

文章编号: 1003-0077(2020)11-0074-10

结合字形特征与迭代学习的金融领域命名实体识别

刘宇瀚, 刘常健, 徐睿峰, 骆旺达, 陈奕, 吉忠晟, 应能涛

(哈尔滨工业大学(深圳) 计算机科学与技术学院, 广东 深圳 518055)

摘要: 针对中文金融文本领域的命名实体识别, 该文从汉字自身特点出发, 设计了结合字形特征、迭代学习以及双向长短时记忆网络和条件随机场的神经网络模型。该模型是一种完全端到端且不涉及任何特征工程的模型, 其将汉字的五笔表示进行编码以进行信息增强, 同时利用迭代学习的策略不断对模型整体预测结果进行改进。由于现有的命名实体识别研究在金融领域缺乏高质量的有标注的语料库资源, 所以该文构建了一个大规模的金融领域命名实体语料库 HITSZ-Finance, 共计 31 210 个文本句, 包含 4 类实体。该文在语料库 HITSZ-Finance 上进行了一系列实验, 实验结果均表明模型的有效性。

关键词: 金融领域命名实体识别; 中文语料库; 深度学习

中图分类号: TP391

文献标识码: A

Utilizing Glyph Feature and Iterative Learning for Named Entity Recognition in Finance Text

LIU Yuhan, LIU Changjian, XU Ruifeng, LUO Wangda,

CHEN Yi, JI Zhongsheng, YING Nengtao

(School of Computer Science, Harbin Institute of Technology (Shenzhen),
Shenzhen, Guangdong 518055, China)

Abstract: To deal with Chinese named entity recognition in finance domain, this paper presents a novel neural network model combining glyph feature and iterative learning. Based on the framework of bidirectional long-short term memory networks and conditional random fields, this model encodes wubi input code of Chinese characters for information enhancement and use iterative learning to continuously update predict results. We manually annotate a large-scale financial named entity corpus named HITSZ-Finance, including 31210 sentences and 4 types of entities. Experiment results on HITSZ-Finance corpus demonstrate the effectiveness of the model.

Keywords: named entity recognition in financial field; Chinese corpus; deep learning

0 引言

命名实体识别(named entity recognition, NER)是自然语言处理中的一项基本任务。它具体指从一段文本中识别出具有意义的实体,命名实体识别的性能优劣会直接影响信息抽取、信息检索、问答系统等下游任务的性能,命名实体识别常常作为序列标注任务来处理。

金融事件的主体是各类金融实体,例如,公司名

称、人物名称、金融机构名称等。随着近些年金融证券行业以及互联网的快速发展,网络信息的重要性逐渐得到人们的重视。企业投资者能否从如此海量的无结构化的文本中获取有效的信息对公司的战略决策、未来走势有着很大的影响。

作为信息抽取的重要环节,命名实体识别最早采用基于规则的方法,然而这种利用人工建立规则、设置词典的方法需要耗费大量人力,并且系统无法在众多领域推广。基于统计机器学习的方法可以有效利用文本的统计特性提取特征,减少人力,但需要大量

收稿日期: 2019-12-11 定稿日期: 2020-02-11

基金项目: 国家自然科学基金(61632011, 61876053); 深圳市基础研究项目(JCYJ20180507183527919, JCYJ20180507183608379); 深圳市技术攻关项目(JSGG20170817140856618)

有标签的数据用于模型训练,且模型存在泛化能力不强的特点。近几年,基于深度学习的命名实体识别模型取得了不错的效果,以卷积神经网络(convolutional neural network, CNN)、循环神经网络(recurrent neural network, RNN)等为代表的端到端学习模型具有较强的泛化能力与迁移能力。

然而,尽管在学术界与实体识别相关的诸多子问题已经有许多模型和方法来处理,在金融领域的实体识别仍面临着各式各样的困难与挑战,具体表现为:①金融领域文本缺乏高质量的有标注的语料库资源,导致底层的文本处理技术缺乏数据支撑;②金融领域专有名词较多,常用的分词工具难以进行准确分词;③现有的命名实体识别方法大多没有结合汉字的字形特征,未能充分利用文本信息。

为了解决以上提到的问题,完善当前金融领域命名实体识别的研究,本文首先构建了金融领域命名实体语料库。语料库的原始数据均来自于各大金融新闻网站的新闻文本,一共包含 1 200 篇新闻,共计 15 960 段文本段落,31 210 个文本句,包含公司、人名、品牌和其他共四类实体。据我们了解,这是首次为金融领域命名实体识别构建的大规模、高质量的人工标注中文数据集。除此之外,本文还提出了一种新的神经网络命名实体识别模型(BiLSTM-CRF combined with Wubi embedding and Label embedding, BCCWL)。该模型利用神经网络提取汉字的五笔信息作为字形嵌入,结合字符本身的字嵌入,并通过迭代学习融入标签信息进行命名实体识别,充分利用了字形特征和文本信息,同时减少了分词错误带来的问题。实验结果表明,本文提出的 BCCWL 模型在我们自标注的实体识别语料库 HITSZ-Finance 上获得了较好的实验性能,在准确率、召回率以及 F_1 值上均有一定提升。

