计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(1):47−54 [doi: 10.15888/j.cnki.csa.008260]
©中国科学院软件研究所版权所有.
E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

简要案情的命名实体识别技术①

陈柱辉[1], 刘　新[1], 张明键[2], 张达为[1]

[1](湘潭大学 计算机学院·网络空间安全学院, 湘潭 411105)
[2](湖南警察学院 信息技术系, 长沙 410138)
通信作者: 刘　新, E-mail: liuxin_new@163.com

摘　要: 简要案情是公安机关为提高“协同办案系统”录入信息质量, 确保信息检索与案件串并工作高效开展而对案情记载的简要描述, 其中各类实体间包含了大量与受害者和作案人相关的案情信息. 因此, 对简要案情文本的深度挖掘是掌握案件始末和分析案情的有效手段之一. 简要案情文本中的实体稠密分布、实体间相互嵌套以及实体简称, 给准确捕捉案件实体带来了巨大的挑战. 针对简要案情文本的特殊性和复杂性, 本文对字符向量生成的方法进行了改进, 提出了 RC-BiLSTM-CRF (Roberta-CNN-BiLSTM-CRF) 网络架构, 相比于主流的“Bert-BiLSTM-CRF”架构, 该架构可以对字符向量特征进行提取, 解决了通过预训练模型带来的字符向量冗长的问题, 通过减少模型的参数量进而提高了模型整体参数的收敛速度. 对比实验选用 5 种主流的架构在湖南省省公安机关提供的简要案情数据集上进行比较, 本文提出的方法在准确率、召回率和 $F1$ 值上均为最优, $F1$ 值达到了 88.02%.

关键词: 简要案情; 卷积神经网络; 双向长短期记忆网络; 条件随机场; 命名实体识别

引用格式: 陈柱辉,刘新,张明键,张达为.简要案情的命名实体识别技术.计算机系统应用,2022,31(1):47–54. http://www.c-s-a.org.cn/1003-3254/8260.html

Named Entity Recognition Technology for Brief Case

CHEN Zhu-Hui[1], LIU Xin[1], ZHANG Ming-Jian[2], ZHANG Da-Wei[1]

[1](School of Computer Science and School of Cyberspace Security, Xiangtan University, Xiangtan 411105, China)
[2](Department of Information Technology, Hunan Police College, Changsha 410138, China)

Abstract: A brief case is a brief description of a case record made by a public security organ to improve the quality of information input in the Collaborative Case Handling System and ensure efficient information retrieval and joint investigation. A large amount of case information related to the victim and the perpetrator is between various entities. Therefore, in-depth excavation of brief case texts is an effective means to grasp the beginning and end of a case and to analyze the case. The dense distribution, inter-nesting, and abbreviation of entities in a brief case text bring great challenges to the accurate capture of the case entities. In response to the particularity and complexity of brief case texts, this study improves the method of character vector generation and proposes a Roberta-CNN-BiLSTM-CRF (RC-BiLSTM-CRF) network architecture. Compared with the mainstream Bert-BiLSTM-CRF architecture, this architecture can extract the character vector features, thereby solving the problem of a lengthy character vector brought by model pre-training. The model parameter number is reduced for a higher overall parameter convergence rate. In the comparative experiment, five mainstream architectures are selected and compared on the brief case dataset provided by the public security organs of Hunan Province. The method proposed in this study is proved to be the best in terms of accuracy, recall rate, and $F1$ value, and its $F1$ value reaches 88.02%.

Key words: brief case; convolutional neural network (CNN); bidirectional long and short term memory (BiLSTM); conditional random field; named entity recognition

① 基金项目: 湖南省自然科学基金 (2018JJ2107); 湖南省科技重大专项 (2017SK1040); 湖南省公安厅科技计划 (2018No.3)
收稿时间: 2021-03-24; 修改时间: 2021-04-21; 采用时间: 2021-04-28; csa 在线出版时间: 2021-12-17

© 中国科学院软件研究所 http://www.c-s-a.org.cn

1 引言

简要案情是指警务人员在接到被害人或者目击者报案时, 使用警务信息系统生成的简短并蕴含重要信息的文本序列, 并便于警务人员管理和存储的警务记录. 简要案情中的案发地点、涉案人员、涉案财产和涉案事件关键词等实体是整个案件的核心信息, 通过这几类实体, 警务人员可以迅速判断出案件的严重程度以及犯罪的类型. 因此, 对简要案情文本的深度挖掘是掌握案件始末和分析案情的有效手段之一. 结合自然语言处理相关技术, 围绕简要案情等警务文本的相关研究, 可为智慧警务、案情问答等场景提供有效的支持与应用.

