计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2019.20180733
56(9):1851-1858,2019

# 结合 GAN 与 BiLSTM-Attention-CRF 的领域命名实体识别

张 晗[1,2] 郭渊博[1] 李 涛[1]
[1](战略支援部队信息工程大学密码工程学院 郑州 450001)
[2](郑州大学软件学院 郑州 450001)
(zhang_han@zzu.edu.cn)

# Domain Named Entity Recognition Combining GAN and BiLSTM-Attention-CRF

Zhang Han[1,2], Guo Yuanbo[1], and Li Tao[1]
[1](*Department of Cryptogram Engineering, Strategic Support Force Information Engineering University, Zhengzhou* 450001)
[2](*Software College, Zhengzhou University, Zhengzhou* 450001)

**Abstract** Domain named entity recognition usually faces the lack of domain annotation data and the inconsistency of entity annotation in the same document due to the diversity of entity names in the domain. To issue the above problems, our work draws on the combination of the generative adversarial network (GAN) which can generate data and the BiLSTM-Attention-CRF model. Firstly, BiLSTM-Attention is used as the generator model of GAN, and CNN is used as the discriminant model. The two models use the crowd annotations and the expert annotations to train respectively, and integrate the positive annotation data consistent with the expert annotation data distribution from the crowd annotations to solve the problem of lack of annotation data in the domain; then we also introduce a new method to obtain the new feature representation of each word in the document through introducing a document-level global feature in the BiLSTM-Attention-CRF model in order to solve the problem of inconsistency of the entity in the same document caused by the diversification of the entity name. Finally, taking the crowd annotations in the information security field as a sample, a comprehensive horizontal evaluation experiment is carried out by learning the common features and applying them to the training BiLSTM-Attention-CRF model for the identification of named entities in the information security field. The results show that compared with the existing models and methods, the model we proposed has made great progress on various indicators, reflecting its superiority.

**Key words** domain named entity recognition; generative adversarial network (GAN); crowd annotations; entity annotations consistent; BiLSTM-Attention-CRF model

**摘 要** 领域内命名实体识别通常面临领域内标注数据缺乏以及由于实体名称多样性导致的同一文档中实体标注不一致等问题.针对以上问题,利用生成式对抗网络(generative adversarial network, GAN)可以生成数据的特点,将生成式对抗网络与 BiLSTM-Attention-CRF 模型相结合.首先以 BiLSTM-Attention 作为生成式对抗网络的生成器模型,以 CNN 作为判别器模型,从众包标注数据集中整合出与专家标注数据分布一致的正样本标注数据来解决领域内标注数据缺乏的问题;然后通过在 BiLSTM-Attention-CRF 模型中引入文档层面的全局向量,计算每个单词与该全局向量的关系得出其新的特征

**收稿日期**:2018-11-01;**修回日期**:2019-04-17
**基金项目**:国家自然科学基金项目(61501515);河南省重点科技攻关项目(172102210002);郑州大学青年骨干教师项目(2017ZDGGJS048)
This work was supported by the National Natural Science Foundation of China (61501515), the Key Scientific and Technological Research Project of Henan Province (172102210002), and the Young Scholar Teachers Project of Zhengzhou University (2017ZDGGJS048).

表示以解决由于实体名称多样化造成的同一文档中实体标注不一致问题;最后,在基于信息安全领域众包标注数据集上的实验结果表明,该模型在各项指标上显著优于同类其他模型方法.

**关键词** 领域命名实体识别;生成式对抗网络;众包标注数据;实体标注一致; BiLSTM-Attention-CRF模型

**中图法分类号** TP183

领域命名实体识别(named entity recognition, NER)[1]旨在从文本中提取出各种类型的实体,其结果可用于领域中后续的其他复杂任务诸如关系提取、领域知识图谱的构建等.与通用领域的命名实体识别任务相比,特殊领域的命名实体识别经常会面临着领域标注数据缺乏以及由于领域内实体名称的多样性而导致的同一文档中实体标注不一致问题[2].

为了解决由于领域内标注数据缺乏而导致的模型性能问题,文献[3]提出使用具有大量标注数据的通用领域数据集(如新闻领域)来训练模型,使得模型可以从中学习到更多特定领域的特征分布,从而提高模型在特定领域数据上的性能.但是显然使用通用领域训练数据中的词汇特征分布来估计专业领域中的数据特征分布会导致偏差太大的问题.为了快速高效且低成本地获取领域内大型标注数据集,文献[4]提出可以使用 Amazon Mechanical Turk 来快速且低成本地收集标签数据,并且证明了众包标注数据对于训练模型来说是可用的.相比于专家标注来说,众包具有低成本、大规模等优点,因此随着众包技术的成熟,国内外众包的应用也越加广泛,其中以亚马逊的 Amazon Mechanical Turk 和维基百科最为著名.目前,国内也出现了不少众包平台,如百度众包、阿里众包和搜狗输入法等.这些众包平台大都属于通用众包平台,虽然准确率经过平台筛选之后还算理想,但是,针对专业领域来说,其缺点也非常明显,这些来自于非专业人员的标注数据远远要比来自于专家标注的数据质量低并且含有极大的噪声,这部分数据如果直接拿来使用,会给模型造成偏差.因此需要针对众包标注数据的质量进行建模,整合出高质量的共识标注.文献[5]采用多数投票法来实现高质量共识标注.这种方法尽管可以获得较高质量的标注数据,但是需要耗费大量的人力,同样的一句话需要交由至少 3 个注释者标注,这样才能通过多数投票法选取出相对准确的标注.而且对于一些本身含义就比较模糊的句子或者实体,标注者之间可能很难达到统一,这样无疑会影响到最终选择结果的正确性.

