doi: 10.3969/j.issn.0490-6756.2019.03.013

基于深度主动学习的信息安全领域命名实体识别研究

彭嘉毅[1], 方 勇[2], 黄 诚[2,3], 刘 亮[2], 姜政伟[3]
(1. 四川大学电子信息学院, 成都 610065; 2. 四川大学网络空间安全学院, 成都 610065;
3. 中国科学院信息工程研究所 & 中国科学院网络测评技术重点实验室, 北京 100093)

摘 要: 针对通用领域模型不能很好地解决信息安全领域的命名实体识别问题,提出一种基于字符特性,双向长短时记忆网络(Bi-LSTM)与条件随机场(CRF)相结合的信息安全领域命名实体识别方法. 该方法不依赖于人工选取特征,通过神经网络模型对序列进行标注,再利用 CRF 对序列标签的相关性进行约束,提高序列标注的准确性. 而且,针对信息安全领域标注数据样本不足的问题,采用主动学习方法,使用少量标注样本达到较好的序列标注效果.

关键词: 信息安全; 命名实体识别; 主动学习; 神经网络; 双向长短时记忆网络; 条件随机场
中图分类号: TP391.1 **文献标识码**: A **文章编号**: 0490-6756(2019)03-0457-06

Cyber security named entity recognition based on deep active learning

PENG Jia-Yi[1], *FANG Yong*[2], *HUANG Cheng*[2,3], *LIU Liang*[2], *JIANG Zheng-Wei*[3]
(1. College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China;
2. College of Cybersecurity, Sichuan University, Chengdu 610065, China; 3. Key Laboratory of Network Assessment Technology, CAS (Institute of Information Engineering, Chinese Academy of Sciences), Beijing 100093, China)

Abstract: To solve the problem of low accuracy in general cyber security named entity recognition (NER) model, a deep active learning method is proposed for NER in general cyber security field, which is based on character feature, Bi-LSTM and conditional random field (CRF). The neural network model is for sequence labeling and CRF is for label dependency constraint, which then improves the accuracy of sequence labeling. Furthermore, as for datasets with the insufficient labeled samples in cyber security field, the proposed active learning method is able to achieve better sequence labeling effect with a small number of labeled samples.

Keywords: Cyber security; Named entity recognition; Active learning; Neural network; Bi-LSTM; CRF

1 引 言

随着互联网产业的发展,各种政企单位都逐步实现了互联网化,大量业务需要直接连接互联网或跨安全域工作,面临着网络入侵,黑客攻击,病毒传播以及 DDos 攻击等各种安全问题. 各种网络攻击事件层出不穷,严重威胁着整个社会的网络空间安全. 为与黑客对抗,安全从业人员会对历次重大网络攻击事件进行总结与复盘,从中汲取经验,并将攻击分析报告通过各种渠道披露. 因此,如何从各种攻击分析报告中自动化地提取关键信息就成了信息安全领域一个重要的研究课题. 其中,命名实体识别是最主要,也是最基本的任务之一.

在自然语言处理领域,命名实体识别技术已有

收稿日期: 2018-11-22
基金项目: 中国科学院网络测评技术重点实验室开放课题基金(NST-18-001)
作者简介: 彭嘉毅(1992-), 男, 硕士研究生, 研究方向为信息系统安全. E-mail: pengjiayi@stu.scu.edu.cn
通讯作者: 黄诚. E-mail: opcodesec@gmail.com

较为深入的研究,并在通用语料库上取得了不错的效果. 常用的命名实体识别方法有基于规则和词典的方法[1],基于统计机器学习的方法[2],基于深度学习的方法[3]. 由于信息安全领域缺乏大规模的专业语料库,通用语料库训练的模型在信息安全领域进行命名实体识别时效果不佳,不能准确标注出一些信息安全术语,漏洞名称,软件名称等.

命名实体识别技术在各类垂直领域得到了大量的尝试,其中邬伦等[4]针对中文地名用字特征,引入了语言学相关知识,设计中文地名特征模板,通过条件随机场模型训练和预测,构建了自然语言文本中的中文地名识别模型. 孙娟娟等[5]针对渔业领域命名实体长度较长的特点,构建了 Character+LSTM+CRF 实体识别模型. 在信息安全领域,娄亮等[6]基于主动学习 CRF 技术,进行了信息安全领域的实体识别研究,但该模型识别效果依赖于附带权重的信息安全领域专业词库.