命名实体识别 (named entity recognition, NER) 是信息抽取和信息检索中一项重要的任务, 其目的是识别出文本中表示命名实体的成分, 并对其进行分类, 也是信息提取过程中的关键技术, 旨在从非结构化文本中抽取各类所需实体, 为语料库的建设和知识图谱的搭建提供了技术支持[1,2]. 通识领域凭借其大量标注数据集, 吸引众多研究人员争相投入其中, 通识领域的命名实体识别技术因此迎来迅速的发展, 然而, 受限于警务领域的简要案情文本的开放, 在简要案情命名实体识别上的研究呈一片空白. 因此, 本文先对小规模简要案情文本进行合理标注, 提取的实体包括案发地点、涉案人员、涉案财产和涉案事件关键词 4 个类别, 为了提高简要案情文本中复杂的专业名词的识别率, 本文对字符向量生成的方法进行了改进, 提出 RC-BiLSTM-CRF 神经网络模型, 通过 Roberta 预训练模型增强训练语料的语义表示并根据上下文特征动态生成字向量, 通过设计合理的卷积神经网络对字向量的局部重要特征进行提炼, 解决了通过预训练模型带来的字符向量冗长的问题, 通过减少模型的参数量进而增加了模型整体参数收敛的速度, 在一定程度上弥补了标注数据集稀缺的缺陷. 本文在湖南省省公安机关提供的简要案情数据集上做了大量的对比实验, 本文提出的网络框架取得了比较理想的实体识别效果.

本文组织结构: 第 2 节介绍相关工作, 包括对通识领域命名实体识别和特定领域命名实体识别的详细阐述; 第 3 节主要介绍本文设计的卷积神经网络, 还对 Roberta 预训练模型, BiLSTM 层和 CRF 层进行详细的介绍; 第 4 节对实验数据集、模型参数设置、模型评估标准和实验结果与分析进行介绍; 第 5 节为结束语.

2 相关工作

近年来, 深度学习技术在命名实体识别上的应用成为新的浪潮. 深度学习方法为科研理论的验证提供了一种新的解决思路, 最典型的深度学习模型为循环神经网络 (RNN), 卷积神经网络 (CNN) 的系列架构[3,4], 本文将对通识领域命名实体识别跟特定领域命名实体识别的研究成果进行介绍.

2.1 通识领域命名实体识别

Huang 等人[5] 提出了 BiLSTM-CRF 模型, 凭借巧妙设计的双向 LSTM 结构, BiLSTM-CRF 模型可以有效地使用过去和未来的输入特性, 该模型通过 CRF 层可以使用句子级标记信息. BiLSTM-CRF 模型可以在 POS、分块和 NER 数据集上产生最先进 (或接近) 的准确性, 并且具有较强的鲁棒性, 对词嵌入的依赖性更小, 可以实现准确的标注精度, 而不需要借助于 word 的嵌入.

Zhang 等人[6] 提出了 Lattice LSTM 模型, 该模型对输入字符序列以及所有与词典匹配的潜在单词进行编码, 与基于字符的方法相比, Lattice LSTM 明确地利用了单词和单词序列信息, 与基于词的方法相比, Lattice LSTM 不存在切分错误. Lattice LSTM 模型使用门控循环单元从一个句子中选择最相关的字符和单词, 以获得更好的实体识别结果. Lattice 方法完全独立于分词, 但由于可以在上下文中自由选择词典单词来消除歧义, 因此在使用单词信息方面更加有效, 在 MSRA 数据集中取得了 93.18% 的 $F1$ 值.

Gui 等人[7] 提出了 LR-CNN 模型, 采取 CNN 对字符特征进行编码, 感受野大小为 2 提取 bi-gram 特征, 堆叠多层获得 multi-gram 信息, 同时采取注意力机制融入词汇信息 (word embed) 以解决 Lattice LSTM 模型[6] 存在不能充分利用 GPU 进行并行化的问题, LR-CNN 最终相比于 Lattice LSTM 快 3.21 倍; LR-CNN 采取 rethinking 机制增加 feedback layer 来调整词汇信息的权值以解决 Lattice LSTM 模型存在无法有效处理词汇信息冲突的问题.

Li 等人[8] 提出了 FLAT 模型, 该模型将其 lattice 结构转换成由跨度 (spans) 组成的平面结构, 每个 span 相当于一个字或者一个词在其原始 lattice 中的位置, 得益于 Transformer 和 position encoding, FLAT 可以充分利用 lattice 信息, 具有出色的并行化能力. FLAT 解决了在中文 NER 中, lattice 模型因为其复杂度和动态性

© 中国科学院软件研究所 http://www.c-s-a.org.cn

问题, 导致其无法很好的利用 GPU, 限制了其运行速度的问题. 在数据集 (OntoNotes、MSRA、Resume 和 Weibo) 上, FLAT 在性能和效率方面均取得了很理想的效果.