除此之外,目前常用来进行命名实体识别任务的模型,如条件随机场(conditional random field, CRF)[6]模型、双向长短记忆网络+条件随机场(bi-directional long short-term memory+conditional random field, BiLSTM-CRF)[7]模型,大都仅仅只是从单词和句子层面来考虑单词的特征表示.但是由于这些模型自身设计的限制,输入序列无论长短都会编码成一个固定长度的向量表示,当输入序列过长时,编码器可能会丢掉上下文中一些比较重要的特征信息,这样就容易造成因为实体名称多样化所带来的同一文档中实体标注不一致问题.例如: Internet Explorer 和 IE 指代的是同一个实体,但是由于表示方法不同以及在文中位置不同,很可能会出现两者标注不一致的情况,从而影响到后续任务的完成效果.为了提高模型的性能,文献[8]在 BiLSTM-CRF 模型的基础上添加了注意力(attention)机制,它打破了传统编码器-解码器结构在编解码时都依赖于内部一个固定长度向量的限制,可以通过训练一个模型来对输入字符串的特征进行重点选择性学习并且在模型输出时将输出序列与各选择特征按照权重进行关联.但是这种模型如果要达到比较理想的效果,在训练时要求用到更加大量准确的标注数据.生成式对抗网络(generative adversarial network, GAN)模型具有生成数据的特点,而 Goodfellow 等人[9]从理论上证明了当 GAN 模型收敛时,生成数据具有和真实数据相同的分布[10].

基于 GAN 的这个特点,本文在众包标注数据集中引入 GAN 模型,以少量专家标注数据作为判别器中的真实数据,以众包标注数据作为生成器所生成的模拟数据,通过对抗学习,整合出众包标注数据中与真实数据分布一致的正样本数据并将其与 BiLSTM-Attention-CRF 模型相结合,提出了一种新的模型 BiLSTM-Attention-CRF-crowd,该模型由 GAN 和 BiLSTM-Attention-CRF 这 2 个子模型组成.首先,通过 GAN 模型在给定的众包标注数据集上寻找出标注数据的共有特征以整合出最佳的

唯一共识标注,解决目前众包数据中质量不高的问题;然后使用通过 GAN 模型所生成的标注数据来训练 BiLSTM-Attention-CRF 模型进行领域命名实体识别,并引入文档层面的全局特征向量,通过计算每个单词与全局向量的关系得出其新的特征表示,以解决同一实体在同一文档中可能出现的标记不一致问题.

## 1　相关工作

### 1.1　生成式对抗网络(GAN)

相对于在计算机视觉领域的应用,GAN 模型在语言处理领域的应用较少,原因在于图像和视频数据的取值是连续的,可直接使用梯度下降法对生成器和判别器进行训练,而文本中的字母、单词都是离散的,无法直接应用梯度下降法,需要对其进行修改.文献[11]所提出的 TextGAN 模型采用一些技巧对离散变量进行处理,例如,采用光滑近似来逼近长短记忆网络(long short-term memory, LSTM)的离散输出,并在生成器训练过程中采用特征匹配技术.由于 LSTM 的参数明显多于卷积神经网络(convolutional neural network, CNN)的参数个数而更难训练,TextGAN 的判别器仅在生成器多次更新后才进行一次更新.文献[12]提出的 SeqGAN 借鉴强化学习处理离散输出问题,将判别器输出的误差视为强化学习中的奖赏值,并将生成器的训练过程看作强化学习中的决策过程,应用于诗句、演讲文本以及音乐生成.文献[13]和文献[14]分别将 GAN 应用于开放式对话文本生成和上下文无关语法(context-free grammar, CFG)[10].本文主要借鉴了文献[11]中对离散变量的处理方法及其进行特征匹配的目标函数.

### 1.2　众包标注数据

文献[15]提出了一种从多个标签中进行学习的条件随机场(conditional random field, CRF)模型,但是 CRF 可以学习的特征是有限的;文献[16]所提 Dawid&Skene 模型假定每一个标注者出现每一类标注错误的概率确定,这样就可以用一个统一的混淆矩阵来描述标注者的标注质量,最后通过最大似然估计就可以得到所有的实体标注概率,其中也包括每个实体的正确标注.这属于比较理想化的情况,在实际应用中,由于众包数据来源的多样性,很难确定每个标注者的错误概率.文献[17-19]虽然所使用的模型不同,但其本质都是对标注者身份进行区别,将标注正确率较高的标注者身份提取出来,从而提高他们标注的可信度,这样做确实提高了模型的性能,但是对标注的选取过于依赖某个标注者的可信度.

