基于 LERT 和 BiTCN 的金融领域命名实体识别

陈雪松[1],王璐瑶[1],王浩畅[2]
(1. 东北石油大学 电气信息工程学院,黑龙江 大庆 163318;
2. 东北石油大学 计算机与信息技术学院,黑龙江 大庆 163318)

摘 要:针对传统的命名实体识别方法难以解决金融文本中一词多义且文本的语义特征提取不够充分的问题,提出了一种基于 LERT-BiTCN-CRF 的金融领域命名实体识别模型。首先,使用 LERT 模型对输入的金融文本进行预训练以生成相对应字符向量;然后,通过在 TCN 内部增加反向卷积层将其改进为 BiTCN,采用 BiTCN 对字符向量进行编码以提取字符向量的全局语义特征;最后,通过 CRF 进行解码以得到最佳的预测标签序列。在公开数据集 ChFinAnn 和自制数据集 FinanceNER 两个金融领域数据集上进行对比实验,该模型在两个数据集上的 F1 值分别达到了 84.16% 和 92.17%。相较于其它模型,该模型在金融领域的命名实体识别任务中效果更好,表明该模型具有一定的有效性。同时又在公开的 Resume 数据集上进行对比实验,该模型 F1 值相较于基线模型 BiGRU-CRF 提升 2.31%,表明该模型具有一定的泛化性。

关键词:LERT 模型;金融领域;命名实体识别;双向时间卷积网络;条件随机场
中图分类号:TP391.1 **文献标识码**:A **文章编号**:1673-629X(2025)03-0125-08
doi:10.20165/j.cnki.ISSN1673-629X.2024.0336

Named Entity Recognition in Finance Field Based on LERT and BiTCN

CHEN Xue-song[1], WANG Lu-yao[1], WANG Hao-chang[2]
(1. School of Electrical and Information Engineering, Northeast Petroleum University, Daqing 163318, China;
2. School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China)

Abstract: In order to solve the problem that the traditional named entity recognition method is difficult to solve the problem of multiple meanings of words in financial texts and insufficient semantic feature extraction of texts, a named entity recognition model in the financial field based on LERT-BiTCN-CRF was proposed. Firstly, the LERT model was used to pre-train the input financial text to generate the corresponding character vectors. Then, by adding a reverse convolutional layer inside the TCN, it was improved into BiTCN, and the BiTCN was used to encode the character vector to extract the global semantic features of the character vector. Finally, CRF was used to decode to obtain the best predicted label sequence. Comparative experiments were carried out on two financial domain datasets, the public dataset ChFinAnn and the self-made dataset FinanceNER, and the F1 values of the model on the two datasets reached 84.16% and 92.17%, respectively. Compared with other models, the proposed model has better effect in the named entity recognition task in the financial field, indicating that the model has certain effectiveness. At the same time, comparative experiments were carried out on the public Resume dataset, and the F1 value of the model was increased by 2.31% compared with the baseline model BiGRU-CRF, indicating that the model has a certain generalization.

Key words: LERT model; financial field; named entity recognition; bi-directional temporal convolutional network (BiTCN); conditional random field (CRF)

0 引 言

命名实体识别(Named Entity Recognition, NER)是信息提取、知识图谱构建等自然语言处理(Natural Language Processing, NLP)任务的基础[1],自提出就得到了国内外研究学者的高度关注。

从 1991 年开始,就有研究者试图从金融新闻文本中抽取出公司名称[2],随着互联网信息技术化和 NLP 技术的不断发展,网民规模持续扩大,这使得互联网数

收稿日期:2024-07-29 **修回日期**:2024-11-29
基金项目:国家自然科学基金资助项目(61402099,61702093)
作者简介:陈雪松(1972-),女,教授,博士,研究方向为信息隐藏、信号与信息处理;通信作者:王璐瑶(1998-),女,硕士研究生,研究方向为自然语言处理。

据呈指数趋势增加,如何挖掘出有用的信息来进行研究,是金融行业亟需解决的重要问题。

命名实体识别的发展可分为:基于规则和字典、基于统计机器学习和基于深度学习的研究方法。

基于规则和字典的方法通过事先定义规则和构建专门的字典来识别实体。在早期研究中的金融实体主要是金融企业和机构名称以及它们的缩写。2002 年,王宁等[3]总结了关于金融机构名和企业名的上下文信息及实体特征,并利用手工创建的知识库进行实体识别。统计机器学习方法主要是通过训练标记好的文本来预测未知的样本数据,它是基于特征工程的方法。Wang 等[4]于 2014 年提出了将条件随机场(Conditional Random Field,CRF)和信息熵相结合的方法进行金融实体的识别,改善了金融文本中的实体识别性能。和之前依靠构建规则和字典的研究方法相比,基于机器学习方法取得了一定进展,但仍需要投入大量人力和时间。