2.2 特定领域命名实体识别

在社交领域, 李源等人[9] 为解决基于词粒度信息或者外部知识的中文命名实体识别方法存在中文分词 (CWS) 和溢出词 (OOV) 的问题, 提出一种基于字符的使用位置编码和多种注意力的对抗学习模型, 联合使用位置编码和多头注意力能够更好地捕获字序间的依赖关系, 而使用空间注意力的判别器则能改善对外部知识的提取效果, 该模型分别在 Weibo2015 数据集和 Weibo2017 数据集上进行了实验, 实验结果中的 $F1$ 值分别为 56.79% 和 60.62%.

在军事领域, 李健龙等人[10] 为了减少传统的命名实体识别需要人工制定特征的大量工作, 通过无监督训练获得军事领域语料的分布式向量表示, 采用双向 LSTM 模型解决军事领域命名实体的识别问题, 并且通过添加字词结合的输入向量和注意力机制对双向 LSTM 网络模型进行扩展和改进, 进而提高军事领域命名实体识别, 提出的方法在军事领域数据集上的 $F1$ 值达到了 87.38%.

在军用软件测试领域, 韩鑫鑫等人[11] 针对字词联合实体识别方法准确率不高的问题, 进行字符级特征提取方法的改进, 提出了 CWA-BiLSTM-CRF 识别框架, 该框架包含两部分: 第一部分构建预训练的字词融合字典, 将字词一起输入给双向长短期记忆网络进行训练, 并加入注意力机制衡量词内各字对特征的语义贡献, 提取出字符级特征; 第二部分将字符级特征与词向量等特征进行拼接, 输入给双向长短期记忆网络进行训练, 再通过条件随机场解决标签结果序列不合理的问题, 识别出文中的实体, 所提出的框架在军用软件测试数据集上的 $F1$ 值达到了 88.93%.

在医疗领域, 宁尚明等人[12] 针对电子病历实体的高密度分布以及实体间关系的交叉互联问题, 提出一种基于多通道自注意力机制的“recurrent+transformer”神经网络架构, 提升对电子病历专有文本特点的学习能力, 同时显著降低模型整体复杂度, 并且在该网络架构下提出带权学习的交叉熵损失函数以及基于权重的位置嵌入的辅助训练方法, 该框架相继在 2010 i2b2/VA 及 SemEval 2013 DDI 医学语料中进行验证, 相较于传统自注意力机制, 多通道自注意力机制的引入在模型整体 $F1$ 指标中最高实现 10.67% 的性能提升, 在细粒度单项对比实验中, 引入类别权重的损失函数在小类别样本中的 $F1$ 值最高提升近 23.55%.

3 本文提出的网络框架

本文针对简要案情文本存在实体稠密分布、实体间相互嵌套以及实体简称的问题, 对字符向量的生成方法进行了改进, 提出了 RC-BiLSTM-CRF 网络框架. RC-BiLSTM-CRF 整体算法框架如图 1 所示, 主要分为输入模块、字符向量生成模块和输出模块, 先对待标注文本进行数据清洗, 用正则方法将待标注的文本的噪声信息过滤掉, 有利于本文所提模型提取出重要特征信息; 将清洗后的数据输入到字符向量生成模块, 字符向量生成模块中首先通过 Roberta 预训练模型将文本生成字符向量, 经过本文合理设计的卷积层能够提取字符向量中的局部关键特征, 并能将冗长的字符向量进行浓缩, 紧随的激活层能够有效提高卷积层的特征学习能力和提升网络的性能; 经过字符向量生成模块后, 将字符向量输入到 BiLSTM 层, BiLSTM 层对字符序列进行上下文特征以及字符间依赖性学习, 通过 Dropout 层随机删掉网络中一定比例的隐藏神经元, 可以有效缓解模型过拟合情况, TimeDistributed 层将所有字符的向量维度进行约束, 使得字符向量的维度等于实体标签数, 最后通过 CRF 层得到输入文本的标注序列.

本文提出的 RC-BiLSTM-CRF 模型的整体结构如图 2 所示, 网络结构包括一个 Roberta 预训练层, 一个 CNN 层和 BiLSTM-CRF 模型. 下面将对 RC-BiLSTM-CRF 网络结构的各个部分进行详细阐述.

