基于深度学习的警情记录关键信息自动抽取

崔雨萌，王靖亚，闫尚义，陶知众
中国人民公安大学信息网络安全学院，北京 100038

摘要
随着智慧警务的兴起，民众报警渠道拓宽，非结构化警情激增，警情实体识别难度增大。针对这一业务痛点，引入BERT模型获取词向量，融合自注意力机制来捕获文字之间的长距离依赖关系，并构建BERT-BiGRU-SelfAtt-CRF警情实体识别模型。为了验证模型的性能和泛化能力，在公开数据集上进行了实验。为了验证模型在警情领域的可行性和效率，在构建的警情数据集上进行了实验。实验结果表明，提出的模型在警情数据集上的精确率达到了82.45%，召回率达到了79.03%，F1值达到了80.72%，优于其他模型。可见，提出的模型可以满足实际公安工作需要，是可行、有效的。

关键词
深度学习；预训练语言模型；自注意力机制；警情实体识别

中图分类号：TP391.1 文献标志码：A doi: 10.11959/j.issn.2096-0271. 2022052

Automatic key information extraction of police records based on deep learning

CUI Yumeng, WANG Jingya, YAN Shangyi, TAO Zhizhong
College of Information Network Security, People's Public Security University of China, Beijing 100038, China

Abstract
With the emergence of intelligent policing, the channels of mass to call police are widened, unstructured police records increase immensely, and the difficulty of police entity recognition is magnified. For this pain point, BERT model was introduced to generate the word vector, the self-attention mechanism was integrated to capture the long-distance dependence between words, and the BERT-BiGRU-SelfAtt-CRF police entity recognition model was constructed. In order to verify the performance and generalization ability of this model, experiments were carried out on public datasets. And to prove the feasibility and efficiency of this model in the police field, experiments were run on the annotated police dataset. Ultimately, the results showed that BERT-BiGRU-SelfAtt-CRF model outperformed other models on the police dataset, with the precision of 82.45%, recall rate of 79.03%, and F1 value of 80.72%. It is concluded that this model can meet the requirements of actual police work, and it is feasible and effective in the field of police entity recognition.

Key words
deep learning, pretrained language model, self-attention mechanism, entity recognition in police records

0 引言

随着基层公安机关对社会管控的增强，群众报案的手段从单一的电话报警扩展到语音留言、短信、手机App和微信等，警方接收大量语音、文本等非结构化信息。快速准确地提取报警信息中的关键信息并进行指挥调度成为公安机关亟须解决的业务痛点。因此，公安机关迫切地需要科技手段辅助提取出关键信息以便进行快速地指挥调度。命名实体识别（named entity recognition，NER）是自然语言处理（natural language processing，NLP）的一个重要分支，它是信息提取、机器翻译、信息检索等技术的关键[1]，也是处理和分析警情数据的基础。命名实体识别主要负责对原始文本中具有特定意义的实体进行提取和分类，然后将非结构化的信息转换成半结构化或结构化的信息，最后将信息提供给其他技术，并用于特定领域[2]。在公安实战中，命名实体识别可以从报警记录中提取报警人姓名、案发地址、涉案机构等实体，并将其应用于后续的工作中，如管理涉疫人员、匹配出警单位、分析区域案件趋势、多次报案提醒、累犯重犯记录等。

近年来，在深度学习的基础上实现命名实体识别已成为主流，通过循环神经网络（recurrent neural network，RNN）、卷积神经网络（convolutional neural network，CNN）或其他神经网络模型提取输入文本的特征，通过非线性激活函数学习特征[2]，然后通过条件随机场（conditional random field，CRF）[3]求解最优标注序列。与英文不同，中文文本没有明显的词边界，依赖传统的字符词向量无法解决一词多义问题，分词方法也无法解决分词错误造成的传递错误。因此，找到一种合适的中文分词方法是实现中文警情命名实体识别任务的一个重要研究方向。此外，报警记录的保密性和敏感性导致当前缺少警情实体识别数据集，并且公安领域缺乏统一的标注标准，这极大地增加了本文实施的难度。

鉴于以上问题，本文对中国某市公安局的300条、包括12 513个汉字的实际报警记录进行人工标注，构建标准化警情命名实体识别数据集PRD-PSB；并提出了一种融合自注意力机制（self-attention mechanism）和BERT-BiGRU-CRF的警情实体识别模型——BERT-BiGRU-SelfAtt-CRF。该模型引入BERT（bidirectional encoder representations from transformers）预训练模型来生成包含丰富语义信息的词向量，使用BiGRU（bidirectional gated recurrent unit）来捕捉文本序列的时序特征和上下文语义，并融合了自注意力机制来挖掘文本间的潜在依赖关系，最后使用CRF完成序列标注。在自行标注的警情数据集上进行实验，结果表明，本模型的精确率（precision，P）、召回率（recall，R）和F1值（F1 value，F1）分别达到了82.45%、79.03%和80.72%，该模型的表现较其他基线模型更优。

1 相关工作

在早期，命名实体识别主要是基于字典和规则的（如规则构建或特征工程），但这些方法开销较大且十分依赖具体知识库。之后，命名实体识别逐渐发展成为基于传统机器学习的方法，其通常被转化为序列标注问题。传统机器学习方法主要基于支持向量机（support vector machine，SVM）[4]、CRF[5-6]、隐马尔可夫模型（hidden Markov model，HMM）[7-8]和最大熵（maximum entropy，ME）[9-10]。

近年来，随着词嵌入技术的提出及算力的发展，神经网络能够有效地处理多种命名实体识别任务。在深度学习的基础上，神经网络模型的训练不再依靠传统的特征工程或流水线模式，而是成为一个端到端的过程。这一特点使命名实体识别能够适用于非线性转换，节约成本开销，并能够构建更复杂的网络。

随着深度学习在命名实体识别各方面的广泛使用，能够获取上下文相关信息的RNN模型也被应用于该领域[11]。与RNN相比，长短期记忆（long short-term memory，LSTM）增强了序列记忆能力，并结合CRF组成LSTM-CRF架构，该架构已被广泛应用于中文命名实体识别领域[12-15]。Huang Z H等人[16]提出用BiLSTM和CRF相结合的方式解决序列标注问题，其中BiLSTM可以高效地使用过去和未来的输入特征，CRF则确保模型可以利用句子级的标签信息。Chen Y等人[17]将基于词特征的BiLSTM-CRF应用于中文不良药品实体提取，发现模型的平均F1值高达94.35%。李一斌等人[18]将基于BiGRU-CRF的识别方法应用在中文包装产品实体识别中，实验结果表明，该方法F1值最高可达81.40%，相较于传统序列标注结构和RNN，有更高的准确率和召回率。在人工神经网络的基础上，参考文献[14-15]引进了中文偏旁信息以提高识别准确率，并且参考文献[15,19-20]还采用了注意力机制来增强实体和标签之间的语义关系，进一步优化模型效果。