### 1.3　实体标记不一致

为解决实体标记不一致问题,通常采用基于规则的后处理方式来强调标记一致性.如文献[20]中设置规则如果某实体在文档中被 NER 模型至少标记了 2 次,则在该文档中其他位置出现的该实体也被标记同样的类别.但是这种后处理方式并不一定能改善模型的性能,反而可能会因为实体被错误地标记而引入更多的噪声.此外,文献[1]和文献[2]通过使用非本地信息来强调标签的一致性从而提高序列模型的性能,但收效并非十分明显.文献[21]提出的模型虽然也考虑了文档层面的特征表示,但该模型应用于化学领域,该领域内有很多公开标注的数据集可以直接进行使用,无需再考虑领域数据缺乏的问题.

本文所提模型不再对标注者身份进行鉴别,而是利用 GAN 模型生成数据的特点,从众包标注数据集中生成与专家标注数据趋于一致的标注数据,从而解决领域内标注数据准确率不高的问题.

## 2　BiLSTM-Attention-CRF-crowd 模型设计

模型所要完成的任务包括 2 个方面.子任务 1:学习众包标注数据的共有特征,以整合出最优的单一共识标注;子任务 2:以子任务 1 所生成的标注数据集作为训练数据,训练 BiLSTM-Attention-CRF 模型进行领域内命名实体识别并解决同一文档内同一实体标记不一致问题.模型图如图 1 所示.

该模型由 2 个子模型组合而成.图 1 左侧描述的是 GAN 模型,由生成器 BiLSTM-Attention 和判别器 CNN 构成.众包标注数据作为生成器的输入在经过 BiLSTM 层和 Attention 层处理之后形成新的特征表示,该特征表示传入到由专家标注数据训练的 CNN 中,由 CNN 来判别这些特征分布与专家标注数据分布是否趋于一致,如果一致则为正样本数据,可以作为图 1 右侧所示模型的训练语料;反之则为负样本数据,重新传递回 BiLSTM 层对该层进行优化.图 1 右侧描述的是进行命名实体识别的 BiLSTM-Attention-CRF 模型,该模型由 GAN 模型所生成的正样本数据训练而成,为了解决由领域实体多样化所带来的实体标注不一致问题,加入了基于整篇文档的全局向量 $\boldsymbol{r}_i^g$,通过 Attention 层来

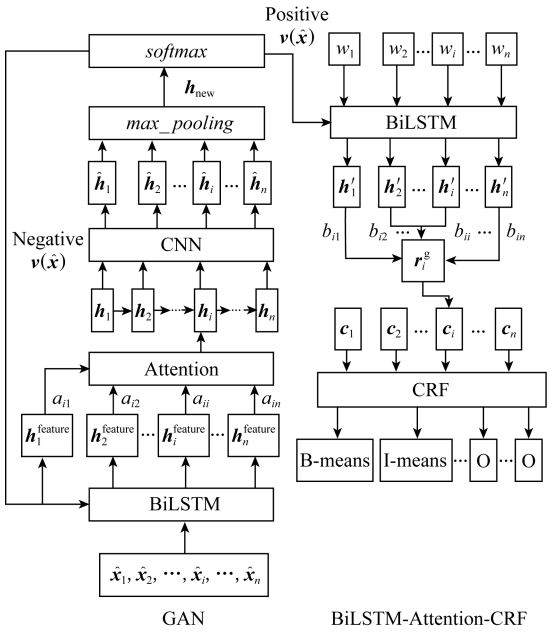


Fig. 1 Model diagram

图 1 模型图

计算 $\boldsymbol{r}_i^{\mathrm{g}}$ 与当前单词的关系权重得出该单词新的特征表表示并传递给 CRF 层进行识别标记.

### 2.1 基于 GAN 模型的共同特征学习

GAN 模型以博弈论中二人零和博弈为核心思想.结构中包含 2 个模型:生成模型和判别模型,将随机变量作为生成模型的输入,经过其非线性映射,输出对应的信号作为判别模型的输入,由判别模型来判断该信号来自于真实数据的概率.因此,二者的目标是截然相反的,判断模型的目标是通过最大化对数似然函数以判断信号的来源,而生成模型的目标则是最小化对数似然函数,使得输出信号的分布逼近于真实数据的分布.

本文将对抗思想应用于寻找众包标注数据中的共同特征.因此将不再对标注者的标注质量进行评价估计,而是利用 GAN 模型可以生成数据的思想,先利用少量的专家标注数据训练出一个判别模型,然后将众包标注数据作为生成模型的输入,输出这些数据的特征分布传递给判别模型,由判别模型来进行判断生成的特征分布与真实数据的特征分布的异同,反复训练模型并生成数据,直到最后判别器无法再对二者进行区别为止,此时判别模型所输出的结果即是标注数据的共同特征也即是我们要整合出的最优单一共识标注.GAN 模型如图 1 左侧框图所示,主要由 2 部分组成:BiLSTM-Attention 构成的生成模型以及 CNN 构成的判别模型.CNN 上面的 *max_pooling* 层和 *softmax* 层主要是对 CNN 层生成的特征图进行最大化选择以及对选择之后的新特征进行归一化,以此判断是否与训练 CNN 的专家标注数据特征分布一致.

