计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(8):173−178 [doi: 10.15888/j.cnki.csa.007609]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 CTD-BLSTM 的医疗领域中文命名实体识别模型①

祝锡永, 吴　炀, 刘　崇

(浙江理工大学 经济管理学院, 杭州 310018)
通讯作者: 吴　炀, E-mail: 379669728@qq.com

摘　要: 为在模型训练期间保留更多信息, 用预训练词向量和微调词向量对双向长短期记忆网络 (Bi-LSTM) 神经模型进行扩展, 并结合协同训练方法来应对医疗文本标注数据缺乏的情况, 构建出改进模型 CTD-BLSTM (Co-Training Double word embedding conditioned Bi-LSTM) 用于医疗领域的中文命名实体识别. 实验表明, 与原始 BLSTM 与 BLSTM-CRF 相比, CTD-BLSTM 模型在语料缺失的情况下具有更高的准确率和召回率, 能够更好地支持医疗领域知识图谱的构建以及知识问答系统的开发.

关键词: 双向长短期记忆网络; 协同训练; 中文命名实体识别; 问答系统; 医疗领域

引用格式: 祝锡永,吴炀,刘崇.基于 CTD-BLSTM 的医疗领域中文命名实体识别模型.计算机系统应用,2020,29(8):173–178. http://www.c-s-a.org.cn/1003-3254/7609.html

Chinese Named Entity Recognition in Medical Field Using CTD-BLSTM Model

ZHU Xi-Yong, WU Yang, LIU Chong

(School of Economics and Management, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: In order to retain more characteristic information in the training process, this study uses pre-training word vector and fine-tuning word vector to extend Bi-directional Long Short-Term Memory network (Bi-LSTM), and combines the co-training semi-supervision method to deal with the feature of sparse annotated text in the medical field. An improved model of Co-Training Double word embedding conditioned Bi-LSTM (CTD-BLSTM) is further proposed for Chinese named entity recognition. Experiments show that compared with the original BLSTM and BLSTM-CRF, the CTD-BLSTM model has higher accuracy and recall rate in the absence of corpora, the proposed method can better support the construction of medical knowledge graph and the development of knowledge answering system.

Key words: Bi-LSTM; co-training; Chinese named entity recognition; question answering system; medical field

随着医疗百科、医疗健康知识问答等平台的发展以及医疗信息管理系统、电子病历档案在医疗单位中的广泛应用, 涌现出大量、多源、冗余并且内容分散的网络医疗数据. 针对这些的海量医疗信息中的数据, 如何高效识别、整合其中的知识实体对搭建医疗知识图谱、提供精准的医疗知识问答和进行医疗知识推理等具有重要的研究意义. 由于大部分的医疗数据为非结构化数据, 这类数据的组织结构虽然能比较方便地表达一些医疗事件和概念, 但在医疗信息统计、知识整合与知识图谱的构建等方面带来很大不便, 因此挖掘和抽取其中的医疗信息成为一项重要任务. 命名实体识别 (Name Entity Recognition, NER) 是一项从给定的一段文本中抽取出诸如人名、物名、机构名等实体的基础技术, 是挖掘和抽取非结构化数据中信息的首要

① 基金项目: 国家自然科学基金 (71501172); 浙江省自然科学基金 (LY15G010010)
Foundation item: National Natural Science Foundation of China (71501172); Natural Science Foundation of Zhejiang Province (LY15G010010)
收稿时间: 2020-01-22; 修改时间: 2020-02-27, 2020-03-24; 采用时间: 2020-04-03; csa 在线出版时间: 2020-07-29

步骤.