除此之外，输入数据应转换为计算机可以识别的格式，而且词向量的训练和生成对整个模型提取效果有显著影响。尽管传统的独热编码方式简单，但产生的向量维度高且稀疏，并不能表达出词之间的关系。Mikolov T等人[21]提出的基于分布表示的Word2vec是词嵌入应用的典型，但它不能解决一词多义和词的多层特征问题。在2018年被提出的BERT预训练语言模型[22]可以通过微调为大量任务提供高级模型，并且针对特定任务，只需要新增一个输出层，而不用对模型结构进行大量修改。在中文命名实体识别任务中，将BERT作为词向量层可以出色地提取单词之间的上下文关系，并为特定的子任务提供支持，因此它已被广泛应用于许多中文命名实体识别任务中[3,23-26]。

2 模型构建

神经网络模型的实现和构建需要综合考虑警情文本的短文本性、中文词语边界的模糊性、实体语境的关联性和警情实体识别的实时性等要求。本文以BiGRU-CRF为基本框架，采用BERT预训练语言模型生成中文词向量，并融合自注意力机制来增加上下文相关的语义信息，捕捉文本之间的潜在语义特征。BERT-BiGRU-SelfAtt-CRF的基本架构如**图1**所示，整体提取模型可分为4层。首先，每个输入的文字由3个词嵌入共同表示，BERT层根据每个文字的3个词嵌入的加和生成对应的词向量。之后，通过BiGRU层（前向GRU和后向GRU）模型可以更好地利用输入的过去和未来的特征。然后，自注意力层可以加强对重要信息的捕捉，更好地获取文本长距离依赖关系。最后，利用CRF层实现序列标注，使模型学习到句子的约束条件，有效地利用句子级别的标记信息。

本文的目标是从电子报警记录中提取出报警人姓名、案发地点和涉案机构3类警情实体。具体的流程如下。

第一，对警情数据集进行预处理。数据集$R = \{r_1, r_2, \cdots, r_n\}$，其中$R$表示整个记录数据集，第$i$个记录$r_i$由$<w_{i1}, w_{i2}, \cdots, w_{in}>$组成，$w_{ik}$表示第$i$个记录中的第$k$个中文文字。

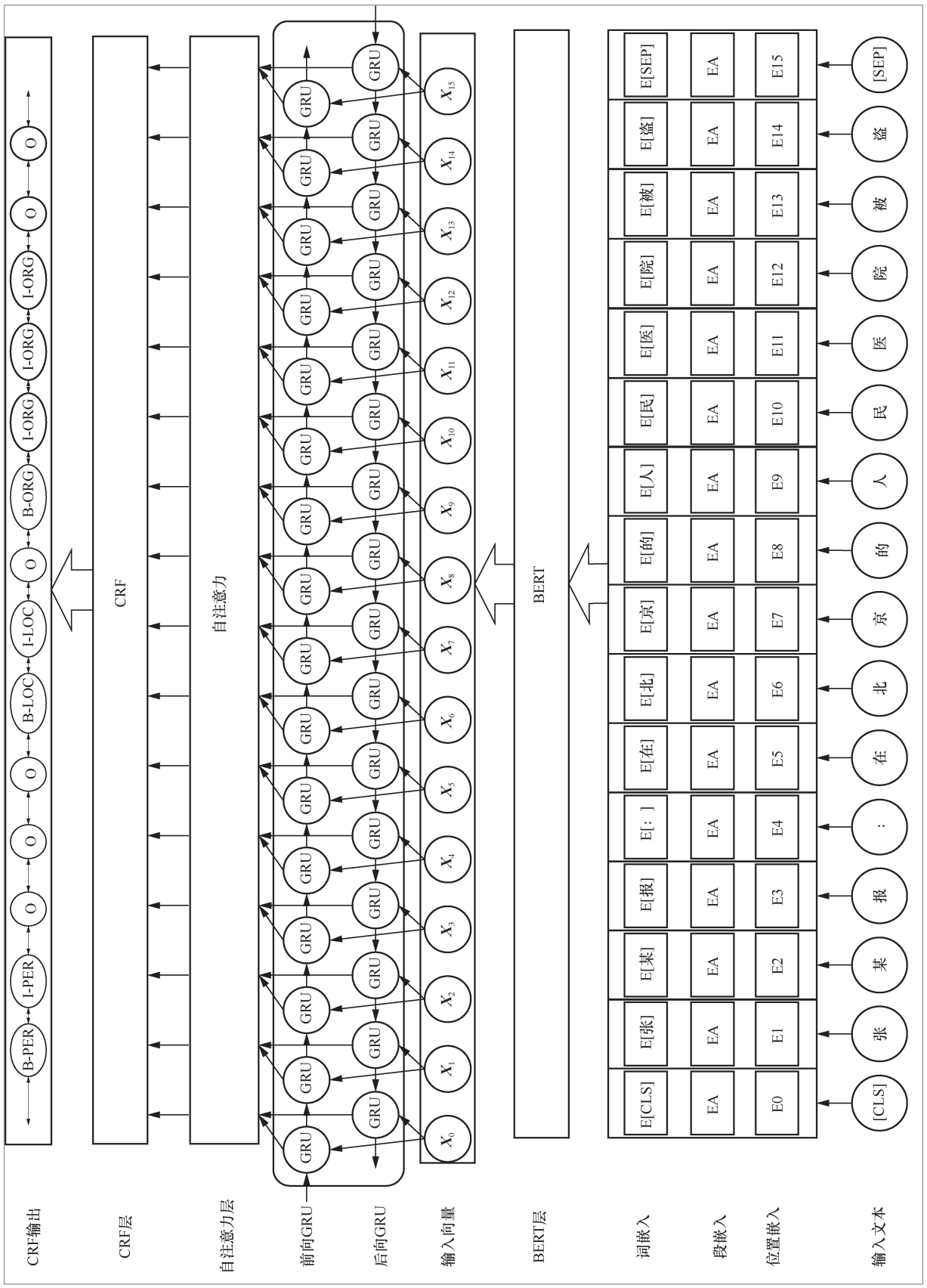


图1 BERT-BiGRU-SelfAtt-CRF的基本架构

第二，构建警情训练数据集。在本文中，采用BIO标记体系来标注训练数据集，其实体类别数据集C={B-PER,I-PER,B-LOC,I-LOC,B-ORG,I-ORG,O}。标记是针对字级别文本进行的，训练数据集中的每一个汉字都用换行符分隔，然后用空格将汉字和对应的标注类别分隔。

