doi:10.6041/j.issn.1000-1298.2021.S0.066

基于 BERT－CRF 模型的生鲜蛋供应链命名实体识别

刘新亮[1,2]　张梦琪[2]　谷　情[2]　任延昭[2]　何东彬[1,3]　高万林[1,3]

(1. 中国农业大学信息与电气工程学院, 北京 100083;
2. 北京工商大学农产品质量安全追溯技术及应用国家工程实验室, 北京 100048;
3. 农业农村部农业信息化标准化重点实验室, 北京 100083)

摘要: 对于生鲜蛋供应链知识图谱构建过程中供应链领域实体名称多样、特征信息提取不充分的问题,提出了一种基于 BERT－CRF 模型(Bidirectional encoder representations from transformers－conditional random field)的命名实体识别方法。该方法使用 BIO(Begin、Internal、Other)标记规则进行序列标注,以字向量和位置向量作为输入,通过 BERT 预训练模型提取输入序列全局特征,并在模型的末端添加 CRF 层引入硬约束,构建适合生鲜蛋供应链领域命名实体识别的模型框架。所提出的模型与其他 3 种命名实体识别模型在自建数据集上进行了对比实验,该数据集包含 12 810 条文本语料数据,5 大类 21 个小类。实验结果表明,本文模型取得了很好的结果,准确率、召回率和 F1 值分别达到 91.82%、90.44%、91.01%,验证了本文模型优于其他 3 种模型。最后本文模型使用自建的食品领域菜谱数据集进行实验,结果表明模型具有一定的泛化能力。

关键词: 生鲜蛋供应链; 命名实体识别; 预训练模型; 条件随机场

中图分类号: TP391.1　　**文献标识码**: A　　**文章编号**: 1000-1298(2021)S0-0519-07

Named Entity Recognition of Fresh Egg Supply Chain Based on BERT－CRF Architecture

LIU Xinliang[1,2]　ZHANG Mengqi[2]　GU Qing[2]　REN Yanzhao[2]　HE Dongbin[1,3]　GAO Wanlin[1,3]

(1. *College of Information and Electrical Engineering*, *China Agricultural University*, *Beijing* 100083, *China*
2. *National Engineering Laboratory for Agri-product Quality Traceability*, *Beijing Technology and Business University*, *Beijing* 100048, *China*
3. *Key Laboratory of Agricultural Informatization Standardization*, *Ministry of Agriculture and Rural Affairs*, *Beijing* 100083, *China*)

Abstract: Recognizing named entities from raw text is the first step to construct a fresh egg supply chain knowledge graph and support a variety of downstream natural language processing tasks. This task can sort out the information in the supply chain and provide a basis for food safety traceability. In the raw text of fresh egg supply chain, there were various types of entities, and feature information extraction was inefficient. In order to solve the problem of fast and accurate identification of the named entities which entity types were pre-defined, a bidirectional encoder representations from transformers－conditional random field (BERT－CRF) architecture was proposed to solve the task of named entity recognition (NER) in the area of fresh egg supply chain. In BERT－CRF architecture, begin, internal and other (BIO) labeling rule was used to label the sequence, and the concatenation of character vector and position vector was used as inputs. The pre-training language model (BERT) was used to obtain the global features of input sequence, and the CRF layer was added at the end of the model to introduce hard constraints. A comparative experiment was conducted with other three NER model on the self-constructed dataset that contained five categories and 21 subcategories. The result showed that the BERT－CRF model was superior to the others and reported a state-of-the-art performance. The precision, recall and F1-score were 91.82%, 90.44% and 91.01%, respectively. Finally, through the comparative

收稿日期: 2021－07－17　修回日期: 2021－09－05

基金项目: 北京市科委科技计划项目(Z191100008619007)

作者简介: 刘新亮(1972—),男,博士生,北京工商大学副教授,主要从事农业领域知识图谱研究,E-mail: liuxinl@btbu.edu.cn

通信作者: 高万林(1965—),男,教授,主要从事农业信息化标准化与大数据管理研究,E-mail: wanlin_cau@163.com

experiments with other self-constructed dataset (dish dataset), the results showed that the model had a certain generalization ability.

