计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.190500063

基于增强 BiLSTM-CRF 模型的推文恶意软件名称识别

古雪梅[1]　刘嘉勇[1]　程芃森[1,2]　何　祥[1]
1 四川大学网络空间安全学院　成都 610000
2 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
(xmg2262@163.com)

摘　要　针对推文中恶意软件名称识别任务存在的文本简短、非正式、实体类别单一以及实体歧义等问题,提出了一种基于BERT-BiLSTM-Self-attention-CRF的实体识别方法,以实现推文中恶意软件名称的自动识别。在BiLSTM-CRF模型的基础上,利用BERT模型编码单词语境信息,提升词嵌入的上下文语义质量,增强原有模型的语义消歧能力;同时,借助Self-attention机制学习单词间关系和句子结构特征,利用加权表征帮助单一类别实体的解码,以提升恶意软件名称实体的识别效果。通过构建包含恶意软件名称实体的推文标记数据集进行实验测试,结果表明,提出的方法可以实现更好的性能,其精确率、召回率、F1值分别为86.38%,84.73%,85.55%,相较于基线模型BiLSTM-CRF,F1值提升了12.61%。

关键词:恶意软件名称识别;实体消歧;动态词嵌入;类别不均;重要性加权
中图法分类号　TP391

Malware Name Recognition in Tweets Based on Enhanced BiLSTM-CRF Model

GU Xue-mei[1],LIU Jia-yong[1],CHENG Peng-sen[1,2] and HE Xiang[1]
1 School of Cybersecurity,Sichuan University,Chengdu 610000,China
2 Key Laboratory of Network Assessment Technology,Institute of Information Engineering,Chinese Academy of Sciences,Beijing 100093,China

Abstract　To address the problems such as short,informal,single entity category and entity disambiguation in the malware name recognition task on Twitter,this paper proposed an entity recognition method based on BERT-BiLSTM-Self-attention-CRF to automatically recognize malware name in tweets. Based on the BiLSTM-CRF model,the BERT is used to encode context information,improve the contextual semantic quality of word embeddings,and enhance the semantic disambiguation ability. At the same time,Self-attention mechanism is used to learn weighted representation to improve the performance of single entity category recognition by learning the long-term relations between words and sentence structure. To evaluate the proposed methods,this paper constructed a labeled dataset in tweets that contains malware name entities. Experimental results show that the proposed method can achieve a better performance,attain 86.38% precision,84.73% recall and 85.55% F-score. The proposed model can outperforms the baseline model,with F-score improved by 12.61%.

Keywords　Malware name recognition,Entity disambiguation,Dynamic word embedding,Class imbalance,Importance weighting

1　引言

随着社交媒体的快速发展,Twitter已成为开源情报的重要来源,被用于恐怖袭击、政府选举和股市预测等[1]。在安全领域,通过自动分析海量推文向安全系统或分析人员提供威胁情报,可减少未知攻击的响应时间。其中,自动识别推文中的恶意软件名称,可减轻人工分析的压力,同时为分析人员快速定位新的安全威胁及关联同类型或不同命名规则的同款恶意软件等提供帮助。恶意软件名称具有独特的语言构成特点,由现有词汇、新词或特殊单词组合构成,因此恶意软件名称的识别需要消除多义词的歧义。

从推文中识别恶意软件名称是一项典型的命名实体识别(Named Entity Recognition,NER)任务。不同于新闻这类规范文本,推文字符长度的限制,导致上下文语义缺乏,使得单词语义消歧困难。此外,推文的非正式性和口语化环境使得推文文本中包含大量不规范的拼写、缩写和不可靠的大小写,造成了句法特征混乱。通用领域中的主流NER系统在Twitter NER任务上表现不佳[2],因此,面向Twitter的NER更具挑战性。