3.1 Roberta 预训练层

预训练模型本质上运用了迁移学习[13] 的思想, 利用大规模训练语料为预训练模型的参数进行训练, 然后将训练好的模型应用到下游任务, 避免了深度学习模型重新训练参数和减少了对标注数据的需求, 缩短了字、词向量训练的耗时. Roberta 预训练模型是 BERT (bidirectional encoder representations from transformers) 预训练模型的改进模型, Liu 等人对 BERT 预训练模型进行精细调参和调整训练集, 训练得到的 Roberta 模型在性能上相较于 BERT 模型提升显著[14–16]. Roberta 预训练模型充分考虑字符级、词语级、句字

© 中国科学院软件研究所 http://www.c-s-a.org.cn

级和句间的关系特征, 增强了字向量的语义表示, 把这些学习到的语义知识通过迁移学习应用到数据规模和标注量较少的简要案情的命名实体识别具体任务上, 能使模型更好的挖掘简要案情文本的特征信息.

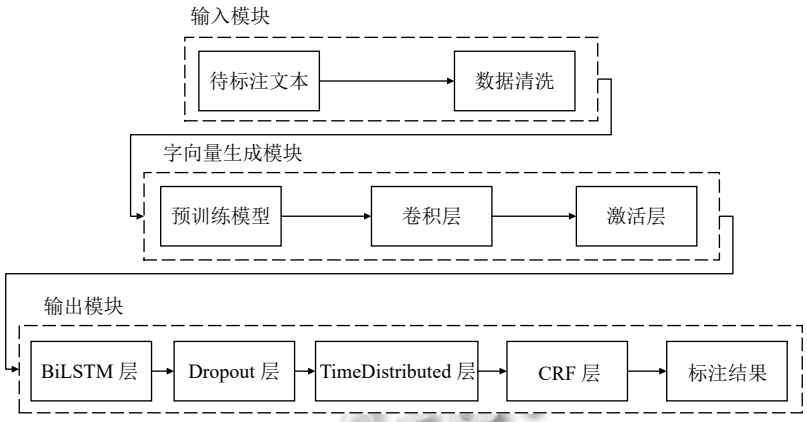

图 1 本文提出的整体算法框架

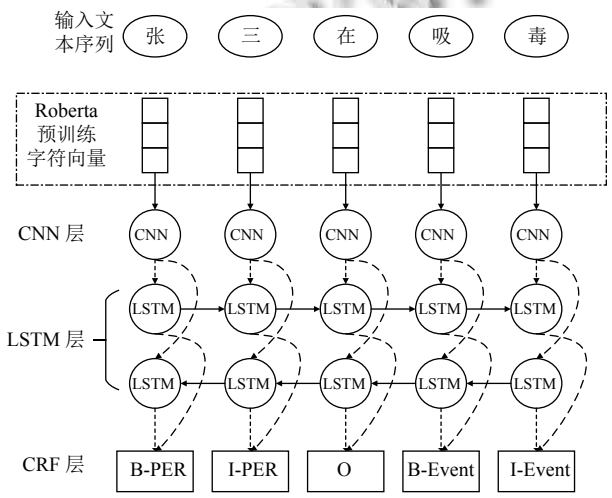

图 2 模型整体结构

将字符序列 chars=(char1, char2,⋯, char$n$), 输入到 Roberta 预训练模型中, Roberta 预训练模型通过在其他大规模语料上训练好的参数将 chars 中所有的字符生成向量, 即 char=(embedding1, embedding2,⋯, embeding$m$). 相较于构建 word2id 词典, 通过 id 匹配 id2vec 词典的方法, 预训练模型可以缩短字符向量的维度, 有效解决文本特征稀疏问题, 学习上下文信息来表征字词的多义性.

3.2 卷积神经网络层

卷积神经网络在本文所提出的字符向量生成方法中起着关键性作用, 卷积神经网络层可以为 Roberta 预训练模型生成的字符向量进一步提炼, 去除冗长字向量中的噪声, 提取出简短并蕴含局部重要特征信息的字符向量. 卷积操作的计算公式如下所示:

$$X_j^l = \sum_i X_i^{l-1} * K_{ij}^l + B_j^l \tag{1}$$

其中, $*$表示卷积计算, X_j^l表示第l层的第j个字符特征向量, X_j^{l-1}表示第$l-1$层的第j个字符特征向量, K_{ij}^l表示用来连接第l层的第i个字符特征向量和第j个字符特征向量的卷积核, B_j^l表示第l层的第j个字符特征向量的偏置量[17].

由于从 Roberta 预训练模型中生成的字符向量是一维向量, 于是本文使用一维卷积层对字符向量进行细粒度特征捕捉操作. 为了合理选择卷积层的滤波器的数量, 本文选取了 10–40 个滤波器进行实验, 实验参数中 epochs 均为 50, batch_size 均为 16, 卷积核大小均为 3, 实验结果如图 3 所示.