Key words: fresh egg supply chain; named entity recognition; pre-training model; conditional random field

0 引言

近年来,部分生鲜蛋品牌推出了溯源功能,将生鲜蛋的生产、包装、仓储和物流等环节的供应链信息放到追溯平台供用户查询使用[1]。生鲜蛋供应链领域数据纷繁复杂[2],如何有效地整合生鲜蛋供应链信息,使供应链各环节的信息联系起来,是食品安全溯源问题的关键[3]。命名实体识别技术应用于生鲜蛋供应链领域,能有效梳理供应链信息,为食品安全溯源提供了一种解决途径。命名实体识别主要是从文本中识别出实体的边界并判断其所属的实体类型,是文本挖掘和信息抽取的核心任务[4]。实体类型需要预先定义。通过命名实体识别,能够实现以最小的人力成本从大量的半结构化和非结构数据中识别出不同的实体,是数据挖掘、问答系统、知识图谱构建等领域中的一项重要技术[5-9]。

近年来,命名实体识别在一些垂直领域已经有了广泛研究,文献[10]提出了一种基于句法语义特征的命名实体识别方法,选择支持向量机(Support vector machine,SVM)作为机器学习的实现途径,以真实旅游领域文本作为语料进行实验,验证了该方法在命名实体识别上的有效性。传统的机器学习方法对特征工程有很大的依赖[11],模型的泛化能力不高。基于深度学习的命名实体识别技术,在特征提取的深度和模型的精确度上已经逐渐超过了传统基于特征的方法[12-14],成为主流方法。文献[15]将CNN-CRF(Convolutional neural network-conditional random field)模型应用于中文电子病历的命名实体识别研究中,准确率、召回率和 F1 值都取得了很好的效果。文献[16]通过基于注意力机制的循环神经网络(Recurrent neural network, RNN)进行命名实体识别,能够很好地获取标签之间的关系。文献[17]在改进标注模式的基础上通过 BiLSTM(Bi-directional long short-term memory)-CRF 模型在命名实体识别上取得良好的效果。在开放域命名实体识别任务中,文献[18-20]在食品、农业等领域也都取得了理想的效果。

上述深度学习与传统机器学习相结合进行命名实体识别的方法,为本文在生鲜蛋供应链领域进行命名实体识别提供了参考。但生鲜蛋供应链领域实体名称多样,特征信息复杂,上述方法依赖于时间序列不能获得较长上下文信息。针对这个问题,本文提出基于 BERT-CRF 模型自动抽取生鲜蛋供应链文本中命名实体的解决框架。

1 数据处理与标注

本文通过数据获取与处理、数据标注两个步骤,建立生鲜蛋供应链领域的语料库。

1.1 数据获取与处理

生鲜蛋供应链命名实体识别目前还没有公开的数据集,本文利用基于 Python 的网页爬虫抓取中国农垦网上正大生鲜蛋供应链各个环节的信息作为文本语料初始数据。首先使用 Requests 框架获取 HTML 数据,再利用 Beautifulsoup 库获取标签中的文本数据,最后通过正则表达式、字符格式规范化等方法,删除非文本数据内容,从而得到规范的生鲜蛋供应链语料库。

1.2 数据标注

本文预先定义好的命名实体类别包括人名、地名、车辆、时间和批次 5 大类,并将 5 大类分为 21 小类,具体实体类型信息统计如表1所示。通过使用

表 1 实体类型信息

Tab. 1 Entity types statistics

实体大类	实体小类	数量
人名	加工负责人	6 405
	养殖负责人	
	包装负责人	
	分拣负责人	
	驾驶员	
地名	养殖地	6 537
	加工场地	
	包装场地	
	仓库	
车辆	车内温度	8 342
	车内湿度	
	车牌号	
时间	养殖时间	9 502
	包装时间	
	入库时间	
	出库时间	
	运输时间	
批次	养殖批次	10 098
	包装批次	
	分拣批次	
	加工批次	

农产品质量安全追溯技术及应用国家工程实验室开发的语料标注系统，对获取的语料库进行了人工标注。本文采用 BIO 规则对语言序列进行标注，其中，B(Begin)描述句子中每个命名实体的开始位置；I(Internal)描述命名实体除起始位置外的其他部分，O(Other)用来描述句子中其他非预先定义好的实体。标注中为了更好地识别命名实体的类别信息，本文将类别信息与 BIO 规则进行融合，类别信息如人名实体用(－PER)表示、地名实体用(－LOC)表示等。以句子“刘春霞是正大蛋业有限公司的仓储负责人”为例，其序列标注如图 1 所示。