第三，训练BERT-BiGRU-SelfAtt-CRF模型。将已标记的训练数据集$D_{\text{training}}=\{w_1,w_2,\cdots,w_n\}$输入模型，其中$w_i$表示训练数据集中的第$i$个中文文字。之后，输出结果集合$P_{\text{predict}}=\{<w_1,c_1>,<w_2,c_2>,\cdots,<w_n,c_n>\}$，其中$c_i$表示第$i$个中文文字的预测类别。然后结合预定义标注类别集合$P_{\text{define}}$，根据精确率、召回率和F1值，对模型进行训练和调整。

2.1 BERT层

BERT是一种深度无监督的双向语言表示模型，在原始未标注文本中，通过对上下文语境进行共同条件化，对所有层进行预训练[22]。如**图1**所示，对于每个给定的汉字，BERT的输入表示由3个词嵌入部分的总和组成，即词嵌入、段嵌入和位置嵌入。**图1**中，Ei为位置嵌入，代表在输入语句中的第i个位置；EA为段嵌入，A表示属于第1句话。此外，Transformer采用了位置编码方式，并加入编码和嵌入数据，从而加入相对位置信息。最终，BERT输出生成的词向量$\boldsymbol{X}_i$。

与传统的单向语言模型或简单地拼接两个单向模型进行预训练不同，BERT采用一种新的掩码语言模型（masked language model，MLM）来生成深层双向语言表征。此外，其采用深度双向Transformer编码器来构建整个模型架构。Transformer[27]采用了自注意力机制，以确保模型的并行计算能力，多头自注意力机制（multi-head self-attention mechanism）使模型能够捕获更丰富的特征，还采用了残差机制来保证计算两个位置之间的相关性所需的操作不会随着距离增加而增加。另外，在预训练阶段，BERT采用了两个训练任务：MLM和下一句预测（next sentence prediction，NSP）。由于其庞大的参数和强大的特征提取能力，BERT可以有效地从大量的语料库中学习到语义信息。

2.2 BiGRU层

GRU是原始RNN的一个改进版本，旨在解决RNN中的梯度消失问题，并且由于其相似的基本概念，它也可以被视为LSTM的一个变体[28]。一般来说，为了保证重要信息在长期传播过程中不会丢失，并解决标准RNN中的梯度消失问题，GRU和LSTM都使用多种门函数来保留关键特征。此外，GRU的结构和组成比LSTM更加简洁，因此其参数更少，训练速度更快。在单向GRU网络中，状态有规律地从前向后传递。然而，在警情实体识别领域，实体与其前后文本具有很强的关联性。因此，本文试图将当前时间的输出与未来的状态结合起来。需要BiGRU来建立这些连接，BiGRU模型结构如**图2**所示。在BiGRU中，输入将同时提供给两个相反方向的GRU，输出由两个单向GRU共同决定。因此，BiGRU的当前隐藏层状态由3个部分决定：当前时刻t输入x_t，$t-1$时刻前向隐藏层状态的输出$\overrightarrow{h_{t-1}}$，$t-1$时刻后向状态的输出$\overleftarrow{h_{t-1}}$。相应计算式如式（1）~式（3）所示。最终状态h_t将是从输入的每个警情记录文字提取出来的特征，b_t表示t时刻隐藏层状态的偏置。

$$\overrightarrow{h_t}=\text{GRU}\left(x_t,\overrightarrow{h_{t-1}}\right) \tag{1}$$

$$\overleftarrow{h_t} = \mathrm{GRU}\left(x_t, \overleftarrow{h_{t-1}}\right) \quad (2)$$

$$h_t = w_t \overrightarrow{h_t} + w_t \overleftarrow{h_t} + b_t \quad (3)$$

2.3 自注意力层

注意力机制最早被应用于视觉图像领域，其思想来源于人类视觉注意力机制，即人类视觉在感知物体的时候会先将注意力放于某个特定最重要的部分。Bahdanau D等人[29]将注意力机制应用于神经机器翻译模型，首次在自然语言处理领域引入了注意力机制。自注意力机制[30]属于一种特殊的注意力机制，其将每一个词都和文本内部的所有词进行缩放点积注意力（scaled dot-product attention）计算，以捕获文本内部结构，学习内部的依赖关系。缩放点积注意力计算式如式（4）所示，其中$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$分别代表查询矩阵、键矩阵和值矩阵，d_K为输入向量的维度。且在自注意力机制中，$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$都等于BiGRU输出的结果向量。

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{soft\,max}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_K}}\right)\boldsymbol{V} \quad (4)$$

单一自注意力机制的性能往往是有限的，因此本文所使用的自注意力机制是多头自注意力机制。其是基于自注意力机制进行改善的，通过多次计算，可以使模型从多个角度提取文本中的隐含依赖关系，在不同的表示子空间中学习到相关信息[27]。多头自注意力机制的结构如**图3**所示，其中h代表多头自注意力机制的头数，$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$首先经过h次不同参数的线性变换，然后分别输入h个缩放点积注意力进行计算，并将结果进行拼接。最后，再进行一次线性变换，得到多头自注意力机制的输出结果。计算式如式（5）和式（6）所示，其中i表示第i个头。$\boldsymbol{W}_i^{\boldsymbol{Q}}$、$\boldsymbol{W}_i^{\boldsymbol{K}}$和$\boldsymbol{W}_i^{\boldsymbol{V}}$分别代表第$i$个头中$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$的参数矩阵，$\boldsymbol{W}^{\mathrm{O}}$代表输出时线性变化的参数矩阵。

$$\mathrm{head}_i = \mathrm{Attention}\left(\boldsymbol{Q}W_i^{\boldsymbol{Q}}, \boldsymbol{K}W_i^{\boldsymbol{K}}, \boldsymbol{V}W_i^{\boldsymbol{V}}\right) \quad (5)$$

$$\mathrm{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{Concat}(\mathrm{head}_1,\cdots,\mathrm{head}_h)W^{\mathrm{O}} \quad (6)$$

