doi:10.6041/j.issn.1000-1298.2021.01.021

基于注意力机制的农业文本命名实体识别

赵鹏飞[1]　赵春江[1,2]　吴华瑞[2,3]　王　维[2,3]

(1.山西农业大学工学院,太谷 030801;2.国家农业信息化工程技术研究中心,北京 100097;
3.北京农业信息技术研究中心,北京 100097)

摘要:针对农业智能问答系统构建过程中传统的农业命名实体识别方法依赖人工特征模板、特征信息提取不充分、实体名称多样导致标注不一致等问题,提出一种基于注意力机制的农业文本命名实体识别方法。采用连续词袋模型(Continuous bag of words,CBOW)对输入字向量进行预训练,丰富字向量特征信息,缓解分词准确度对性能的影响;引入文档级的注意力(Attention)机制,获取实体间相似信息,保证实体在不同语境下的标签一致性;基于双向长短期记忆网络(Bi-directional long-short term memory,BiLSTM)和条件随机场(Conditional random field,CRF)模型,构建适合农业领域实体识别的模型框架。选取 4 604 篇农业文本,针对病害、虫害、农药、农作物品种 4 类实体进行了识别实验。结果表明,模型能有效地辨别农业文本中的实体,缓解实体标记不一致的问题,在农业语料上达到了较好的结果,识别的准确率、召回率、F 值分别为 93.48%、90.60%、92.01%。与其他 3 种识别方法相比,模型在不同规模语料库的准确率均有一定提高,具有明显的性能优势。

关键词:农业文本;命名实体识别;注意力机制;神经网络;深度学习

中图分类号:TP391.1　　**文献标识码**:A　　**文章编号**:1000-1298(2021)01-0185-08　　**OSID**:

Named Entity Recognition of Chinese Agricultural Text Based on Attention Mechanism

ZHAO Pengfei[1]　ZHAO Chunjiang[1,2]　WU Huarui[2,3]　WANG Wei[2,3]

(1. *College of Engineering*, *Shanxi Agricultural University*, *Taigu* 030801, *China*
2. *National Engineering Research Center for Information Technology in Agriculture*, *Beijing* 100097, *China*
3. *Beijing Research Center for Information Technology in Agriculture*, *Beijing* 100097, *China*)

Abstract: Agricultural named entity recognition is a fundamental tasks for natural language processing in the agricultural field. More importantly, it is the key basic step of constructing agricultural knowledge graph and intelligent question answering system. Traditional named entity recognition (NER) methods based on CRF model which relies on large amounts of hand-crafted features, cannot extract more effective features and solve the inconsistency of entity tagging caused by the diversity of entity names. To issue the above problems, an Att-BiLSTM-CRF framework was proposed based on deep learning. Firstly, the CBOW model was used to pre-train character embedding on a large number of unlabeled agricultural corpora, and alleviate the impact of segmentation accuracy on the performance of the model. Then, the document-level attention mechanism was introduced to obtain the similar information between entities in the text, so as to ensure the consistency of entity tagging in different contexts. Finally, based on BiLSTM-CRF benchmark model, a model framework suitable for agricultural named entity recognition was constructed. Totally 4 604 agricultural texts were chosen to identify diseases, pests, pesticides and crop varieties. The experimental results showed that the model can effectively identify the entities in the agricultural text and alleviate the problem of inconsistent entity tagging. The model achieved good result in the agricultural corpus, and the recognition precision, recall, and F-score were respectively 93.48%, 90.60% and 92.01%. Compared with other models, such as LSTM model, LSTM-CRF model and BiLSTM-CRF model, Att-BiLSTM-CRF had obvious advantages in different size corpus, and it can effectively identify entities for agricultural texts.

Key words: agricultural text; named entity recognition; Attention mechanism; neural network; deep learning

收稿日期:2020-04-13　修回日期:2020-05-25

基金项目:国家自然科学基金项目(61871041)、国家重点研发计划项目(2019YFD1101105)和北京市科技计划项目(Z191100004019007)

作者简介:赵鹏飞(1987—),男,助理研究员,博士生,主要从事农业信息化研究,E-mail:zhaopf@nercita.org.cn

通信作者:赵春江(1964—),男,中国工程院院士,博士生导师,主要从事农业人工智能与智能系统研究,E-mail:zhaocj@nercita.org.cn

0 引言

随着农业信息化技术的快速发展,农户可通过农技服务平台进行在线问答咨询。面对海量的问题数据,快速而准确地定位关键词、挖掘深层的语义关系是农业智能问答系统亟需解决的问题[1]。农业命名实体识别作为一种智能化信息抽取方法,其主要任务是从非结构化的问答数据中识别不同类型的实体,如农作物病虫害、作物品种、农药名称等,这是构建智能问答系统的关键技术环节,是农业文本信息挖掘领域的热点研究方向。