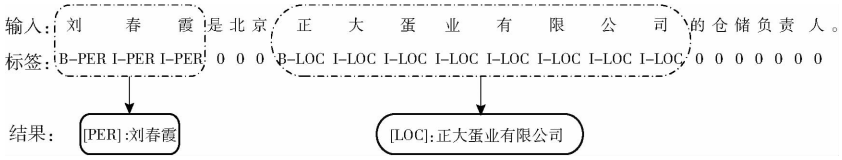

图 1　语料序列标注示例

Fig. 1　Corpus tagging sample

2　模型框架

传统的命名实体识别模型通常是 RNN(循环神经网络)[21]、LSTM(长短期记忆)[22]等神经网络模型与隐马尔可夫模型(Hidden Markov model, HMM)、条件随机场(Conditional random field, CRF)等概率图模型进行组合。BERT 模型通过双向全自注意力网络(Transformer)既能解决 RNN 无法解决的长期依赖问题又能解决 RNN 和 LSTM 必须要根据时间推移进行顺序处理的问题。本文在字嵌入层的基础上，采用 BERT－CRF 模型进行命名实体识别，模型结构如图 2(图中 CLS 表示起始 Token，SEP 表示结束 Token，$\boldsymbol{I}$ 表示初始向量，$\boldsymbol{T}$ 表示通过 BERT 模型得到的向量)所示。

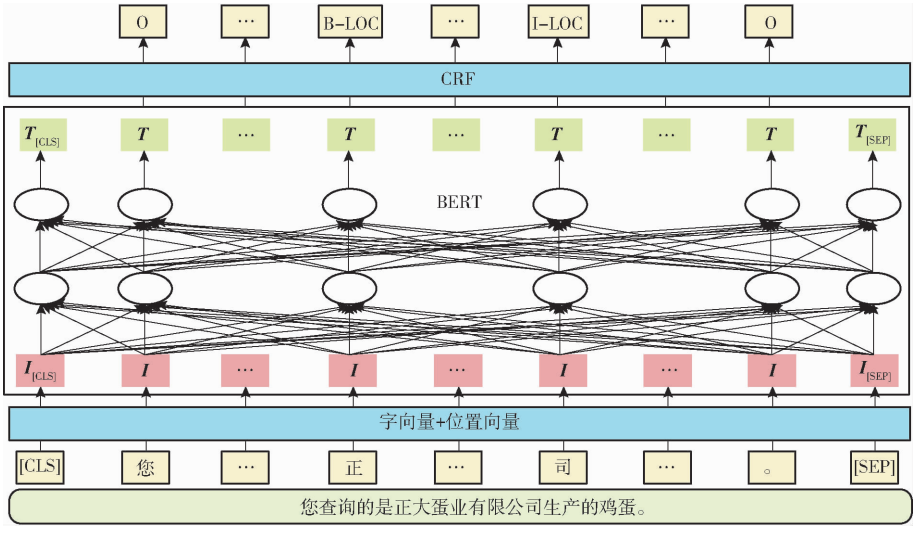

图 2　BERT－CRF 模型结构

Fig. 2　Model structure of BERT－CRF

2.1　字向量表示

在进行命名实体识别时，通过 Word2vec 技术将生鲜蛋供应链上的文本语料转换为计算机能够处理的向量形式[23]。生鲜蛋供应链上的文本语料多为常用词，Word2vec 中的连续词袋(CBOW)模型在常用词上有更高的精确性和更快的训练速度，故本文采用 CBOW 模型。CBOW 模型通过背景词预测中心词的方式，将语料中的每个中文汉字转换为计算机能够处理的字向量。如图 3 所示，CBOW 模型包括输入层、映射层和输出层。该模型通过不断优化训练得出每个词语的词向量，然后根据词向量之间的余弦值来计算词语之间的相似度，相似度越高的词语语义越接近，从而达到预测的效果。给定 CBOW 模型的窗口大小为 $2m$，背景词 w_{t-m}、w_{t-m+1}、

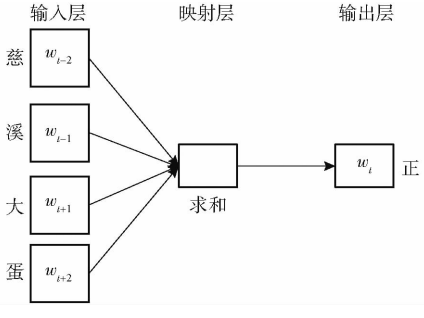

图 3　CBOW 模型结构

Fig. 3　CBOW model structure

…、w_{t+m-1}、w_{t+m}通过一层映射层的加权求和来预测中心词 w_t 的概率。实验结果表明,在窗口大小为 2 的情况下数据集取得了良好的识别效果。表 2 给出了以“慈溪正大蛋业有限公司”为例,窗口大小为 2 时,CBOW 模型的输入词和期望输出。