恶意软件名称面临语义消歧的问题,如"Everything you

到稿日期:2019-05-14　返修日期:2019-07-17　本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。
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)
This work was supported by Open Research Fund of the Key Laboratory of Network Assessment Technology of Chinese Academy of Sciences (NST-18-001).
通信作者:刘嘉勇(ljy@scu.edu.cn)

need to know about the Anna Kournikova worm…"中的 Anna Kournikova 指代一款蠕虫病毒,而它在"Hello My name is Anna Kournikova You killed…"中则指代一个人名。同时,推文的恶意软件名称识别只关注单一实体类别,即使是大型标记语料,也仅包含少量实体样例,实体与非实体的类别不均衡导致恶意软件名称实体难以被正确识别。

为了解决上述问题,本文提出了一种基于 BERT-Bi-LSTM-Self-attention-CRF 的恶意软件名称识别模型(简称 ehBiLSTM-CRF 模型)。该模型利用 BERT 生成的基于当前输入语境特征的动态词嵌入向量来提升词嵌入的语义质量,消除多义词间的歧义;使用自注意力(Self-attention)机制学习句子的内部结构和依赖关系,有助于解决类别不均衡的问题,提升识别效果,从而在不手动构建特征的情况下实现推文的恶意软件名称识别。

2 相关研究

为解决 Twitter NER 中缺乏上下文语义,包含大量噪音、口语化和非正式单词等问题,许多研究者通过引入大量特征来辅助识别命名实体。Le 等[3]提出了一种基于词法化语义网络和本体的多义计数的新特征提取方法。该系统引入拼写(如词缀、大写、标点符号和数字等)、词汇、句法、词性标注、一词多义计数和最长 n-gram 长度等特征,使用 CRF 训练序列标注模型。Mahmood 等[4]采用了监督学习和字典相结合的方式,首先使用 CRF 模型标注实体,然后使用精确和模糊搜索字典的结果辅助模型分类,以提高命名实体的识别效果。Liu 等[5]提出了结合 KNN 分类器和 CRF 模型的混合方法,将人工标记推文作为种子,使用 Bootstrapping 半监督学习方法解决推文 NER 中缺乏训练语料的问题。Ritter 等[6]提出了一种使用主题模型的远程监督新方法,将 CRF 与多特征相结合来实现实体分割;然后将从 Freebase 收集的大量实体及其类型的列表作为远程监督源,结合 LabeledLDA 对大量未标记数据进行学习,从而实现实体分类。

传统机器学习方法中的特征构建是一项耗时的工作,为减小模型对传统特征工程的依赖,许多学者将深度学习方法应用于 Twitter NER 任务中。Okur 等[7]采用了一种基于神经网络的半监督学习方法,使用 Word2vec 训练未标记语料的词嵌入,将其作为特征来实现土耳其语推文的命名实体识别。Zhang 等[8]提出了一种使用 CRF 和自适应 co-attention 网络扩展 BiLSTM 的方法,该方法使用自适应 co-attention 网络在 Twitter NER 中融合文本和视觉信息。Limsopatham 等[9]使用 C,c,n,p 替换单词中的大写、小写、数字和标点,生成单词的拼写表示,然后将原句子、拼写表示的句子的词嵌入和字符嵌入作为 Bi-LSTM 模型的输入,自动学习拼写特征,实现推文的命名实体识别。该方法在 WNUT 2016 数据集的 10 类实体识别任务上取得了最优的效果,$F1$ 值为 52.41%。

Devlin 等[10]提出的 BERT 模型,是一个基于 Transformer 的多层双向编码器语言模型。该模型在学习过程中充分考虑双向上下文特征,能有效地学习词汇的语境信息,有助于实体的识别。通过在 BERT 模型后加入分类层,$BERT_{LARGE}$ 在 CoNLL 2003 命名实体识别数据集上取得了 92.8%的 $F1$ 值,相较于 CVT+Multi 模型的 $F1$ 值提升了 0.2%。

Vaswani 等[11]、Shen 等[12]和 Tan 等[13]使用 Self-attention 机制获取句子中的长距离依赖,并在机器翻译、语言理解、语义角色标注等任务中获得了优异的效果。Cao 等[14]在中文命名实体识别中引入自注意力机制,学习句子的内部结构和单词间依赖关系。自注意力机制通过选择性地关注重要信息,学习句子中任意词间的关系,捕获更多层面的特征,能在一定程度上帮助解决类别不平衡问题。