2.4 CRF层

在BiGRU层，对BiGRU网络的最终隐藏状态进行拼接和计算，以获得每个

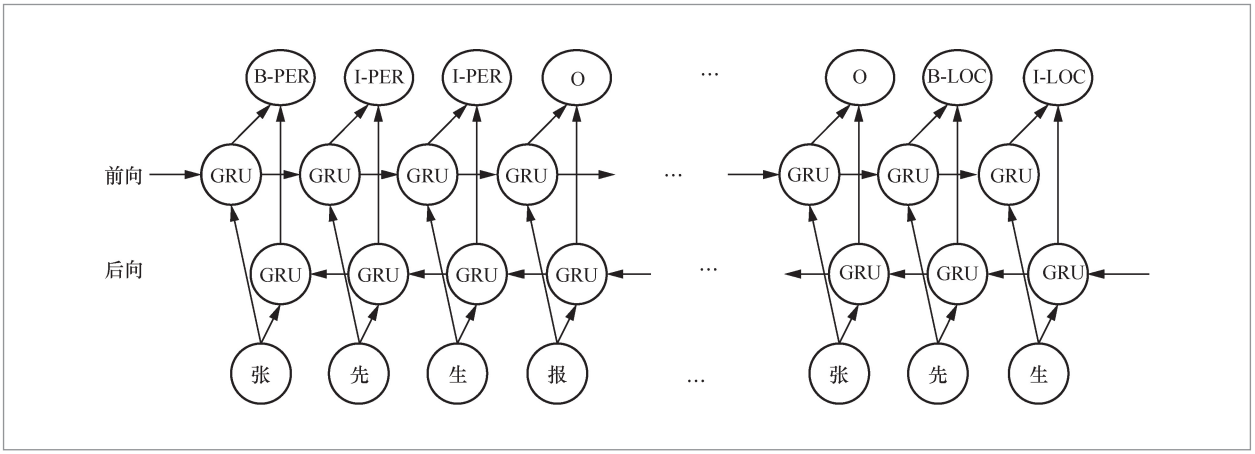


图 2 BiGRU 模型结构

文字属于各个标签的分数。命名实体识别可以被视为序列标记问题，如果没有CRF层，直接选择BiGRU层中得分最高的标签也是可以理解的。然而，BiGRU只考虑警情记录中的上下文信息，而不考虑标签之间的依赖关系，因此无法保证能够输出有意义的标签序列。CRF[30]是一种判别式无向图机器学习模型，其可以添加很多约束条件，以确保最终的预测是有价值的。CRF层的输入是报警记录序列$x=(x_1, x_2,\cdots,x_t)$，输出是最佳标签序列$y=(y_1, y_2,\cdots,y_t)$。首先，式（7）用于计算标签序列位置分数。在式（7）中，$\boldsymbol{P}$是BiGRU层的输出矩阵，$\boldsymbol{A}$是转移分数矩阵，其中$\boldsymbol{A}_{i,j}$表示从标签i到标签j的转移分数。

$$\text{score}(x,y)=\sum_{i=1}^{m}\boldsymbol{P}_{i,y_i}+\sum_{j=1}^{m+1}\boldsymbol{A}_{y_{j-1},y_j} \quad (7)$$

预测序列y的归一化概率如式（8）所示。此外，对于每个训练样本，将通过式（9）计算对数似然函数。

$$P(y\mid x)=\frac{\mathrm{e}^{\text{score}(x,y)}}{\sum_{y'=1}^{k}\mathrm{e}^{\text{score}(x,y')}} \quad (8)$$

$$\log(y^x\mid x)=\text{score}(x,y^x)-\log\sum_{y'}\exp(\text{score}(x,y')) \quad (9)$$

最终，通过最大化对数似然函数和式（10）中的维特比算法，将得分最高的标签序列作为预测结果。

$$y^*=\text{argmax}_{y'}\text{score}(x,y') \quad (10)$$

3 实验

3.1 实验环境

在本文中，BERT-BiGRU-SelfAtt-CRF模型的开发语言是Python 3.7，该模型是在深度学习框架Tensorflow的基础上实现的。硬件环境采用6核Intel Xeon E5-2620 v3 2.40 GHz CPU，64 GB RAM和Windows Server 2012 R2 64位操作系统。该模型相关参数设置见**表1**。

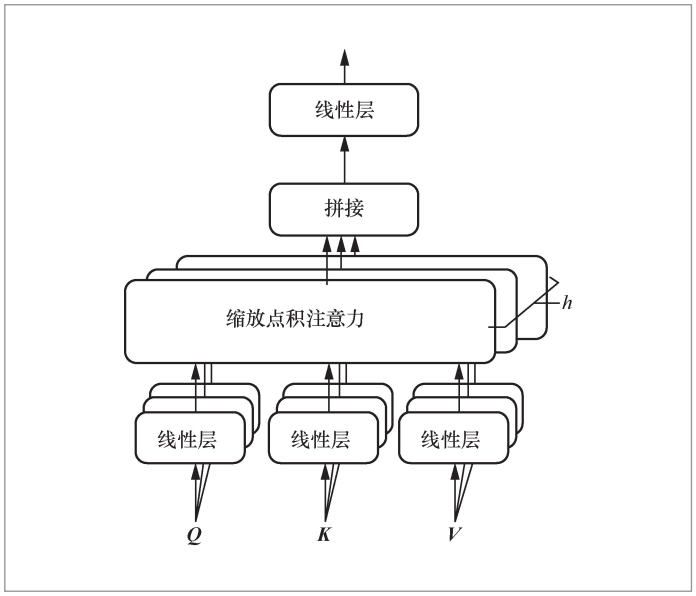


图 3 多头自注意力机制的结构

3.2 数据准备

本文的实验数据由两组构成。第一组是两个公开数据集，目的是更全面地验证本模型在大数据集上的性能和可提升空间

表 1 BERT-BiGRU-SelfAtt-CRF 模型参数设置

参数类型	参数名称	参数值
BERT参数	layer_nums	4
	head_num	12
	hidden_size	768
BiGRU-CRF超参数	BiGRU units	128
	max_seq_length	100
	dropout_rate	0.4
	SelfAtt_head	12
BiGRU-CRF训练参数	Epochs	5
	batch_size	64

以及在不同领域上的泛化能力。其中一个公开数据集是北京大学根据1998年《人民日报》数据建立的语料库，并从其中抽取出20 864条训练样本、4 636条测试样本和2 318条验证样本；另一个数据集为微软亚洲研究院提供的MSRA数据集[31]，并从其中抽取46 364句文本组成训练集，抽取4 365句文本组成测试集。

第二组是本文基于中国某市公安局的300条电子警情记录进行人工标注而构建的报警记录数据集PRD-PSB，目的是验证本模型在警情领域小数据集上的可行性和可推广性。由于在公安实战中，不同城市的地名和机构名存在很大的差异，因此在实际应用中，部署模型前需要根据当地警情记录进行标注和训练。而且，公安领域缺少标准的警情实体识别数据集和统一的实体数据标注规范，在实体数据标注中需要消耗很大的成本。因此，为了便于实战应用，本文在小规模警情数据集上进行验证，虽然小规模语料训练会在一定程度上限制模型的表现，但符合一线公安工作需求，便于各地普及应用。而且，在小规模数据集上满足基本实体提取需求后，各地公安机关前期只需要花费很少的标注成本就可以实际应用此模型，并可以在后期针对性地对实体进行扩充。