随着各种神经网络模型的发展与应用,基于深度学习的实体识别方法成为主流。2015 年,Huang 等[5]提出将双向长短期记忆神经网络(Bidirectional Long Short-Term Memory,BiLSTM)与 CRF 相结合的方法用于序列标注任务。2018 年,Jiao 等[6]又将双向门控循环单元(Bidirectional Gated Recurrent Unit,BiGRU)与 CRF 相结合进行实体识别。之后,BiLSTM/BiGRU-CRF 模型成为 NER 研究的主流模型,很多研究者将其应用到医学[7]、农业[8]等领域,并作为基线模型并针对其不足来进行改进。

近年来,随着 BERT 等预训练模型的出现,将预训练模型与 BiLSTM-CRF 模型相结合使得 NER 研究迎来了新的发展机遇。2019 年,彭小钰[9]提出了 BERT-BiLSTM-CRF 模型,BERT 模型通过在大规模语料上进行自监督学习,与传统方法相比,能够更好地捕捉上下文之间的关联信息,因此,引入预训练模型后可以提升模型的性能表现。2021 年,徐泽蕙[10]提出了融合 BERT 模型多层输出的金融实体识别模型,相较于只使用最后一层获取的字符向量特征,性能亦有所提升。2023 年,白旭等[11]构建了基于 BERT 的实体识别系统,在自行构建金融领域嵌套实体数据集上取得了不错的效果。

此外,为更充分地提取到有用的信息,研究人员又相继提出各种方法。2020 年,刘宇瀚等[12]提出的结合字形特征与迭代学习的模型相较于基线模型 BiLSTM-CRF 的 F1 值提升 1.52%。2022 年,焦樵[13]、Zhi 等[14]、朱家成[15]针对传统的 NER 模型特征提取能力不够充分的问题,设计了多级别特征融合的方法来提取金融实体的上下文信息。盛金兰[16]在 BiLSTM-CRF 模型的基础上引入了自注意力机制来进一步提取金融票据特征。Zhang 等[17]通过引入基于门的多通道注意力机制来学习增强的汉字特征,以描绘金融实体边界。2024 年,李淦[18]提出基于迁移语料库训练的 ELMo-BiLSTM-CRF 模型,相较于基线模型性能也有所提升。

虽然 BERT 模型强大的迁移学习能力,然而金融实体较为复杂,其难以学习到正确的边界信息,因此需要学习更多的语言特征。针对传统模型对金融实体特征提取不够充分的问题,大多数研究都是采用 BiLSTM 与 IDCNN 进行特征融合或引入注意力机制的方式进行改进,但是也增加了模型的复杂度。此外,目前可用于金融领域进行命名实体识别研究的开源数据集较少。针对上述问题,该文提出一种基于 LERT-BiTCN-CRF 的金融实体识别模型,主要贡献如下:

(1)收集了大量金融文本数据,通过 doccano 平台进行实体标注,构建了可用于 NER 任务的金融领域数据集 FinanceNER。

(2)以 BiLSTM/BiGRU-CRF 作为基线模型进行研究,针对该模型不能解决金融文本中一词多义且文本特征提取不够充分的问题,采用基于语言学信息增强的预训练模型 LERT 作为嵌入层。

(3)在减少模型复杂度的同时能够增强特征提取性能的情况下,改进 TCN 模型将 BiTCN 模型作为特征提取层,在 TCN 模型内中增加后向卷积改善其只能单向提取特征的问题,并在两个金融领域数据集以及公开的通用领域数据集 Resume 上进行实验,验证了该模型的有效性和泛化性。

1 模型架构

模型的整体结构如图 1 所示。首先,利用 LERT 模型得到具有语言信息增强的金融文本动态字符向量 $\boldsymbol{X}$,之后通过 BiTCN 对字符向量进行编码,分别得到前向序列特征向量 $\boldsymbol{F}$ 和后向序列特征向量 $\boldsymbol{B}$,并进行特征融合获取全局特征向量 $\boldsymbol{C}$,最后通过 CRF 对向量解码,从而得到预测实体结果。

1.1 LERT 嵌入层

LERT 模型[19]是由哈工大讯飞联合实验室提出的一种语言学信息增强模型,旨在将语言特征融入预训练模型。大多数预训练模型(如 BERT 模型)都是在文本表面进行预训练的,并未考虑到语言特征。LERT 预训练语言模型则是在这种预训练的基础上进行了改进,它结合了掩码语言模型(Masked Language Model,MLM)和 3 种语言学信息预训练(Linguistically Informed Pretraining,LIP)的训练机制,以提高模型对语言特征的学习效果,丰富模型的表征能力。它的模