#### 2.1.1 BiLSTM-Attention 生成模型

给定众包标注数据集中的一个句子 $s=\{w_1,w_2,\cdots,w_n\}$,$w_i$ 表示句子中识别出的命名实体单词,$y=\{y_1,y_2,\cdots,y_n\}$为标注集,其中 $y_i$ 表示命名实体 $w_i$ 所对应的类别标签.用向量 $\hat{\boldsymbol{x}}_i$ 表示单词 $w_i$ 和标注 $y_i$ 通过 word2vec 训练出来的联合向量,以此作为生成模型 BiLSTM 的输入,则得到特征 $\boldsymbol{h}_1^{\mathrm{feature}},\boldsymbol{h}_2^{\mathrm{feature}},\cdots,\boldsymbol{h}_n^{\mathrm{feature}}$ 为

$$\boldsymbol{h}_1^{\mathrm{feature}}\boldsymbol{h}_2^{\mathrm{feature}}\cdots\boldsymbol{h}_n^{\mathrm{feature}}=BiLs(\hat{\boldsymbol{x}}_1,\hat{\boldsymbol{x}}_2,\cdots,\hat{\boldsymbol{x}}_n),\tag{1}$$

其中,$BiLs$ 作为模型 BiLSTM 的缩写.为了获取到更加重要的特征,此时,在判别模型 CNN 之前再加上一层 Attention 机制以获得新的特征 $\boldsymbol{h}_i$:

$$f(\boldsymbol{h}_i,\boldsymbol{h}_j)=(\boldsymbol{h}_i^{\mathrm{feature}})^{\mathrm{T}}\boldsymbol{W}_a\boldsymbol{h}_j^{\mathrm{feature}},$$

$$\boldsymbol{h}_i=\sum_{j=1}^{n}a_{ij}\boldsymbol{h}_j^{\mathrm{feature}},$$

$$a_{ij}=softmax(f(\boldsymbol{h}_{i-1},\boldsymbol{h}_j^{\mathrm{feature}})),\tag{2}$$

其中,$\boldsymbol{W}_a$ 为模型参数.

#### 2.1.2 CNN 判别模型

利用 CNN 卷积神经网络作为判别器,将窗口大小设置为 5[18],可以得出 $\boldsymbol{h}_i$ 新的特征表示 $\hat{\boldsymbol{h}}_i$:

$$\hat{\boldsymbol{h}}_i=\tanh(\boldsymbol{W}_c[\boldsymbol{h}_{i-2},\boldsymbol{h}_{i-1},\cdots,\boldsymbol{h}_{i+2}]).\tag{3}$$

其中,$\boldsymbol{W}_c$ 为 CNN 模型参数,在接下来的池化层,我们选用 max-over-time pooling[18] 的方法选取出最大值 $\boldsymbol{h}_{\mathrm{new}}$:

$$\boldsymbol{h}_{\mathrm{new}}=\max\{\hat{\boldsymbol{h}}_1,\hat{\boldsymbol{h}}_2,\cdots,\hat{\boldsymbol{h}}_n\}.\tag{4}$$

该池化方法认为具有最高值的特征才是最重要的特征,有效地过滤掉信息量较少的单词组合,并且可以保证提取的特征与输入句子的长度无关.

通过一个 *softmax* 层来将特征 $\boldsymbol{h}_{\mathrm{new}}$ 映射到输出 $D(\boldsymbol{h}_{\mathrm{new}})\in[0,1]$,以此来判断输入特征是否与专家标注数据特征分布一致.

#### 2.1.3 模型目标

我们采用了文献[11]中类似特征匹配的方法,设 $S$ 为专家标注数据集,迭代优化生成式对抗网络模型目标函数为

minimizing:

$$L_D=-E_{s,S}\ln(D(s))-E_{\hat{\boldsymbol{x}},p(\hat{\boldsymbol{x}})}[\ln(1-D(G(\hat{\boldsymbol{x}})))],\tag{5}$$

minimizing:

$$L_G=\mathrm{tr}(\boldsymbol{\Sigma}_s^{-1}\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}}+\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}}^{-1}\boldsymbol{\Sigma}_s)+(\boldsymbol{\mu}_s-\boldsymbol{\mu}_{\hat{\boldsymbol{x}}})^{\mathrm{T}}(\boldsymbol{\Sigma}_s^{-1}+\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}}^{-1})(\boldsymbol{\mu}_s-\boldsymbol{\mu}_{\hat{\boldsymbol{x}}}),\tag{6}$$

其中,$\boldsymbol{\Sigma}_{\hat{x}}$ 和 $\boldsymbol{\Sigma}_s$ 分别表示专家标注数据和生成模型处理后的数据特征向量的协方差矩阵;$\boldsymbol{\mu}_s$,$\boldsymbol{\mu}_{\hat{x}}$ 分别表示专家标注数据和生成模型处理后的数据的平均特征,它们的值是从小批量数据里根据经验估算出来的.这里,$L_G$ 表示 2 个多元高斯分布 $dtri(\boldsymbol{\mu}_s,\boldsymbol{\Sigma}_s)$ 和 $dtri(\boldsymbol{\mu}_{\hat{x}},\boldsymbol{\Sigma}_{\hat{x}})$ 之间的 Jensen-Shannon 差异.这样做的目的主要是从技术上为修改生成模型提供更强的信号,使得生成模型输出的特征分布与判别模型中的专家标注数据特征分布更加趋于一致[11].

