融合注意力机制的电子病历命名实体识别

陈 琛,刘小云,方玉华
(厦门医学院信息中心,福建 厦门 361023)

摘 要:命名实体识别是自然语言处理中的一项基础性关键任务,基于电子病历命名实体识别是临床决策支持和医疗知识图谱构建等任务的基础。针对传统的双向长短时记忆神经网络(bi-directional long short-term memory,BiLSTM)结合条件随机场(conditional random field,CRF)的 BiLSTM-CRF 模型在处理医疗文本命名实体识别问题时面临的文本特征提取不够充分和未登录词不能充分识别等问题,引入注意力机制(attention mechanisms),提出一种基于注意力机制的 BiLSTM-CRF 命名实体识别模型。该模型以字向量作为神经网络的输入,BiLSTM 层建模上下文信息,捕捉双向的语义依赖;ATTENTION 层重点关注输入数据中显著的与当前输出相关的特征,抑制无用信息;CRF 层充分考虑了句子级别的标签依赖信息,对整个句子进行解码预测输出。实验结果表明,在电子病历的命名实体识别中,该模型较传统模型提升了一定的识别效果。

关键词:命名实体识别;注意力机制;电子病历;双向长短期记忆神经网络;条件随机场
中图分类号:TP391.1 **文献标识码**:A **文章编号**:1673-629X(2020)10-0216-05
doi:10.3969/j.issn.1673-629X.2020.10.038

Named Entity Recognition in Electronic Medical Record Introducing Attention Mechanisms

CHEN Chen,LIU Xiao-yun,FANG Yu-hua
(Information Technology Center,Xiamen Medical College,Xiamen 361023,China)

Abstract:Named entity recognition is a basic key task in natural language processing,and named entity recognition based on electronic medical records is the basis of tasks such as clinical decision support and medical knowledge graph. For the problems that the traditional bi-directional BiLSTM-CRF model combined long-short-term memory(BiLSTM) and conditional random field (CRF) is faced with insufficient extraction of text features and insufficient recognition of unregistered words when dealing with the recognition of medical text naming entities,attention mechanisms are introduced to propose a BILSTM-CRF named entity recognition model based on the attention mechanism. The model takes the word vector as the input of the neural network,modeling context information on the BiLSTM layer to capture the bidirectional semantic dependence. The ATTENTION layer focuses on the salient features of the input data related to the current output and suppresses the useless information. The CRF layer fully considers the label dependency information at the sentence level to decode the prediction output for the whole sentence. The experiment shows that this model is better than the traditional model in the recognition of named entity identification in electronic medical records.

Key words:named entity recognition;attention mechanisms;electronic medical records;BiLSTM;CRF

0 引 言

命名实体识别又称专名识别(named entity recognition,NER),是自然语言处理中的一项基础性关键任务,应用广泛。其一般是指从非结构化文本中识别出有特定意义的实体,通常指人名、地名、组织机构名称等专有名词,为实现关系抽取、自动问答系统、知识图谱等任务做基础。

近年来,随着国内医疗信息化程度的显著提升,医疗领域内大量使用电子病历代替传统医生手写病历,累积了海量的包含患者临床医疗、诊疗、个人信息电子的病历数据。应用自然语言处理、信息抽取等技术对累积的电子病历文本数据进行数据挖掘获取医疗知识进行临床决策支持的研究受到广泛关注[1-2];从电子病历里自动挖掘、自动识别电子病历文本中与患者健康密切相关的各类命名实体以及类型也可为将来进行医疗领域知识图谱构建、医疗问答系统和医疗信息检

收稿日期:2019-12-23 **修回日期**:2020-04-27
基金项目:福建省中青年教师教育科研项目(JAT170697)
作者简介:陈 琛(1984-),女,硕士,讲师,研究方向为医疗大数据、自然语言处理、数据挖掘。

索等诸多自然语言技术处理任务打下良好基础[3]。

传统的命名实体识别方法主要有三种。第一种是基于规则和字典,主要采用语言学专家手工构造规则模板,这类方法依赖专家、规则定义复杂、系统可移植性差。第二种是基于统计的学习方法,如支持向量机(SVM)[4]、隐马尔可夫模型(HMM)[5]、最大熵(maximum entropy)[6]和条件随机场(CRF)[7]等。此类方法特征工程复杂,需要大量的标注数据。第三种方法是基于神经网络的深度学习方法,此类方法不依赖特征模板,数据驱动,具有较好的泛化性。如 Collobert[8]最早利用深度神经网络进行命名实体研究;Lample[9]使用双向长短时记忆神经网络(bi-directional long short-term memory,BiLSTM)结合条件随机场(conditional random field,CRF)的 BiLSTM-CRF 模型在命名实体识别中获得了较好结果,并成为学界主流使用的进行命名实体识别的模型。BiLSTM-CRF 模型虽然能考虑到上下文信息,但并没有考虑到不同词语、字符在句子中的重要性不同,识别结果仍有进一步提升的空间。