型架构与 BERT 模型一致,由 12 层双向的 Transformer 编码器堆叠而成。

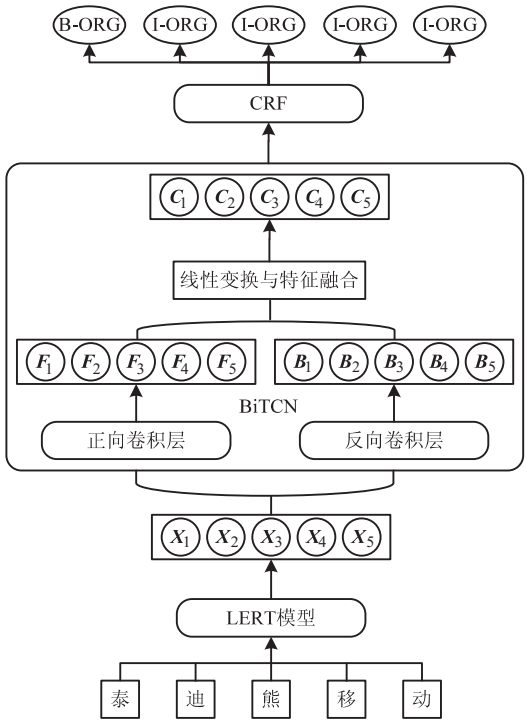

图 1　LERT-BiTCN-CRF 的模型结构

该文使用 LERT 模型将输入的金融文本转换成具有语言信息增强的动态字符向量。假定输入的金融文本的定义为 $E=(E_0,E_1,\cdots,E_n)$,LERT 模型对输入的文本进行序列化,之后经过 LERT 模型中多层 Transformer 编码器的训练,最终将向量转换成输出文本字符向量,其向量表示如下:

$$\boldsymbol{X}=(\boldsymbol{X}_0,\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n) \tag{1}$$

其中,E_i 表示输入金融文本中的第 i 个字,$\boldsymbol{X}_i$ 表示输入金融文本中第 i 个字所对应的字符向量。

1.2　BiTCN 编码层

用于时间序列建模的神经网络模型:时间卷积网络(Temporal Convolutional Network,TCN)[20] 具有并行化计算的优势。相较于存在梯度爆炸的循环神经网络(Recurrent Neural Network,RNN),TCN 能够更有效地捕捉序列中的长期依赖关系。它的核心是通过一维卷积操作对时间序列数据进行特征提取和建模。近年来,该模型在许多文本序列任务中得到应用。TCN 主要包括因果卷积、空洞卷积和残差链接三部分,它的结构如图 2 所示,其中,d 为空洞卷积的空洞因子,$\{x_n\}$ 为输入序列,$\{y_n\}$ 为输出序列。

(1)因果卷积。因果卷积的特征在于模型中的信息只能单向流动。与传统卷积运算不同,因果卷积对序列数据的处理具有严格的时序性,仅能从历史数据中抽取信息,而不能对未来的数据进行观测。

(2)空洞卷积,又称膨胀卷积。传统的 CNN 网络使用固定大小的卷积核和池化层,这限制了模型对不同尺度和大小特征的处理能力。为了获得更大的感受野,研究人员开始对 CNN 增加池化层,但也带来了信息丢失的问题。尽管因果卷积在处理复杂数据时展现出了独有的优势,但是它的建模速度仍然受卷积核大小限制,且难以捕获到金融领域文本间的长距离依赖关系。为解决这一问题,TCN 引入了空洞卷积这一概念。

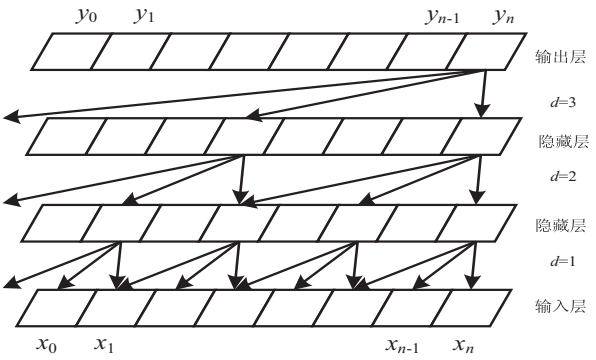

图 2　TCN 的结构

空洞卷积对输入信号进行卷积操作是以间隔为单位进行采样的,采样率由空洞因子 d 控制。在图 2 的结构中,最底层($d=1$)表示在输入过程中对每个点都进行采样;中间层($d=2$)表示每隔一定间距采样一个点作为输入。随着层级的增加,每个卷积核能够观察到更远距离的输入信息,即具有了更大的感受野,从而使得整个网络能够捕捉到更大范围的上下文信息。

(3)残差链接。虽然使用了空洞卷积,但有时模型网络结构仍然很深,这会引起梯度爆炸等问题。为解决这一问题,TCN 在每一层卷积后都加上残差链接模块,残差链接模块如图 3 所示。每个残差模块中都含有两层膨胀因果卷积和非线性映射(使用 ReLU 激活函数),并在此基础上添加了权重归一化和丢弃率对网络进行正则化,防止模型过拟合。