现有的 Twitter NER 通过使用深度学习的方法减少了对传统特征工程的依赖。Limsopatha 等[9]的研究证明了 BiLSTM-CRF 模型解决推文的噪音、非正式等问题的有效性。但是,恶意软件名称识别还面临着语义消歧和单一实体类别不均衡的问题,对此,本文结合 BERT 的语义消歧能力和自注意力机制解决实体类别不平衡的问题,提出了基于 ehBiLSTM-CRF 的推文恶意软件名称识别方法。

3 恶意软件名称识别

不同于传统 NER 任务,推文中的恶意软件名称识别任务存在多义词歧义和实体类间不均衡的问题,往往导致典型的 BiLSTM-CRF 模型在该任务上表现不佳。针对以上问题,本文将多义词语义过滤和单一类别实体识别融入到基于 BiLSTM-CRF 模型的学习中,利用 BERT 生成的动态词嵌入向量辅助消除同义词歧义,同时结合 Self-attention 学习的单词对的加权表征来缓解实体类间不均衡的问题,进而提出了包含 BERT 层、BiLSTM 层、Self-attention 层以及 CRF 层的 ehBiLSTM-CRF 模型。

3.1 多义词语义过滤

不同语义环境下,Anna Kournikova 这类多义词在典型实体识别任务中可能是不同类别的待识别实体,因而在推文恶意软件名称识别任务中需要充分利用单词上下文的语义信息准确识别出具有恶意软件名称语义类别的实体,忽略其他语义类别实体。

one-hot 是一种直观的词表示方法,它使用以词表大小为维度的向量表征单词,词表索引对应的位置为 1,其他位置为 0。one-hot 向量的维度随词表的增大而增大,且词与词是相互独立的,而词嵌入方法可以有效地降低词向量空间的维度,并将语义融入到词表示中。Word2vec, FastText 和 Glove[15-17]是 3 种广泛使用的词嵌入方法,它们使用同一个词嵌入向量表征不同上下文中的同一单词,在本文的任务中无法很好地体现不同上下文的语义差异;同时,推文中包含更多由拼写变化产生的未登录词汇,导致预训练词的向量覆盖率不高,应用于推文的恶意软件名称识别时效果不佳。而 BERT 是一种基于 Transformer 的深层双向编码器语言模型,可以充分学习语料的字符级、词级、句子级及句间关系特征,为输入文本的每个单词生成基于当前语境上下文的动态词嵌入向量,解决了传统词嵌入方法将不同语境中的同一单词映射到相同语义空间的问题,提升了语义消歧能力。

3.2 单一类别实体识别

在推文的文本中，恶意软件实体类和非实体类的样例数量存在严重的不均衡问题，由于实体类训练样例太少，大量恶意软件名称实体被错误地标记为非实体，严重影响了识别效果。

为了解决类间不均的问题，本文在推文恶意软件名称识别任务中引入多头自注意力机制来学习更丰富的文本局部特征。多头自注意力机制是指将查询 Q、键 K、值 V 进行 h 次不同的线性转换，其中 $Q=K=V$；然后将 h 次的变换结果输入到放缩点积 *attention*（scaled dotproduct attention）中进行并行运算，连接放缩点积 *attention* 结果后再进行一次线性转换；最终得到加权的特征向量。

通过为不同影响程度的词分配不同的权重，多头自注意力机制可以集中关注句子中的重要信息而忽略其他不重要的部分，从而减小不重要信息对结果的影响，有助于正确识别命名实体。