文中提出在主流的 BiLSTM-CRF 模型中引入注意力(attention)机制,建立了一个基于注意力机制的 BiLSTM-CRF 的命名实体模型,将其命名为 ATTENTION-BiLSTM-CRF 模型,应用在医疗领域的命名实体识别中。注意力机制模仿人类的注意力机制,重点关注有效信息,提升文本命名实体识别的 F_1 值,在实验中获得了更好的结果。该模型无需特征工程,可以达到较 BiLSTM-CRF 模型更好的识别效果。

1　基于 ATTENTION-BiLSTM-CRF 的命名实体模型

该模型由输入层、BiLSTM 层、注意力机制层和 CRF 层构成,整体架构见图 1。

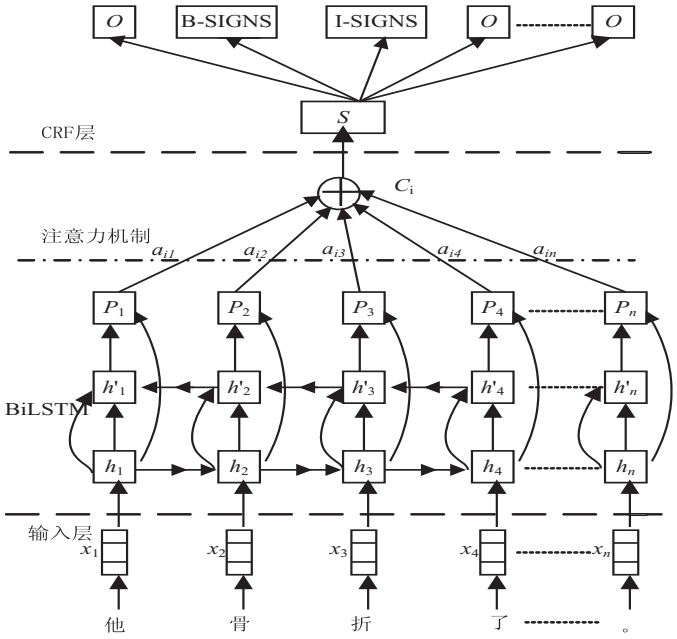

图 1　基于 ATTENTION-BiLSTM-CRF 模型整体架构

1.1　输入层

输入的句子 x 包括 n 个字符,使用随机初始化的嵌入矩阵将查找向量表后获得的 one-hot 向量映射为低维稠密的字向量作为文中模型的输入。

$$\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n) \tag{1}$$

其中,$x_i \in \mathbb{R}^d$,表示 i 时刻神经网络的输入向量,d 为维度。

1.2　双向长短期记忆神经网络(BiLSTM)层

(1)长短时记忆神经网络(LSTM)。

长短期记忆神经网络[10]模型是 RNN 的一种改进模型,通过输入门、遗忘门和输出门三个门结构概念,通过门控状态控制传输,忘记不重要信息,保留需长时间记忆的信息,整合后在当前状态下产生输出状态。实现了可以长期记忆一个状态,解决了长距离依赖问题,如图 2 所示[11]。