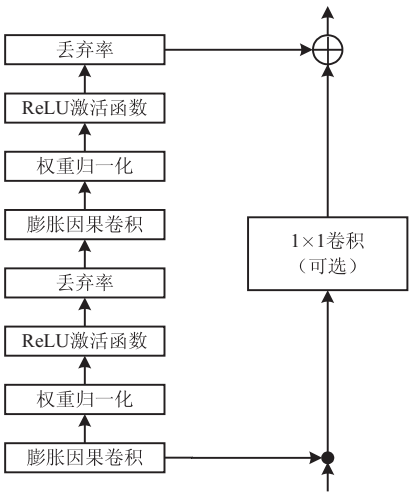

图 3　TCN 中的残差链接模块

在 NLP 领域中,文本序列中某些字词可能会受上下文的影响而表现出不同的表达方式,若只针对单向的序列进行建模,无法有效提取文本的语义特征,对文本进行双向建模可以获取更加充分全面的特征,从而提高模型对文本的理解和表达能力。因此,该文将 TCN 模型改进成 BiTCN 模型,在 TCN 模型内部增加后向卷积,改善 TCN 模型中因果卷积只能单向提取特征的问题。将 LERT 模型转换的字符向量 $\boldsymbol{X}$ 输入到 BiTCN 模型中,通过前向卷积层和后向卷积层分别提取每个字符向量 $\boldsymbol{X}_i$ 的前向特征向量 $\boldsymbol{F}_i$ 和后向特征向量 $\boldsymbol{B}_i$,其向量表示如下:

$$\boldsymbol{F}_i = \sum_{j=0}^{k-1} \boldsymbol{W}_j \times \boldsymbol{X}_{i-j} \tag{2}$$

$$\boldsymbol{B}_i = \sum_{j=0}^{k-1} \boldsymbol{W}_j \times \boldsymbol{X}_{n-(i-j)} \tag{3}$$

其中,$\boldsymbol{W}_j$ 是第 j 个位置的卷积核权重,k 为滤波器大小,$\boldsymbol{X}_{i-j}$ 是输入序列中与当前卷积核对应的值,n 为序列长度。然后,分别将经过多层膨胀因果卷积计算后得到的前向特征向量 $\boldsymbol{F}$ 和后向特征向量 $\boldsymbol{B}$ 进行归一化处理与线性变换。

$$\boldsymbol{F}' = \boldsymbol{F} \times \boldsymbol{V}^{n\times m} \tag{4}$$

$$\boldsymbol{B}' = \boldsymbol{B} \times \boldsymbol{V}^{n\times m} \tag{5}$$

其中,$\boldsymbol{V}^{n\times m}$ 为线性层的权重矩阵,n 为变换前特征向量的维度,m 为变换后特征向量的维度。最后经过特征融合的向量 $\boldsymbol{C}$ 表示为:

$$\boldsymbol{C} = \mathrm{concat}(\boldsymbol{F}', \boldsymbol{B}') \tag{6}$$

1.3 CRF 解码层

CRF 是一种概率图模型,用于建模标记序列的条件概率分布。它是给定输入序列的情况下,对输出序列进行建模的一种判别式模型。

在构建模型中,使用 CRF 作为解码层,计算标签序列 $Y=(Y_0, Y_1, \cdots, Y_n)$ 的条件概率,预测序列 Y 的得分为:

$$S(E,Y) = \sum_{i=0}^{n} \boldsymbol{A}_{Y_i, Y_{i+1}} + \sum_{i=1}^{n} \boldsymbol{P}_{i, Y_i} \tag{7}$$

其中,$\boldsymbol{A}$ 为转移矩阵,$\boldsymbol{P}$ 为发射矩阵,$\boldsymbol{A}_{Y_i, Y_{i+1}}$ 是标签 Y_i 转移成 Y_{i+1} 的转移得分,$\boldsymbol{P}_{i, Y_i}$ 是第 i 个字被标记为的概率得分,n 为序列的长度。当 $S(E,Y)$ 越高时,模型对序列的预测越准确。对于给定的输入序列和输出序列,条件概率表示为:

$$P(Y \mid E) = \frac{\exp(S(E,Y))}{\sum_{Y\in Y_E} S((E,Y'))} \tag{8}$$

其中,Y_E 为所有可能的标签序列,Y' 为正确的标签序列。CRF 在解码过程中通过 Viterbi 算法选择预测序列中得分最高的标签序列 $\hat{Y}$,如下:

$$\hat{Y} = \arg\max_{Y\in Y_E} S(E,Y) \tag{9}$$

模型训练过程中的最小化损失函数为:

$$L = \sum_{i=1}^{n} \log(P(Y_i \mid E_i)) \tag{10}$$

2 实 验

2.1 数据集

实验中共用到三个数据集。两个金融领域的数据集:公开数据集 ChFinAnn 和自制数据集 FinanceNER,两个数据集均按照 7∶1.5∶1.5 随机划分成训练集、验证集和测试集;同时为验证该模型在不同领域的泛化性,选择了公开的 Resume 数据集进行对比实验。