医疗领域的中文命名实体识别不同于开放领域的中文命名实体识别, 文本中的实体单元不再是人名、地名等, 而是诸如疾病名、临床症状、药名等实体. 例如, “核磁共振结果显示患者脊柱内有胶质瘤”中“胶质瘤”就是疾病名称; “患者逐渐肌肉萎缩”中“肌肉萎缩”就是临床症状; “静脉注射硝苯吡啶控制”中“硝苯吡啶”就是药品名称. 这些疾病名称、临床症状以及药品名称实体反映了患者的疾病情况以及治疗手段. 医疗领域的中文命名实体识别最主要的任务就是识别文本中的专业术语, 为医疗领域信息的抽取、检索以及构建问答系统等提供支持. 但目前中文医疗标注语料数据缺乏, 并且中文中词与词间没有明显的边界, 相较于英文命名实体识别来说更加困难, 为应对该情况, 本文提出一种面向医疗领域的高效中文命名实体识别模型: 基于协同训练半监督方法的双词向量双向长短期记忆神经网络算法 (Co-Training Double word embedding conditioned Bi-LSTM, CTD-BLSTM).

1 相关研究

早期的命名实体识别主要是针对开放领域, 但近年来, 越来越多的学者关注到了特定领域, 例如医疗领域. 与开放领域的命名实体识别一样, 早期医疗实体识别也是通过领域专家和语言学家手工制定规则的方法, 比如国外著名的电子病历命名实体识别系统 MedKAT (Medical Knowledge Analysis Tool) 和 cTAKES (Clinical Text Analysis and Knowledge Extraction System), 但该类方法的构建需要掌握大量语言学知识, 并且对于不规范语料处理效果也不理想, 具有很大的局限性. 随后, 由于标注语料的增加, 医疗实体识别也渐渐开始运用传统机器学习和词典规则相结合方法的方法. Murugeasan 等[1] 提出的 BCC-NER 方法在 BioCreative II GM 语料库中对基因的识别取得了较好的成果; 张金龙等[2] 则将筛选规则与外部上下文特征结合入 CRF 中, 使得模型对中文医疗机构实体体现了显著的识别效率. 除此之外, Lei JB 等[3] 将中文临床病历作为主要研究对象, 分别采用了最大熵、CRF、SVM、结构化的 SVM 等 4 种方法对医疗实体进行识别, 其中结构化的 SVM 识别效果最佳, 使病程记录和出院小结的 F 值达到 90% 以上.

然而上述方法大多是以机器学习为基础结合词典或规则等构建模型, 需要完成繁重的规则构建、设定特征与标注语料等预先工作. 近年来, 随着深度学习的发展与具有提取自我特征的特性, 出现了大量基于神经网络的模型用于完成命名实体识别任务, 神经网络方法在命名实体识别方面表现出更强的泛化性和更低的特征工程依耐性[4], 减少特征提取所需代价, 有效地提高了模型对实体识别的效率. Liu 等[5] 通过设计多组对照组对 I2B2 数据集中英文电子病历进行实验, 得出 LSTM 模型的识别效果优于 CRF 模型; Almgren 等[6] 提出一种基于字符的深度 Bi-LSTM 的医疗实体识别模型, 其 F 值比经典模型高出 60%; Xu 等[7] 将 Bi-LSTM 与条件随机场进行结合用于医学命名实体识别, 实验结果该 BLSTM-CRF 模型优于两者单一模型. 目前大多数实体识别任务均以 BLSTM-CRF 模型进行, 但医疗领域的中文命名实体任务复杂, 医疗领域标注语料不足和中文分词困难等都会影响模型准确率, 如杨红梅等[8] 利用 BLSTM-CRF 模型对中文电子病历实体进行识别, 其中出院小结的识别率偏低.

总结上述研究及难点, 本文在双向长短期记忆网络的基础上, 通过双词向量扩展 BLSTM 神经网络结构, 并结合协同训练协同学习方法, 最终构建出一种基于协同训练半监督方法的双词向量双向长短期记忆神经网络算法 (CTD-BLSTM), 用以提高中文医疗命名实体的识别效率.