1 相关工作

在语料库资源上,常见的中文命名实体识别数据集较为缺乏,这也是制约中文命名实体识别研究的一大瓶颈。目前主流的中文数据集包括:①MSRA 数据集^[1],该数据集发布于 2006 年,时间较早,主要为一些时政类新闻,标注了人名、地名和组织名三大类实体;②Weibo 数据集^[2],该数据集来源于社交平台微博,标注了人名、地名、行政区名与机构名,缺点是数据量较少;③OntoNotes 数据集^[3],该数据集的数据来源比较混杂。为了有效填

补命名实体识别在金融新闻领域的空白,并为中文命名实体语料库做出一定的贡献,本文构建并标注了面向金融领域的命名实体语料库 HITSZ-Finance。

在命名实体识别任务上,我们的工作与现有的方法一致,均采用神经网络的形式进行。命名实体识别作为序列标注问题的一种,早期采用隐马尔可夫模型(hidden Markov model, HMM)和最大熵马尔科夫模型^[4](maximum-entropy Markov model, MEMM)进行序列标注。2003 年 Lafferty 等人提出了统计机器学习模型条件随机场^[5](conditional random field, CRF)。CRF 模型采用全局归一化的方式进行模型的学习训练,在一定程度上优于 MEMM 模型的局部归一化,并在序列标注问题上取得了很好的效果。2011 年 Collobert 等^[6]提出了 CNN-CRF 的模型架构,并取得了相比于其他统计模型更好的效果。近几年,循环神经网络在处理序列问题上展现出了天然的优势,其中,研究者多采用 LSTM 作为基本单元进行实体识别。Huang 等^[7]在 2015 年首次提出 LSTM-CRF 网络架构,并使用手工构建的英文拼写特征作为输入,通过双向长短期记忆网络(bidirectional long-short time memory, BiLSTM)提取特征,然后结合条件随机场网络进行序列标注。

为了减少分词错误带来的负面影响,中文命名实体识别多采用字符级序列标注处理。Hovy 等^[8]在 2016 年采用字符级卷积结合词向量的方式来增强词的表示,而 Lample 等^[9]则完全采用字符级 LSTM 的形式提取各个词的特征。结合分词信息与字表示的方法被证明能对实体识别有一定的提升。Peng 等^[10]提出了一种结合分词与实体识别的多任务学习方法,联合学习的方式使得两个任务的性能都有一定的提升。而 Dong 等^[11]则提出结合汉字的偏旁部首信息的 LSTM-CRF 网络结构以增强汉字的表示信息。Zhang 等^[12]在 2018 年提出 Lattice-LSTM 的新型网络结构,通过对 LSTM 内部结构进行改造达到融入词信息与短语信息的目的。受以上工作启发,我们发现汉字的字形也是一种强特征,Wu 等^[13]利用汉字字形进行语义增强表示,而汉字五笔表示也能在一定程度上反映汉字的构造特点,并由 Nikolov 等^[14]证明是非常有效的。除此之外,受 Seq2Seq^[15]与迭代学习^[16-17]的启发,本文在结合五笔字形嵌入与字嵌入的命名实体识别模型基础上融入标签信息进行进一步信息增强以完成实体识别。

2 金融领域命名实体语料库构建

2.1 数据收集与清洗

本文使用的语料均来源于各大金融数据网站,包括 21 世纪经济报道^①、财新网^②、每经网—公司板块^③、生意社^④、人民网^⑤等,具体方法为通过 scrapy 网络爬虫对以上网站的数据进行爬取,并采用 json 格式存储数据中新闻的标题、内容、时间与网页链接,最终收集新闻文本共计 22 681 篇。

对原始新闻文本的数据清洗主要从以下几个方面进行:

- (1) 剔除特殊符号,例如“©※▲”等;
- (2) 剔除文本中 url 标签和网站链接等;
- (3) 剔除因网页
 等换行或者空格标签导致的大空格;
- (4) 剔除文本长度小于 20 的新闻文本;
- (5) 利用正则匹配剔除一些无关信息,例如“实习编辑”“记者报道”“编者按”等与新闻本身无关的新闻文案编辑内容;
- (6) 对整条新闻文本进行自然段分段。

2.2 标注系统搭建与标注准则

完成原始数据的收集与清洗后,为了方便标注,我们利用 larvel 框架搭建了多用户数据标注系统。该标注系统如图 1 所示,它具有良好的并发性,能够允许许多用户进行多任务标注,同时具有数据增删查改的功能。

文本内容:

同样在今年4月, **广州市从化区环境保护局** 当月18日测得太平污水处理厂出现进水异常,调查后发现先强药业含二氯甲烷的废水未经防范措施而导致污染,被处以10万元整罚款;4月29日,在 **从化区环境保护局** 对于 **先强药业** 的执法检查中发现, **先强药业** 排放的废水中,废水污染物中的化学需氧量浓度超过相关标准,责令 **先强药业** 停产整治十日,处罚款20万元整。

实体	实体类型	位置
先强药业	<input checked="" type="radio"/> 公司 <input type="radio"/> 人名 <input type="radio"/> 品牌 <input type="radio"/> 其他 <input type="radio"/> 有歧义	56 57 58
从化区环境保护局	<input type="radio"/> 公司 <input type="radio"/> 人名 <input type="radio"/> 品牌 <input checked="" type="radio"/> 其他 <input type="radio"/> 有歧义	51 52 53 54
先强药业	<input checked="" type="radio"/> 公司 <input type="radio"/> 人名 <input type="radio"/> 品牌 <input type="radio"/> 其他 <input type="radio"/> 有歧义	65 66 67
广州市从化区环境保护局	<input type="radio"/> 公司 <input type="radio"/> 人名 <input type="radio"/> 品牌 <input checked="" type="radio"/> 其他 <input type="radio"/> 有歧义	5 6 7 8 9
先强药业	<input checked="" type="radio"/> 公司 <input type="radio"/> 人名 <input type="radio"/> 品牌 <input type="radio"/> 其他 <input type="radio"/> 有歧义	86 87 88