表 2 CBOW 模型示例

Tab. 2 Example of CBOW

输入词 w_{t-2}	输入词 w_{t-1}	输入词 w_{t+1}	输入词 w_{t+2}	输出期望 w_t
		溪	正	慈
	慈	正	大	溪
慈	溪	大	蛋	正
溪	正	蛋	业	大
正	大	业	有	蛋
大	蛋	有	限	业
蛋	业	限	公	有
业	有	公	司	限
有	限	司		公
限	公			司

2.2 BERT 层

BERT[24]模型近年来在自然语言处理领域取得了良好的效果,它通过双向全自注意力网络并行计算同时获取整个输入序列的特征,而不依赖于时间序列,具有获得较长上下文信息的优势,可以更好地提升特征提取水平。本文只考虑单一句子级别的输入,将输入到 BERT 模型的 Token(本文中一个汉字为一个 Token)序列在下文统称为序列。BERT 模型训练通常分为两步,首先对模型进行预训练(Pre-training BERT),其次在预训练的基础上,根据下游任务进行模型微调(Fine-tuning BERT)。与传统语言模型相比 Pre-training BERT 作出了较大的创新,传统的模型从左到右进行语言训练,而 BERT 模型则通过遮盖语言模型(Masked LM)和下一句子预测(Next sentence prediction)两种方法分别完成词语和句子级别的描述。本文只考虑句子级别的命名实体识别,故只使用 Masked LM 进行模型预训练。Masked LM 在进行训练时,随机遮罩一定比例的 Token,然后对这些 Token 进行预测。通常情况下,Masked LM 随机选取 15% 的 Token 进行遮罩,以第 i 个 Token(Token_i)作为需要被遮罩的字符为例,遮罩替换规则如下:有 80% 概率用[MASK]替换 Token_i;10% 概率通过随机产生的词替换 Token_i;10% 概率 Token_i 不进行替换。在本文的命名实体识别任务中,Fine-tuning BERT 只需要考虑模型的输入和输出,通过多头自注意力机制(Multi-head self-attention mechanism)进行参数调整。

本文以单个序列作为模型输入,每个序列均以特殊分类标记(Classification Token,[CLS])作为起始 Token,特殊令牌(Special Token,[SEP])作为结束 Token。由于 BERT 模型对每个 Token 位置的感知能力存在缺陷,故本文将 Token 对应的字向量和位置向量求和得到模型的输入向量 $\boldsymbol{I}$。$\boldsymbol{I}$ 的数学模型为

$$\boldsymbol{I} = \boldsymbol{E}_{char} + \boldsymbol{E}_{POS} \tag{1}$$

式中 $\boldsymbol{E}_{char}$——字向量 $\boldsymbol{E}_{POS}$——位置向量

以序列“您查询的是正大蛋业有限公司生产的鸡蛋。”为例,序列表示和模型的输入如图 4 所示。

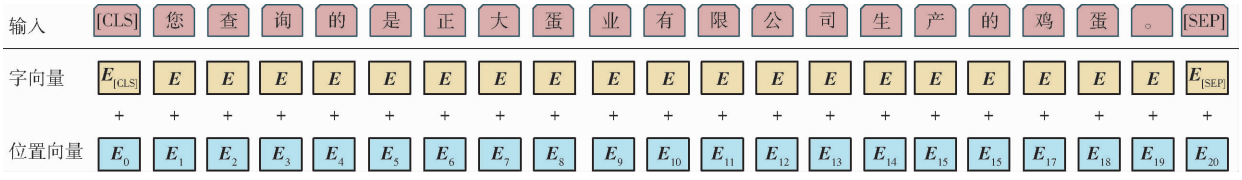

图 4 BERT 模型输入表示

Fig. 4 BERT input representation

BERT 模型的核心是由 2 个全自注意力网络分别作为编码器(Encoder)和解码器组成(Decoder),能够并行考虑序列中每个 Token 的特征,而不依赖于序列中 Token 顺序。注意力机制本质上是计算余弦相似度,两个向量的余弦值越大,这两个向量越相似。在 Encoder 模型中,每个模块都加入了全连接前馈网络以及多头自注意力机制,通过拼接各个注意力获得多头注意力信息,使得每个注意力机制都能考虑到序列中各个信息。Encoder 模型处理过程为:通过线性层得到查询矩阵 $\boldsymbol{Q}$、值矩阵 $\boldsymbol{K}$、索引矩阵 $\boldsymbol{V}$,其计算公式为