2 基于 CTD-BLSTM 的命名实体识别

2.1 Bi-LSTM 神经网络

自然语言处理 (NLP) 中的命名实体识别是典型的文本序列标准问题, 而循环神经网络 (RNN) 能有效地解决序列标注任务, 但处理长序列数据时可能会发生梯度消失或梯度爆炸等现象, 为此在长短期记忆网络 (LSTM) 中通过引入“门”限制机制解决这一问题. 此外, 由于在实体识别中不仅需要当前词的上文信息, 也需要下文信息, 因此本文选择在能同时包含上下文的信息[9] 的双向长短期记忆网络 (Bi-LSTM) 的基础上提出一种新的模型算法改进. Bi-LSTM 输入层中包含 2 个方向相反且相互独立的 LSTM, 输出层结果包含了 2 个 LSTM 的结果, 这种方式保证了整个 Bi-LSTM 是非循环的, 并且使得整个网络能兼顾上下文信息并自动提取句子特征, 从而获取更好的特征信息.

2.2 双词向量扩展 BLSTM (D-BLSTM)

词向量在 Bi-LSTM 网络中训练的过程中进行不

断微调可以包含更多有用的信息, 但与此同时也会逐渐丢失原本的句法与语义等特征, 针对这种情况, 本文参照双词向量卷积 2E-CNN 模型[10], 利用预训练的词向量与微调后的词向量对神经单元结构进行扩展, 提出一种双词向量 BLSTM 神经网络算法 (Double word vector conditioned BLSTM, D-BLSTM)[11], 将两种词向量同时作为 BLSTM 网络的输入进行训练, 其网络模型结构如图 1 所示.

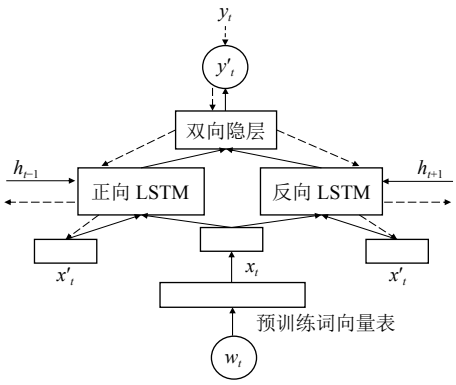

图 1 双词向量扩展的 BLSTM 网络模型图 (D-BLSTM)

通过预训练的词向量表, 对语料中的单词 w_t 进行唯一向量化表示, 将其作为输入词向量 x_t 和 x'_t 的初始值进行训练, 在整个模型的训练过程中保持预训练词向量 x_t 不变, 对 x'_t 不断进行微调, 每个时刻都有 x_t、x'_t 和 h_{t-1} 等 3 个变量输入 LSTM 神经单元, 其单元内输入门、遗忘门、输出门的计算公式见式 (1) 至式 (3), 存储块更新的计算公式见式 (4).

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t, x'_t] + b_i) \quad (1)$$

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t, x'_t] + b_f\right) \quad (2)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t, x'_t] + b_o) \quad (3)$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t, x'_t] + b_c) \quad (4)$$

2.3 融入半监督框架的协同训练

为解决医疗领域标注语料稀缺的问题, 在 D-BLSTM 中引入半监督学习方法, 协同训练 (co-training) 是一种非常广泛使用且有效的半监督学习框架[12], 该方法将大量未标注语料加入到少量已标注语料库中, 并反复训练从而得到识别能力良好的模型, 但必须将语料集自然分割成两个在给定条件下相互独立的特征集, 并且由特征集训练的两个模型都是有效的[13], 训练步骤如下:

(1) 提出两个相互独立特征集, 并对两者的算法模型进行构建;

(2) 分别使用两个特征集中少量已标注语料对模型进行训练;

(3) 将未标注的语料分别输入两个模型中进行标注预测, 并对预测结果进行置信度评价;

(4) 挑选若干置信度 (即对未标注语料赋予正确标记的置信度) 高的样本加入到另一方模型的训练语料库中, 用扩充后的语料库重新对模型进行训练;

(5) 重复上述步骤 (3)、(4), 直至达到设定条件要求.

协同训练半监督学习方法可以在医疗领域已标注语料不充足的情况下大大提高模型实体识别的性能.