异常报告

提交

图 1 标注系统

金融新闻文本往往不仅含有大量与金融相关的实体,包括金融机构、公司、品牌、政府组织等,也含

有大量与金融不相关的实体,如地区、国家、会议等。在数据标注的过程中,我们只关注四大类实体:公司名、品牌名、人名和其他类别。其中媒体机构、金融机构、政府机构、公司下属部门、民间组织、金融产品以及其他数据平台标注均归为其他类别。我们首先安排 4 名标注人员预标注 2 000 个文本段,在标注过程中对各自的标注结果进行比对,收集差异与有歧义的地方,针对模糊和有冲突的语境制定相应的标注准则,部分标注准则如表 1 所示。

表 1 部分标注准则

准则 1: ××公司+国家、地区名称或者地区名称+××公司,根据长匹配原则,应该整体标注

具体实例:

- 德国戴姆勒与奔驰中国达成一系列商业协议
- Swisse 中国区执行总裁安玉婷表示,要从购买国外品牌到通过收购的国外品牌打开国际市场

准则 2: 金融机构、政府机构、民间组织标注为其他

具体实例:

- 对此,5月17日深交所就此下发重组问询函,询问三起交易是否为一揽子交易安排

准则 3: 金融数据平台、社交平台、其他数据平台标注为其他

具体实例:

- Wind 数据显示,截至11月12日,沪深两市共有51只股票处于停牌状态
- 富士康也推出自己的工业互联网平台 BEACON、徐工集团也发布了工业联网平台 Xrea
-

接下来,本文根据已制定的标注准则对余下语料进行标注。在标注过程中,每一段新闻文本首先由至少两名标注者独立标注,即标注过程中标注者完全依赖先前制定好的标注准则,彼此之间没有交流。由于标注过程是按照词级别进行标注的,因此可能存在分词错误的情况,故还需要同时记录分词错误的新闻文本并后期进行人为修改。

独立标注完成后,对于有差异或有错误的标注结果,会有一名额外的标注者参与讨论,直到所有的标注者意见统一后,利用标注系统通过查询新闻文本编号的方式对已标注数据进行人为修改,最终完成标注。

① <http://news.21so.com/chanye/>

② <http://companies.caixin.com/news/>

③ <http://www.nbd.com.cn/columns/346>

④ <http://news.toocle.com/list/c-3511-1.html>

⑤ <http://industry.people.com.cn/>

2.3 语料库构成与统计分析

我们从爬取的 22 681 篇新闻文本中随机选择了 1 200 篇新闻进行标注,按照自然段进行切分,平均每篇新闻文本约 13.3 个自然段,语料库具体信息如表 2、表 3 所示。

表 2 HITSZ-Finance 语料库统计信息 1

段落数	句子数	句平均长度	字符数
15 960	31 210	55.60	1 735.28k

表 3 HITSZ-Finance 语料库统计信息 2

公司	人名	品牌	其他
36 766	7 123	1 753	7 186

在该数据集上,不同标注人员对同一金融文本的标注一致性,我们通过卡帕系数进行衡量。在最终的数据集中,同一金融文本的两个实体标注之间的卡帕一致性^[18]达到了 87.6%,充分说明实体标注的高度一致性,侧面反映了数据集的可靠性。

3 模型

本节主要描述模型的基本模块。我们的基本架构采用 BiLSTM-CRF^[7] 框架,整个模型包含三个模块:字形嵌入模块、迭代学习模块以及条件随机场模块。汉字的字形特征包含汉字的很多信息,模型首先通过字形嵌入模块将汉字的五笔输入转换成五笔向量表示,并将其与每个汉字的字向量表示结合作为模型的整体输入,然后将得到的输入分别送入迭代学习模块(由若干子模块构成)中进行迭代学习,再将最后一个子模块的输出送入条件随机场网络得到最终的标签序列结果。标注方面,我们采用 BIO 标注准则进行字级别序列标注。

3.1 字形嵌入

汉字的五笔字型依据笔画和字形特征对汉字进行编码,是典型的形码输入法。融入汉字的字形特征可有效增强汉字的表示信息。例如,图 2 中,“泰”的五笔表示为“dw i u”,代表 4 种笔画信息。这些笔画可以有效地体现出汉字的构造特征。