这里存在一个问题,即当生成模型的输出值为离散值时,判别模型的误差梯度无法利用反向传播算法回到生成模型,我们采用文献[11]所提到的方法将返回给生成模型 BiLSTM 的特征向量 $\boldsymbol{v}(\hat{\boldsymbol{x}})$ 表示为

$$\boldsymbol{v}(\hat{\boldsymbol{x}})=\boldsymbol{W}_e\, softmax(\boldsymbol{V}\boldsymbol{h}_{\text{new}}\circ L), \tag{7}$$

其中,$\circ$ 表示元素积运算,$\boldsymbol{V}$ 是用于计算单词分布的权重矩阵,$\boldsymbol{W}_e$ 为模型参数,当 $L\to\infty$ 时,该公式近似于 BiLSTM 的默认输入向量计算公式.

### 2.2　BiLSTM-Attention-CRF 子模型

由于领域内实体名称的多样性,以单词特征及句子特征作为特征学习的模型在进行命名实体识别任务时,可能会出现同一实体在同一文档内标记不一致问题.

子模型 BiLSTM-Attention-CRF 在经典 BiLSTM-CRF 模型的基础上加入 Attention 机制来关注当前实体与文档中其他所有单词的相关性,得到该单词在文档层面的特征表示,以解决实体标记不一致问题.模型结构图如图 1 右侧所示.

用 $D=\{s_1,s_2,\cdots,s_m\}$ 表示文档,文档中的每一个句子 $s_i=(w_1,w_2,\cdots,w_n)$,这里 $w_i$ 表示单词.将文档 $D$ 中所包含的 $N$ 个单词作为 BiLSTM 的输入,得到单词 $w_i$ 的新的表示 $\boldsymbol{h}'_i$(此处 $\boldsymbol{h}'_i$ 与式(1)中 $\boldsymbol{h}_i^{\text{feature}}$ 的计算方法相同),并将其作为 Attention 层的输入,这里 Attention 层主要用来计算当前单词 $w_i$ 与文档中其他单词 $w_j$($j=1,2,\cdots,i-1,i+1,\cdots,N$)的相关性,该 Attention 权重值 $b_{ij}$ 可表示为

$$b_{ij}=\frac{\exp(f(w_i,w_j))}{\sum_{k=1}^{N}\exp(f(w_i,w_k))}, \tag{8}$$

$f(w_i,w_j)$ 的计算方法如式(2)所示.

此时,可以得出单词 $w_i$ 基于文档层面的一个全局特征表示 $\boldsymbol{r}_i^{\text{g}}$:

$$\boldsymbol{r}_i^{\text{g}}=\sum_{j=1}^{N}b_{ij}\boldsymbol{h}'_j. \tag{9}$$

单词 $w_i$ 在 Attention 层的输出 $\boldsymbol{c}_i$ 可表示为

$$\boldsymbol{c}_i=\tanh(\boldsymbol{W}_{\text{g}}[\boldsymbol{r}_i^{\text{g}},\boldsymbol{h}_i]). \tag{10}$$

将 $\boldsymbol{c}_i$ 传递给更上层的 CRF 作为输入,这里 CRF 主要用来预测 2 个部分:一是计算每个 $\boldsymbol{c}_i$ 对应标签的得分 $o_i$;二是通过转换矩阵 $\boldsymbol{T}$(用于定义 2 个连续标签的分数)和 $o_i$ 采用维比特算法来计算出最佳标注序列,其计算过程为

$$o_i=\boldsymbol{W}\boldsymbol{c}_i, \tag{11}$$

$$score(D,y)=\sum_{i=1}^{N}(o_{i,y_i}+T_{y_{i-1},y_i}), \tag{12}$$

$$y^{\text{result}}=\arg\max(score(D,y)), \tag{13}$$

其中,函数 $score()$ 是用来计算输入文档 $D$ 的标签序列 $y=\{y_1,y_2,\cdots,y_N\}$ 的分数,$y^{\text{result}}$ 是最终输出的标签序列的结果(即 BIO 标签),$\boldsymbol{W}$ 表示模型参数.

本文所提出的 BiLSTM-Attention-CRF 子模型除了可以用于识别新文档中的实体之外,还可以将其用于众包标注数据集中,对实体进行再次识别以提高众包标注数据集的质量.

## 3　实验及结果分析

本文的实验目标主要是从 2 个方面来验证 BiLSTM-Attention-CRF-crowd 模型的有效性,在此我们将基线模型根据其要比较的任务不同分为 2 组:1)将本文所提出的模型和第 1 组基线模型应用在信息安全领域的众包标注数据上,来验证 BiLSTM-Attention-CRF-crowd 模型在整合共识标注方面优于其他基线模型的能力;2)将本文所提出的模型和第 2 组基线模型应用在信息安全领域的相关文献上进行特定实体识别,来验证 BiLSTM-Attention-CRF-crowd 模型对同一实体在同一文档中的标注一致能力,并且对该模型和第 2 组基线模型在领域命名实体识别任务上的性能做了对比.