在农业领域,许多研究者利用机器学习进行实体识别研究。文献[2]提出基于条件随机场的识别方法,通过添加词性、左右指界词等模板特征,对农作物、病虫害及农药 3 类实体进行识别。文献[3]采用 BIO 和 BMES 两种实体标注方式,基于 CRF 模型对数据集中农作物、家禽、病虫害等实体进行识别。文献[4]将农业本体概念作为子特征加入 CRF 模型中,对涉农商品名称进行抽取和类别标注。但是,传统的基于机器学习的方法依赖手工设计的特征模板,在提高模型性能的同时也导致整个模型的鲁棒性和泛化能力下降[5]。

农业实体构词复杂、种类众多,导致农业领域实体识别研究更具有挑战性,主要体现在:由于缺乏规范的农业词典,采用分词工具对农业语料进行分词出现分词错误的现象,影响了模型性能;同一实体在文本中所处位置不同,以单句为处理单元的识别方法无法聚焦全文语境,存在实体标注不一致问题。

随着深度学习算法的改进,网络模型能够自动学习到更深层次的特征信息,在很多领域实体识别任务取得了理想的效果[6-12]。

近年来,注意力机制在自然语言处理领域得到了广泛的应用[13-15]。文献[16]基于 BiLSTM - CRF 框架,通过添加注意力机制学习有效的字符特征向量。文献[17]提出基于双向注意机制的循环神经网络(Recurrent neural network, RNN)模型,该模型能更好地获取标签之间的关系。文献[18]提出了多注意力模型,在阿拉伯语实体识别任务中取得较好的结果。文献[19]利用卷积神经网络提取汉字分解后的特征信息,基于自注意力机制识别医学电子病历的相关实体。

上述基于深度学习的方法为农业领域开展命名实体识别研究提供了参考依据,但在农业文本向量化表示方面并未提出有效的方法来获取字符之间丰富的语义特征,并且相关模型在农业领域数据集上没有进行验证,不足以说明农业领域命名实体识别的相关问题。

本文在农业领域命名实体识别任务中,基于深度学习方法,在 BiLSTM - CRF 网络模型基础上,有针对性地引入大量无标注农业语料,通过预训练方式对农业实体字符分布式表示进行扩充,并引入文档级注意力机制重点关注实体关键字信息,通过余弦距离相似度得分获取文本中实体之间的相关系数,进一步对模型结构和训练参数进行优化和改进,构建基于注意力机制的 Att - BiLSTM - CRF 混合网络模型,以期实现农业文本命名实体的精准识别。

1 数据采集与预处理

1.1 数据采集

农业命名实体识别缺少公开的语料数据集,本文通过数据采集、数据预处理、数据标注 3 个步骤,建立农业领域实体识别语料库。本文的语料数据主要通过爬虫框架,抓取各大农业网站(中国农业信息网、中国农业知识网、中国作物种质资源信息网、国家农业科学数据中心等)关于农作物病虫害和农作物品种的文本语料。其中,标注语料库作为实验数据集,包含 4 604 篇农业文本,共 33 096 个句子;未标注语料库作为预训练数据集,包含 26 025 条语料,共 300 万个中文字符。

1.2 数据预处理

通过爬虫抓取的语料数据,包含大量的网站标签、链接、特殊字符等非文本的结构数据,不利于数据标注。通过 Python 正则表达式、字符格式规范化等操作,删除非文本数据,获取规范化的农业语料库。

1.3 数据标注

本文采用人工标注的方式进行语料库的标注,语料库包含实体共 26 309 个,其中,病害名称 4 129 个,虫害名称 4 275 个、农药名称 11 952 个、农作物品种名称 5 953 个,不同类型实体统计如表 1 所示。使用 BIEO 标记方案表示命名实体,B 表示实体名称的开始,I 和 E 分别表示实体的内部和实体的结束标记,O 表示语料中的非实体。语料库注释示例如图1所示。为更好地识别实体所属类别,将类别

表 1 语料库统计信息

Tab. 1 Corpus statistics

类型	训练集	测试集	验证集
病害名称	2 890	827	412
虫害名称	2 993	856	426
农药名称	8 360	2 391	1 201
农作物品种名称	4 166	1 190	597