目前,利用 CNN 进行文本特征提取已经成为

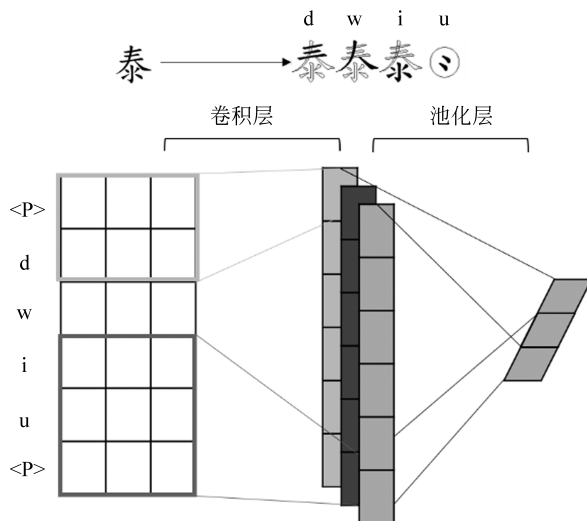


图 2 利用卷积神经网络提取汉字特征

一种通用的处理手段。受 TextCNN^[19] 和文献 [8,20] 的启发,我们首先通过五笔向量查找转换函数得到汉字字符对应的五笔字形嵌入,然后通过卷积和池化操作进行进一步特征提取,得到最终的五笔特征向量。具体过程如下,给定一段文本序列 $S = \{w_1, w_2, \dots, w_i, \dots, w_n\}$, 对于每个汉字 w_i , 我们都可以获取该字的五笔输入表示 $\text{Wubi}(w_i) = \{c_{i,1}, \dots, c_{i,k}\}$ 。设五笔向量查找转换函数为 e^{wubi} , 则通过对该字符五笔输入表示进行字嵌入可以得到对应的五笔向量矩阵 $C \in \mathbb{R}^{k \times d}$, 如式(1)所示。

$$C = \begin{bmatrix} e^{\text{wubi}}(c_{i,1}) \\ \vdots \\ e^{\text{wubi}}(c_{i,k}) \end{bmatrix} \quad (1)$$

其中, d 代表五笔向量的维度。

接下来我们采用尺寸为 $H_j = \{2, 3, 4\}$, $j = 1, 2, 3$ 的三种卷积核 $F_j \in \mathbb{R}^{d \times H_j}$ 在五笔向量矩阵上滑动,可以提取到各自对应的特征图。然后通过最大池化操作可以获得汉字 w_i 的最终五笔表示 v^{w_i} , 具体如式(2)~式(4)所示。

$$m_{j,i} = F_j * C_{[i:i+H_j-1]} + b \quad (2)$$

$$o_j = \text{MaxPool}\{\alpha([m_{j,1} \oplus \dots \oplus m_{j,k}])\} \quad (3)$$

$$v^{w_i} = [o_1 \oplus o_2 \oplus o_3] \quad (4)$$

其中, MaxPool 表示最大池化, $\alpha(\cdot)$ 代表激活函数, b 代表偏置项, \oplus 表示拼接。

另一种提取五笔特征的方法采用双向长短时记忆网络 BiLSTM, 通过将各五笔向量 $e^{\text{wubi}}(c_{i,1}), \dots, e^{\text{wubi}}(c_{i,k})$ 作为 BiLSTM 的输入, 我们可以进一步学习到五笔字符的隐状态表示 $\overrightarrow{h_{i,1}}, \dots, \overrightarrow{h_{i,k}}$ 和

$\overleftarrow{h_{i,1}}, \dots, \overleftarrow{h_{i,k}}$, 则最后的五笔特征 v^{w_i} 可以表示为式(5):

$$v^{w_i} = [\overrightarrow{h_{i,1}}; \overleftarrow{h_{i,k}}] \quad (5)$$

3.2 迭代学习

基于 RNN 的序列标注模型通常存在着无法对序列全局信息建模的问题。因此,我们提出使用迭代学习(iterative learning)策略不断对模型整体预

测结果进行改进,以增强模型对文本全局的建模能力,从而提高实体标注性能。

基于迭代学习的序列标注模型由网络结构相同,但参数不共享的 s 个子模型组成,记做 M_1, M_2, \dots, M_s , 如图 3 所示。相邻子模型相互串联,即 M_{t-1} 的类标输出 $Y^{t-1} = \{y_1^{t-1}, y_2^{t-1}, \dots, y_{|Y|}^{t-1}\}$ 作为文本全局特征输入到 M_t 中进行迭代预测,这里 $|Y|$ 表示序列长度。