2.4 基于协同训练的 D-BLSTM (CTD-BLSTM)

本文在 D-BLSTM 算法中引入协同训练半监督学习方法, 最终构建出一种结合协同训练的双词向量 BLSTM 神经网络算法 (CTD-BLSTM), 其流程如图 2 所示.

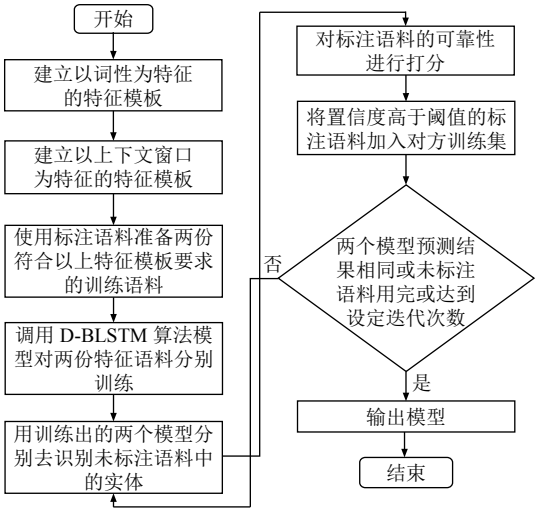

图 2 CTD-BLSTM 算法流程示意图

(1) 根据实体识别任务具体场景, 选择两个独立的特征并制定相应特征模板;

(2) 将采集的文本语料进行少量标注, 并依据两套不同的特征模板整理出训练集;

(3) 用两套语料训练集训练得到两个不同的 D-BLSTM 模型, 并分别用其对各自未标注语料进行标注;

(4) 将标注的预测结果进行置信度评价, 并与设定的阈值进行对比, 若置信度高于阈值, 则将语料加入到

对方的训练语料中, 同时迭代计数加 1;

(5) 循环执行步骤 (3), 直到两个模型对未标注语料标注结果基本一致, 或用尽测试集中未标注语料, 或达到预先设定的循环次数.

3 实验

3.1 实验数据

为验证 CTD-BLSTM 算法对医疗领域实体的识别提取效果, 本文通过爬取如寻世界医疗网、健康大全网、好大夫在线网等医疗网站中的药物信息、疾病信息以及医患问答信息, 经整理后得到上述 3 种信息的有效数据分别为 2665 条、7359 条和 157 090 条.

由于需要将数据以基于词典的形式进行分词标注, 实验过程选取搜狗细胞词库中的“医学词汇大全”作为基础词典, 并通过爬取百度百科中“药理学”、“疾病”、“解剖学”、“传统医学”、“卫生保健”等几个与医疗领域相关标签内的词条来进一步扩充词典, 最终对两者进行对齐、消歧、合并等操作共得 98 647 条词条.

3.2 特征选择

协同训练半监督学习方法需建立两个相互独立的特征集, Chen 等[14] 提取分词、词性等特征完成临床命名实体识别任务, 得到选取分词特征和词性特征作为训练特征识别效果最佳; Peters 等[15,16] 将上下文窗口的特征融入基于预训练的 BLSTM-CRF 语言模型中, 提高已有模型的效果, 因此本文选用上下文窗口特征与词性特征作为两个相互独立的特征构建特征集.

考虑到研究主体为医疗领域文本内的中文实体, 本文参考中文电子病历词性标注规范[17] 将爬取的数据定义为“症状”、“部位”、“药品”、“疾病”、“检查”、“科室”与“治疗”7 类实体, 词性标注见表 1.