$$\boldsymbol{Q} = \text{Liner}_Q(\boldsymbol{I}) = \boldsymbol{W}_Q^{\mathrm{T}}\boldsymbol{I} \tag{2}$$

$$\boldsymbol{K} = \text{Liner}_K(\boldsymbol{I}) = \boldsymbol{W}_K^{\mathrm{T}}\boldsymbol{I} \tag{3}$$

$$\boldsymbol{V} = \text{Liner}_V(\boldsymbol{I}) = \boldsymbol{W}_V^{\mathrm{T}}\boldsymbol{I} \tag{4}$$

式中 $\boldsymbol{W}_Q$、$\boldsymbol{W}_K$、$\boldsymbol{W}_V$——权重矩阵

计算 $\boldsymbol{Q}$ 与 $\boldsymbol{K}$ 的余弦相似性,通过 Softmax 强化最大值后与 $\boldsymbol{V}$ 作运算得到一个注意力的值。其公式为

$$\text{att}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{5}$$

式中 d_k——矩阵 $\boldsymbol{K}$ 的维度

然后把一个 att 加权后的值记为一个头

(head),att 加权的计算公式为

$$\boldsymbol{h}_i = \mathrm{att}(\boldsymbol{Q}\boldsymbol{W}_i^{\mathrm{Q}}, \boldsymbol{K}\boldsymbol{W}_i^{\mathrm{K}}, \boldsymbol{V}\boldsymbol{W}_i^{\mathrm{V}}) \tag{6}$$

式中　$\boldsymbol{W}_i^{\mathrm{Q}}$、$\boldsymbol{W}_i^{\mathrm{K}}$、$\boldsymbol{W}_i^{\mathrm{V}}$——对应矩阵的权重

多头注意力由各个头加权运算获得,计算公式为

$$M_h = \mathrm{Concat}(\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_h)\boldsymbol{W}^{\mathrm{O}} \tag{7}$$

式中　Concat——求和运算函数

$\boldsymbol{W}^{\mathrm{O}}$——权重矩阵

Decoder 模型与 Encoder 模型类似,两个模型之间通过残差连接。Decoder 模型的输入为 Encoder 模型的输出。

2.3　CRF 层

将命名实体识别定义为一个序列标注任务,BERT 得出的序列标签只受到词本身和上下文特征的影响。但在 BIO 标注模式中,I 只能跟在 B 后面,这说明标签之间是相互依赖的。所以命名实体识别模型中增加 CRF 层引入一些硬性约束,如:实体起始标签必须为 B,I 标签只能跟在 B 标签后等。在 CRF 层进行命名实体识别时,首先,输入转移方阵 $\boldsymbol{A}$,转移方阵 $\boldsymbol{A}$ 用于表示标签与标签之间的关系,$\boldsymbol{A}_{i,j}$表示从第 i 个标签到第 j 个标签的概率值。每一个输入序列 $\boldsymbol{X} = (x_1, x_2, \cdots, x_n)$,有与其长度相等的标签序列 $\boldsymbol{Y} = (y_1, y_2, \cdots, y_n)$ 相对应,n 表示输入 Token 的个数,输入序列 $\boldsymbol{X}$ 得到输出序列 $\boldsymbol{Y}$ 的分数计算公式为

$$\mathrm{score}(x, y) = \sum_{i=1}^{n} P_{i,y_i} + \sum_{i=1}^{n+1} A_{y_{i-1},y_i} \tag{8}$$