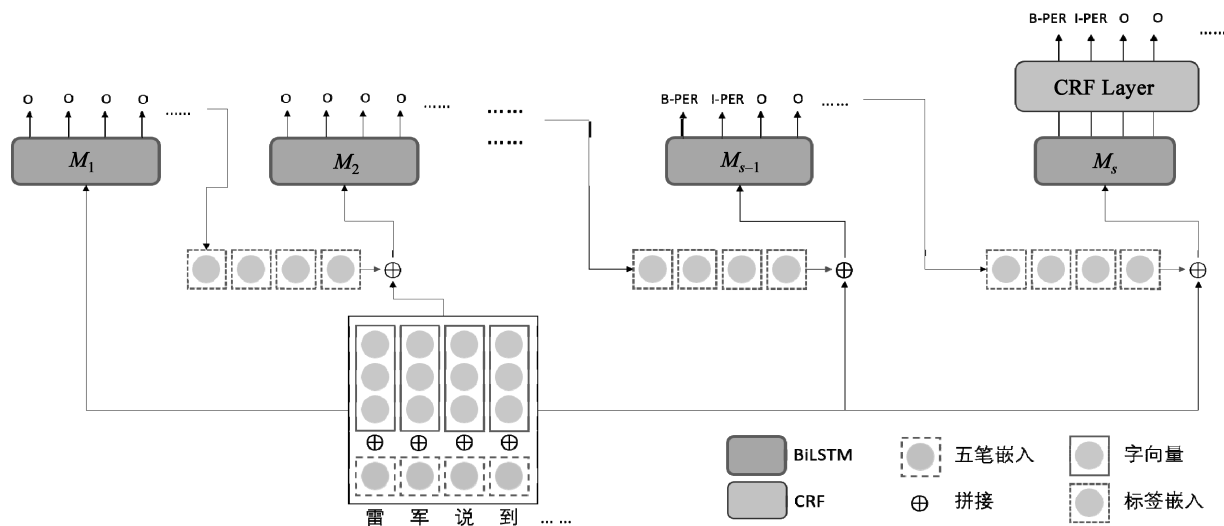


图 3 结合字形信息与迭代学习的模型结构图

为了更好地对序列上下文进行全局建模,我们使用单层双向 LSTM 网络作为子模型。对于第 t 层的子模型 $M_t (t=1, 2, \dots, N)$, 在处理字符 w_i 时, 其输入为 I_i^t , 如式(6) 所示。

$$I_i^t = [e^{\text{char}}(w_i); v^{w_i}; u_i^{t-1}] \quad (6)$$

其中, e^{char} 代表字向量转换函数, v^{w_i} 表示五笔向量, u_i^{t-1} 为上一层子模型 M_t 输出的全局特征 ($t=1$ 时, u_i^{t-1} 使用全零向量作为填充)。在经过双向 LSTM 门函数运算后, 得到本层字符 w_i 的表示 h_i^t , 如式(7) 所示。

$$h_i^t = \text{biLSTM}(I_i^t, h_{i-1}^t) \quad (7)$$

为了得到本层全局表示向量 u_i^t , 我们使用带有 softmax 函数的全连接网络将 h_i^t 映射到标签预测概率 $p_i^t \in \mathbb{R}^{|Y|}$, 如式(8) 所示。

$$p_i^t = \text{softmax}(W^t h_i^t + b^t) \quad (8)$$

我们取概率最大标签作为本层预测结果 y_i^t , 在此基础上, 定义标签嵌入查询表为 e^{label} , 进而我们可以对获取到的对应标签进行向量化表示, 作为本层输出的全局特征向量 u_i^t , 如式(9) 所示。

$$u_i^t = e^{\text{label}}(y_i^t) \quad (9)$$

由于 u_i^t 是通过双向 LSTM 模型对整个序列全

局建模后得到的表示, 将 u_i^t 输入下一层子模型中可以使使得深层模型对全局信息进行感知与处理。同时在预测类标过程中, 模型能通过 u_i^t 提供的信息考虑到字符上下文类标之间的潜在关系, 以此提高模型对序列的建模能力。从另一种角度, 我们可将 u_i^t 看作使用模型中间表示预测的初步类标, 将预测得到的中间预测类标信息不断通过迭代学习进行改进, 我们期望预测结果不断修正自身错误, 最终得到与正确类标相近的结果。

3.3 条件随机场

序列标注任务常常利用相邻序列节点对应的标签相关性来解码出最好的标签序列。我们采用标准的线性条件随机场(linear conditional random field)进行求解。通过主模块 Main 我们可以得到输入的文本序列的最终表示 $S = \{h_1, \dots, h_i, \dots, h_n\}$, 其中 h_i 表示第 i 个字所提取的特征表示, 我们希望求解出最优标签序列 $Y = \{y_1, y_2, \dots, y_n\}$ 。

由于每个字都有 q 种标签可能, 则整个文本序列一共有 q^n 种标签序列。我们定义所有标签序列集合为 $\gamma(Z)$, 则在给定条件 S 的情况下 Y 的最优

条件概率分布为 $P(Y | S)$ ，如式(10) 所示。

$$P(Y | S) = \frac{\prod_{i=1}^n \varphi_i(y_{i-1}, y_i, S)}{\sum_{y' \in \mathcal{Y}(Z)} \prod_{i=1}^n \varphi_i(y'_{i-1}, y'_i, S)} \quad (10)$$

其中， $\varphi_i(l', l, S) = \exp(W_{l', l}^T h_i + b_{l', l})$ 为特征函数，用于表示转移特征与状态特征。这里 $W_{l', l}$ 与 $b_{l', l}$ 分别为与标签对 (l', l) 相对应的权重矩阵与偏置项。

我们采用最大似然估计的方法进行 CRF 训练，对于给定的训练数据集 $\{Y_i, S_i\} |_{i=1}^N$ ，式(10)的对数形式的最大似然估计值可以表示为 Loss^{crf} ：

$$\text{Loss}^{\text{crf}} = \sum_{i=1}^N \log(P(Y_i | S_i)) \quad (11)$$

CRF 的解码过程为寻找最优的标签序列，使得条件概率值 $P(Y | S)$ 最大，我们采用维特比算法对解码过程进行有效计算。

3.4 模型训练与优化

整个模型的子模块 M_1, M_2, \dots, M_s 每次都会有对应的标签输出。各子模块的 BiLSTM 输出通过全连接层与 softmax 激活函数得到标签的概率分布，因此采用交叉熵损失函数作为子模块的目标函数 Loss_j^M ，如式(12)所示。

$$\text{Loss}_j^M = -\frac{1}{N} \sum_{i=1}^N y_i \log \hat{y}_i \quad (12)$$

其中， $0 \leq j \leq m, y_i$ 为真式标签，这里以 one-hot 形式表示。 \hat{y}_i 表示预测的标签概率分布。

结合条件随机场的优化目标，我们可以得出整个模型的最终损失函数 Loss ，如式(13)所示。

$$\text{Loss} = -\text{Loss}^{\text{crf}} + \lambda * \sum_{j=0}^m \text{Loss}_j^M \quad (13)$$