### 3.1　数据来源

本文实验所使用的数据集主要来自于信息安全领域,包括来自于 we live security,threatpost 等处的博客文章、CVE(common vulnerabilities and exposures)描述、微软安全公告以及信息安全类文章摘要,我们从中摘取了 10 187 条句子(其中连续段落包括 20 篇摘要、45 篇博客文章、59 段 CVE 描述以及 50 篇微软安全公告),将每条句子分配给 3 个注释者来完成,以此作为众包标注数据集.标注者只需要从句子中标注出 4 种类型的命名实体:product,vulnerability,attacker,version.此外,由 2 名

专家来标注其中随机抽取的 1 000 条句子,以训练 GAN 模型中的判别模型.

### 3.2 基线模型分组

将基线模型分为 2 组进行实验.

第 1 组.学习众包标注数据的共同特征,我们考虑使用下面 2 个模型作为比较模型.

1) 多数投票(majority vote, MV).即文献[5]提到的一种方法.

2) Dawid&Skene 模型.即文献[16]用到的模型.

第 2 组.在未标注文本上预测命名实体序列,考虑使用下面 4 个模型作为比较模型.

1) BiLSTM-Attention-CRF.即文献[8]中用到的模型,使用未加处理的标注者标注数据直接训练.

2) BiLSTM-Attention-CRF-VT.即文献[8]中用到的模型,使用通过多数投票法选择出的可用标注数据训练.

3) Dawid&Skene-LSTM.即文献[18]用到的模型.

4) CRF-MA.即文献[5]用到的模型,我们使用了该文作者提供的源代码.

### 3.3 实验结果评价

实验中所采用的评价指标分别为准确率(用 $P$ 表示)、召回率(用 $R$ 表示)以及 $F1$ 值.

#### 3.3.1 整合众包标注数据模型性能比较

1) 本模型在众包标注数据集中对不同类型实体标注整合性能评价.

为了验证本文提出模型对不同类型实体标注上的整合性能,针对 3.1 节提出的 4 种类型实体在众包标注数据集中的整合结果进行了对比.从图 2 中可以看出,BiLSTM-Attention-CRF-crowd 模型在类型 product 和类型 version 上表现性能较好,主要原因在于类型 product 表示的类别是产品,一般情况下,这种类型的实体虽然属于领域内实体,但是对于大众来说认知度比较高,因此标注的准确率相对也比较高;而类型 version 有相对固定的模式,比如出现在产品名之后,通常用数字表示,对于这类有固定模式的实体,因为其特征更容易学习,因此模型表现的性能最好.类型 vulnerability 和类型 attacker 属于信息安全领域内相对较为专业的实体类型,在众包标注数据集中标注的准确率较低,并且没有固定的模式可以遵循,因此性能表现稍弱.

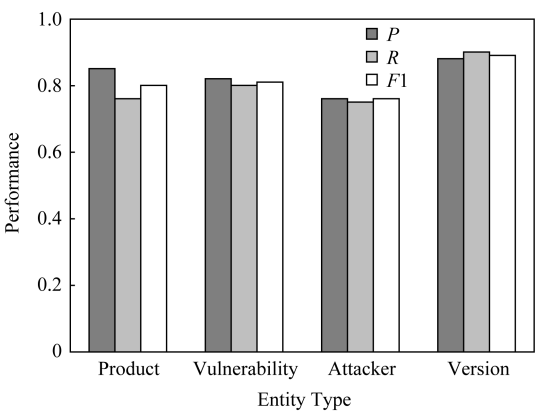


Fig.2 Performance comparison of models with different types of entities

图 2 模型关于不同类型实体的性能对比

2) 本模型与其他比较模型的性能比较

考虑使用各个模型从训练语料中获得的正确标注语句的正确率作为评价标准,正确率可计算为

$$\text{正确率(accuracy)}=\frac{\text{正确标注语句条数}}{\text{训练语句总条数}}.$$

实验结果如表 1 所示:

**Table 1 Model Performance Comparison**

**表 1 模型性能对比表** %

| Method | Accuracy |
|---|---|
| MV[5] | 65.3 |
| Dawid&Skene[16] | 72.5 |
| BiLSTM-Attention-CRF-crowd | 78.9 |

由表 1 可以看出,多数投票(MV)的性能相对较差,这是因为领域的专业性以及标注者标注水平的高低分布不均,对于一些模糊性的实体识别很难达到统一;本文所提出的模型表现最好,该模型除了可以获得训练语料库中原有的正确标注语句之外,还可以将训练语料库中一些原本不正确的标注语句通过反复优化之后的生成模型生成与专家标注数据特征分布趋于一致的正样本数据,并且通过 BiLSTM-Attention-CRF 子模型的再次提取,提高了训练语料库中正确标注语句的正确率.