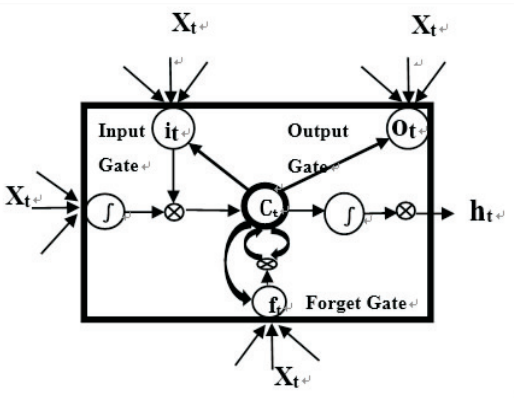

图 2　LSTM 单元内部结构

公式如下：

遗忘门：

$$f_t = \sigma(W_{hf}h_{t-1} + W_{xf}x_t + W_{cf}c_{t-1} + b_f) \tag{2}$$

输入门：

$$i_t = \sigma(W_{hi}h_{t-1} + W_{xi}x_t + W_{ci}c_{t-1} + b_i) \tag{3}$$

当前输入的单元状态 $\tilde{c}_t$，根据上一次的输出和本次输入来计算：

$$\tilde{c}_t = \tanh(W_{hc}h_{t-1} + W_{xc}x_t + b_c) \tag{4}$$

当前时刻的单元状态 c_t 为：

$$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t \tag{5}$$

输出门，控制了长期记忆对当前输出的影响：

$$o_t = \sigma(W_{ho}h_{t-1} + W_{xo}x_t + W_{co}c_t + b_o) \tag{6}$$

LSTM 单元最终的输出由输出门和单元状态共同确定：

$$h_t = o_t \circ \tanh(c_t) \tag{7}$$

其中，符号∘表示按点乘运算，$W_{hf}, W_{xf}, W_{cf}, W_{hi}, W_{xi}, W_{ci}, W_{hc}, W_{xc}, W_{ho}, W_{xo}, W_{co}$ 分别是权重，b_f, b_i, b_c, b_o 为偏置项，σ 表示 sigmoid 函数。

(2)双向长短期记忆神经网络(BiLSTM)[12]。

模型中针对单层 LSTM 模型只能获得过去时刻的信息的问题，使用双向 LSTM 模型，获得过去和未来时刻的信息。BiLSTM 模型[12]同一时刻包含两个分别按前向和后向顺序进行记忆的记忆单元(LSTM unit)，最后将该时刻两个方向的输出进行拼接，即：

$$h_t = [\overrightarrow{h}_t; \overleftarrow{h}_t] \in \mathbb{R}^m \tag{8}$$

将结果从 m 维映射到 k 维，k 为标注集的标签数，从而得到 BiLSTM 层输出结果矩阵 $\boldsymbol{P} = (\boldsymbol{P}_1, \boldsymbol{P}_2, \cdots, \boldsymbol{P}_n) \in \mathbb{R}^{n \times k}$。

1.3 注意力机制

双向长短时记忆网络在计算过程中，已将所有上下文信息考虑在内，取得不错的识别效果。文中引入注意力机制[13]，参考人类对注意力焦点的处理方式，使模型更专注于找到输入数据中需要关注的目标信息和与当前输出相关信息，抑制无用信息，提高输出的质量和效率。为了使输出更为准确，利用注意力机制为 BiLSTM 层的输出分配不同的权重，新的输出向量则是由各特征向量与对应权重的乘积相加后获得。对于 i 时刻的模型输出向量，模型利用注意力权重分布向量对编码的源序列的隐藏层输出进行加权求和计算，得到针对当前输出的源序列编码结果，公式如下：

$$\boldsymbol{c}_i = \sum_{j=1}^{n} a_{ij}\boldsymbol{P}_j \tag{9}$$

其中，$\boldsymbol{c}_i$ 表示利用注意力机制输出新的字特征向量，它是由前序模型输出的各特征向量 $\boldsymbol{P}_j$ 与对应权重 a_{ij} 的乘积和计算得到。a_{ij} 由前一时刻字特征向量 $\boldsymbol{c}_{i-1}$ 与 $\boldsymbol{P}_j$ 通过式(10)和式(11)计算得出。ATTENTION 层即对所有时刻的输出乘上对应的权重相加作为最终输出。随后设置 dropout[14]，避免深度神经网络训练小数据集时产生的过拟合问题。

$$a_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{n}\exp(e_{ik})} \tag{10}$$

$$e_{ij} = v_a \tanh(w_a c_{i-1} + w_b P_j) \tag{11}$$

其中，v_a, w_a, w_b 为权重。

1.4 条件随机场层

该层进行句子级别的序列标注。CRF 模型[7]在标注过程中可以利用句子级别的标签之间的依赖信息，进而预测标签与标签之间的关系。文中定义一个状态转移矩阵 $\boldsymbol{A}$ 作为参数随模型一起训练，A_{ij} 表示的是从第 i 个标签转移到第 j 个标签的转移概率。设待预测的标签序列为 $y = (y_1, y_2, \cdots, y_n)$，则模型对于序列 y 的预测概率由注意力层输出的字特征向量 $\boldsymbol{c}_i$ 和 CRF 的参数矩阵 $\boldsymbol{A}$ 共同决定，为各个位置的概率之和，公式为：

$$S(x, y) = \sum_{i=1}^{n} c_{i,y_i} + \sum_{i=1}^{n+1} A_{y_{i-1}, y_i} \tag{12}$$