其中，负号表示统一采用梯度下降的形式， λ 为交叉熵损失函数的权重占比。

4 实验

实验所用到的 CRF 模型采用 crfsuite 工具包进行编写，神经网络模型采用 PyTorch 深度学习框架进行编写。我们采用上文构建的 HITSZ-Finance 语料库作为实验数据集，采用准确率 P (Precision)、召回率 R (Recall) 和 F_1 值 (F_1 score) 作为评价指标。实验过程中，选取数据集的 70% 作为训练集，30% 作为测试集，从训练集中抽取 10% 作为验证集，所有实验均采用单独测试 10 次后的平均值作为

最终结果以避免实验误差与偶然性，实验代码详见：<https://github.com/A-Rain/NER-with-Wubi-embedding-and-iterative-learning>。

4.1 参数设置

我们采用文献[21]所提出的在各大金融新闻网站上爬取的语料训练好的 300 维词向量进行实验，并在训练时更新参数。五笔嵌入与标签嵌入的初始参数均从 $\left[-\sqrt{\frac{3}{\text{dim}}}, +\sqrt{\frac{3}{\text{dim}}}\right]$ 的均匀分布中随机采样^[8]，这里 dim 为向量维度。

参数优化采用 Adam^[22] 优化算法，对词向量矩阵以及不同 LSTM 层之间的连接采用 dropout^[23]。整个训练期间通过观测模型在验证集的表现进行 early stopping^[24]。根据反复观测与实验，大概经过 25~30 个 epoch 会选出最优参数。具体数值如表 4 所示。

表 4 各超参数取值

超参数名	参数值
Dropout	0.1~0.4
Kernel window size	[2, 3, 4]
Number of filters	24
Learning rate	0.001
Weight decay rate	0.0001
Submodule number	[2, 3]
LSTM layer	2
Hidden size	200
Embedding size	300
Early stop patience	7
Lambda	10
Batch size	16
Label embedding	30
Wubi embedding	24

4.2 实验结果与分析

为了验证本文所提模型的有效性，我们设置了不同模型间的对比实验和模型本身的消融实验，进一步，还探究了子模块数量以及字/五笔向量是否进行训练对于实验结果的影响。

4.2.1 对比实验

本文采用 CRF、BiLSTM 以及 BiLSTM-CRF

作为基线模型和我们自己的模型进行对比。各模型的基本信息如下:

- **CRF**: 以人为构建的特征模板提取文本统计信息的条件随机场模型。
- **BiLSTM**: 采用 BiLSTM 提取特征并通过全连接层直接对标签进行分类的模型。
- **BiLSTM-CRF**: 结合 BiLSTM 与条件随机场网络的模型。
- **BCCWL(CNN)**: 采用卷积神经网络提取五笔特征并通过迭代学习结合标签信息的 BiLSTM-CRF 模型。
- **BCCWL(LSTM)**: 采用 BiLSTM 提取五笔特征并通过迭代学习结合标签信息的 BiLSTM-CRF 模型。

对比实验结果如表 5 所示,由表中数据可知: CRF 的 P 值在基线模型中表现最优,达到了 88.03%。BiLSTM 的效果最差,在 F_1 值上比 CRF 差了 7.25 个百分点,但加上 CRF 结构后,BiLSTM-CRF 模型的 R 值与 F_1 值都达到了基线模型中的最优值,分别为 86.95%和 87.20%。

表 5 对比实验结果(%)

模型	P	R	F_1
CRF	88.03	82.96	86.75
BiLSTM	77.10	82.00	79.50
BiLSTM-CRF	87.00	86.95	87.20
BCCWL(CNN)	89.24	88.28	88.72
BCCWL(LSTM)	89.06	87.63	88.40

由于融合了五笔向量后汉字的表示信息得到有效的增强,并且迭代学习使得模型预测结果得以不断改进,结合字形特征和与迭代学习后的 BCCWL(CNN)/BCCWL(LSTM)模型取得了在当前数据集上最好的效果。其中 BCCWL(CNN)模型相比三个基线模型在 P 值上平均有 5.19%的提升,在 R 值上平均有 4.31%的提升,在 F_1 值上平均有 4.23%的提升。而 BCCWL(LSTM)模型在 P 、 R 与 F_1 值上则分别平均有 5.01%、3.66%与 3.91%的提升。单独对比最好的基线模型 BiLSTM-CRF,在 F_1 值上 BCCWL(CNN)/BCCWL(LSTM)模型分别提高了 1.52%与 1.20%。

4.2.2 消融实验

消融实验模型设置具体细节如下:

- **BCCWL(CNN) w/o Wubi**: 仅仅采用迭代学习的 BiLSTM-CRF 模型。

- **BCCWL(CNN) w/o iterative**: 仅仅采用卷积神经网络提取五笔特征的 BiLSTM-CRF 模型。
- **BCCWL(LSTM) w/o Wubi**: 仅仅采用迭代学习提取五笔特征的 BiLSTM-CRF 模型。
- **BCCWL(LSTM) w/o iterative**: 仅仅采用 BiLSTM 神经网络提取五笔特征的 BiLSTM-CRF 模型。