考虑报警信息文本的结构、各实体出现的频率以及实际警务工作的需要，将报警信息文本的内容分为4类：报警人姓名、案发地址、涉案机构和非实体。在模型训练之前，本文对数据进行了预处理，包括去除非法字符、无效空格、无意义的换行符等。最终从300条电子警情记录中筛选出395个句子和12 513个字。经过统计，PRD-PSB数据集的非实体文字共8 290个，中文实体文字共4 223个，其中案发地址3 447个字，报警人姓名585个字，涉案机构191个字。PRD-RSB数据集的占比分布如**图4**所示。

接下来，处理过的数据被逐字标记并分类到单独的训练文本文档中。其次，将所有数据按照8:1:1的比例拆分为训练、测试和验证集。警情数据标注格式如**图5**所示，数据按照这种格式进行处理和标注，文字之间用换行符分隔，文字和标签之间用空格分隔。本文采用BIO标注方案，有7个标签：B-LOC、I-LOC、B-PER、I-PER、B-ORG、I-ORG和O。BIO机制各个标签的文本实例和含义见**表2**。

3.3 评价指标

在可靠性方面，本文将精确率、召回率和F1值作为评价指标。此外，考虑到模型的性能和应用价值，本文还统计了每个模型训练所消耗的时间。精确率和召回率均保持在较高水平是最理想的，但实际上，两者在某些情况下是矛盾的。在不同的情况下，要判断需要高准确率还是高召回率。因此，评估方法中引入了F1值作为另一个评估指标，它同时考虑了准确率和召回率，可以看作二者的加权平均值。

4 结果与分析

在实验阶段，本模型对比了CNN-LSTM、BiLSTM-CRF和BiGRU-CRF，测试了三者在公开数据集和PRD-PSB数据集上的性能。另外，实验部分还分别对比了Word2vec和BERT两种词嵌入方法对每个模型表现的影响以及引入自注意力机制的效果。**表3**展示了8个模型在公开数据集上的结果。很明显，在大型公开数据集中，未引入自注意力机制的基线模型中，除CNN-LSTM之外，其余5个模型均表现优良，且BiLSTM-CRF和BiGRU-CRF的3个评价指标基本上高于其他模型。在F1

值大致相同的情况下，BiGRU-CRF的时间成本远低于BiLSTM-CRF。虽然BiGRU-CRF和BiLSTM-CRF在公开数据集上的评价指标差别不大，但BiGRU-CRF的训练时间却比BiLSTM-CRF缩短了153 min，原因可能是BiGRU的模型结构比BiLSTM简单，参数较少，因此BiGRU-CRF最适合公开数据集。因此，在BiGRU-CRF的基础上，本文对比了自注意力机制的效果，但在语料规模较大的数据集中，自注意力机制的引入对模型的性能提升不是十分明显。

表4描述了基于PRD-PSB数据集的实验结果，由于PRD-PSB数据集的样本量远小于公开语料库，因此训练的时间大大降低。BERT的引入会极大地提高模型的性能，虽然加载BERT可能会花费时间，但引入BERT后，模型可以以较短的训练周期获得更出色的识别效果。另外，在小数据集上，自注意力机制能帮助模型更好地捕获文本潜在的语义信息，在F1值方面，对BiGRU-CRF模型提升了2.23个百分点，对BERT-BiGRU-CRF模型提升了2.86个百分点。实验结果显示，在所有基线模型

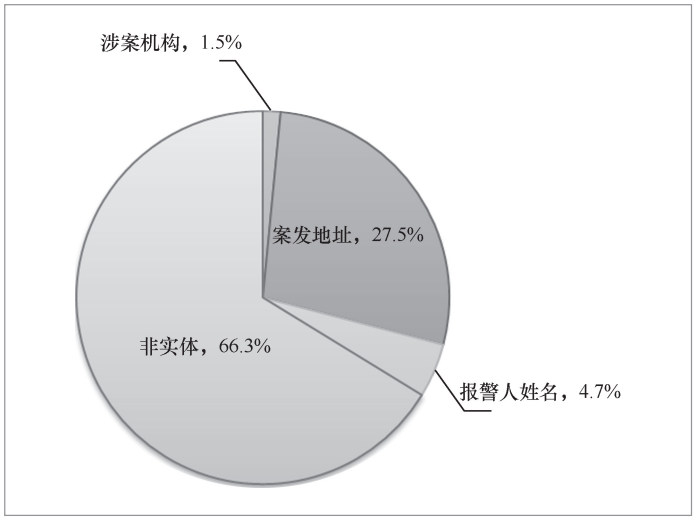


图 4 PRD-PSB 数据集的占比分布

张	B-PER
三	I-PER
报	O
：	O
在	O
北	B-LOC
京	I-LOC
市	I-LOC

图 5 警情数据标注格式

表 2 BIO 机制实例

文本实例	标签	含义
北	B-LOC	案发地址实体的首部
京	I-LOC	案发地址实体的中间部分
的	O	非实体
李	B-PER	报警人姓名实体的首部
先	I-PER	报警人姓名实体的中间部分
生	I-PER	报警人姓名实体的中间部分
在	O	非实体
人	B-ORG	涉案机构实体的首部
民	I-ORG	涉案机构实体的中间部分
医	I-ORG	涉案机构实体的中间部分
院	I-ORG	涉案机构实体的中间部分