表 1 扩展词性标注表

标注代码	代码诠释	例子
nh(名词)	n和“health”的组合	
nhd(疾病名)	nh和“disease”的组合	肠胃炎/nhd
nhs(症状名)	nh和“symptom”的组合	腹痛/nhs
nhi(检查名)	nh和“inspection”的组合	意识模糊/nhi
nht(治疗名)	nh和“treatment”的组合	[低盐/n低脂/n饮食/n]Nht
nhm(药品名)	nh和“medicine”的组合	罗红霉素胶囊/nhm
nhp(部位)	nh和“position”的组合	肺/nhp
nhde(科室)	nh和“department”的组合	妇产科/nhde

在词性特征的构建上, 首先需构造词性表 $D=\{d_1,d_2,d_3,\cdots,d_n\}$($d_i$ 表示一个词性, $i\in[1,n]$), 并构建每个词的词性特征向量 $V=\{v_1,v_2,v_3,\cdots,v_n\}$($v_i$ 代表该词词性是否对应词性表 D 中的 d_i), 假设某词的词性为 *nhd*, 则该 v_i 的取值见式 (5), 计算后求得该词的特征向量.

$$v_i=\begin{cases}1, d_i = nhd\\ 0, d_i \neq nhd\end{cases} \tag{5}$$

在设置上下文窗口特征时, 为提取当前词的前后两个词, 将窗口大小设为 5, 对每个词建立词性特征向量以及词特征向量, 并将不存在的上文窗口词与下文窗口词设为空值. 以“患者之前有低血糖病史, 有高血压病史啊”为例, 上下文窗口处理过程见表 2. 注: 通过删除语句中如“吗”、“吧”、“啊”等语气词以及一些数词或状态词来提升运行效率.

表 2 上下文窗口处理过程

词	上文窗口词1	上文窗口词2	下文窗口词1	下文窗口词2
患者	NULL	NULL	之前	有
之前	NULL	患者	有	低血糖
有	患者	之前	低血糖	病史
低血糖	之前	有	病史	,
病史	有	低血糖	,	有
,	低血糖	病史	有	高血压
有	病史	,	高血压	病史
高血压	,	有	病史	NULL
病史	有	高血压	NULL	NULL

表 3 实体标注实例

原句	患者, 男, 50岁, 有二型糖尿病病史15年, 规律服用盐酸二甲双胍片, 近期CT显示患者左胫骨髓内存在病变.
实体标注结果	患者/O 男/O 50/O 岁/O 有/O 二型/B-Nhd 糖尿病/E-Nhd 病史/O 15/O 规律/O 服用/O 盐酸/B-Nhm 二甲双胍片/E-Nhm 近期/O CT/S-Nhi 显示/O 患者/O 左胫骨/B-Nhp 髓内/E-Nhp 存在/O 病变/O

3.3 实验测评

待选定特征向量后, 采用 CTD-BLSTM 算法完成命名实体识别. 首先通过上文构造的词典对少量语料中实体进行 O-S-B-I-E 标注, 然后再经人工进行审查与修正, 标注实例见表 3.

待实验语料标注完毕后, 将数据中 70% 作为训练集分别对基于词性特征的 CTD-BLSTM 模型和基于上下文窗口特征的 CTD-BLSTM 模型进行迭代训练, 30% 作为测试集对模型进行检验, 并将模型训练迭代次数设为 40 次. 实验得到的模型性能变化如图 3 和图 4 所示, 观察可得两个模型的 F 值、准确率和召回率均随迭代次数的增加呈上升趋势, 证实了 CTD-BLSTM 算法的有效性.