#### 3.3.2 领域命名实体识别模型性能比较

由于本文所提出的模型在进行命名实体识别任务时主要针对同一实体在同一文档中前后标记的不一致问题,因此我们需要从未标注语料库中选择段落或者文档作为输入.为保证实验的客观性,我们随机选择了 50 篇摘要、20 篇博客以及 20 篇微软安全公告作为测试语料.

1) 各模型对实体标注一致性的性能比较

本文从各文档中选取出 4 种信息安全领域内比较常见的实体,如表 2 所示:

**Table 2　Entity Name and Their Abbreviations**

**表 2　实体名及对应缩写表**

| Full Name | Abbreviation |
|---|---|
| Advanced Persistent Threat | APT |
| FireWall | FW |
| Zero day attack | 0 Day Attack |
| Intrusion Detection Systems | IDS |

由图 3 中可以看出 BiLSTM-Attention-CRF-crowd 模型的性能相对来说要优于其他模型,其中以在实体 APT 和实体 IDS 上性能表现最为突出,主要原因在于实体 APT 和实体 IDS 对应的实体在文档中出现次数相对较多,并且通常与整个文档的内容有密切的联系,此时通过 Attention 层所计算出的该单词与文档层面全局向量的关系更为密切,从而得出的特征表示更为显著,此时模型的性能相较于其他模型来说有明显提高.

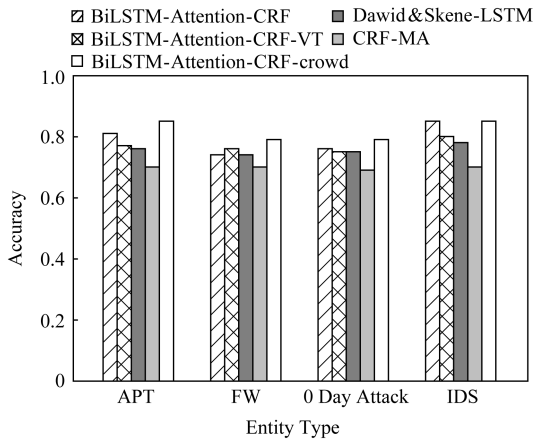


Fig. 3　Model performance comparison

图 3　各模型性能比较图

2) 各模型对领域内命名实体识别综合性能比较

由表 3 可以看出,就正确率而言,直接使用不加处理的众包标注数据作为训练数据所训练出的 BiLSTM-Attention-CRF 模型和 BiLSTM-Attention-CRF-crowd 模型的表现较为出色,这也意味着使用众包标注数据来对 NER 模型进行训练确实是可取的,而以多数投票法选择语料标注作为训练数据的 BiLSTM-Attention-CRF-VT 模型以及通过混淆矩阵和最大似然估计来确定正确标注的 Dawid&Skene-LSTM 模型表现较为平庸.BiLSTM-Attention-CRF-VT 模型表现不佳的原因可能是由于信息安全领域内文本语句的复杂性和专业性,很多实体是无法通过投票选取出来的,另外还因为投票选择所丢掉的上下文信息很可能是重要的,丢掉了这些上下文信息很可能就丢掉了重要的特征信息.Dawid&Skene-LSTM 模型则更注重对标注者标注质量的估计,但是事实上标注者的标注质量通常并非是稳定的,会受到各种情况的影响,另外该模型对初始值的设置也并不理想,影响最后结果收敛到最优解.

**Table 3　Comprehensive Performance Evaluation**

**表 3　各模型综合性能评价表**　　%

| Model | *P* | *R* | *F*1 |
|---|---|---|---|
| BiLSTM-Attention-CRF[8] | 87.7 | 80 | 83.67 |
| CRF-MA[5] | 81.97 | 76 | 78.87 |
| Dawid&Skene-LSTM[18] | 79.23 | 74.48 | 76.78 |
| BiLSTM-Attention-CRF-VT[8] | 80.7 | 75.1 | 77.8 |
| BiLSTM-Attention-CRF-crowd | 89.3 | 85.2 | 87.2 |

结合以上实验,本文所提出的对抗式学习模型 BiLSTM-Attention-CRF-crowd 的性能要优于本文所引用的其他模型,有较为出色的表现能力.

## 4　结　　论

本文的主要工作包括 3 个方面:1)通过 GAN 模型在给定的众包标注数据集上寻找出标注数据的共有特征以生成最佳的唯一共识标注,解决目前众包数据中准确率不高、标注的不一致性等问题;2)将这些通过 GAN 模型所生成的注释数据来作为训练数据,训练模型进行命名实体识别任务;3)提出 BiLSTM-Attention-CRF-crowd 模型解决统一实体在同一文档中的标注不一致问题.我们在信息安全领域的数据集上评估了本文所提出的模型,结果表明:其性能优于本文中所提到的其他作为基线的模型.

目前,BiLSTM-Attention-CRF-crowd 模型主要应用于对名词及名词性短语类型的实体进行识别,对于识别领域内一些有特殊要求的类型实体,例如安全领域本体 UCO[22] 中的 consequence 类,该类别下的实体类型通常为动词短语(如 steal login credentials,control the system 等),模型尚待进一步研究.