表3 在《人民日报》和MSRA数据集上的实验结果

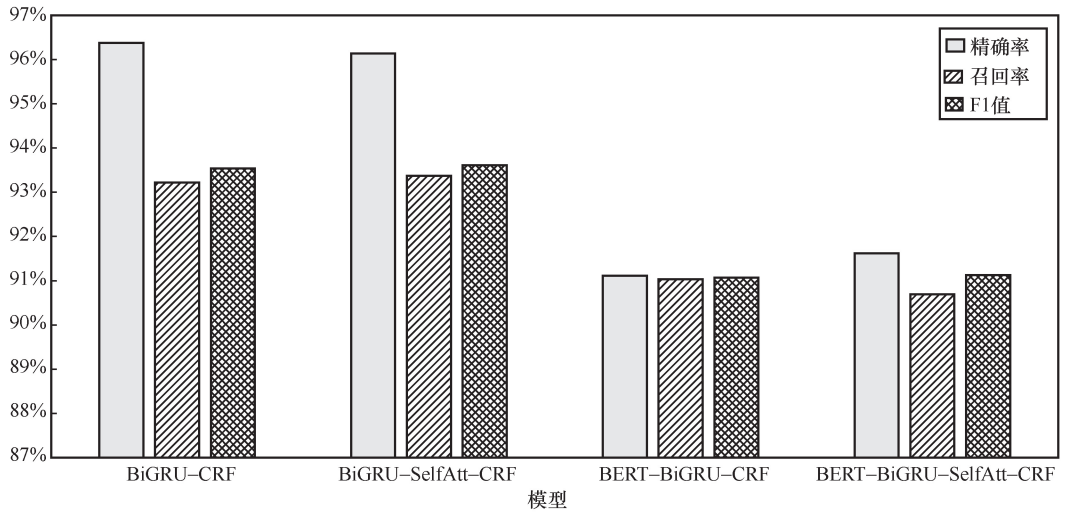
模型	训练周期	精确率	召回率	F1值	消耗时间/min
CNN-LSTM	40	72.87%	74.94%	73.98%	91
BiLSTM-CRF	40	95.50%	92.09%	93.76%	398
BiGRU-CRF	40	96.38%	93.22%	93.54%	245
BiGRU-SelfAtt-CRF	40	96.14%	93.37%	93.61%	252
BERT-CNN-LSTM	40	90.38%	93.98%	93.53%	309
BERT-BiLSTM-CRF	40	92.68%	90.31%	91.48%	443
BERT-BiGRU-CRF	40	91.11%	91.03%	91.07%	441
BERT-BiGRU-SelfAtt-CRF	40	91.62%	90.69%	91.13%	459

表4 在PRD-PSB数据集上的实验结果

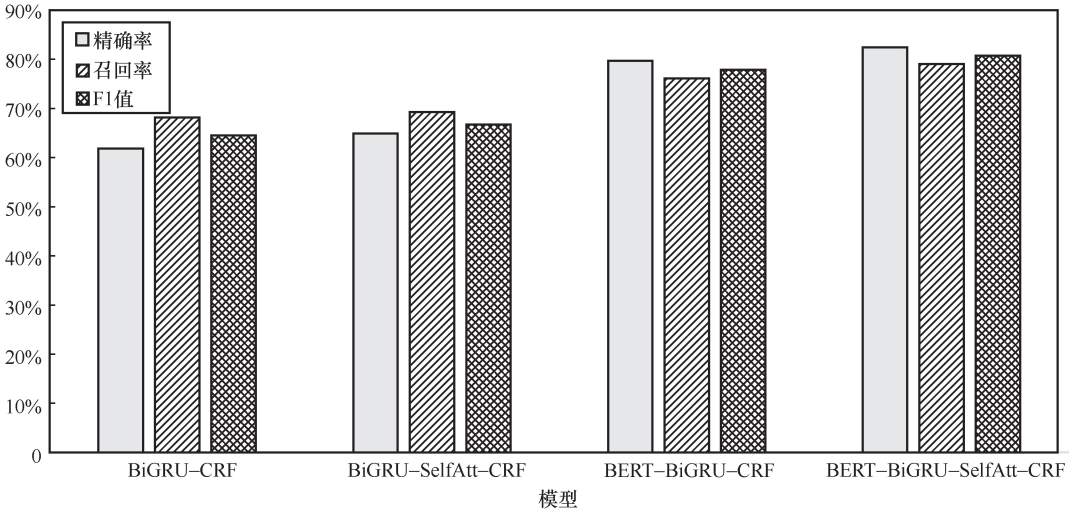
模型	训练周期	精确率	召回率	F1值	消耗时间/min
CNN-LSTM	50	30.95%	19.70%	24.07%	0.68
BiLSTM-CRF	50	68.92%	71.21%	69.79%	7.15
BiGRU-CRF	50	61.83%	68.18%	64.51%	2.62
BiGRU-SelfAtt-CRF	50	64.90%	69.27%	66.74%	3.27
BERT-CNN-LSTM	10	78.57%	65.67%	71.54%	10.28
BERT-BiLSTM-CRF	10	78.12%	74.63%	76.34%	17.22
BERT-BiGRU-CRF	10	79.69%	76.12%	77.86%	16.10
BERT-BiGRU-SelfAtt-CRF	10	82.45%	79.03%	80.72%	17.23

中，BERT-BiGRU-CRF的精确率、召回率和F1值最高，其时间成本也可以接受，仅需10个训练周期。因此，BERT-BiGRU-CRF是所有基线模型中最适合警情实体识别任务的。本文在此基础上，引入自注意力机制构建了BERT-BiGRU-SelfAtt-CRF模型，对模型的效果有了进一步的提升。

图6分别比较了在两种数据集中引入自注意力机制对模型整体表现的影响。首先，如**图6（a）**所示，在公开数据集上引入自注意力机制对模型效果的提升较为有限。结合**表3**可知，引入自注意力机制，BiGRU-CRF和BERT-BiGRU-CRF的F1值仅提高了0.07个百分点和0.06个百分点。这可能是由于大规模语料库中存在大量样本、丰富的语义信息和充足的词特征，并且BERT在大量数据中可以有效地生产包含丰富语义的词向量。因此，自注意力机制的帮助不是特别明显。而由**图6（b）**可知，在小规模的警情数据集



(a) 在公开数据集上对比模型的表现



(b) 在PRD-PSB数据集上对比模型的表现

图 6 在数据集中引入自注意力机制对模型整体表现的影响

上，引入自注意力机制的效果较为明显。结合**表4**可知，BiGRU-CRF在引入自注意力机制后，F1值提高了2.23个百分点。对于BERT-BiGRU-SelfAtt-CRF模型，较引入自注意力机制之前，精确率、召回率和F1值分别提升了2.76个百分点、2.91个百分点和2.86个百分点，并且精确率和F1值都提升到了80%以上。因此，当模型被应用于规模较小的数据集时，引入自注意力机制是很有必要的，模型的整体表现都有较为明显的提升，对警情实体识别任务有重要的意义。

本文基于PRD-PSB数据集，对3个模型进行了训练，训练过程中未引入BERT的模型准确率随训练周期的变化如**图7(a)**所示。综合来看，BiGRU-CRF的准确率最高，其次是BiLSTM-CRF，最后是CNN-LSTM。因此，将BiGRU-CRF作为本文警情实体识别模型的基本架构进行改进。此外，**图7(b)**展示了在BiGRU-SelfAtt-CRF模型中引入BERT在30个训练周期