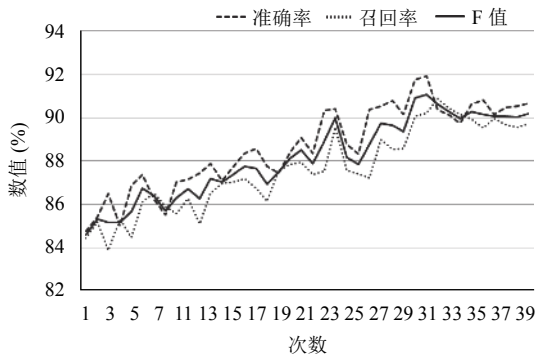

图 3 基于词性特征的 CTD-BLSTM 模型

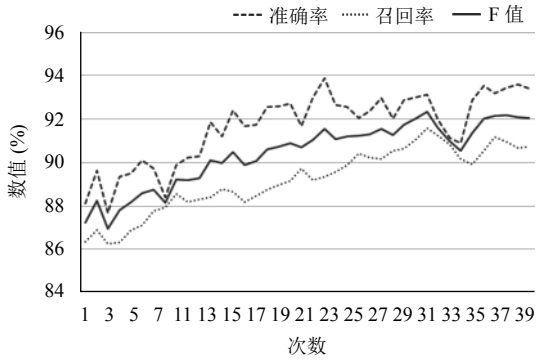

图 4 基于上下文窗口特征的 CTD-BLSTM 模型

为进一步验证 CTD-BLSTM 模型对实体的识别效果, 随机抽取数据集中 5000 条数据进行标注作为模型的训练集, 并用 20 000 条数据对模型进行训练. 除此之外, 并用 5000 条数据和 25 000 条数据分别对 BLSTM、BLSTM-CRF 与 D-BLSTM 模型进行训练得到 12 个对照模型组, 实验结果见表 4 所示.

从表 4 中可以得出以下结论:

(1) 上下文窗口特征比于词性特征包含更多信息, 因此在实验阶段其模型识别率有显著的提高.

(2) 对比特征与规模大小相同训练集下的模型, 得出 D-BLSTM 模型性能优于基础的 BLSTM 模型, 证明了利用双词向量进行扩展的有效性. 例如, 对比模型 7 和模型 11, 模型的 F 值从原来的 88.43% 提升到了 90.41%, 并且由于精简了数据集, 模型的性能比原来提高了 25% 左右, 能做到快速且准确的实体识别.

(3) 对比模型 3、5、9、11、13, 发现协同训练方法的引入使得改进模型 CTD-BLSTM 在准确率、召回率及 F 值上虽略低于用 25000 条数据训练的 D-BLSTM 模型与 BLSTM-CRF 模型, 但其识别效果明显高于经 5000 条数据训练的 D-BLSTM 模型与 BLSTM-CRF 模型, 体现引入协同训练半监督学习方法是有效的, 并在一定程度上大大减少了语料的标注工作.

表 4 实体识别模型对比测评结果 (单位: %)

实验编号	模型类别	F值	准确率	召回率
1	基于词性特征的 BLSTM(5000条)	84.05	84.51	83.60
2	基于上下文窗口特征的 BLSTM(5000条)	85.10	85.84	84.37
3	基于词性特征的 BLSTM-CRF(5000条)	86.36	87.23	85.52
4	基于上下文窗口特征的 BLSTM-CRF(5000条)	87.46	88.64	86.32
5	基于词性特征的 D-BLSTM(5000条)	85.00	85.12	84.89
6	基于上下文窗口特征的 D-BLSTM(5000条)	85.86	86.32	85.41
7	基于词性特征的 BLSTM(25 000条)	88.43	89.31	87.57
8	基于上下文窗口特征的 BLSTM(25 000条)	90.97	91.37	90.57
9	基于词性特征的 BLSTM-CRF(25 000条)	90.43	91.02	89.86
10	基于上下文窗口特征的 BLSTM-CRF(25 000条)	92.19	93.14	91.26
11	基于词性特征的 D-BLSTM(25 000条)	90.41	91.24	89.60
12	基于上下文窗口特征的 D-BLSTM(25 000条)	92.38	93.26	91.51
13	基于词性特征的 CTD-BLSTM	90.18	90.65	89.71
14	基于上下文窗口特征的 CTD-BLSTM	92.07	93.43	90.74

4 结论与展望

本文针对医疗领域的中文命名实体识别任务进行了研究. 为了提高实体识别效果, 在双向长短期神经网络的基础上, 用预训练词向量和微调词向量对神经网络结构单元进行扩展, 并引入协同训练学习方法, 最终提出了 CTD-BLSTM 模型. 经过实验证明, 改进后的模型在准确率、F 值和召回率上均有明显提高, 是一个适用于医疗领域的中文命名实体识别模型.