本文提出了一种基于深度主动学习的信息安全领域命名实体识别方法,由于大多数安全分析文章使用英文撰写,本文的命名实体识别也是基于英文处理. 该方法不依赖于人工选取特征,而是通过 Bi-LSTM 的深度学习模型对序列进行标注,再利用 CRF 对序列标签的相关性进行约束,提高序列标注的准确性. 此外,针对信息安全领域标注数据样本不足的问题,本文采用了主动学习方法,使用少量标注样本达到较好的序列标注效果.

2 信息安全命名实体识别模型

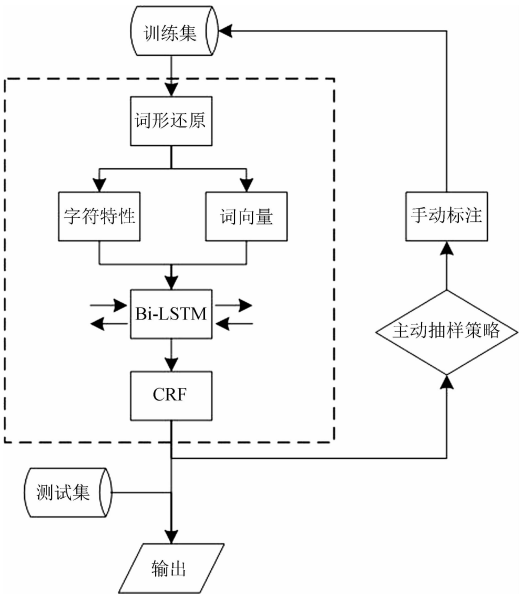

图 1 信息安全命名实体识别模型结构

Fig. 1 The flow chart of cyber security NER model

本文提出的基于深度主动学习的信息安全命名实体识别模型如图 1 所示.

本模型的具体流程如下:1) 对输入文本进行词形还原,主要是针对名词单复数和动词时态等的还原;2) 对输入文本进行文本编码,从字符特性与词向量两个层面对文本进行向量化编码;3) 使用已编码的文本序列训练 Bi-LSTM 神经网络模型. 通过此模型,将向量序列转换为标注概率矩阵;4) 利用 CRF 对序列标注的相关性进行约束,提高序列标注的准确性;5) 根据主动抽样策略,筛选出需要人工标注的样本. 经标注后放入已标注数据集,重新训练序列标注模型,以此循环迭代.

2.1 词形还原

词形还原主要是针对名词和动词在不同的语境中不同的形态进行还原,如名词的单复数变换,动词的时态变换等. 目前主要有基于规则的方法,基于词典的方法,基于机器学习的方法等[7]. 本文词形还原任务主要使用 Python 语言的 NLTK 库[8]实现,NLTK 中的词形还原原理是和 WordNet 词典结合,通过查询 WordNet 词典进行词缀转换与删除.

2.2 文本编码

本文从字符特性和词向量两个层面对输入文本进行向量化编码.

2.2.1 字符特性 针对信息安全领域命名实体经常出现的大小写混用和字母数字混用的特点,对输入的每个词进行了字符特性提取. 本文提取的字符特性如表 1 所示.

2.2.2 词向量 词向量是一种将词的语义映射到向量空间中去的自然语言处理技术. 即将一个词用特定的向量来表示. 本文使用的是 GloVe[9]词向量,一种基于共现矩阵分解的词向量,由 2014 年的英文维基百科数据训练得到. 每个词用 100 维向量表示,预训练的词语数达 400 k 个. 对于 GloVe 中未收录的词,本文中做随机初始化处理.

表 1 字符特征编码

Tab. 1 Character-level feature embedding

字符特征	特征编码
所有字符均为小写字母	0
所有字符均为大写字母	1
词首字母大写,其余小写	2
单词大小写混用	3
所有字符均为数字	4
词中包含数字	5
其他	6

2.3　Bi-LSTM 神经网络

长短时记忆网络[10]（Long Short-Term Memory，LSTM）是一种特殊的循环神经网络（Recurrent Neural Networks，RNN）类型. 不同于人工神经网络[11]与卷积神经网络[12]，循环神经网络可以学习长期依赖信息. LSTM 与 RNN 一样，通过链式形式将重复的神经网络模块相连接. 在标准的 RNN 中，重复的神经网络模块通常是由简单的神经网络层构成，而 LSTM 中有四个神经元，以一种非常特殊的方式进行交互. LSTM 将每一个重复的神经网络模块视为一个细胞，通过门（gate）结构来控制细胞状态，门是一种让信息选择性通过的方法. 他们包含一个 Sigmoid 层和一个乘法操作. Sigmoid 层的输出决定了每个部分有多少量可以通过. 计算过程如下所示.