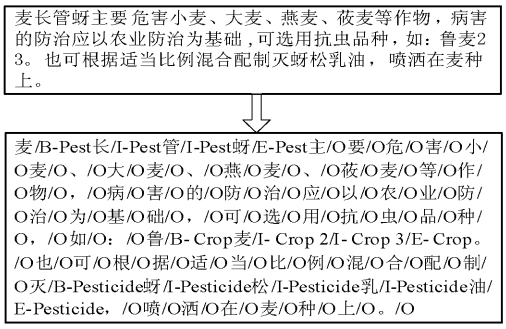

图 1　语料库注释示例

Fig. 1　Corpus tagging results

信息添加在实体标签上，实体类型描述如下：病害名称实体-Disease、虫害名称实体-Pest、农药名称实体-Pesticide、农作物品种名称实体-Crop。其中，B-Disease和 B-Crop 分别表示病害和农作物品种的命名实体的开始。

2　模型框架

本文模型包含字嵌入层、BiLSTM 层、Attention 层和 CRF 层 4 部分，模型结构如图 2 所示。

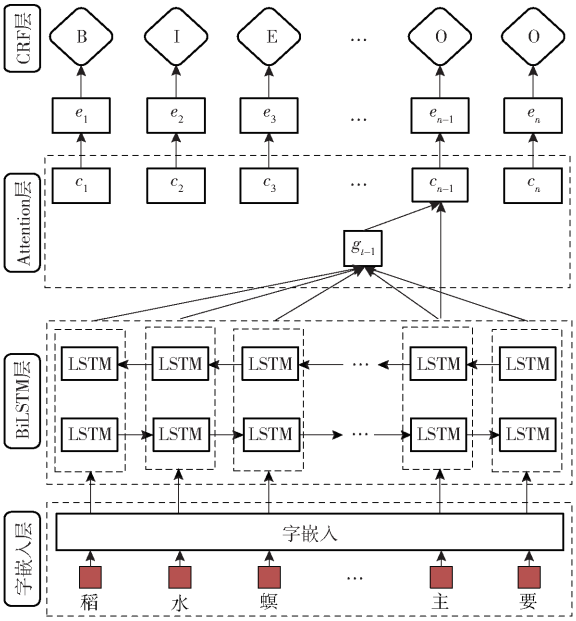

图 2　Att－BiLSTM－CRF 模型结构

Fig. 2　Model architecture of Att－BiLSTM－CRF

2.1　字嵌入层

2.1.1　预处理

在英文 NER 任务中，由于每个单词被空格分隔，很多研究将词向量与字符向量拼接作为模型输入，提高模型的性能。与英文单词不同，中文词语之间没有明显的分隔标记，而且词语具有较强的领域性。因此，为更好地处理中文实体识别任务，多数方法都将分词作为语料处理的基本步骤。但是，现有分词技术不能准确地进行切分，会产生各种各样的错误[20]。

例如，病害实体“水稻细菌性褐条病”分词结果为“水稻/细菌性/褐/条/病”，农作物品种实体“两优培九”分词结果为“两/优/培九”。这些实体被错误地拆分，从而导致模型不能正确获取实体的特征表示，基于字的实体识别可以有效地避免这类问题。

本文使用字向量作为模型初始输入，采用预训练方式，以字为单位进行切割，获取特征表示，缓解分词准确度对性能的影响。

2.1.2　字向量表示

农业文本数据需进行文本向量化，将相应字符映射为一定维度的实数向量，才能被计算机处理。本文采用 Word2vec 的 CBOW 模型[21－22]，在模型架构基础上，针对字向量维度，进一步优化和验证，通过对这些无标注的语料进行无监督训练，得到相应的分布式表示，最终生成特定维度的字向量，构建字向量表。CBOW 模型的框架如图 3 所示，主要有输入层、映射层和输出层 3 层。

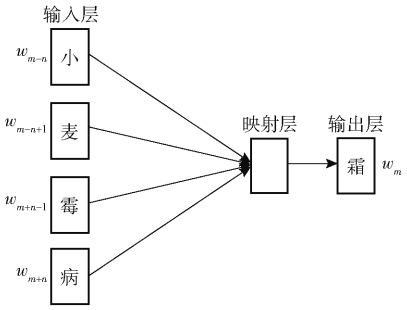

图 3　CBOW 模型结构

Fig. 3　CBOW model structure

在 CBOW 模型中，目标字由上下文推测得到，已知当前字 w_m，利用周围 $2n$（n 为窗口尺寸）个字 w_{m-n}、w_{m-n+1}、…、w_{m+n-1}、w_{m+n} 预测 w_m 当前字出现的概率。以病害实体“小麦霜霉病”为例，通过字“霜”的上下文“小”、“麦”、“霉”、“病”4 个字，来预测所有字出现的概率，其中目标字“霜”出现的概率最大。