3.3 ehBiLSTM-CRF 模型结构

ehBiLSTM-CRF 模型由 BERT 层、BiLSTM 层、Self-Attention 层及 CRF 层组成，模型结构如图 1 所示。

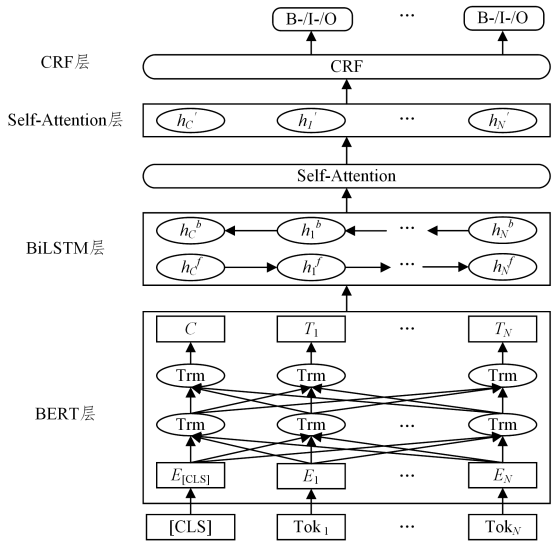

图 1 ehBiLSTM-CRF 模型结构

Fig. 1 Structure of ehBiLSTM-CRF

BERT 层包含 token 嵌入、分段嵌入和位置嵌入 3 部分输入。为了适应 BERT 的输入，每个词在输入 BERT 层前需要经过 wordpiece 操作，并在句子首尾分别嵌入[CLS]和[SEP]两个特殊的 token，如图 2 所示。wordpiece 是指将单词分割成更小的单元，其目的是压缩词表大小，更好地处理未登录词。token 嵌入会将预处理后的 token 转换为固定维度的向量；由于命名实体识别的对象是一个句子，因此在分段嵌入时全部使用 0 进行填充；位置嵌入主要用于解决 Transformer 没有编码序列顺序的问题，该层会学习一个位置向量来表征序列的顺序信息。BERT 层通过联合调节内部的双向 Transformer，利用 Self-attention 在单词编码过程中学习上下文中其余词对当前词的贡献程度，从而增强上下文语义信息的提取。最终，编码生成基于当前语境上下文的词嵌入向量。

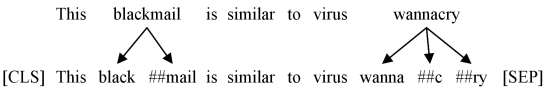

图 2 预处理示例

Fig. 2 Example of preprocessing

BiLSTM（Bi-directional Long Short-Term Memory）层包含一个前向 LSTM 和一个后向 LSTM，分别用于学习序列中各个 token 的左右上下文信息。BiLSTM 层接收 BERT 层的词嵌入向量作为输入。在每个时间步 t 中，前向隐藏状态 $\overrightarrow{h_t}$ 由前一时刻的隐藏状态 $\overrightarrow{h_{t-1}}$ 和当前输入 x_t 计算得到，后向隐藏状态 $\overleftarrow{h_t}$ 由后一时刻的隐藏状态 $\overleftarrow{h_{t+1}}$ 和当前输入 x_t 计算得到，连接 $\overrightarrow{h_t}$ 和 $\overleftarrow{h_t}$ 得到 t 时刻的隐藏状态 $h_t=[\overrightarrow{h_t};\overleftarrow{h_t}]$。当输入句子 $S=\{w_1,w_2,\cdots,w_n\}$ 时，输出句子的隐藏状态序列 $H=\{h_1,h_2,\cdots,h_n\}$。

Self-attention 层接收 BiLSTM 层的隐藏状态序列作为输入。该层包含查询 Q、键 K、值 V 3 个输入，其中 $Q=K=V=H=\{h_1,h_2,\cdots,h_n\}$，$H$ 是 BiLSTM 层的输出。首先对输入进行多次不同的线性转换，然后将变换后的结果输入到放缩点积 *attention* 中进行并行运算，连接放缩点积 *attention* 结果并再次变换得到新的状态序列，将其输出到 CRF 层中。通过 Self-attention 的词对加权表征，挖掘词间的关联关系，缓解实体类别过少引起的误识别问题，有助于提升恶意软件名称实体类的解码效果。

CRF 层使用条件随机场（Conditional Random Filed，CRF）模型[18]对 Self-attention 层的输出进行解码，通过学习标签间的约束条件提升标签预测的准确性，得到最终的预测标签序列。

4 实验及分析

本文以 NER 系统中广泛使用的精确率（Precision，P）、召回率（Recall，R）和 F1 值（F-score）作为评价指标。其中，精确率指实体边界预测正确的个数与预测结果中实体数的比值；召回率指实体边界预测正确的个数与测试集中标注实体数的比值；F1 值指精确率和召回率的调和平均数。