$$f_t=\sigma(W_f\cdot[x_t,h_{t-1}]+b_f) \tag{1}$$

$$i_t=\sigma(W_i\cdot[h_{t-1},x_t]+b_i) \tag{2}$$

$$\tilde{C}_t=\tan h(W_c\cdot[h_{t-1},x_t]+b_C) \tag{3}$$

$$C_t=f_t\cdot C_{t-1}+i_t\cdot\tilde{C}_t \tag{4}$$

$$o_t=\sigma(W_o\cdot[h_{t-1},x_t]+b_o) \tag{5}$$

$$h_t=o_t\cdot\tan h(C_t) \tag{6}$$

其中，i,f,o 分别表示输入门，忘记门和输出门；σ 代表 Sigmoid 层；W 和 b 分别表示 Sigmoid 层的权重和常数参数；C 是由忘记门和输入门控制的细胞状态参数.

LSTM 能够有效地过滤和记忆文本上文的信息，但在命名实体识别中，文本的下文信息也同样重要. Schuster 等[13]提出双向 RNN 模型，通过前向与后向两层循环神经网络充分利用文本的上下文信息. Graves 等[14]在此基础上，使用 LSTM 记忆单元替换了双向 RNN 中的神经元，构建了 Bi-LSTM 神经网络模型，其网络结构如图 2 所示.

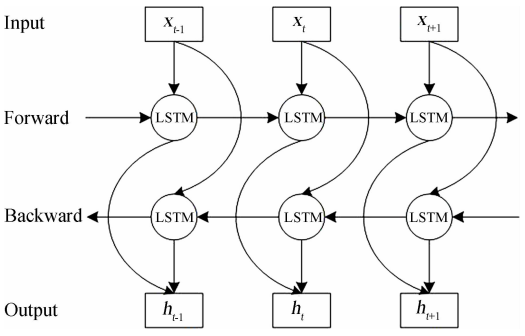

图 2　Bi-LSTM 神经网络结构

Fig. 2　The structure of Bi-LSTM

在 Bi-LSTM 神经网络模型中，记 t 维度的前向隐向量为$\overrightarrow{h_t}$，后向隐向量为$\overleftarrow{h_t}$. 则 t 维度的输出为

$$h_t=\overrightarrow{h_t}+\overleftarrow{h_t} \tag{7}$$

2.4　条件随机场

条件随机场是一种判别式概率无向图模型，在给定输入随机变量的情况下，计算输出随机变量的条件概率分布. 条件随机场在自然语言处理领域已有广泛运用，如序列标注，中文分词等.

本文使用的是线性链条件随机场，其条件概率为 $P=(Y|X)$，其中，Y 为输出的标记序列；X 为输入的观测序列. 令 $X=(X_1,X_2,\ldots,X_n)$，$Y=(Y_1,Y_2,\ldots,Y_n)$ 则条件概率 $P=(Y|X)$ 为

$$P(y|x)=\frac{1}{Z(x)}\exp(A_{y_{i-1},y_i}+p_{y_i}) \tag{8}$$

$$A_{y_{i-1},y_i}=\sum_k\lambda_k\sum_i t_k(y_{i-1},y_i,x,i) \tag{9}$$

$$p_{y_i}=\sum_l\mu_l\sum_i s_l(y_i,x,i) \tag{10}$$

其中，$Z(x)$ 是规范化因子；t_k 为转移特征函数；λ_k 为其对应权重；s_l 为状态特征函数；μ_l 为其对应权重. 条件随机场在训练过程中使用极大似然估计方法对参数进行估计，序列标注过程中使用 Viterbi 算法进行路径求解.

2.5　主动抽样策略

监督学习需要大量完整标注的数据样本，但由于标注成本过高，实际中往往很难得到大量的标注样本. 主动学习算法利用大量未标记样本，通过筛选出信息量大的样本进行标注和训练. 再使用尽可能少的标注样本的情况下，达到与监督学习类似的分类效果. 主动学习算法可以抽象为以下的五元模型[15].

$$A=(C,U,L,Q,S) \tag{11}$$

其中，C 为分类算法；L 为已标注样本；U 为未标记样本；Q 为主动抽样策略；S 为人工标注者. 主动学习算法过程可概括为：首先从 U 中随机标记少量样本加入 L 作为初始训练样本集，C 使用 L 训练分类模型，同时 Q 根据分类模型从 U 中筛选出一定数量未标注数据，由 S 进行人工标注，并将这部分已标注数据加入 L，C 再次使用 L 训练分类模型，迭代上述过程直到达到停止条件.