消融实验结果如表 6 所示,根据表中数据可以发现当不添加五笔特征或者不采取迭代学习时,模型的性能有所下降。以 BCCWL(CNN)模型为例,当不结合五笔特征时,其 P 、 R 与 F_1 值分别下降了 0.91、0.24 与 0.55 个百分点。当不采用迭代学习结合标签信息时,三个评价指标分别下降了 0.61、0.75 与 0.66 个百分点。对于模型 BCCWL(LSTM),当不添加这两个因素时,其 P 、 R 与 F_1 指标也呈现了不同程度的下降。这些都说明同时结合五笔特征与标签信息有利于提升模型性能的有效性。

表 6 消融实验结果(%)

模型	P	R	F_1
BCCWL(CNN)	89.24	88.28	88.72
w/o Wubi	88.33	88.04	88.17
w/o iterative	88.63	87.53	88.06
BCCWL(LSTM)	89.06	87.63	88.40
w/o Wubi	88.59	87.61	88.10
w/o iterative	88.36	87.49	87.97

4.2.3 子模块数量对实验结果的影响

为了探讨子模块数量对于实验结果的影响,本文分别在 BCCWL(CNN)模型与 BCCWL(LSTM)两个模型的基础上采用 1、2、3、4、5 五种数量的子模块进行验证,实验结果如表 7、表 8 所示。

表 7 BCCWL(CNN)模型实验结果(%)

子模块数量	P	R	F_1
1	87.70	87.02	87.38
2	88.38	87.45	87.93
3	89.24	88.28	88.72
4	89.47	87.75	88.62
5	88.92	87.24	88.15

表 8 BCCWL(LSTM)模型实验结果(%)

子模块数量	P	R	F_1
1	87.60	87.60	87.60
2	88.67	87.60	88.07
3	89.20	87.42	88.30
4	89.06	87.63	88.40
5	88.83	87.92	88.31

实验结果表明,对于 BCCWL(CNN)模型,在子模块数量为 3 时取得了 R 值和 F_1 值的最好结果,随后这两个评价指标都呈下降趋势, P 值在子模块数量为 4 时取得最优值,但随后开始下降。

对于 BCCWL(LSTM) 模型, 当子模块数量取值为 4 时可以获得 R 值与 F_1 值的最好结果, 随后这两个评价指标均呈下降趋势。 P 值在子模块数量为 3 时取得最优值, 随后开始下降。以上分析表明当子模块堆叠到一定程度的时候可能存在过拟合的情况, 模型性能反而会下降。

4.2.4 字/五笔向量对实验结果的影响

本节主要探寻字向量与五笔向量训练对于模型性能的影响,我们分别用 char-freeze 与 wubi-freeze 表示是否对字向量与五笔向量进行参数更新(表中“√”表示进行参数更新,“×”表示不进行参数更新)。具体实验结果如表 9、表 10 所示。

从表中可以看出,字向量与五笔向量是否参与训练对于模型性能具有一定的影响。其中字向量进行参数更新对结果影响较大。对于 BCCWL(CNN)

表 9 BCCWL(CNN)实验结果(%)

char-freeze	wubi-freeze	P	R	F_1
✓	×	88.21	87.10	87.40
✓	✓	89.15	87.36	88.24
×	✓	88.99	87.60	88.28
×	×	89.24	88.28	88.72

表 10 BCCWL(LSTM)实验结果(%)

char-freeze	wubi-freeze	P	R	F1
✓	×	88.05	86.88	87.46
✓	✓	87.93	86.83	87.43
×	✓	89.14	87.40	88.26
×	×	89.06	87.63	88.40

和 BCWWL(LSTM),字向量不参与训练相比字向量参与训练其 F_1 值分别下降了 1.32 个百分点与 0.94 个百分点。

4.3 案例分析

我们以 BCCWL(CNN)模型为例,选取子模块数量为 3 的模型来验证迭代学习能够不断对模型整体预测结果进行改进,以增强模型对文本全局的建模能力。如表 11 所示,其中 M_1, M_2, M_3 分别代表三个子模块的输出序列结果,我们选取了三个例子,在模型训练过程中可以发现迭代学习具有一定的修正作用。

表 11 不同子模块的实体识别结果

[illegible]

5 总结与展望

在本文工作中,为解决语料缺乏的问题,我们首先构建了金融领域命名实体语料库 HITSZ-Finance,其中共包含 1 200 篇新闻,共计 15 960 段文本段落,31 210 个文本句,四类实体平均长度 55.6 个字符,并划分了训练集与测试集。同时针对命名实体识别这个问题,本文设计了结合五笔字形嵌入与字嵌入的神经网络模型 BCCWL。汉字的五笔表示能够代表汉字的笔画与字段信息,而标签也可以作为文本信息的一个强特征。我们将五笔信息与标签信息嵌入到 BiLSTM-CRF 框架中并在 HITSZ-Finance 数据集上进行了一系列实验,取得了不错的效果。

在 BCCWL 模型中,BCCWL(CNN)的实验结果最佳,几个子实验也验证了迭代学习与添加字形特征的有效性,然而模型也存在收敛较慢、对部分存在歧义的实体难以准确识别的问题。未来的工作将会考虑结合知识库、领域知识及语言模型这几个方面对模型做进一步改进。