如图 3 所示, 使用 28 个滤波器的卷积层在本模型中的效果是最好的, 所以本文针对简要案情命名实体识别设计了包含 28 个滤波器, 卷积核大小为 3 的卷积层. 由于池化层是一个下采样的过程, 在减小特征向量长度的同时, 会使得部分案情实体特征信息丢失, 从而降低下一步 BiLSTM 进行上下文特征提取的性能, 收敛速度变得缓慢, 从而影响模型最终的实体标注的准确率, 因此在本文的网络结构中取消了池化层的使用. 经 Roberta 预训练模型处理得到每个字符向量的维度为 3072 维, 通过综合考虑设计的包含 28 个滤波器, 卷积核大小为 3 的卷积层, 对字符向量的特征进行提取,

© 中国科学院软件研究所 http://www.c-s-a.org.cn

使得字符向量序列从 100×3072 降维到 100×28 维, 解决了预训练模型带来的字符向量冗长的问题, 框架参数量的减少促使模型整体参数收敛的速率提高了 9.46%, 同时 $F1$ 值提高了 1.73%.

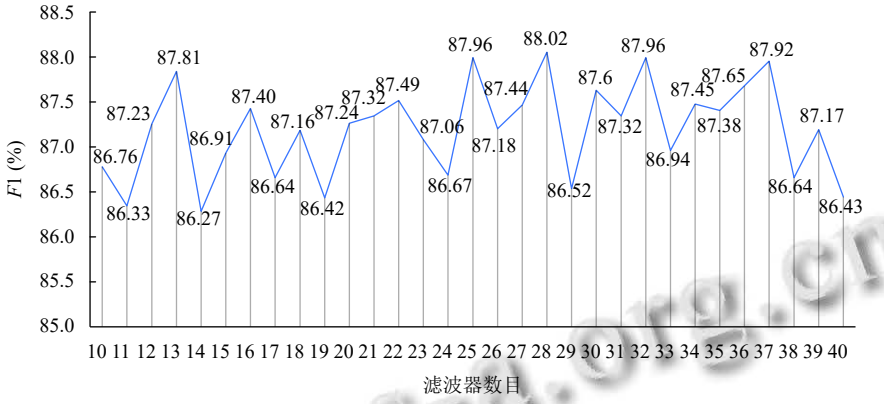

图 3 滤波器实验对比

3.3 BiLSTM-CRF 模型

BiLSTM-CRF 模型拉开了命名实体识别深度学习时代的序幕, 使得命名实体识别模型更加简洁高效, 鲁棒性更强. 本文针对简要案情本文构建基于 BiLSTM-CRF 的实体识别模型, 模型分为两部分, 接下来进行详细介绍.

3.3.1 BiLSTM 层

将已标注训练文本输入到上文提及的字符向量生成方法中, 生成字符向量. 将字符向量输入到 BiLSTM 层中. BiLSTM 包含了前向和后向的长短期记忆 (LSTM), 通过 BiLSTM 可以更好地学习上下文信息以及捕捉双向的语义依赖, 弥补了 LSTM 不能向前编码信息的能力. 在 LSTM 中, 有两个状态向量 C 和 h, 其中 C 作为 LSTM 的内部状态向量, 可以理解为 LSTM 的内存状态向量 Memory, 而 h 表示 LSTM 的输出向量. 相对于基础的 RNN 来说, LSTM 把内部 Memory 和输出分开为两个变量, 同时利用 3 个门控: 输入门 (input gate)、遗忘门 (forget gate) 和输出门 (output gate) 来控制内部信息的流动, 公式如式 (2)–式 (7) 所示:

$$f_t = \sigma\left(W_f * [h_{t-1}, x_t] + b_f\right) \tag{2}$$

$$i_t = \sigma(W_i * [h_{t-1}, x_t] + b_i) \tag{3}$$

$$\tilde{C}_t = \tanh(W_C * [h_{t-1}, x_t] + b_C) \tag{4}$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{5}$$

$$o_t = \sigma(W_o * [h_{t-1}, x_t] + b_o) \tag{6}$$

$$h_t = o_t * \tanh(C_t) \tag{7}$$

其中, W、b分别表示 LSTM 的隐藏层权重矩阵和偏置向量, f_t, i_t, o_t分别表示时间戳t的遗忘门、输入门和输出门, σ是 Sigmoid 激活函数, tanh 是 tanh 激活函数, h_t和C_t分别表示时间戳 t 的输出和细胞单元状态. 正向 LSTM 的输出值为$\vec{h} = \left(\vec{h}_1, \vec{h}_2, \cdots, \vec{h}_n\right)$, 反向 LSTM 的输出值为 $\overleftarrow{h} = \left(\overleftarrow{h}_1, \overleftarrow{h}_2, \cdots, \overleftarrow{h}_n\right)$, BiLSTM 则是将正向 LSTM 跟反向 LSTM 所得的向量进行拼接, 即 $h_i = \left(\vec{h}_i, \overleftarrow{h}_i\right)$.