使用 Softmax 函数进行归一化计算得到真实标签序列 $y = (y_1, y_2, \cdots, y_n)$ 的概率为：

$$P(y \mid x) = \frac{\exp(S(x, y))}{\sum_{y'}\exp(S(x, y'))} \tag{13}$$

使用 Adam[15]来训练文中模型参数。在预测时对输入的每个句子使用维特比算法(Viterbi algorithm)[16]解码来得到使目标函数最大化的最佳标签序列。

$$y^* = \arg\max_{y'} S(x, y') \tag{14}$$

2 实验结果和分析

2.1 实验数据及标注方式

文中在 CCKS2017Task2 公开数据集上进行了实验，取数据中的 80% 作为训练集，10% 作为验证集，10% 作为测试集。

该数据集为 1 200 份经人工标注数据，人工标注的标签共计 29 866 个，其中身体部位10 719个，约占 36%；检查和检验共计 9 546 个，约占 32%；症状和体征 7 831 个，约占 26%；治疗 1 048 个，约占 4%；疾病和诊断 722 个，约占 2%。

使用 BIO 标注方式对语料中的字符进行标注。即 B-、I-代表实体首字、实体非首字，O 代表该字不属于命名实体的一部分。在此基础上，为了将实体进行分类，还在实体标注的时候以"-type"形式添加类别。在数据集中标注如表 1 所示。

表 1　实体标注

标注	意义
B-CHECK/ I-CHECK	检查和检验首字/非检查和检验首字部分
B-SIGNS/ I-SIGNS	症状和体征首字/非症状和体征首字部分
B-DISEASE/ I-DISEASE	疾病和诊断首字/非疾病和诊断首字部分
B-TREATMENT/I-TREATMENT	治疗首字/非治疗首字部分
B-BODY/ I-BODY	身体部位首字/非身体部位首字部分
O	非实体

2.2　评价指标

模型的实体识别效果使用3 个指标,准确率 P 、召回率 R 和 F1 值来评价。公式如下:

$$P = \frac{\text{模型正确识别出的实体的个数}}{\text{模型预测是实体的总个数}} \times 100\% \quad (15)$$

$$R = \frac{\text{模型真正识别出的实体的个数}}{\text{模型识别出的所有实体的总个数}} \times 100\% \quad (16)$$

$$\text{F1} = \frac{2PR}{P + R} \quad (17)$$

2.3　实验环境和超参数设置

实验的环境为 Windows10,显卡为 GTX1080Ti,内存为16 G,Python 版本为 Python3.7.0,TensorFlow 版本为 Tensorflow1.4.0。

经过多次实验后,模型表现最好的超参数设置如表2 所示。

表 2　ATTENTION-BiLSTM-CRF 模型的超参数设置

参数	值
学习率	0.01
向量维数	300
Dropout	0.5
Batch_size	32
隐藏层节点数	128
Epoch	100

2.4　实验结果

该模型在 CCKS2017Task2 语料库上做了4 组对照实验,比较了与 ATTENTION-LSTM-CRF 模型、BiLSTM-CRF 模型和 LSTM-CRF 模型分别进行命名实体识别的效果。

结果如表3 所示。

表 3　命名实体类别模型效果总体比较　%

Model	P	R	F1
ATTENTION-BiLSTM-CRF	89.48	92.25	90.84
ATTENTION-LSTM-CRF	89.43	88.74	89.08
BiLSTM-CRF	87.78	91.70	89.70
LSTM-CRF	87.14	88.19	87.66

其中,对各类别实体识别效果如表4 所示。

2.5　分析与讨论

由表3 可以看出,引入 ATTENTION 机制后,F 值较传统 BiLSTM-CRF 模型所获得的 F 值提高 1.14%,可见加入 ATTENTION 层能够有效选择更有价值的样本,提高模型性能。ATTENTION-BiLSTM-CRF 模型的结果相较 ATTENTION-LSTM-CRF 模型提升了1.76%,BiLSTM-CRF 较 LSTM-CRF 提升 2.04%,说明双向获取全面信息的 BiLSTM 模型较单向 LSTM 模型能获得更好的识别效果。

表 4　各类别实体识别的准确率、召回率、F1 值　%

Model	实体类型	P	R	F1
ATTENTION-BiLSTM-CRF	检查和检验	92.83	92.74	92.79
	身体部位	85.61	92.32	88.84
	症状和体征	96.76	97.55	97.15
	治疗	95.81	94	94.9
	疾病和诊断	66.23	77.38	71.37
ATTENTION-LSTM-CRF	检查和检验	93.97	83.28	88.3
	身体部位	85.11	87.23	86.15
	症状和体征	97.02	95.59	96.3
	治疗	91.79	91.95	91.87
	疾病和诊断	69.37	74.81	71.99
BiLSTM-CRF	检查和检验	93.27	91.8	92.53
	身体部位	86.63	90	88.28
	症状和体征	96.33	97.43	96.87
	治疗	94.33	95.68	95
	疾病和诊断	56.51	75.84	64.76
LSTM-CRF	检查和检验	92.09	83.82	87.76
	身体部位	87.2	86.14	86.67
	症状和体征	96.25	95.26	95.75
	治疗	91.45	90.2	90.82
	疾病和诊断	56.46	74.16	64.11