(1)ChFinAnn 数据集。该数据集共包含 10 类实体,分别为:股票代码、股票简称、企业名、日期、股权持有人、质权人、股份、比率、价格和机构。实体总数为 32 007 个,其中训练集有 21 132 个实体,验证集有 5 566个实体,测试集有 5 309 个实体。

(2)FinanceNER 数据集。该数据集为自制的数据集,语料来源于两部分,一部分是由中国计算机学会举办的 2019CCF 大数据与计算智能大赛中的互联网金融新实体发现竞赛中提供的训练集语料,另一部分是由 SmoothNLP 团队所收集的有关金融资讯新闻和上市公司的文本语料。以上两部分都是公开的数据文本,共选取了其中 1 000 条金融新闻语料并使用 doccano 标注平台进行实体标注,包含人名、地名、金融企业及组织机构名称、金融产品名及时间 5 类实体,实体总数共计 44 064 个,其具体的实体数量分布如表 1 所示。

表 1 FinanceNER 数据集实体数量

实体类别	训练集	验证集	测试集
人名	5 556	884	918
地名	5 548	970	966
金融企业及组织机构名	8 993	1 794	1 794
产品名	5 557	1 014	967
时间	6 572	1 281	1 250
总数	32 226	5 943	5 895

(3)Resume 数据集。该数据集为简历数据集,其统计信息如表 2 所示。

表 2 Resume 数据集统计

类型	训练集	验证集	测试集
句子	3.8k	0.5k	0.5k
字符	124.1k	13.9k	15.1k

2.2 实验环境和参数设置

实验中所用到的环境及相关配置如下:操作系统为 Windows 10,CPU 为 Intel(R) Xeon(R) Platinum 8255C CPU @ 2.50 GHz,GPU 为 NVIDIA GeForce

RTX 3080(10 GB),开发语言使用 Python3.7,开发框架使用 Pytorch1.10.0。采用的 LERT 预训练语言模型作为嵌入层。在训练过程中使用 Adam 优化器来更新参数。模型的主要参数设置见表 3。

表 3　实验参数设置

参数	值
BERT 学习率	5e-5
LSTM/GRU 学习率	0.001
隐藏层维度	128
训练批次	16
训练轮数	16
TCN 的卷积核	5
TCN 的卷积层	6(Finance NER)/5(ChFanAnn)
TCN 的空洞因子	2^n+1

2.3　评价指标

为了能够有效地评价模型的性能,采用精确率 P 、召回率 R 和 F1 值来评估模型的实验效果。其具体计算公式如下:

$$P = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}) \tag{11}$$

$$R = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \tag{12}$$

$$\mathrm{F1} = 2 \times P \times R/(P + R) \tag{13}$$

其中,TP 表示预测实体标签和实际实体标签都为正例的数据量,FP 表示预测实体标签为正例而实际实体标签为负例的数据量,FN 表示预测实体标签为负例而实际实体标签为正例的数据量。

2.4　实验设计

2.4.1　对比实验

为验证 LERT-BiTCN-CRF 算法的有效性和泛化能力,该文对多个实体识别算法进行对比。主要方法如下:

BiLSTM-CRF:Huang 等[5] 于 2015 年提出的模型,通过 BiLSTM 对输入的文本进行编码,再通过 CRF 对向量解码。

BiGRU-CRF:Jiao 等[6] 于 2018 年提出的模型,通过 BiGRU 对输入的文本进行编码,再通过 CRF 对向量解码。

BERT-IDCNN-GAM-CRF:BERT-IDCNN-CRF 为 2020 年李妮等[21] 提出的方法,但 IDCNN 只能提取局部特征,容易忽略上下文信息,因此引入全局注意力机制(Global Attention Mechanism,GAM)[22] 与该模型相结合作为对比算法。利用 BERT 模型得到输入的字符向量,之后通过 IDCNN 提取局部信息,再通过 GAM 捕获全局信息增强特征提取能力,最后通过 CRF 对向量解码。

BERT-BiGRU-CRF:2021 年 Chen 等[23] 提出的命名实体识别方法,利用 BERT 模型得到输入的字符向量,之后通过 BiGRU 对字符向量进行编码,再通过 CRF 对向量解码。

BERT-BiTCN-CRF:将 BERT-BiGRU-CRF 中的 BiGRU 替换为 BiTCN,以对比两种特征提取模块。利用 BERT 模型得到输入的字符向量,之后通过 BiTCN 对字符向量进行编码,分别得到前向序列特征向量和后向序列特征向量,并进行线性变换和特征融合,以获取全局特征向量,最后通过 CRF 对向量解码。