3.3.2 CRF 输出层

CRF 在整个模型中起着至关重要的作用, 因为经过 BiLSTM 层处理, 得到的字符向量是字符对应的所有标签的概率, 最终输出的结果是每个字符对应的最大概率值的标签, 这样会导致输出的标签序列可能不符合命名实体识别规范. CRF 的维特比算法在解码时候拥有较高的效率, 通过 CRF 层的约束, 让输出标签序列符合实体规则. 标签序列的最终标注由 Emission-Score (发射状态矩阵) 跟 TransitionScore (转移分数) 决定. 当输入序列$x = (x_1, x_2, \cdots, x_n)$, 标注序列$y = (y_1, y_2, \cdots, y_n)$, 公式如式 (8) 和式 (9) 所示:

$$S_i = EmissionScore + TransitionScore \tag{8}$$

$$P(y|x) = \frac{e^{S_i}}{\sum_{n=1}^{m} e^{S_n}} \tag{9}$$

其中, $EmissionScore$表示 BiLSTM 输出标签的分数,

© 中国科学院软件研究所 http://www.c-s-a.org.cn

*TransitionScore*表示标签之间转移的分数, e^{S_i}表示当前标签序列分数, $\sum_{n=1}^{m} e^{S_n}$是所有标签序列的分数的总计, 最大$P(y|x)$值对应的y为序列x的正确标注序列.

4 实验及分析

4.1 实验数据集

本文所使用的数据集来自湖南省省公安机关的简要案情数据集 (JW_data), 数据的格式为 tsv, 本文先用正则方法将待清洗的数据集的噪声信息过滤掉, 有利于本文模型提取出重要特征信息, 清洗后的数据集中训练集占 60%, 测试集占 20%, 验证集占 20%. 数据统计信息如表 1 所示.

表 1 数据实体统计

数据集	训练集	测试集	验证集	总计
JW_data	1 545	515	516	2 576

清洗后的简要案情数据集在标注平台 doccano 进行人工标注, 将案发地点、涉案人员、涉案财产和涉案事件关键词 4 类实体作为本实验的标注实体, 对简要案情数据集采用 BIO 标注方式, B (Begin) 对应字符序列中实体的起始位置, I (Intermediate) 对应字符序列中实体的中间位置或者结束位置, O (Other) 对应字符序列中非实体的字符, 案发地点的标签包括 (B-LOC, I-LOC), 涉案人员的标签包括 (B-PER, I-PER), 涉案财产的标签包括 (B-Property, I-Property), 涉案事件关键词的标签包括 (B-Event, I-Event), 非实体字符的标签为 (O). 实体统计如表 2 所示.

表 2 数据实体标签统计

数据	PER	LOC	Event	Property	总计
训练集	2 712	2 839	722	747	7 020
测试集	1 153	1 016	217	325	2 711
验证集	1 000	984	244	295	2 523

4.2 模型参数设置

本文选择 Roberta 预训练模型生成简要案情数据集的字符向量, 字符序列长度设置为 100, 所生成的字符向量维度设定为 3 072, 所以输出字符向量序列的维度为 100×3072. 本文实验使用的 Batch_size 设定为 16, epochs 设定为 50, 学习率为 0.001. 卷积神经网络中卷积层的过滤器数量为 28, kernel_size 为 3, padding 为 “same”, 激活函数为 ReLU 函数, 卷积层的权重初始化方法为“glorot_uniform”, 偏移初始化方法为“zeros”, 输出的字符序列维度为 100×28; 双向长短期记忆的 units 设定为 128, 故输出的字符序列维度为 100×256, dropout 为 0.4.

4.3 模型评估标准

本文采用准确率 (*precision*), 召回率 (*recall*), *F*1 值作为模型的评价标准, 对简要案情数据的实体识别结果进行全方面的评价. 精确度、召回率和 *F*1 值的公式如式 (10) 和式 (12) 所示:

$$precision = \frac{Tp}{Tp + Fp} \tag{10}$$

$$recall = \frac{Tp}{Tp + Fn} \tag{11}$$

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{12}$$

其中, Tp表示实际为正被预测为正的实体数量, Fp表示实际为负但被预测为正的实体数量, Fn表示实际为正但被预测为负的实体的数量.

4.4 实验结果与分析

本文采用 BiGRU、BiBRU-CRF、BiLSTM、BiLSTM-CRF 及 CNN-LSTM 作为基线模型与本文所提的模型进行对比, 各模型的基本信息如下:

(1) BiGRU: 采用 BiGRU 提取特征并通过全连接层直接对字符向量序列进行标注的模型.