由表4 可以观察出,加入 ATTENTION 机制后,普遍提升了各类别实体的召回率(R 值),即说该模型较其他模型能够获取更多信息,提高了查全率,即更好地处理了未登陆词问题。对于训练数据量明显过少的疾

病和诊断部分(仅占 2%),注意到在该类型实体存在未登陆词较多,且多为长词可能存在实体嵌套等情况下,加入 ATTENTION 机制后明显提升准确率、识别率和召回值,说明该模型在训练数据较少的情况下仍能显著提高模型识别效果,ATTENTION 机制有利于去除噪音,凸显重要信息,提升模型识别效果。但需要进一步提升仍然需要增加语料或者改进模型,进一步挖掘语义之间的关系。

3 结束语

从实验结果来看,提出的 ATTENTION-BiLSTM-CRF 模型在医疗文本命名实体识别上能显著提升训练数据量较少情况下的实体的识别效果,且在不添加任何人工特征,也无复杂的后续处理的情况下,取得了较当前学界主流 BiLSTM-CRF 模型更好的结果,充分显示了该模型的优越性。深度学习在医疗文本的文本挖掘中仍然有很大的提升空间,Attention 机制对于提升识别效果有一定效果,未来可以考虑引进谷歌提出的 BERT 等模型应用于医疗文本命名实体识别方向等方式,进一步改进模型,提升命名实体识别的效果。

参考文献:

[1] DEMNER-FUSHMAN D, CHAPMAN W W, MCDONALD C J. What can natural language processing do for clinical decision support? [J]. Journal of Biomedical Informatics, 2009, 42(5): 760-772.

[2] WASSERMAN R C. Electronic medical records (EMRs), epidemiology, and epistemology: reflections on EMRs and future pediatric clinical research[J]. Academic Pediatrics, 2011, 11(4): 280-287.

[3] 杨锦锋,关 毅,何 彬,等. 中文电子病历命名实体和实体关系语料库构建[J]. 软件学报, 2016, 27(11): 2725-2746.

[4] LUO G, HUANG X, LIN C, et al. Joint entity recognition and disambiguation[C]//Empirical methods in natural language processing. Lisbon, Portugal: The Association for Computational Linguistics, 2015: 879-888.

[5] PASSOS A, KUMAR V, MCCALLUM A. Lexicon infused phrase embeddings for named entity resolution[C]//Proceedings of the eighteenth conference on computational natural language learning. Ann Arbor, Michigan, USA: Association for Computational Linguistics, 2014: 78-86.

[6] LONG J, ZHANG J, XIANG N, et al. An iterative maximum entropy thresholding algorithm[C]//2016 international conference on cyberworlds (CW). Los Alamitos, Calif: IEEE Computer Society Conference Publishing Services (CPS), 2016: 171-174.

[7] 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.

[8] COLLOBERT R, WESTON J. A unified architecture for natural language processing: deep neural networks with multitask learning[C]//Proceedings of the 25th international conference on machine learning. New York, NY, United States: Association for Computing Machinery, 2008: 160-167.

[9] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition[C]//Proceedings of NAACL 2016. [s. l.]: Association for Computation Linguistics, 2016: 87-98.

[10] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[11] ZHANG S, ZHENG D, HU X, et al. Bidirectional long short-term memory networks for relation classification[C]//29th Pacific Asia conference on language, information and computation. Shanghai: [s. n.], 2015: 73-78.

[12] GRAVES A, SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks, 2005, 18(5-6): 602-610.

[13] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[C]//ICLR. [s. l.]: [s. n.], 2014: 124-135.

[14] HINTON G E, SRIVASTAVA N, KRIZHEVSKY A, et al. Improving neural networks by preventing co-adaptation of feature detectors[C]//CoRR. [s. l.]: [s. n.], 2012: 212-223.

[15] KINGMA D, BA J. Adam: a method for stochastic optimization [C]//3rd international conference on learning representations. San Diego, CA, USA: [s. n.], 2015: 1-15.

[16] BENEDETTO S, BIGLIERI E. Viterbi algorithm[M]//Principles of digital transmission. US: Springer, 2002.