## 参　考　文　献

[1] Ratinov L, Roth D. Design challenges and misconceptions in named entity recognition [C] //Proc of the 13th Conf on Computational Natural Language Learning. Stroudsburg: ACL, 2009: 147-155

[2] Finkel J R, Grenager T, Manning C. Incorporating non-local information into information extraction systems by Gibbs sampling [C] //Proc of the 43rd Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2005: 363-370

[3] Hal D III. Frustratingly easy domain adaptation [C] //Proc of the 45th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2007: 256-263

[4] Snow R, O'Connor B, Jurafsky D, et al. Cheap and fast-but is it good? evaluating non-expert annotations for natural language tasks [C] //Proc of the Conf on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2008: 254-263

[5] Rodrigues F, Pereira F, Ribeiro B. Sequence labeling with multiple annotators [J]. Machine Learning, 2014, 95(2): 165-181

[6] Zhao Hai, Chunyu K. Unsupervised segmentation helps supervised learning of character tagging for word segmentation and named entity recognition [C] //Proc of the 6th SIGHAN Workshop on Chinese Language Processing. Stroudsburg: ACL, 2008: 106-111

[7] Huang Zhiheng, Xu Wei, Yu Kai. Bidirectional LSTM-CRF models for sequence tagging [J/OL]. arXiv preprint arXiv: 1508.0191 2015

[8] Ramamoorthy S, Murugan S. An attentive sequence model for adverse drug event extraction from biomedical text [J]. arXiv preprint arXiv: 1801.0625 2018

[9] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets [C] //Proc of Neural Information Processing System. New York: Curran Associates, 2014: 2672-2680

[10] Wang Wanliang, Li Zhuorong. Advance in generative adversarial network [J]. Journal on Communications, 2018, 39(2): 135-148 (in Chinese)
(王万良, 李卓蓉. 生成式对抗网络研究进展[J]. 通信学报, 2018, 39(2): 135-148)

[11] Zhang Yizhe, Gan Zhe, Lawrence C. Generating text via adversarial training [C] //Proc of the 30th Neural Information Processing Systems Workshop on Adversarial Training. New York: Curran Associate, 2016: 2852-2858

[12] Yu Lantao, Zhang Weinan, Wang Jun, et al. SeqGAN: Sequence generative adversarial nets with policy gradient [C] //Proc of the 31st AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2017. arXiv: 1609.05473

[13] Li Jiwei, Will M, Shi Tianlin, et al. Adversarial learning for neural dialogue generation [J]. arXiv preprint arXiv1701.06547, 2017

[14] Kusner M J, Hernandeslobato J M. GANS for sequences of discrete elements with the gumbel-softmax distribution [J]. arXiv preprint arXiv1611.04051, 2016

[15] Dredze M, Talukdar P P, Crammer K. Sequence learning from data with multiple labels [C/OL] //Proc of ECML/PKDD Workshop on Learning from Multi-Label Data. 2009: 39-48 [2018-08-15]. https://www.academia.edu/2809854/Sequence_learning_from_data_with_multiple_labels

[16] Dawid A P, Skene A M. Maximum likelihood estimation of observer error-rates using the EM algorithm [J]. Applied Statistics, 1979, 28(1): 20-28

[17] Kajino H, Tsuboi Y, Kashima H. A convex formulation for learning from crowds [C] //Proc of the 26th AAAI on Artificial Intelligence. Menlo Park, CA: AAAI, 2012: 73-79

[18] Nguyen A T, Wallace B C, Li J J, et al. Aggregating and predicting sequence labels from crowd annotations [C] //Proc of the Association for Computational Linguistics. Stroudsburg: ACL, 2017: 299-309

[19] Yang Yaosheng, Zhang Meishan, Chen Wenliang, et al. Adversarial learning for Chinese NER from crowd annotations [C] //Proc of the 32nd AAAI Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2018: 1627-1634

[20] Robert L, Chih-Hsuan W, Lu Zhiyong. tmChem: A high performance approach for chemical named entity recognition and normalization [J]. Journal of Cheminformatics, 2015, 7(1): S1-S3

[21] Luo Ling, Yang Zhihao, Yang Pei, et al. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition [J]. Bioinformatics, 2017, 34(8): 1381-1388

[22] Syed Z, Padia A, Finin T, et al. UCO: A unified cybersecurity ontology [C] //Proc of the 36th AAAI Workshop on Artificial Intelligence for Cyber Security. Menlo Park, CA: AAAI, 2016: 1381-1388

**Zhang Han**, born in 1985. PhD candidate in software engineering at Strategic Support Force Information Engineering University. Lecturer in Zhengzhou University. Her main research interest is natural language processing.

**Guo Yuanbo**, born in 1975. Received his PhD degree in computer science from Xidian University, China, in 2005. Professor in Strategic Support Force Information Engineering University. Member of IEEE, senior member of Chinese Institute of Electronics and senior member of CCF. His main research interests include insider threat detection, security analytics and security architectures. (yuanbo_g@hotmail.com)

**Li Tao**, born in 1992. PhD candidate in computer science and technology at Strategic Support Force Information Engineering University. His main research interest is knowledge graph. (1527105421@qq.com)