(2) BiBRU-CRF: 采用 BiGRU 提取特征并结合 CRF 对输入字符向量序列进行标注的模型.

(3) BiLSTM: 采用 BiLSTM 提取上下文特征并通过全连接层直接对字符向量序列进行标注的模型.

(4)BiLSTM-CRF: 采用 BiLSTM 提取上下文特征并结合 CRF 对输入字符向量序列进行标注的模型.

(5) CNN-LSTM: 采用本文设计的卷积神经网络对字符.

向量的局部特征进行捕捉, 再通过 LSTM 层对字符向量序列的正向特征进行学习, 最后通过全连接层直接对字符向量序列进行标注的模型.

为验证本文所提模型加入 Roberta 预训练模型的必要性, 本文对所提出的模型和基线模型进行了验证, 性能对比如表 3 所示.

从表 3 的结果可以看出以上 6 种模型通过加入 Roberta 预训练模型训练数据集的字符向量, 准确率、召回率和 *F*1 值都能大幅度提升, Roberta 预训练模型

© 中国科学院软件研究所 http://www.c-s-a.org.cn

通过在大规模训练语料训练模型参数, 一定程度上减少了对本文实验标注数据的依赖性, 避免了本文实验数据较少导致模型效果不理想的情况. 以上 6 种模型, 相较于未加入 Roberta 预训练模型的框架, 准确率提高了 5.94%~10.75%, 召回率提高了 5.56%~9.36%, *F*1 值提高了 5.78%~10.12%, 由此可见在本文所提的模型中加入 Roberta 预训练模型是必要的.

表 3 对比实验结果 (%)

模型	未加入Roberta预训练模型			加入Roberta预训练模型		
	精确率	召回率	*F*1值	精确率	召回率	*F*1值
BiGRU	75.60	79.55	77.50	83.13	86.25	84.64
BiBRU-CRF	78.26	80.57	79.37	84.20	86.13	85.15
BiLSTM	75.50	78.80	77.08	84.22	87.28	85.71
BiLSTM-CRF	78.02	80.06	79.02	85.38	87.24	86.29
CNN-LSTM	71.68	75.78	73.63	82.43	85.14	83.75
本文模型	79.41	79.47	79.44	87.23	88.82	88.02

通过表 3, 可知基于本文的简要案情数据, 加入 Roberta 预训练模型可以全方面提升模型的性能, 于是本文将 R-BiGRU、R-BiBRU-CRF、R-BiLSTM、R-BiLSTM-CRF 和 R-CNN-LSTM 这 5 种模型相互之间进行性能对比, 模型性能对比如表 4 所示.

表 4 模型性能对比 (%)

模型	精确率	召回率	*F*1值
R-BiGRU	83.13	86.25	84.64
R-BiBRU-CRF	84.20	86.13	85.15
R-BiLSTM	84.22	87.28	85.71
R-BiLSTM-CRF	85.38	87.24	86.29
R-CNN-LSTM	82.43	85.14	83.75

由表 4 所示, 在加入 Roberta 预训练模型后, 以上 5 种模型在简要案情文本上的准确率、召回率和 *F*1 值上都表现出了不错的性能, 其中 R-BiLSTM-CRF 模型的 precision 值为 85.38% 和 *F*1 值为 86.29%, 相对于其它 4 种模型来说有较大的领先优势.

为了验证本文设计的卷积神经网络能大幅度提升模型的效率跟模型的性能, 将 CNN-BiLSTM-CRF 模型与 BiLSTM-CRF 进行实验对比, RC-BiLSTM-CRF 与加入 Roberta 预训练模型的 BiLSTM-CRF 模型 (R-BiLSTM-CRF 模型) 进行实验对比, 多方面的实验对比结果如表 5 所示.

由表 5 可知, CNN-BiLSTM-CRF 模型相较于 BiLSTM-CRF 模型, 准确率提高了 1.39%, *F*1 值提高了 0.42%, 以及耗时减少了 7.10%. RC-BiLSTM-CRF 模型相较于 R-BiLSTM-CRF 模型, 准确率提高了 1.85%, 召回率提高了 1.58%, *F*1 值提高了 1.73%, 以及耗时减少了 9.46%. 由此可见, 本文所提出的卷积神经网络能大幅度提升模型的效率跟模型的性能.