参考文献

- [1] Levow G. The third international Chinese language processing bakeoff: Word segmentation and named entity recognition[C]//Proceedings of the 3rd International Chinese Language Processing Bakeoff, 2006: 108-117.
- [2] Peng N, Dredze M. Named entity recognition for Chinese social media with jointly trained embeddings [C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015: 548-554.
- [3] Ralph W, Sameer P, Lance R, et al. 2011. Ontonotes release 4.0. LDC2011T03. [DB/CD] Philadelphia: Linguistic Data Consortium, 2011.
- [4] Andrew M, Dayne F, Fernando P. Maximum entropy markov models for information extraction and segmentation.[C]//Proceedings of the 17th International Conference on Machine Learning. 2000: 591-598.
- [5] Lafferty J D, McCallum A, Pereira F, et al. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]//Proceedings of the 18th International Conference on Machine Learning, 2001: 282-289.
- [6] Collobert R, Weston J, Bottou L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011,12: 2493-2537.
- [7] Huang Z, Xu W, Yu K, et al. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv: Computation and Language, 2015.
- [8] Ma X, Hovy E H. End-to-end sequence labeling via Bidirectional LSTM-CNNs-CRF[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL, 2016: 1064-1074.
- [9] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition [C]//Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2016: 260-270.
- [10] Peng N, Dredze M. Improving named entity recognition for Chinese social media with word segmentation representation learning[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL, 2016: 149-155.
- [11] Dong C, Zhang J, Zong C, et al. Character-based LSTM-CRF with radical-level features for Chinese named entity recognition [C]//Proceedings of the 2016 International Conference on Computer Processing of Oriental Languages. 2016: 239-250.
- [12] Zhang Y, Yang J. Chinese NER using lattice LSTM [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 1554-1564.
- [13] Wu W, Meng Y, Han Q, et al. Glyce: Glyph-vectors for Chinese character representations[J]. arXiv Preprint arXiv: 1901.10125v1, 2019.
- [14] Nikolov N I, Hu Y, Tan M X, et al. Character-level Chinese-English translation through ASCII encoding [C]//Proceedings of the 3rd Conference on Machine Translation, 2018: 10-16.
- [15] Sutskever I, Vinyals O, Le Q V, et al. sequence to sequence learning with neural Networks[J]. neural information processing systems, 2014: 3104-3112..
- [16] Gui L, Du J, Zhao Z, et al. An End-to-End scalable iterative sequence tagging with multi-task learning [C]//Proceeding of the 7th CCF International Conference on Natural Language Processing and Chinese Computing, 2018: 288-298.
- [17] Yin X, Zheng D, Lu Z, et al. Neural entity reasoner for global consistency in NER[J]. arXiv preprint arXiv: 1810.00347, 2018
- [18] Kilem L G. Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters [M]. Gaithersburg, MD: Advanced Analytics, LLC, 2014.
- [19] Kim Y. Convolutional neural networks for sentence

- Classification[C]//Proceeding of the 19th conference on Empirical Methods in Natural Language Processing, 2014: 1746-1751.
- [20] 陈钊, 徐睿峰, 桂林, 等. 结合卷积神经网络和词语情感序列特征的中文情感分析[J]. 中文信息学报, 2015, 29(6): 172-178.
- [21] Li S, Zhao Z, Hu R, et al. Analogical reasoning on chinese morphological and semantic relations[C]//Proceeding of the 56th Annual Meeting of the Association for Computational Linguistics, 2018: 138-143.
- [22] Kingma D P, Ba J. Adam: A method for stochastic optimization[C]//Proceeding of the 3rd International Conference for Learning Representations, 2015.
- [23] Srivastava N, Hinton G E, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.
- [24] Caruana R, Lawrence S, Giles C L, et al. Overfitting in neural nets: backpropagation, conjugate gradient, and early stopping[C]//Proceedings of the 13th International Conference on Neural Information Processing Systems, 2000: 381-387.



刘宇瀚(1997—), 硕士研究生, 主要研究领域为自然语言处理、情感分析和主题模型。

E-mail: liuyuhan_hitsz@163.com



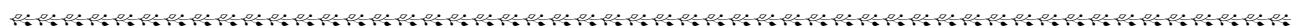
刘常健(1996—), 硕士研究生, 主要研究领域为自然语言处理、情感分析和对话系统。

E-mail: cjliux@163.com



徐睿峰(1973—), 通信作者, 博士, 教授, 主要研究领域为自然语言处理、文本情感计算和社交媒体分析。

E-mail: xuruifeng@hit.edu.cn



(上接第 73 页)

- [14] Kingma D P, Ba J. Adam: a method for stochastic optimization[C]//Proceedings of the 3rd International Conference for Learning Representations, San Diego, 2015.
- [15] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting [J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.



慈祯嘉措(1989—), 讲师, 博士, 主要研究领域为计算语言学、藏文信息处理、机器翻译。

E-mail: 543819011@qq.com



桑杰端珠(1986—), 博士研究生, 主要研究领域为计算语言学、藏文信息处理、机器翻译。

E-mail: sangjeedondrub@live.com



孙茂松(1962—), 通信作者, 教授, 博士生导师, 主要研究领域为自然语言理解、中文信息处理、机器翻译等。

E-mail: sms@mail.tsinghua.edu.cn