4.1 数据集

由于缺乏公开的推文恶意软件名称标注数据集，我们人工建立了一个 Twitter 平台上的恶意软件名称标记数据集。构建方式如下：

1）数据采集。以 wikipedia（Timeline of computer viruses and worms）和互联网安全报告为源，获取了 156 个恶意软件名称的列表，并以该列表内的条目为关键词，使用 Twitter Intellifence Tool（TWINT）工具以搜索方式获取可能含有恶意软件名称的推文。

2）数据预处理。对采集的数据进行去重及语言过滤，参考 Belainine 等[19]的预处理方法，除去 hashtags 前面的 # 标记，同时使用 userid，http://url.com，1 分别替换推文中的@username、URL 和数字。

3）数据标注。使用 IOB2 标记策略，通过正则匹配自动

标注推文中的恶意软件名称，其中 O 表示非实体，B-malware 表示单字实体或多字实体首字，I-malware 表示多字实体的非首字。然后由人工校验标注结果，生成数据集。

4)数据集基本信息。该数据集共包含 1319 个恶意软件命名实体和 461 种恶意软件名称。

本文采取 5 折交叉验证法来验证模型的有效性，将标记数据集随机均分成 5 个子集$\{p_1, p_2, p_3, p_4, p_5\}$，然后依次选择其中 1 个子集作为测试集，剩余 4 个子集作为训练集(见图 3)，最后计算不同测试、训练集下各评估指标的均值作为模型的最终效果。

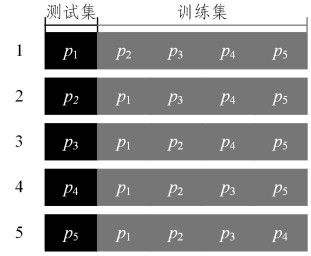

图 3 5 折交叉验证

Fig. 3 Five-fold cross-validation

4.2 实验环境及参数

实验采用 Python 编程，操作系统为 Ubuntu 16.04.4 LTS，64 位，处理器为 32 Intel(R) Xeon(R) CPU E5-2620 V4 @ 2.10GHz，内存大小为 47GB。

实验中预训练词向量的维数为 300，各方向 LSTM 隐藏节点数均为 100，Self-attention 层头数为 8。训练过程中使用 Adam 作为参数优化算法，学习率为 $5 * 10^{-5}$。为避免过拟合，在 BiLSTM 的输入、输出以及 Self-attention 层中使用了 Dropout 率为 0.5 的正则化方法。

4.3 实验分析

为了验证多义词语义过滤、单一类别实体识别及本文所提方法的性能，设置了 3 组对比实验。

1)多义词语义过滤实验。为了验证 BERT 结构的有效性，进行了不同向量/词嵌入方式的对比实验。将 BiLSTM+CRF 模型作为基准模型，分别使用 one-hot 向量、Word2vec、FastText、Glove 以及 BERT 这 4 种词嵌入方式作为 BiLSTM 的输入，测试不同向量/词嵌入方法在本文的推文恶意软件名称标记数据集上的性能。

2)单一类别实体识别实验。在 BERT+BiLSTM+CRF 模型的 BiLSTM 层后引入 Self-attention 层，对比 Self-attention 层对推文恶意软件名称识别性能的影响。

3)与现有工作的对比实验。将本文方法与现有其他工作进行对比，验证所提方法在推文恶意软件名称识别任务中的有效性。

由于缺少不同词嵌入方法在同一语料上训练的预训练向量，为控制无关变量的一致性，多义词语义过滤实验和单一类别实体识别实验使用英文维基百科(32M tokens)语料作为预训练语料，分别训练了 Word2vec，FastText，Glove 3 种预训练词嵌入，以及 BERT 的预训练模型，相关参数参考 Bojanowski 等[20]、Pennington 等[21]、Devlin 等[10]的工作。在与现有工作的对比实验中，为使模型达到更好的效果，ehBiLSTM-CRF 模型使用了 Devlin 等提供的在大规模语料上训练的预训练模型 BERT-Base[10]。