LERT-BiGRU-CRF:将 BERT-BiGRU-CRF 中的 BERT 替换为 LERT,以对比两种字符嵌入模块。利用 LERT 模型得到具有语言信息增强的金融文本的字符向量,之后通过 BiGRU 对字符向量进行编码,再通过 CRF 对向量解码。

将 LERT-BiTCN-CRF 模型与上述模型分别在 FinanceNER 和 ChFinAnn 数据集上进行对比实验,实验结果分别如表 4 和表 5 所示。

表 4　FinanceNER 数据集上的对比实验　%

模型	P	R	F1
BiLSTM-CRF	79.42	73.90	76.56
BiGRU-CRF	77.53	77.27	77.40
BERT-IDCNN-GAM-CRF	80.18	84.65	82.35
BERT-BiGRU-CRF	81.53	84.88	83.17
BERT-BiTCN-CRF	81.57	85.32	83.40
LERT-BiGRU-CRF	82.24	85.07	83.63
LERT-BiTCN-CRF	82.67	85.71	84.16

由表 4 和表 5 可以看出:

(1)通过对比 BiLSTM-CRF 和 BiGRU-CRF,BiGRU-CRF 在两个数据集上的 F1 值都比 BiLSTM-CRF 的 F1 值要高,可见,BiGRU-CRF 能够更好地处理上下文信息。该模型参数更少,结构更简单,计算效率更高。

(2)后续的实验都加入了预训练模型来获取金融文本动态的字符向量,相较于未使用预训练模型的 BiGRU-CRF,F1 值均在一定程度上有所提升,说明加入预训练模型可以提升 NER 任务性能。通过对比 BERT-IDCNN-GAM-CRF、BERT-BiGRU-CRF、BERT-BiTCN-CRF 和 LERT-BiGRU-CRF、LERT-BiTCN-CRF 在两个数据集上的实验,说明使用 BiTCN 作为特征提取层的模型效果要更好,IDCNN 加 GAM 可以提取局部和全局信息,BiGRU、BiTCN 都可以从前后两个方向来获取上下文信息,但是,BiGRU 在处理长序列时可能存在梯度消失或者梯度爆炸的问题,而 BiTCN 具有更少的参数量,通过卷积操作可以在更大范围内传播信息,更有利于捕捉序列中较远位置的依赖关系。

表 5 ChFinAnn 数据集上的对比实验 %

模型	*P*	*R*	F1
BiLSTM-CRF	88.41	83.95	86.12
BiGRU-CRF	89.96	83.71	86.72
BERT-IDCNN-GAM-CRF	87.62	91.67	89.60
BERT-BiGRU-CRF	89.72	92.00	90.85
BERT-BiTCN-CRF	89.50	92.69	91.07
LERT-BiGRU-CRF	90.91	93.08	91.98
LERT-BiTCN-CRF	90.45	93.96	92.17

(3)通过对比 BERT-BiGRU-CRF、LERT-BiGRU-CRF 和 BERT-BiTCN-CRF、LERT-BiTCN-CRF,验证了 LERT 模型更适合用于金融文本的实体识别。在两个数据集上的实验结果表明,基于 LERT-BiTCN-CRF 的实体识别模型在金融领域数据集上的效果更好、泛化能力更强。

2.4.2 与其他模型对比

为进一步证明该方法的有效性,将该模型与其他先进实体识别模型进行对比,表 6 为两个数据集上对比的 F1 值。对比模型如下:

Lattice LSTM:Zhang 等[24]于 2018 年提出的词汇增强模型,使用 Lattice 结构动态地将词汇信息融入到字符向量中。

BERT-BiLSTM-CRF:2019 年彭小钰[9]提出的模型,通过 BERT 模型获取字符向量后使用 BiLSTM 进行特征提取,最后通过 CRF 进行解码。

BERT-BiLSTM-IDCNN-CRF:2022 年焦樵[13]提出的金融领域实体识别模型,通过 BERT 模型获取输入文本的字符向量,之后使用 BiLSTM 和 IDCNN 分别提取局部特征和上下文信息,最后通过 CRF 进行解码。

MFF-CNER:Zhi 等[14]于 2023 年提出的方法,该模型在 BERT-BiLSTM-IDCNN-CR 模型的基础上引入了加权注意力机制来进一步提取金融文本的上下文特征。

FLAT:LIX 等[25]于 2020 年提出的方法,将词汇信息通过 Transformer 动态融入模型。

表 6 与其他模型对比的 F1 值 %

模型	FinanceNER	ChFinAnn
Lattice LSTM	80.77	88.65
BERT-BiLSTM-CRF	82.62	90.84
BERT-BiLSTM-IDCNN-CRF	83.26	90.86
MFF-CNER	83.42	91.20
FLAT	83.82	91.40
LERT-BiTCN-CRF	84.16	92.17

