人工智能辅助构建知识图谱大赛[15]所提供的糖尿病相关标注文档集构成，并按照 80%训练和 20%验证的方式来划分数据集。此外，选取竞赛第一轮 test_B 中公布的 58 份预测文档作为模型的独立测试集。所有的数据集标注工作都是基于 brat[16]软件完成。标注信息包含了以 T 开头的实体标记，后接实体序号，实体类别，起始和结束位置以及实体在文档中所对应的原始内容五个部分。整个数据集总共包含了 15 种不同的实体类别，具体类别名称如表 2 所示。

表 2 实体类别与对应的符号表示

**Table 2** Entity categories and corresponding symbolic representations

| 实体类别 | 标识符 | 案例 | 实体类别 | 符号表示 | 案例 |
|---|---|---|---|---|---|
| 疾病名称 | Disease | 糖尿病 | 病因 | Reason | 胰岛素抵抗 |
| 临床表现 | Symptom | 血脂紊乱 | 检查方法 | Test | 总胆固醇 |
| 检查指标值 | Test Value | 下降 5% | 药品名称 | Drug | 视黄醇 |
| 用药频率 | Frequency | 1 日 1 次 | 用药剂量 | Amount | 500 μmol/L |
| 用药方法 | Method | 注射 | 非药治疗 | Treatment | 血液透析 |
| 手术 | Operation | 甲状腺手术 | 部位 | Anatomy | 肾脏 |
| 不良反应 | SideEff | 输液反应 | 程度 | level | 轻度 |
| 持续时间 | Duration | 6 个月后 | — | — | — |

所有实体对应的类别皆采用 BIO 三级标注体系按实体序列出现的先后顺序来完成实体标注(B 表示实体的头部,I 表示实体中间及尾部,O 表示非实体),并以标点符号作为字符串切分边界。

为了减少中文分词所带来的误差,在训练数据集中采用频率大于 2 的字来构建相应的字向量,它可以减少因分词不准确而引起的误识别现象。为了进一步提升识别的准确性,在数据预处理阶段对文档内容中的特殊字符、空格等做了处理。

## 3 实验结果及分析

为验证和测试模型的性能,在第 2 节所述的数据集上,实现了多组不同模型的对比实验,从而完成模型的对比验证和对新数据的预测。在验证数据集上,采用所有类别的平均准确率来评估模型性能,在独立的测试数据集上,采用所有类别的平均 $F_1$ 值来作为评测指标。

在文本集合里,令 $D=\{d_1,d_2,\cdots,dn\}$,其中 $di=<wi_1,wi_2,\cdots,\mathrm{win}>$,为第 i 篇文档。预定义实体类别:$C=\{c_1,c_2\cdots cm\}$,实体和所属类别对的集合:$\{<m_1,c_2>,<m_2,c_2>,\cdots,<m_p,c_p>\}$,其中 $m_i=<d_i,b_i,e_i>$ 是出现在文档 $d_i$ 中的实体,$b_i$ 和 $e_i$ 分别表示 $m_i$ 在 $d_i$ 中的起止位置,$C_{\mathrm{mi}}$ 表示实体 $\mathrm{m}_i$ 的预定义类别,要求实体之间不重叠,即 $e_i<b_i+1$。

采用 $F_1$ 值作为评测模型的指标。模型的输出结果集合记为 $S=\{s_1,s_2,\cdots,s_m\}$,人工标注的结果集合记为 $G=\{g_1,g_2,\cdots,g_n\}$。集合元素为 1 个实体,表示为四元组$<d,\mathrm{pos_b},\mathrm{pos_e},C>$,$d$ 表示文档,$\mathrm{pos_b}$和 $\mathrm{pos_e}$分别对应实体在文档 $d$ 中的起止下标,$C$ 表示实体所属预定义类别。按照如下指标进行评价。定义 $s_i\in S$ 与 $g_j\in G$ 等价,当且仅当:$s_i.d=g_j.d$,$s_i.pos_b=g_j.pos_b$,$s_i.pos_e=g_j.pos_e$,$s_i.C=g_j.C$。基于以上等价关系,取集合 $S$ 与 $G$ 的严格交集并结合所有预定义类别 $Ci$ 中预测正确的数量 $TP(C_i)$ 来进行评测。相关评测指标计算公式见式(1)和式(2):