4.3.1 多义词语义过滤实验

为验证 BERT 层对语义消歧能力的提升，对比了使用 one-hot 向量作为 BiLSTM+CRF 模型输入和使用不同词嵌入方式作为 BiLSTM+CRF 模型输入时的实验结果，如表 1 所列。

表 1 不同向量/词嵌入方法的实验结果

Table 1 Results of different vector/word embedding

(单位：%)

向量/词嵌入方法	精确率	召回率	$F1$ 值
one-hot	77.62	68.80	72.94
Word2vec	61.28	37.58	46.59
FastText	59.64	48.21	53.32
Glove	60.15	42.55	49.84
BERT	77.48	78.23	77.85

通过实验结果可知，使用 one-hot 向量的模型的效果优于使用 Word2vec，FastText，Glove 3 种预训练词嵌入的模型的效果，这可能是因为在小规模预训练语料上训练的词嵌入质量不高，并且过多的未登录词导致预训练词向量的覆盖率不高。而基于 BERT 的方法在少量预训练语料上的精确率仅比 BiLSTM+CRF 低了 0.14%；而召回率和 $F1$ 值有了较大改善，分别提升了 9.63%和 4.91%。

与 Word2vec，FastText，Glove 预训练词嵌入方法相比，基于 BERT 的方法的 $F1$ 值分别提升了 31.26%，24.53%，28.01%。下面对比分析不同词嵌入模型的预测样例。如图 4 所示，相比其他模型无法识别或漏识别实体的情况，BERT+BiLSTM+CRF 模型正确识别出了推文中所有的 LoveLetter 和 Melissa 实体，表明了 BERT 的双向结构和动态向量表征能有效地学习更丰富和准确的语义信息，有助于提升推文恶意软件名称识别任务的消歧能力。

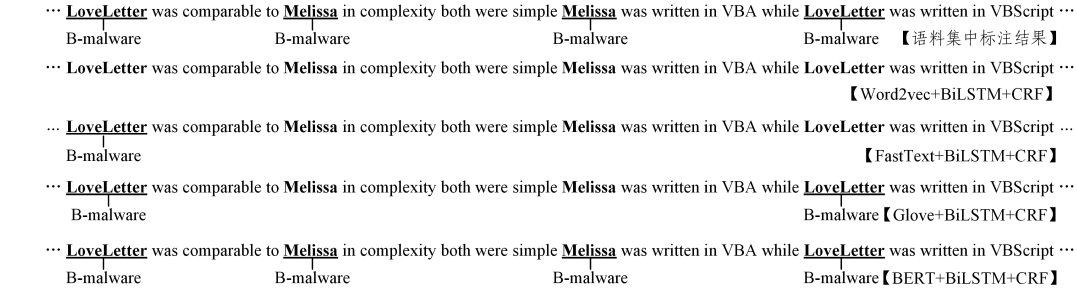

图 4 消歧实例

Fig. 4 Example of disambiguation

4.3.2 单一类别实体识别实验

为验证 Self-attention 对单一类别实体识别性能的提升，对比了未引入 Self-attention 层的 BERT＋BiLSTM＋CRF 模型和引入了 Self-attention 层的 BERT＋BiLSTM＋Att＋CRF 模型的识别效果。如表 2 所列，加入 Self-attention 层后，模型的精确率、召回率、$F1$ 值分别提升了 1.25%，1.98%，1.61%。实验结果表明，引入 Self-attention 机制能够识别出更多的实体，有助于解决实体被错误识别成非实体的问题。

表 2 Self-attention 层的对比结果

Table 2 Comparison results of Self-attention layer

（单位：%）

方法	精确率	召回率	$F1$ 值
BERT＋BiLSTM＋CRF	77.48	78.23	77.85
BERT＋BiLSTM＋Att＋CRF	78.73	80.21	79.46

通过对比分析 BERT＋BiLSTM＋CRF 模型的错误样例可知，引入 Self-attention 有助于改善以下两种情况。

1）具有并列关系的实体

在图 5 所示的实例中，对于 Stuxnet，Industroyer，CrashOverride 这 3 个并列恶意软件名称实体，BERT＋BiLSTM＋CRF 模型只识别出了 Stuxnet，漏识别了 Industroyer 和 CrashOverride，而 BERT＋BiLSTM＋Att＋CRF 模型成功识别出了 3 个并列实体，这表明 Self-attention 能更好地捕获词对间的关系。