通过表 6 可以看出结合 BERT 后的 F1 值要高于 Lattice LSTM 模型;使用 IDCNN 和 BiLSTM 对局部和全局特征进行特征融合后实体识别效果要优于只使用 BiLSTM 提取特征;添加加权注意力机制后的 MFF-CNER 模型效果进一步提升;FLAT 模型在对比模型中的 F1 值最高;而文中方法在两个数据集上的 F1 值均优于其他模型,进一步表明文中方法具有一定的有效性。

2.4.3 不同领域泛化性和鲁棒性对比实验

为进一步验证该方法的泛化能力及鲁棒性,选用公开且不属于金融领域的 Resume 数据集进行模型泛化性与鲁棒性实验。

采用 BiGRU-CRF、BERT-BiGRU-CRF、BERT-BiTCN-CRF 与 LERT-BiTCN-CRF 模型进行对比,实

验中 TCN 的卷积核和卷积层均设为 5,其他参数与表 2 一致,实验结果如表 7 所示。

表 7　Resume 数据集上的对比实验　%

模型	*P*	*R*	F1
BiGRU-CRF	92.88	93.68	93.28
BERT-BiGRU-CRF	94.51	95.74	95.12
BERT-BiTCN-CRF	94.74	95.74	95.24
LERT-BiTCN-CRF	94.72	96.48	95.59

通过四组实验验证,LERT-BiTCN-CRF 模型在该数据集上的 F1 值均优于其他基线模型,表明该模型具有良好的泛化性和鲁棒性。

2.4.4　预训练模型对比实验

为探究不同预训练模型对金融领域文本中实体的识别性能,在两个数据集上对比了 BERT 模型、BERT-wwm 模型、RoBERTa-wwm 模型和 LERT 模型对实验的影响。BERT 模型是谷歌发布的预训练模型,通过双向编码器和 Transformer 结构来学习文本表示。BERT-wwm 模型是对原始 BERT 模型的改进,采用了全词掩码(whole word masking,wwm)的方式,在训练过程中选择整个词作为掩码的单位,而不是原始的单词片段。这种方法可以更好地处理中文等没有天然分词特性的语言,提高模型在 NER 任务中的表现。RoBERTa-wwm 同样采用了 wwm 的方式,在 BERT 模型的基础上使用了更多的无监督语料进行训练,改进了训练批次和动态掩码策略。LERT 模型则是在 BERT 模型的基础上融合了多种语言学知识后所得到的预训练模型。对比不同的预训练模型对金融实体识别性能的影响时,只需替换嵌入层,其他层保持不变,在 FinanceNER 和 ChFinAnn 数据集上的 F1 值如表 8 所示。

表 8　不同预训练模型的 F1 值　%

模型	FinanceNER	ChFinAnn
BERT	83.40	91.07
BERT-wwm	83.70	91.20
RoBERTa-wwm	84.11	91.61
LERT	84.16	92.17

从表 8 中可以看出,LERT 模型在两个数据集上的 F1 值都高于其它预训练模型,说明融合语言学的 LERT 模型更适合作为嵌入层来处理金融文本数据。

2.4.5　消融实验

为探究不同模块的有效性,在 FinanceNER 数据集上进行了消融实验,结果如表 9 所示。

表 9　FinanceNER 数据集上的消融实验　%

模型	*P*	*R*	F1
BiTCN-CRF	77.82	80.37	79.07
LERT-CRF	80.60	85.24	82.86
LERT-BiLSTM-CRF	81.83	84.82	83.30
LERT-TCN-CRF	82.08	86.02	84.00
LERT-BiTCN-CRF	82.67	85.71	84.16

实验结果表明,在对金融领域的文本进行实体识别时,引入 LERT 模型和 BiTCN 都会提升对金融实体的识别效果。当去除 BiTCN 后,模型 F1 值下降 1.3 百分点,当去除 LERT 模型后,F1 值下降 5.09 百分点。另因 BiLSTM-CRF 模型为 NER 任务中常见的基线模型,所以分别将 BiTCN 和 BiLSTM 两个神经网络模型作为编码层进行消融对比,实验结果进一步表明使用 BiTCN 可以提升对金融实体的识别性能。

3　结束语

针对金融领域文本实体识别困难的问题,提出 LERT-BiTCN-CRF 模型。该模型利用 LERT 模型得到具有语言信息增强的金融文本字向量表示,使用 BiTCN 从前、后两个方向学习金融文本的全局语义特征,由于该网络具有并行化计算的特点,可以弥补 RNN 等网络存在的梯度爆炸等问题,最后通过 CRF 获取预测结果。经过在两个金融领域文本的数据集和公开的 Resume 数据集上的实验验证,该模型更能准确地获取金融文本的语义特征,提升对金融实体的识别性能,且该模型具有较好的泛化性。由于金融领域文本存在实体边界识别模糊的问题,在后续的工作中,会收集更多金融文本来构建金融词典进行研究。