主动抽样策略是主动学习算法的关键所在，也是区别于其他学习模型的最大不同点. 主动抽样策略主要有以下两种算法[16]：基于流抽样算法（Stream-Based Selective Sampling）和基于池抽样算法（Pool-Based Sampling）. 本文采用的是基于流抽样算法，其主要思路是将未标记样本池的数据送入

模型,根据模型标记结果,由抽样策略决定哪些样本需要手工标记. 具体实现方面,常用的有最小置信度算法(Least Confidence)与最大归一化对数概率算法(Maximum Normalized Log-Probability)等.

最小置信度算法的原理是求出标注模型输出概率积的最大值,即标注序列的概率积. 概率积越小说明不确定度越大,以此抽样出需要手工标记的样本,如式(12)所示. 最小置信度算法会倾向于选出长度更长的序列.

$$\max_{y_1,y_2,\dots,y_n} P(y_1,y_2,\dots,y_n|\{X_{ij}\}) \tag{12}$$

最大归一化对数概率算法(Maximum Normalized Log-Probability)原理是计算预测中最大概率序列的概率值对数和,并进行归一化,如公式(13)所示. 该方法可以避免序列长度变化的影响.

$$\max_{y_1,y_2,\dots,y_n} \frac{1}{n}\sum_{i=1}^{n}\log P(y_i|y_1,y_2,\dots,y_n,\{X_{ij}\}) \tag{13}$$

本文采用的主动抽样算法是最大归一化对数概率算法. 具体的主动学习步骤为:首先随机选取2%的训练集数据进行手工标注,训练上文构建的基于字符特性,Bi-LSTM 与 CRF 的信息安全领域命名实体识别模型. 然后利用此模型并采用最大归一化对数概率算法抽样出 2%的训练集数据进行手工标注,并再次训练模型. 迭代此过程,直到标注数据达到训练集的 30%时,达到本文设置的主动学习停止条件.

3 实验分析

3.1 信息安全领域语料库

为构建信息安全领域语料库,本文采集国外几个主要安全公司技术博客发布的文章. 文章主要针对各种黑客攻击行为进行剖析复盘,也对攻击中所使用的各种资源进行了描述和深度分析. 本文采集的安全公司博客及其文章数如表 2 所示.

表 2 本文所收集的信息安全文章来源统计

Tab. 2 The source of our dataset

安全公司名	文章数(篇)
Cloudflare	297
Naked Security	1837
Microsoft Security	1333
Cisco Security	517
McAfee	775
Fireeye	219
Krebson Security	964
Palo Alto Networks	716
总计	6 658

首先需要对采集的博客文章进行格式化清洗,提取文章正文,去除掉 HTML 标签及代码段,这部分工作使用 Python 的 BeautifulSoup 库实现. 然后去除掉句子长度过小(小于 4 个单词)或句子长度过大(大于 100 个单词)的样本. 数据清洗后,得到信息安全领域语料 127709 条. 将语料库按8∶1∶1的比例随机分为训练集,测试集和验证集.

使用主动学习算法时,只需要对测试集和验证集数据提前进行标注,但本文出于对照组实验需要,对所有数据集进行提前标注. 本文的数据集标注有以下三个步骤:

1) 首先使用训练好的通用领域命名实体识别模型进行命名实体预标注,这部分工作使用 Stanford NLP 的命名实体识别工具[17]实现.

2) 本文收集了众多信息安全领域命名实体相关字典,具体数目如表 3 所示. 其中信息安全相关公司名如 CloudFlare, Rapid7, Trend Micro 等,信息安全领域术语如 CSRF, RBAC, Rootkit 等,常见软件名与服务器常见组件名如 Chrome, Nginx, Struts. 黑客工具如 Metasploit, Burp Suite, Hydra 等. 使用这些字典在命名实体预标注的基础上对文本进行字符匹配标注,即文本中出现字典包含的实体时就标注为命名实体.

3) 由于前两步存在一定程度的误报与漏报,所以需要再进行人工校验,修正自动标注不正确的情况. 使用此方法对数据集进行标注的工作量显著低于完全人工标注.