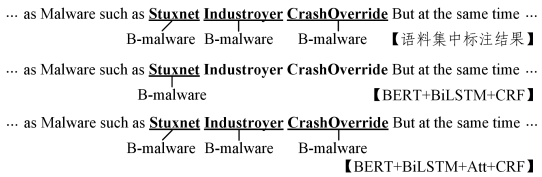

图 5 具有并列关系的实体实例

Fig. 5 Entity instance with parallel relationship

2）具有伴随词的实体

从图 6 的实例中可以发现，BERT＋BiLSTM＋CRF 没有很好地学习到实体后的 Trojan，virus，worms 等伴随词特征，造成模型对实体的判别出现错误。融合了词对加权表征的 BERT＋BiLSTM＋Att＋CRF 模型能够更好地学习伴随词特征，正确识别出伴随词前的实体。

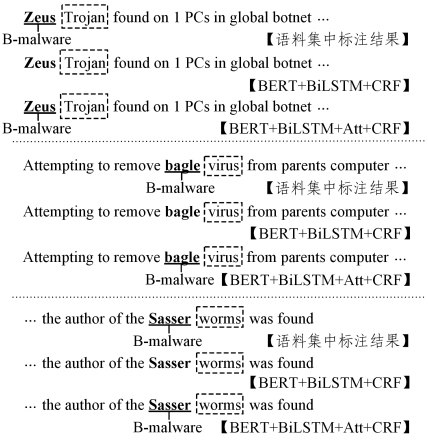

图 6 具有伴随词的实体实例

Fig. 6 Entity instance with accompanying word

4.3.3 与现有其他工作的对比实验

为了验证本文提出的 ehBiLSTM-CRF 模型的性能，将其与现有其他工作在推文恶意软件名称数据集上的效果进行对比。BiLSTM＋CRF 是目前 NER 任务上主流的神经网络模型，该模型使用 one-hot 向量作为输入，未加入其他特征。2016 年，Limsopatham 等提出了 CNN＋BiLSTM＋CRF 模型[9]，该模型在 BiLSTM＋CRF 模型的基础上使用卷积神经网络（CNN）学习单词的字符级嵌入，然后连接原句子，使用符号替换（C-大写，c-小写，n-数字，p-标点）的句子的字符级嵌入和预训练词嵌入作为模型输入，在 WNUT（Workshop on Noisy User-generated Text）的 Twitter NER 任务上取得了较好的成绩，相较于第二名，$F1$ 值提升了 6.2%。

对比实验结果如表 3 所列。与 BiLSTM＋CRF 相比，本文方法在精确率、召回率和 $F1$ 值上分别提升了 8.76%，15.93%，12.61%；与 CNN＋BiLSTM＋CRF 相比，无论是精确率还是召回率，本文提出的模型都取得了更好的效果，$F1$ 值提升了 10.18%。模型复杂度和运行时间的对比结果如表 4 所列。可以看到，虽然 ehBiLSTM-CRF 模型的识别效果有了明显提升，但由于 BERT 层复杂的网络结构，其在参数规模和运行时间两方面较已有工作开销更大。

表 3 本文方法与现有其他工作的对比结果

Table 3 Comparison results between proposed method and other existing works

（单位：%）

方法	精确率	召回率	$F1$ 值
BiLSTM＋CRF	77.62	68.80	72.94
CNN＋BiLSTM＋CRF	80.29	71.02	75.37
ehBiLSTM-CRF	**86.38**	**84.73**	**85.55**

表 4 与现有其他工作的效率对比结果

Table 4 Efficiency results of proposed method and other existing works

Method	Train times	Test time/s	Params memory/MB
BiLSTM＋CRF	5.92	27	1.22
CNN＋BiLSTM＋CRF	9.87	39	1.26
ehBiLSTM-CRF	20.16	82	～110