表 5 多方面对比实验

模型	精确率 (%)	召回率 (%)	*F*1值 (%)	耗时 (s)
BiLSTM-CRF	78.02	80.06	79.02	2796.9348
CNN-BiLSTM-CRF	79.41	79.47	79.44	2598.3346
R-BiLSTM-CRF	85.38	87.24	86.29	3478.7318
RC-BiLSTM-CRF	87.23	88.82	88.02	3149.6232

综合表 3–表 5 所述, 本文针对简要案情的实体识别方法, 与基线模型的对比之下, 在准确率、召回率和 F_1 值均表现出了突出的性能优势. 得益于本文合理设计的卷积神经网络, 使得本文所提出的 RC-BiLSTM-CRF 模型相较于 R-BiLSTM-CRF 模型, 在大幅度提高模型识别性能的同时, 还降低了训练模型所耗费的时间.

5 结束语

本文主要研究了面向简要案情的命名实体识别任务, 考虑到目前尚无针对该领域命名实体识别的研究, 本文首次尝试对该方向进行了学习和探讨, 构建了用于命名实体识别的简要案情文本的标注数据集, 并在前人研究的基础之上提出了一种改进的识别框架 (RC-BiLSTM-CRF), 通过改进的字符向量生成方法对简要案情数据的字符进行了有效的表示, 生成字符向量, 通过该方法中合理设计的卷积神经网络层对字符向量的局部细粒度特征进行提取, 降低了字符向量维度, 解决了预训练模型带来的字符向量冗长的问题, 框架参数量的减少促使模型整体参数收敛的速率大幅度提高, 为弥补一维卷积层在字符序列上下文特征和依赖关系提取的缺陷, 在模型中引入 BiLSTM 层, 最后利用 CRF 层对文本序列标签进行约束输出. 本文提出的 RC-BiLSTM-CRF 网络框架, 相对于未加入本文设计的卷积神经网络的网络框架, 在准确度、召回率和 *F*1 值上分别提高了 1.85%、1.58% 和 1.73%, 总耗时减少了 9.46%, 与其它 4 种模型相比较, 在准确率、召回率和 *F*1 值 3 个评价标准上均取得了最好的效果. 由于本实验是在标注量少的简要案情数据集上进行的, 在接下来的工作中, 可拓展简要案情的数据规模, 使得模型的鲁棒性更好.

© 中国科学院软件研究所 http://www.c-s-a.org.cn

参考文献

1 Li J, Sun AX, Han JL, *et al*. A survey on deep learning for named entity recognition. arXiv: 1812.09449, 2018.

2 刘浏, 王东波. 命名实体识别研究综述. 情报学报, 2018, 37(3): 329–340. [doi: 10.3772/j.issn.1000-0135.2018.03.010]

3 Žukov-Gregorič A, Bachrach Y, Coope S. Named entity recognition with parallel recurrent neural networks. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Melbourne: Association for Computational Linguistics, 2018. 69–74.

4 Pham NQ, Kruszewski G, Boleda G. Convolutional neural network language models. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin: Association for Computational Linguistics, 2016. 1153–1162.

5 Huang ZH, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv: 1508.01991, 2015

6 Zhang Y, Yang J. Chinese NER using lattice LSTM. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018. 1554–1564.

7 Gui T, Ma RT, Zhang Q, *et al*. CNN-based Chinese NER with lexicon rethinking. Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, 2019. 4982–4988.

8 Li XN, Yan H, Qiu XP, *et al*. FLAT: Chinese NER using flat-lattice transformer. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 6836–6842.

9 李源, 马磊, 邵党国, 等. 用于社交媒体的中文命名实体识别. 中文信息学报, 2020, 34(8): 61–69. [doi: 10.3969/j.issn.1003-0077.2020.08.008]

10 李健龙, 王盼卿, 韩琪羽. 基于双向 LSTM 的军事命名实体识别. 计算机工程与科学, 2019, 41(4): 713–718. [doi: 10.3969/j.issn.1007-130X.2019.04.019]

11 韩鑫鑫, 贲可荣, 张献. 军用软件测试领域的命名实体识别技术研究. 计算机科学与探索, 2020, 14(5): 740–748. [doi: 10.3778/j.issn.1673-9418.1906031]

12 宁尚明, 滕飞, 李天瑞. 基于多通道自注意力机制的电子病历实体关系抽取. 计算机学报, 2020, 43(5): 916–929. [doi: 10.11897/SP.J.1016.2020.00916]

13 Pan SJ, Yang Q. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345–1359. [doi: 10.1109/TKDE.2009.191]

14 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2017. 6000–6010.

15 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics, 2019. 4171–4186.

16 Liu YH, Ott M, Goyal N, *et al*. RoBERTa: A robustly optimized BERT pretraining approach. arXiv: 1907.11692, 2019.

17 沈军, 廖鑫, 秦拯, 等. 基于卷积神经网络的低嵌入率空域隐写分析方法. 软件学报, 2020, 32(9): 2901–2915. [doi: 10.13328/j.cnki.jos.005980]

© 中国科学院软件研究所 http://www.c-s-a.org.cn