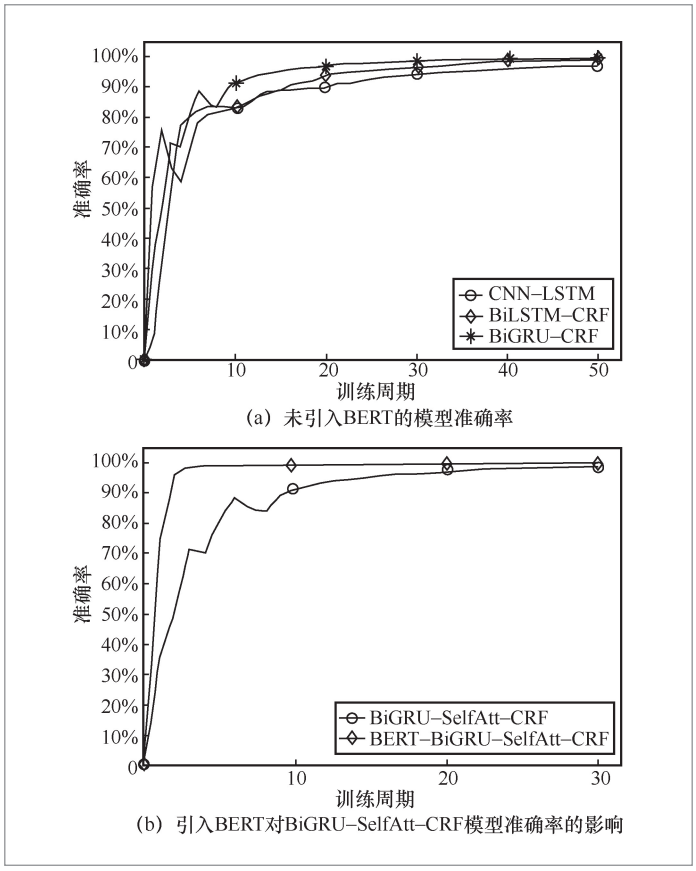


图 7 在 PRD–PSB 数据集上引入 BERT 对模型准确率变化的影响

对识别准确率的影响。由**图7(b)**可知，引入BERT后，模型可以在3个训练周期内将模型准确率快速提高到95%以上，而未引入BERT的模型则需要15个训练周期才能将准确率稳定在95%以上。并且，引入了BERT的模型，整体准确率要更高。另外，根据**图6(b)**及**表4**可知，在PRD–PSB数据集上，引入BERT可以将BiGRU–SelfAtt–CRF的F1值提高13.98个百分点。因此，在小规模的警情数据集上，引入BERT可以使模型的准确率提高得更快，而且可以明显优化模型整体效果。

5 结束语

为了准确有效地提取电子报警记录中的报警人姓名、案发地址和涉案机构，本文构建了BERT–BiGRU–SelfAtt–CRF模型来完成报警信息的命名实体识别任务。此外，本文还比较了3种经典的命名实体识别框架：CNN–LSTM、BiLSTM–CRF和BiGRU–CRF。在《人民日报》语料库、MSRA和PRD–PSB数据集上，BiGRU–CRF和BiLSTM–CRF具有相似的识别效果，并且比CNN–LSTM的效果更好。另外，本文还探究了引入BERT和自注意力机制对实验效果的影响。最终，本文通过实验得出如下结论。

(1)在大规模公开数据集上，由于数据量充足，语义信息丰富，BERT并没有提高模型性能，反而增加了时间成本。而在小规模警情数据集中，BERT能在很短的训练周期内显著提升各项指标。在PRD–PSB数据集上的实验结果表明，对于BiGRU–CRF模型来说，引入BERT将其F1值提高了13.35个百分点。因此，在数据集有限的情况下，BERT可以生成包含更丰富语义信息的词向量，提高后续实体识别的性能。

(2)类似地，自注意力机制也是在小规模警情数据集上的效果更加明显。对于BERT–BiGRU–CRF模型来说，在PRD–PSB数据集中引入自注意力机制后，精确率、召回率和F1值分别提升了2.76、2.91和2.86个百分点。多头自注意力机制可以从多个方向、多个表示子空间中提取文本的隐藏依赖关系，捕捉文本结构，提高模型识别的表现。

(3)BiGRU模型在保证BiLSTM模型效果的基础上，结构更加简单，参数更少。本文模型采用BiGRU模型，能加快模型的收敛速度，降低时间成本，符合实际公安工作的需求。另外，本文提出的BERT–BiGRU–SelfAtt–CRF模型在标注体量有限的警情数据集上，实体提取的精确率

和F1值都达到了80%以上，可以满足公安实战中的准确率要求。并且在小规模警情数据集上进行验证，可以证明模型的可行性，并证明在实战部署中具备可推广性，不需要消耗大量的标注成本。此外，也在大规模的公开数据集上验证了此模型的性能，其各方面指标都可以达到90%以上，可以泛化到不同领域，随着数据集的增大，模型有提升的空间。

综上所述，BERT模型中的多头自注意力机制与BiGRU模型中的双向结构保证了该模型能够充分考虑报警信息中的上下文关系，解决中文词边界模糊的问题，从而增加实体提取准确性。自注意力机制可以保证模型学习到文本内部结构，捕获文本中的长距离依赖关系。另外，BiGRU模型结构简单，参数较少，节约了模型的训练时间。最后，CRF层可以从实际训练数据中学习约束条件。在标签层面，其考虑了标签之间的顺序，优化了提取效果。该项目总体上能够满足公安实战工作的需要，填补了当前警务工作信息化的空白。

但实际警情数据中也存在着各类实体比例不均衡等问题，在未来的工作中，笔者将在数据集方面丰富实体类别，着重增加稀疏实体数量。在模型方面，笔者将尝试构建更优秀的深度学习模型来完成警情命名实体识别任务，探索出效果更优的模型。

参考文献:

[1] 张晓艳，王挺，陈火旺. 命名实体识别研究[J]. 计算机科学，2005，32(4)：44-48.
ZHANG X Y，WANG T，CHEN H W. Research on named entity recognition[J]. Computer Science，2005，32(4)：44-48.

[2] 何玉洁，杜方，史英杰，等. 基于深度学习的命名实体识别研究综述[J]. 计算机工程与应用，2021，57(11)：21-36.
HE Y J，DU F，SHI Y J，et al. Survey of named entity recognition based on deep learning[J]. Computer Engineering and Applications，2021，57(11)：21-36.

[3] 王月，王孟轩，张胜，等. 基于BERT的警情文本命名实体识别[J]. 计算机应用，2020，40(2)：535-540.
WANG Y，WANG M X，ZHANG S，et al. Alarm text named entity recognition based on BERT[J]. Journal of Computer Applications，2020，40(2)：535-540.

[4] ISOZAKI H，KAZAWA H. Efficient support vector classifiers for named entity recognition[C]//Proceedings of the 19th International Conference on Computational Linguistics. Morristown：Association for Computational Linguistics，2002.