参考文献:

[1]　祁鹏年,廖雨伦,覃　飙. 基于深度学习的中文命名实体识别研究综述[J]. 小型微型计算机系统,2023,44(9):1857-

1868.
[2] RAU L F. Extracting company names from text[C]//Proceedings of the seventh IEEE conference on artificial intelligence applications. Miami Beach:IEEE,1991:29-32.
[3] 王 宁,葛瑞芳,苑春法,等. 中文金融新闻中公司名的识别[J]. 中文信息学报,2002,16(2):1-6.
[4] WANG S,XU R,LIU B,et al. Financial named entity recognition based on conditional random fields and information entropy[C]//2014 international conference on machine learning and cybernetics. Lanzhou:IEEE,2014:838-843.
[5] HUANG Z,XU W,YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. Computer Science,2015,63(3):32-42.
[6] JIAO Z,SUN S,SUN K. Chinese lexical analysis with deep Bi-GRU-CRF network[EB/OL]. (2018-07-05)[2024-04-12]. https://arxiv.org/pdf/1807.01882v1.
[7] DENG N,FU H,CHEN X. Named entity recognition of traditional Chinese medicine patents based on BiLSTM-CRF[J]. Wireless Communications and Mobile Computing,2021,2021:1-12.
[8] QIAN Y,CHEN X,WANG Y,et al. Agricultural text named entity recognition based on the BiLSTM-CRF model[C]//Fifth international conference on computer information science and artificial intelligence (CISAI 2022). Chongqing:SPIE,2023:525-530.
[9] 彭小钰. 面向金融领域的命名实体识别算法的设计与实现[D]. 武汉:华中科技大学,2019.
[10] 徐泽蕙. 基于 BERT 的金融领域命名实体识别方法研究[D]. 南昌:江西财经大学,2021.
[11] 白 旭,周琳娜,杨忠良,等. 金融嵌套命名实体识别系统的实现与应用[J]. 网络安全技术与应用,2023(10):52-56.
[12] 刘宇瀚,刘常健,徐睿峰,等. 结合字形特征与迭代学习的金融领域命名实体识别[J]. 中文信息学报,2020,34(11):74-83.
[13] 焦 樵. 基于深度学习的金融领域命名实体识别方法[D]. 武汉:中南财经政法大学,2022.
[14] ZHI Y,TAO X,JI Y. MFF-CNER:a multi-feature fusion model for Chinese named entity recognition in finance securities[J]. Academic Journal of Science and Technology,2023,7(3):40-49.
[15] 朱家成. 基于特征融合的互联网金融领域命名实体识别算法研究[D]. 西安:西安电子科技大学,2022.
[16] 盛金兰. 面向国际贸易中金融票据命名实体识别方法研究[D]. 南京:南京林业大学,2022.
[17] ZHANG H,WANG X,LIU J,et al. Chinese named entity recognition method for the finance domain based on enhanced features and pretrained language models[J]. Information Sciences,2023,625:385-400.
[18] 李 淦. 面向金融新闻的命名实体识别方法[J]. 电脑知识与技术,2024,20(18):4-6.
[19] CUI Y,CHE W,WANG S,et al. LERT:a linguistically-motivated pre-trained language model[EB/OL]. (2022-11-10)[2024-04-12]. https://arxiv.org/pdf/2211.05344.pdf.
[20] BAI S,KOLTER J Z,KOLTUN V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling[EB/OL]. (2018-04-19)[2024-04-12]. https://arxiv.org/pdf/1803.01271.
[21] 李 妮,关焕梅,杨 飘,等. 基于 BERT-IDCNN-CRF 的中文命名实体识别方法[J]. 山东大学学报:理学版,2020,55(1):102-109.
[22] LIU Y,SHAO Z,HOFFMANN N. Global attention mechanism:retain information to enhance channel-spatial interactions[EB/OL]. (2021-12-10)[2024-04-12]. https://arxiv.org/pdf/2112.05561.
[23] CHEN X,QIU Z. Research on core function of adjacency pairs prediction based on BERT-BIGRU-CRF[C]//Proceedings of the 2021 2nd international conference on control,robotics and intelligent system. New York:Association for Computing Machinery,2021:117-121.
[24] ZHANG Y,YANG J. Chinese NER using lattice LSTM[C]//Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1:long papers). Melbourne:Association for Computational Linguistics,2018:1554-1564.
[25] LIX N,YAN H,QIU XP,et al. FLAT:Chinese NER using flat-lattice transformer[C]//Proceedings of the 58th annual meeting of the association for computational linguistics.[s.l.]:[s.n.],2020:6836-6842.