本文选择词性特征与上下文窗口特征构建相互独立的特征向量集对 CTD-BLSTM 算法模型进行训练, 而可选的特征不仅仅限于词性特征与上下文窗口特征, 因此, 在未来的研究中, 需要进一步测试不同特征模板的选择对模型效果的影响. 此外, 我们将进一步检测此

模型在其它特定领域的效果, 来探索此模型算法对不同领域的适用性, 以进一步提高识别准确率, 获得更精准的实体识别结果.

参考文献

1 Murugesan G, Abdulkadhar A, Bhasuran B, *et al*. BCC-NER: Bidirectional, contextual clues named entity tagger for gene/protein mention recognition. EURASIP Journal on Bioinformatics and Systems Biology, 2017, 2017: 7. [doi: 10.1186/s13637-017-0060-6]

2 张金龙, 王石, 钱存发. 基于 CRF 和规则的中文医疗机构名称识别. 计算机应用与软件, 2014, 31(3): 159–162, 198. [doi: 10.3969/j.issn.1000-386x.2014.03.042]

3 Lei JB, Tang BZ, Lu XQ, *et al*. A comprehensive study of named entity recognition in Chinese clinical text. Journal of the American Medical Informatics Association, 2014, 21(5): 808–814. [doi: 10.1136/amiajnl-2013-002381]

4 杨可心, 桑永胜. 基于 BP 神经网络的 DDoS 攻击检测研究. 四川大学学报 (自然科学版), 2017, 54(1): 71–75.

5 Liu ZJ, Yang M, Wang XL, *et al*. Entity recognition from clinical texts via recurrent neural network. BMC Medical Informatics and Decision Making, 2017, 17(2): 67.

6 Almgren S, Pavlov S, Mogren O. Named entity recognition in Swedish health records with character-based deep bidirectional LSTMs. Proceedings of the 5th Workshop on Building and Evaluating Resources for Biomedical Text Mining. Osaka, Japan. 2016. 30–39.

7 Xu K, Zhou ZF, Hao TY, *et al*. A bidirectional LSTM and conditional random fields approach to medical named entity recognition. In: Hassanien AE, Shaalan K, Gaber T, *et al*, eds. Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2017. Cham: Springer, 2018. 355–365.

8 杨红梅, 李琳, 杨日东, 等. 基于双向 LSTM 神经网络电子病历命名实体的识别模型. 中国组织工程研究, 2018, 22(20): 3237–3242. [doi: 10.3969/j.issn.2095-4344.0302]

9 冯艳红, 于红, 孙庚, 等. 基于 BLSTM 的命名实体识别方法. 计算机科学, 2018, 45(2): 261–268. [doi: 10.11896/j.issn.1002-137X.2018.02.045]

10 李民强. 微博文本情感分类与观点挖掘研究及实现 [硕士学位论文]. 成都: 电子科技大学, 2018.

11 刘崇. 基于知识图谱的医疗知识搜索研究 [硕士学位论文]. 杭州: 浙江理工大学, 2018.

12 Blum A, Mitchell T. Combining labeled and unlabeled data with co-training. Proceedings of the 11th Annual Conference on Computational Learning Theory. Madison, WI, USA. 1998. 92–100.

13 周志华, 王珏. 机器学习及其应用. 北京: 清华大学出版社, 2007.259–275.

14 Xia Y, Wang Q. Clinical named entity recognition: ECUST in the CCKS-2017 shared task 2. Proceedings of CEUR Workshop Proceedings. Chengdu, China. 2017. 43–48.

15 Peters ME, Ammar W, Bhagavatula C, *et al*. Semi-supervised sequence tagging with bidirectional language models. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada. 2017. 1756–1765.

16 Peters ME, Neumann M, Iyyer M, *et al*. Deep contextualized word representations. Proceedings of 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, LA, USA. 2018. 2227–2237.

17 Che WX, Li ZH, Liu T. LTP: A Chinese language technology platform. Proceedings of the 23rd International Conference on Computational Linguistics. Beijing, China. 2010. 13–16.