结束语 本文针对推文文本简短且表达不规范、实体类型分布不均以及实体歧义等问题，提出了一种 ehBiLSTM-CRF 模型，用于推文中恶意软件名称的自动识别。区别于传统的将不同语境上下文的同一单词映射到相同词嵌入空间的方法，本文使用 BERT 作为嵌入层，有效地学习单词基于当前语境的双向上下文词向量，以解决语义消歧问题；同时引入自注意力机制捕获序列中任意单词间的关系及内部结构，选择性地关注重要的信息，以提升实体识别的效果。实验结果显示，提出的模型取得了更好的识别效果，但模型结构也更加复杂。在未来的研究工作中，我们将进一步简化模型结构，提升模型运行效率，同时实现恶意软件其他实体类别的识别。

参 考 文 献

[1] MITTAL S，DAS P K，MULWAD V，et al. Cybertwitter：Using twitter to generate alerts for cybersecurity threats and vulnera-

bilities[C]// Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE Press, 2016: 860-867.
[2] DERCZYNSKI L, MAYNARD D, RIZZO G, et al. Analysis of named entity recognition and linking for tweets[J]. Information Processing & Management, 2015, 51(2): 32-49.
[3] LE N T, MALLEK F, SADAT F. Uqam-ntl: Named entity recognition in twitter messages[C]// Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT). 2016: 197-202.
[4] MAHMOOD T, MUJTABA G, SHUIB L, et al. Public bus commuter assistance through the named entity recognition of twitter feeds and intelligent route finding[J]. IET Intelligent Transport Systems, 2017, 11(8): 521-529.
[5] LIU X, ZHANG S, WEI F, et al. Recognizing named entities in tweets[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011: 359-367.
[6] RITTER A, CLARK S, ETZIONI O. Named entity recognition in tweets: an experimental study[C]// Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011: 1524-1534.
[7] OKUR E, DEMIR H, ÖZGÜR A. Named entity recognition on twitter for turkish using semi-supervised learning with word embeddings[J]. arXiv: 1810. 08732, 2018.
[8] ZHANG Q, FU J, LIU X, et al. Adaptive co-attention network for named entity recognition in tweets[C]// Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
[9] LIMSOPATHAM N, COLLIER N H. Bidirectional LSTM for named entity recognition in Twitter messages[C]// Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT). 2016: 197-202.
[10] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv: 1810. 04805, 2018.
[11] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. 2017: 5998-6008.
[12] SHEN T, ZHOU T, LONG G, et al. Disan: Directional self-attention network for rnn/cnn-free language understanding[C]// Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
[13] TAN Z, WANG M, XIE J, et al. Deep semantic role labeling with self-attention[C]// Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
[14] CAO P, CHEN Y, LIU K, et al. Adversarial Transfer Learning for Chinese Named Entity Recognition with Self-Attention Mechanism[C]// Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. 2018: 182-192.
[15] MEFTAH S, SEMMAR N. A neural network model for part-of-speech tagging of social media texts[C]// Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018). 2018.
[16] JANSSON P, LIU S. Distributed representation, lda topic modelling and deep learning for emerging named entity recognition from social media[C]// Proceedings of the 3rd Workshop on Noisy User-generated Text. 2017: 154-159.
[17] GUPTA D, EKBAL A, BHATTACHARYYA P. A Deep Neural Network based Approach for Entity Extraction in Code-Mixed Indian Social Media Text[C]// Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018). 2018.
[18] LAFFERTY J, MCCALLUM A, PEREIRA F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]// Proceedings of International Conference on Machine Learning. 2001: 282-289.
[19] BELAININE B, FONSECA A, SADAT F. Named entity recognition and hashtag decomposition to improve the classification of tweets[C]// Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT). 2016: 102-111.
[20] BOJANOWSKI P, GRAVE E, JOULIN A, et al. Enriching word vectors with subword information[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 135-146.
[21] 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). 2014: 1532-1543.

GU Xue-mei, born in 1995, postgraduate, is not member of China Computer Federation (CCF). Her main research interests include information content security and so on.

LIU Jia-yong, born in 1962, Ph.D, professor, Ph.D supervisor, is not member of China Computer Federation (CCF). His main research interests include information security, and network communication and security.