式中　P_{i,y_i}——第 i 字预测为标签 y_i 的分数

A_{y_{i-1},y_i}——标签 y_{i-1}转移到标签 y_i 的分数

2.4　实验参数配置及评价指标

本文在 Tensorflow 1.2 框架下搭建 BERT－CRF 模型进行生鲜蛋命名实体识别任务,通过验证集进行超参数调整。其中,在 CBOW 模型中,窗口大小设置为 2;BERT 预训练模型参照文献[24]的实验设置,使用 12 头注意力机制的全自注意力网络,将字嵌入的向量设置成 300 维;本文的命名实体识别任务预先定义 21 小类实体,即将识别出的实体划分到 21 类中,故 CRF 层的全连接层参数为 21;在处理批次选择上,批次设置越大,识别速度越快,对内存的要求也越高,本实验批次选用 32 是实验机器所允许的最大批次;使用自适应时刻估计方法进行算法优化和交叉熵损失函数;其次,将学习效率设为 0.001,Dropout 设置为 0.3,处理批次设置为 32,均为一般深度学习常用的参数选择。

本实验通过准确率、召回率和 F1 值来评估模型的性能。

3　实验与结果分析

本实验使用的数据集包含 12 810 条文本语料,按照 7∶1∶2的比例将数据集划分为训练集、验证集和测试集,数据集各语料之间无重合。训练集在模型训练阶段使用,用于模型训练;验证集对训练好的模型进行验证,并调整超参数使模型在验证集上效果最好;通过测试集对模型进行最终评估。数据集中训练集数据 8 967 条,测试集 2 562 条,验证集 1 281 条。

3.1　不同模型性能比较

为了验证 BERT－CRF 模型在生鲜蛋供应链领域语料上的命名实体识别性能,分别使用 RNN－CRF、LSTM－CRF、BiLSTM－CRF 模型与本文提出的 BERT－CRF 模型进行对比实验。实验结果如表 3 所示。

表 3　不同模型的实验结果对比

Tab.3　Experimental results of different models　%

模型	准确率	召回率	F1 值
RNN－CRF	82.72	80.90	83.80
LSTM－CRF	85.65	83.78	86.19
BiLSTM－CRF	88.45	86.52	88.47
BERT－CRF	91.82	90.44	91.01

由表 3 可知,RNN－CRF 模型因梯度消失问题导致无法处理长距离依赖,对于命名实体识别任务而言准确率较低,为 82.72%。LSTM－CRF 模型相比于 RNN－CRF,通过增加输入门、遗忘门和输出门,获得句子前后的关系,准确率 85.65%。BiLSTM－CRF 模型通过双向 LSTM,其准确率为 88.45%。BERT 预训练模型能够并行获取全局信息,该模型与其他 3 个模型相比,效果最好,其准确率、召回率和 F1 值都是最高的,分别为 91.82%、90.44% 和 91.01%,比 3 个模型中效果最佳的 BiLSTM－CRF 模型在准确率、召回率和 F1 值上分别提升了 3.37、3.92、2.54 个百分点。

为了更直观地看出不同模型对不同类别实体的识别效果,本文进行了对比实验。通过准确率、召回率和 F1 值 3 个指标,对比不同模型对不同类别实体的识别效果,结果如图 5 所示。

由图 5 可看出,总体来说,4 种模型对人名和地名实体识别的准确率、召回率和 F1 值较高,而对车辆、批次和时间的实体识别的准确率、召回率和 F1 值偏低。本文提出的 BERT－CRF 模型相比于 RNN－CRF、LSTM－CRF 和 BiLSTM－CRF,对生鲜蛋供应链各类实体的识别准确率、召回率和 F1 值均是最高的。

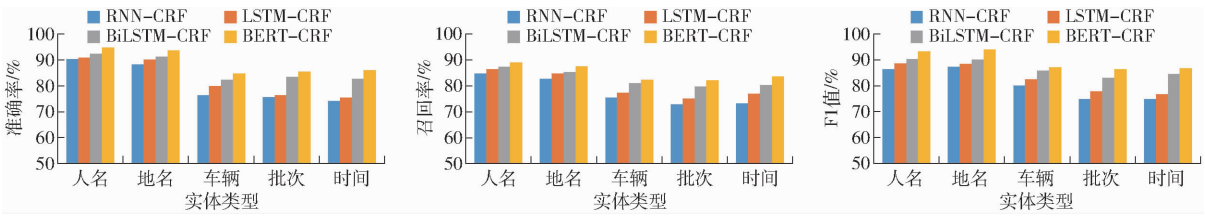

图 5 实体实验结果对比

Fig. 5 Experimental results of entities

在车辆、批次和时间 3 类实体类别中,4 种模型的识别率都低于人名和地名。据分析可知,由于人名和地名实体的命名规则比较规范,全部由汉字组成,不含其它数字或者字符,因此识别率比这 3 类实体高,若想提升这 3 类实体的识别率,还需要设计更复杂的模型来获取更加丰富的语义特征。