$$P_S=\frac{S\cap G}{|S|},R_S=\frac{S\cap G}{|G|},F_{1S}=\frac{2P_SR_S}{P_S+R_S} \tag{1}$$

$$\mathrm{val\_Accu}=\frac{\sum_{i=1}^{15}TP(C_i)}{Total\_number\_of\_val\_Samples} \tag{2}$$

在对应的数据集中，分别使用了传统的条件随机场 CRF 模型、结合了深度学习的单层 Bi-LSTM-CRF、双层 Bi-LSTM-CRF 和双层 GRU-CRF 这 4 种不同的模型进行实验。相应的实验结果如图 2 所示。

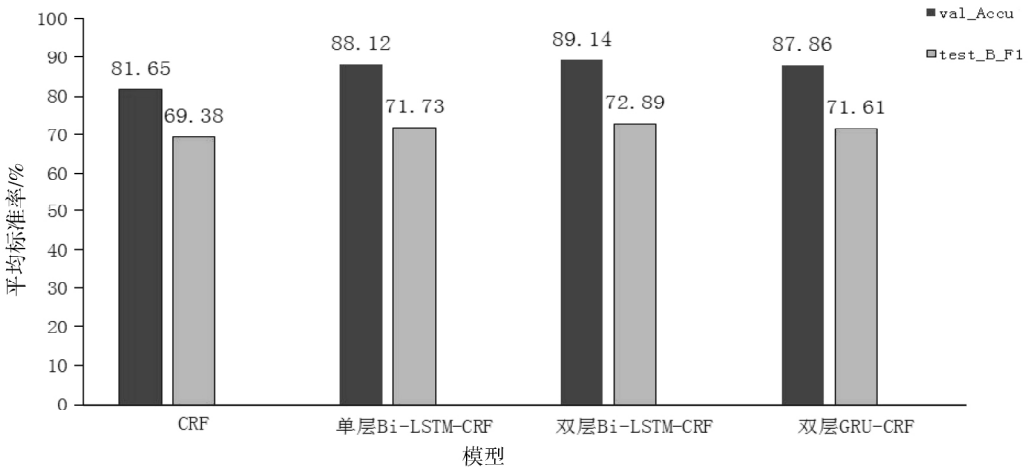


**图 2　不同模型的实验结果**

**Fig.2**　Experimental results of different models

图 2 展示了 4 个不同的模型在验证集和独立测试集 test_B 上的实验结果。训练过程中，从结果中可以看出 CRF 模型效果一般，只有 81.65%的平均准确率，双层 Bi-LSTM-CRF 模型取得了 89.14%的平均准确率，单层 Bi-LSTM-CRF 和双层 GRU-CRF 模型平均准确率分别是 88.12%和 87.86%。在预测过程中，双层 Bi-LSTM-CRF 模型预测的 $F_1$ 值达到了 72.89%，而其他模型的 $F_1$ 值处于 69.38%到 71.73%之间。由图 2 结果可知：双层 Bi-LSTM-CRF 模型相对基线模型而言，在实验中的平均准确率和 $F_1$ 均属最高。此外，为了进一步验证模型的泛化能力，调用已经训练好的双层 Bi-LSTM-CRF 模型，对引言中所述的样例片段内容进行预测，得出的预测结果见表 3。

**表 3　文中所述样例片段所对应的实体识别结果**

**Table 3**　Entity recognition results corresponding to the sample fragments described in this paper

| 实体编号 | 实体类型 | 开始下标 | 结束下标 | 实体内容 |
|---|---|---|---|---|
| T1 | Drug | 13 | 17 | 肠促胰素 |
| T2 | Drug | 55 | 75 | 胰高血糖素样肽 1(GLP－1)受体激动剂 |
| T3 | Drug | 78 | 94 | 二肽基肽酶Ⅳ(DPP－4)抑制剂 |