[5] LIU K X，HU Q C，LIU J W，et al. Named entity recognition in Chinese electronic medical records based on CRF[C]//Proceedings of 2017 14th Web Information Systems and Applications Conference. Piscataway：IEEE Press，2017：105-110.

[6] HAN A L F，WONG D F，CHAO L S. Chinese named entity recognition with conditional random fields in the light of Chinese characteristics[C]//Proceedings of the Language Processing and Intelligent Information Systems. [S.l.：s.n.]，2013：57-68.

[7] MORWAL S. Named entity recognition using hidden Markov model (HMM)[J]. International Journal on Natural Language Computing，2012，1(4)：15-23.

[8] FU G H，LUKE K K. Chinese named entity recognition using lexicalized HMMs[J]. ACM SIGKDD Explorations Newsletter，2005，7(1)：19-25.

[9] BENDER O，OCH F J，NEY H. Maximum entropy models for named entity recognition[C]//Proceedings of the 7th Conference on Natural Language Learning

at HLT-NAACL 2003. Morristown: Association for Computational Linguistics, 2003: 148-151.

[10] CHIEU H L, NG H T. Named entity recognition: a maximum entropy approach using global information[C]//Proceedings of the 19th International Conference on Computational Linguistics. Morristown: Association for Computational Linguistics, 2002.

[11] 吴超, 王汉军. 基于GRU的电力调度领域命名实体识别方法[J]. 计算机系统应用, 2020, 29(8): 185-191.
WU C, WANG H J. Named entity recognition in electric power dispatching field based on GRU[J]. Computer Systems & Applications, 2020, 29(8): 185-191.

[12] DONG C H, WU H J, ZHANG J J, et al. Multichannel LSTM-CRF for named entity recognition in Chinese social media[C]//Proceedings of the Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. [S.l.:s.n.], 2017: 197-208.

[13] WU F Z, LIU J X, WU C H, et al. Neural Chinese named entity recognition via CNN-LSTM-CRF and joint training with word segmentation[C]//Proceedings of World Wide Web Conference (WWW 2019). New York: ACM Press, 2019: 3342-3348.

[14] DONG C H, ZHANG J J, ZONG C Q, et al. Character-based LSTM-CRF with radical-level features for Chinese named entity recognition[C]//Proceedings of the Natural Language Understanding and Intelligent Applications. [S.l.:s.n.], 2016: 239-250.

[15] TANG B Z, WANG X L, YAN J, et al. Entity recognition in Chinese clinical text using attention-based CNN-LSTM-CRF[J]. BMC Medical Informatics and Decision Making, 2019, 19(Suppl 3): 74.

[16] HUANG Z H, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint, 2015, arXiv:1508.01991.

[17] CHEN Y, ZHOU C J, LI T X, et al. Named entity recognition from Chinese adverse drug event reports with lexical feature based BiLSTM-CRF and tri-training[J]. Journal of Biomedical Informatics, 2019, 96: 103252.

[18] 李一斌, 张欢欢. 基于双向GRU-CRF的中文包装产品实体识别[J]. 华东理工大学学报(自然科学版), 2019, 45(3): 486-490.
LI Y B, ZHANG H H. Chinese packaging product entity recognition based on bidirectional GRU-CRF[J]. Journal of East China University of Science and Technology, 2019, 45(3): 486-490.

[19] WU G H, TANG G G, WANG Z R, et al. An attention-based BiLSTM-CRF model for Chinese clinic named entity recognition[J]. IEEE Access, 2019, 7: 113942-113949.

[20] ZHONG Q, TANG Y. An attention-based BILSTM-CRF for Chinese named entity recognition[C]//Proceedings of 2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics. Piscataway: IEEE Press, 2020: 550-555.

[21] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of the Advances in Neural Information Processing Systems. [S.l.:s.n.], 2013: 3111-3119.

[22] DEVLIN J, CHANG M. W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint, 2018, arXiv:1810.04805.

[23] LI X Y, ZHANG H, ZHOU X H. Chinese clinical named entity recognition with variant neural structures based on BERT methods[J]. Journal of Biomedical

Informatics, 2020, 107: 103422.

[24] 尹学振, 赵慧, 赵俊保, 等. 多神经网络协作的军事领域命名实体识别[J]. 清华大学学报(自然科学版), 2020, 60(8): 648-655.
YIN X Z, ZHAO H, ZHAO J B, et al. Multi-neural network collaboration for Chinese military named entity recognition[J]. Journal of Tsinghua University (Science and Technology), 2020, 60(8): 648-655.

[25] GU L, ZHANG W J, WANG Y, et al. Named entity recognition in judicial field based on BERT-BiLSTM-CRF model[C]// Proceedings of 2020 International Workshop on Electronic Communication and Artificial Intelligence. Piscataway: IEEE Press, 2020: 170-174.

[26] NIE Y Y, TIAN Y H, WAN X, et al. Named entity recognition for social media texts with semantic augmentation[C]//Proceedings of 2020 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2020: 1383-1391.

[27] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Proceedings of the Advances in Neural Information Processing Systems. [S.l.:s.n.], 2017: 5998-6008.

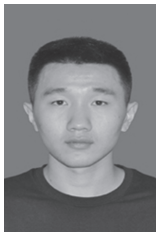
[28] CHO K, VAN MERRIENBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2014: 1724-1734.

[29] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint, 2018, arXiv:1409.0473.

[30] LAFFERTY J, MCCALLUM A, PEREIRA F. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the 18th International Conference on Machine Learning. [S.l.:s.n.], 2001, 3(2): 282-289.

[31] GINA A L. The third international Chinese language processing bakeoff: word segmentation and named entity recognition[C]//Proceedings of the 5th SIGHAN Workshop on Chinese Language Proceeding. [S.l.:s.n.], 2006: 548-554.

作者简介



崔雨萌(1998-), 男, 中国人民公安大学信息网络安全学院硕士生, 主要研究方向为命名实体识别。

王靖亚(1966-), 女, 中国人民公安大学信息网络安全学院教授, 主要研究方向为自然语言处理、样本对抗。

闫尚义（1998- ），男，中国人民公安大学信息网络安全学院硕士生，主要研究方向为自然语言处理、文本分类。

陶知众（1997- ），男，中国人民公安大学信息网络安全学院硕士生，主要研究方向为人工智能、图像风格转换。

收稿日期： 2022-03-14
通信作者： 王靖亚，wangjingya@ppsuc.edu.cn
基金项目： 国家社会科学基金资助项目（No.20AZD114）
Foundation Item: The National Social Science Foundation of China (No.20AZD114)