以上实验结果表明,本文提出的 BERT - CRF 模型更加能够考虑输入序列全局特征,从而提高命名实体识别效果,且准确率、召回率和 F1 值均达到 90% 以上。

3.2 不同数据集性能比较

为了验证 BERT - CRF 模型在食品领域其他数据集上的表现效果,本文与本课题组构建的菜谱数据集行了对比实验,菜谱数据集共包含语料 2 961 条,实体 5 922 个,6 种实体类型。表 4 为生鲜蛋供应链和菜谱两个数据集模型性能对比。

从表 4 可以看出,BERT - CRF 模型在菜谱数据集上也有一定的效果。说明 BERT - CRF 模型在命名实体识别任务中具有良好的泛化能力。

表 4 不同数据集的实验结果对比

Tab. 4 Experimental results of different datasets %

数据集	准确率	召回率	F1 值
生鲜蛋供应链	91.82	90.44	91.01
菜谱	89.79	88.64	90.23

4 结论

(1)针对正大生鲜蛋供应链领域实体名称多样、特征信息提取不充分的问题,提出一种基于 BERT - CRF 模型的命名实体识别方法,充分考虑输入序列的全局特征,提升了模型性能,F1 值为 91.01%。

(2)利用 BERT 预训练模型获得每个 Token 的向量化表示,CRF 层的添加充分考虑了标签之间的依赖关系,提高了命名实体识别的结果,该模型的准确率和召回率分别为 91.82% 和 90.44%。

(3)所提出的 BERT - CRF 模型,在课题组自建的食品领域其他数据集上评价指标均能达到 90% 以上,验证了该模型具有良好的泛化能力。

参 考 文 献

[1] 宋宝娥. 食品供应链质量安全可追溯系统构建研究[J]. 物流工程与管理,2017,39(3):57 - 61.
SONG Baoe. Research on food supply chain safety traceability system[J]. Logistics Engineering and Management, 2017, 39(3):57 - 61. (in Chinese)

[2] 徐慧馨,生吉萍,徐红. 基于供应链风险管理的食品流通数据库研究[J]. 食品安全质量检测学报,2020,11(18):6475 - 6481.
XU Huixin, SHENG Jiping, XU Hong. Research on food circulation database based on supply chain risk management[J]. Journal of Food Safety and Quality, 2020, 11(18):6475 - 6481. (in Chinese)

[3] 兰洪杰,黄锋权,林自葵. 2008 年北京奥运会食品可追溯系统设计[J]. 中国储运,2008(5):86 - 89.
LAN Hongjie, HUANG Fengquan, LIN Zikui. Design of food traceability system for 2008 Beijing Olympic Games[J]. China Storage & Transport, 2008(5):86 - 89. (in Chinese)

[4] GOLSHAN P N, DASHTI H R, AZIZI S. A study of recent contributions on information extraction [J]. arXiv preprint arXiv: 1803.05667, 2018.

[5] 张帆,王敏. 基于深度学习的医疗命名实体识别[J]. 计算技术与自动化,2017,36(1):123 - 127.
ZHANG Fan, WANG Min. Medical named entity recognition based on deep learning [J]. Computing Technology and Automation, 2017, 36(1):123 - 127. (in Chinese)

[6] LIU X H, ZHANG S D, WEI F R, et al. Recognizing named entities in tweets[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies,2011:359 - 367.

[7] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition [J]. arXiv preprint arXiv: 1603.01360, 2016.

[8] 刘峤, 李杨, 段宏,等. 知识图谱构建技术综述[J]. 计算机研究与发展, 2016, 53(3): 582 - 600.
LIU Qiao, LI Yang, DUAN Hong, et al. Knowledge graph construction techniques [J]. Journal of Computer Research and

Development, 2016, 53(3): 582－600. (in Chinese)

[9] 徐增林,盛泳潘,贺丽荣,等. 知识图谱技术综述[J]. 电子科技大学学报,2016,45(4):589－606.
XU Zenglin, SHENG Yongpan, HE Lirong, et al. Review on knowledge graph techniques [J]. Journal of University of Electronic Science and Technology of China, 2016, 45(4):589－606. (in Chinese)