表 3 信息安全领域命名实体字典数目

Tab. 3 Cyber security named entity dictionary

类别	数量
信息安全相关公司名	496
信息安全领域术语	366
常见软件名与服务器组件名	287
黑客工具名	125

本文的信息安全命名实体识别模型的标注类型为{PER, LOC, ORG, SEC, O}分别表示人名,地名,组织机构名,安全术语和非实体. 采用 BOI 的标注体系,B 代表实体开头单词;I 代表实体后续单词;O 代表非实体.

3.2 实验环境和评价指标

实验环境的软硬件配置信息如下:CPU:Intel

(R) Core(TM) i7-7700 3. 60 GHz；内存：16 G；GPU：NVIDIA GeForce GTX 1060 6 GB；操作系统：Ubuntu 16. 04. 4 LTS. Bi-LSTM 算法与 CRF 算法使用 Python 语言的 PyTorch 库实现.

本文的命名实体识别任务可以看作一个共 9 类的多分类问题，这 9 个类别分别是：B-PER(人名实体开头单词)，I-PER(人名实体后续单词)，B-LOC(地名实体开头单词)，I-LOC(地名实体后续单词)，B-ORG(组织机构实体开头单词)，I-ORG(组织机构实体后续单词)，B-SEC(安全术语开头单词)，I-SEC(安全术语后续单词)，O(非实体). 评价时可以转换为多个二分类问题. 如将某一类看作正类时，其他类即为负类，以此求出单个二分类问题的评价指标，再求出各类别的平均值. 对标注结果有以下分类：

1) TP(真阳性，Ture Positive)预测值为正类并且样本为正类；

2) FP(假阳性，False Positive)预测值为正类但样本为负类；

3) TN(真阴性，True Negative)预测值为负类且样本为负类；

4) FN(假阴性，False Negative)预测值为负类但样本为正类.

实验评价指标采用精确率 P，召回率 R 和调和平均数 F_1. 其计算公式分别为

$$P=\frac{TP}{TP+FN} \tag{14}$$

$$R=\frac{TP}{TP+FP} \tag{15}$$

$$F_1=\frac{2PR}{P+R} \tag{16}$$

其中，精确率 P 表示正确识别的正类占所有预测为正类的比值；召回率 R 表示正确识别的正类占总的正类样本的比值；调和平均数 F_1 表示精确率与召回率的综合值.

3.3　实验设计与结果分析

本文基于字符特性，Bi-LSTM 和 CRF 的命名实体识别模型每次采用主动抽样策略后的训练轮数为 20，初始学习率为 0. 01，优化算法采用随机梯度下降法(SGD). 为了验证本文的基于深度主动学习的信息安全领域命名实体识别模型，设计了如下 2 组实验.

3. 3. 1　实验一　采用主动学习算法与不采用主动学习算法时，本文提出的信息安全领域命名实体识别模型的识别效果比较.

表 4　主动学习与非主动学习算法效果比较

Tab. 4　Exam comparison of active learning and non-active learning algorithms

算法	精确率 P	召回率 R	F_1 值	标注数据比例
非主动学习算法	0. 8928	0. 8954	0. 8941	100%
主动学习算法	0. 8925	0. 8816	0. 8872	30%

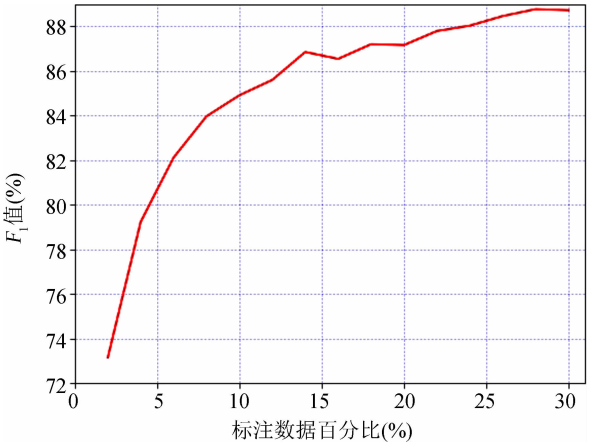

图 3　深度主动学习模型 F_1 值曲线

Fig. 3　F_1 curve of deep active learning model

从图 3 和表 4 可以分析出，本文提出的基于深度主动学习的信息安全领域命名实体识别模型的识别效果随标注数据的增加而显著提升. 并且使用主动学习算法在标注数据量达到训练集数据总数 30%时，与使用整个已标注训练集的非主动学习算法标注效果相差无几，也由此说明了基于主动学习算法的模型可以显著降低标注代价.