表 3 所示的命名实体识别结果中共包含了 3 个不同的实体，分别标记为 T1，T2 和 T3。其中每个实体对应 5 列，分别为实体标签、实体类别、实体在原始文档中的起始下标位置、结束下标位置、实体内容。根据表 3 的实验结果可以看出 T1，T2 和 T3 都是属于药物类的实体，这与真实的标注结果相吻合，这也进一步揭示了双层 Bi-LSTM-CRF 模型的优越性。

## 4　结论与展望

在糖尿病领域命名实体识别任务中，鉴于目前缺乏成熟的自动化技术来支撑实体识别任务；提出了基于双层 Bi-LSTM-CRF 模型来识别糖尿病领域的命名实体。该模型在实验数据集上取得了较好的效果，它的平均准确率达到了 89.14%，在外部测试集上的 $F_1$ 值为

72.89%。提出的双层Bi-LSTM-CRF命名实体识别模型,在公开的糖尿病领域实体识别数据集上已经取得了较高的识别准确率,未来会进一步提升模型的性能,并将改进后的双层Bi-LSTM-CRF模型应用于大规模糖尿病领域命名实体识别任务,从而形成结构化知识为辅助决策提供数据支持。

参考文献:

[1]刘雪莲,李春卉.糖尿病痛苦对社区老年糖尿病患者生活质量的影响研究[J].吉林医药学院学报,2018,39(1):14-16.

[2]唐咸玉,曾慧妍,何柳,等.近20年中医药防治肥胖2型糖尿病研究趋势可视化分析[J].中国中医药信息杂志,2018,25(4):96-101.

[3]郁小玲,张铁山,吴彤,等.基于两位一体的中文电子病历命名实体识别[J].中国卫生信息管理杂志,2017,14(4):552-556.

[4]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.

[5]栗伟,赵大哲,李博,等.CRF与规则相结合的医学病历实体识别[J].计算机应用研究,2015,32(4):1082-1086.

[6]何炎祥,罗楚威,胡彬尧.基于CRF和规则相结合的地理命名实体识别方法[J].计算机应用与软件,2015,32(1):179-185,202.

[7]翟菊叶,陈春燕,张钰,等.基于CRF与规则相结合的中文电子病历命名实体识别研究[J].包头医学院学报,2017,33(11):124-125,130.

[8]史海峰.基于CRF的中文命名实体识别研究[D].苏州:苏州大学,2010.

[9]叶枫,陈莺莺,周根贵,等.电子病历中命名实体的智能识别[J].中国生物医学工程学报,2011,30(2):256-262.

[10]YING Q,YING F Z.Research of clinical named entity recognition based on Bi-LSTM-CRF[J].Journal of Shanghai Jiaotong University(Science),2018,23(3):392-397.

[11]MA X,HOVY E.End-to-end sequence labeling via bi-directional lstm-cnns-crf[J].arXiv preprint arXiv:1603.01354,2016.

[12]DONG X,CH S,QIAN L,et al.Transfer bi-directional LSTM RNN for named entity recognition in Chinese electronic medical records[C]//IEEE International Conference on E-health Networking,IEEE,2017:1-4.

[13]LYU C,CHEN B,REN Y,et al.Long short-term memory RNN for biomedical named entity recognition[J].BMC Bioinformatics,2017,18(1):462.

[14]曲春燕.中文电子病历命名实体识别研究[D].哈尔滨:哈尔滨工业大学,2015.

[15]ALIBABAl CLOUD.A libaba Cloud Labeled Chinese Dataset for diabetes[EB/OL].https://tianchi.aliyun.com/dataset/dataDetail?dataId=22288,2019-03-15/2019-03-20.

[16]STENETORP P,PYYSALO S,TOPIć G,et al.BRAT:a web-based tool for NLP-assisted text annotation[C]//Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics.Association for Computational Linguistics,2012:102-107.