[10] 甘丽新,万常选,刘德喜,等. 基于句法语义特征的中文实体关系抽取[J]. 计算机研究与发展,2016, 53(2):284－302.
GAN Lixin, WAN Changxuan, LIU Dexi, et al. Chinese named entity relation extraction based on syntactic and semantic features [J]. Journal of Chinese Information Processing, 2016, 53(2):284－302. (in Chinese)

[11] 刘浏,王东波. 命名实体识别研究综述[J]. 情报学报,2018,37(3):329－340.
LIU Liu, WANG Dongbo. A review on named entity recognition[J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(3): 329－340. (in Chinese)

[12] 马孟铖,杨晴雯,艾斯卡尔·艾木都拉,等. 基于词向量和条件随机场的中文命名实体分类[J]. 计算机工程与设计,2020,41(9):2515－2522.
MA Mengcheng, YANG Qingwen, HANDULLA Aisikaer, et al. Chinese named entity classification based on word2vec and conditional random fields[J]. Computer Engineering and Design, 2020, 41(9):2515－2522. (in Chinese)

[13] 陈曙东,欧阳小叶. 命名实体识别技术综述[J]. 无线电通信技术,2020,46(3):251－260.
CHEN Shudong, OUYANG Xiaoye. Overview of named entity recognition technology[J]. Radio Communications Technology, 2020,46(3):251－260. (in Chinese)

[14] 鄂海红,张文静,肖思琪,等. 深度学习实体关系抽取研究综述[J]. 软件学报,2019,30(6):1793－1818.
E Haihong, ZHANG Wenjing, XIAO Siqi, et al. Survey of entity relationship extraction based on deep learning [J]. Journal of Software, 2019, 30(6):1793－1818. (in Chinese)

[15] 曹依依,周应华,申发海,等. 基于 CNN－CRF 的中文电子病历命名实体识别研究[J]. 重庆邮电大学学报(自然科学版),2019,31(6):869－875.
CAO Yiyi, ZHOU Yinghua, SHEN Fahai, et al. Research on named entity recognition of Chinese electronic medical record based on CNN－CRF [J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2019, 31(6):869－875. (in Chinese)

[16] PANDEY C, IBRAHIM Z, WU H, et al. Improving RNN with attention and embedding for adverse drug reactions[C]// Proceedings of the 2017 International Conference on Digital Health,2017: 67－71.

[17] 曹明宇,杨志豪,罗凌,等. 基于神经网络的药物实体与关系联合抽取[J]. 计算机研究与发展,2019,56(7):1432－1440.
CAO Mingyu, YANG Zhihao, LUO Ling, et al. Joint drug entity and relations extraction based on neural network [J]. Journal of Computer Research and Development, 2019, 56(7):1432－1440. (in Chinese)

[18] 秦兵,刘安安,刘挺. 无指导的中文开放式实体关系抽取[J]. 计算机研究与发展,2015,52(5):1029－1035.
QIN Bing, LIU An'an, LIU Ting. Unsupervised Chinese open entity relation extraction[J]. Journal of Computer Research and Development, 2015, 52(5):1029－1035. (in Chinese)

[19] 郑丽敏,齐珊珊,田立军,等. 面向食品安全事件新闻文本的实体关系抽取研究[J]. 农业机械学报,2020,51(7):244－253.
ZHENG Limin, QI Shanshan, TIAN Lijun, et al. Entity relation extraction of news texts for food safety events[J]. Transactions of the Chinese Society for Agricultural Machinery,2020,51(7):244－253. (in Chinese)

[20] 赵鹏飞,赵春江,吴华瑞,等. 基于注意力机制的农业文本命名实体识别[J]. 农业机械学报, 2021,52(1):185－192.
ZHAO Pengfei, ZHAO Chunjiang, WU Huarui, et al. Named entity recognition of Chinese agricultural text based on attention mechanism [J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(1):185－192. (in Chinese)

[21] ZAREMBA W, SUTSKEVER I, VINYALS O. Recurrent neural network regularization [J]. ArXiv: 1409.2329 [cs], 2015.

[22] HOCHREITER S. Long short-term memory [J]. Neural Computation, 1997, 9(8):1735－1780.

[23] MIKOLOV T, CHEN K, GREG C, et al. Efficient estimation of word representations in vector space[C]// Proceedings of Workshop at ICLR, 2013.

[24] JACOB D, CHANG M, KENTON L, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C]// Processing of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019:4171－4186.