3. 3. 2　实验二　本文采用基于 CRF 的命名实体识别模型作为对照组，在使用相同训练集的情况下与本文提出的深度主动学习模型的识别效果比较.

表 5　不同模型效果比较

Tab. 5　Exam comparison of two models

模型	精确率 P	召回率 R	F_1 值
深度主动学习模型	0. 8925	0. 8816	0. 8872
CRF 模型	0. 8403	0. 8082	0. 8238

从表 5 可以分析出，本文提出的基于深度主动学习的信息安全领域命名实体识别模型的识别效果会优于基于 CRF 的命名实体识别模型，也由此说明了本文提出模型的有效性.

4　结　论

本文提出了一种基于深度主动学习的信息安全领域命名实体识别模型，将字符特性，Bi-LSTM 与 CRF 相结合构建神经网络模型，并使用本文采

集的信息安全领域语料进行训练. 解决了通用领域模型不能准确识别信息安全领域命名实体的问题,通过实验证明了本文提出模型的有效性. 同时利用主动学习算法,有效地避免了信息安全领域标注语料不足的问题. 在较少的标注语料上取得了与监督学习算法基本相同的标注效果.

参考文献:

[1] Rizzo G, Troncy R. NERD: a framework for unifying named entity recognition and disambiguation extraction tools [C]//Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. [s. l.]: ACM Press, 2012.

[2] 冯艳红, 于红, 孙庚, 等. 基于词向量和条件随机场的领域术语识别方法 [J]. 计算机应用, 2016, 36: 3146.

[3] Chiu J P C, Nichols E. Named entity recognition with bidirectional LSTM-CNNs [J/OL]. Comput Sci, (2015-11-26) [2018-03-25]. https://arxiv.org/abs/1511.08308.

[4] 邬伦, 刘磊, 李浩然, 等. 基于条件随机场的中文地名识别方法 [J]. 武汉大学学报: 信息科学版, 2017, 42: 150.

[5] 孙娟娟, 于红, 冯艳红, 等. 基于深度学习的渔业领域命名实体识别 [J]. 大连海洋大学学报, 2018, 33: 265.

[6] 娄亮, 周安民. 基于主动学习 CRF 的信息安全领域命名实体识别研究 [J]. 通信与信息技术, 2016, 46: 61.

[7] 吴思竹, 钱庆, 胡铁军, 等. 词形还原方法及实现工具比较分析 [J]. 数据分析与知识发现, 2012, 28: 27.

[8] Bird S, Loper E. NLTK: the natural language toolkit [C]//Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. Stroudsburg: ACM Press. 2004.

[9] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation [C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). [s. l.]: EMNLP, 2014.

[10] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Comput, 1997, 9: 1735.

[11] 杨可心, 桑永胜. 基于 BP 神经网络的 DDoS 攻击检测研究 [J]. 四川大学学报: 自然科学版, 2017, 54: 71.

[12] 李勤, 师维, 孙界平, 等. 基于卷积神经网络的网络流量识别技术研究 [J]. 四川大学学报: 自然科学版, 2017, 54: 959.

[13] Schuster M, Paliwal K K. Bidirectional recurrent neural networks [J]. IEEE Process Soc, 1997, 45: 2673.

[14] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. [J]. Neural Networks, 2005, 18: 602.

[15] 刘康, 钱旭, 王自强. 主动学习算法综述 [J]. 计算机工程与应用, 2012, 48: 1.

[16] 胡峰, 周耀, 王蕾. 基于邻域粗糙集的主动学习方法 [J]. 重庆邮电大学学报: 自然科学版, 2017, 29: 776.

[17] Finkel J R, Grenager T, Manning C. Incorporating non-local information into information extraction systems by gibbs sampling [C]//Proceedings of the 43rd annual meeting on association for computational linguistics. Stroudsburg: ACM Press, 2005.

引用本文格式:

中 文: 彭嘉毅, 方勇, 黄诚, 等. 基于深度主动学习的信息安全领域命名实体识别研究 [J]. 四川大学学报: 自然科学版, 2019, 56: 457.

英 文: Peng J Y, Fang Y, Huang C, *et al*. Cyber security named entity recognition based on deep active learning [J]. J Sichuan Univ: Nat Sci Ed, 2019, 56: 457.