在预训练过程中，CBOW 模型字级窗口设置为 2，构建字向量表，每个字对应唯一的向量表示。本文验证了不同维度字向量对模型性能造成的影响，维度设置为 50、100、150 和 200，经过实验对比发现，字向量维度设置为 100 时模型的性能最优。因此，通过预训练方式，获取农业文本 100 维度的字向量特征表示，适用于农业领域命名实体识别。

2.2　BiLSTM 层

LSTM 是一种特殊的循环网络模型，克服了 RNN 模型在训练过程存在的梯度爆炸问题[23]。农业实体的构词方式复杂多样，针对目标实体的识别，需要考虑实体不同位置的上下文信息，来获取更深

层次的特征表示。LSTM 是单向的循环神经网络，只能获取目标词过去的文本信息。例如，病害实体“玉米根腐病”，LSTM 只能访问“腐”的前一个字“根”的特征信息，不能预测下一个字“病”的出现。目标词的上下文信息对实体识别具有不同程度的影响，为了准确识别出农业命名实体，构建了双向 LSTM（BiLSTM）网络模型，进行正向和反向 2 个不同方向的文本表示，充分获取目标词过去和将来的特征信息。

LSTM 网络的主要结构可以形式化地表示为

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_i \boldsymbol{h}_{t-1} + \boldsymbol{U}_i \boldsymbol{x}_t + \boldsymbol{b}_i) \tag{1}$$

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_f \boldsymbol{h}_{t-1} + \boldsymbol{U}_f \boldsymbol{x}_t + \boldsymbol{b}_f) \tag{2}$$

$$\tilde{\boldsymbol{c}}_t = \tan(\boldsymbol{W}_c \boldsymbol{h}_{t-1} + \boldsymbol{U}_c \boldsymbol{x}_t + \boldsymbol{b}_c) \tag{3}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t \tag{4}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o \boldsymbol{h}_{t-1} + \boldsymbol{U}_o \boldsymbol{x}_t + \boldsymbol{b}_o) \tag{5}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t) \tag{6}$$

式中 σ——sigmod 激活函数

tanh——双曲正切激活函数

$\boldsymbol{i}_t$、$\boldsymbol{f}_t$、$\boldsymbol{o}_t$、$\boldsymbol{c}_t$——在 t 时刻的输入门、忘记门、输出门、记忆细胞

$\boldsymbol{U}_i$、$\boldsymbol{U}_f$、$\boldsymbol{U}_c$、$\boldsymbol{U}_o$、$\boldsymbol{W}_i$、$\boldsymbol{W}_f$、$\boldsymbol{W}_o$、$\boldsymbol{W}_c$——不同控制门对应的权重矩阵

$\boldsymbol{b}_i$、$\boldsymbol{b}_f$、$\boldsymbol{b}_o$、$\boldsymbol{b}_c$——偏置向量

$\tilde{\boldsymbol{c}}_t$——输入的中间状态

$\boldsymbol{x}_t$——t 时刻的输入向量

$\boldsymbol{h}_t$——t 时刻的输出结果

$\odot$——点乘运算符

字嵌入层的向量 $\boldsymbol{x}$，将作为 t 时刻 BiLSTM 层的输入，通过正向 LSTM 输出特征序列和反向输出序列，得到隐藏层拼接的向量，经过 tanh 激活函数进行加权得到最终的输出结果 $\boldsymbol{h}_t$，将作为 Attention 层的输入。

2.3 Attention 层

在命名实体识别任务中，由于中文构词方式灵活多变，同一实体具有多种表述方式，实体在文本不同位置可能多次出现。以单句为训练单元的识别模型，关注实体在该句的上下文表示，忽略全文的语境信息，容易造成同一文本实体标注不一致的问题。

例如，水稻稻瘟病的描述如下：水稻又见“【火烧瘟】”，早稻警惕【稻瘟病】流行，一定要早做预防。当前江西早稻，……禾苗都可以点火烧了，名符其实的“【火烧瘟】”。【水稻稻瘟病】又称【稻热病】、【火烧瘟】，症状表现为中央呈灰白色病斑，边缘呈显著褐色，且发病部位在潮湿的环境下会产生灰色的霉状物。

文本中，水稻稻瘟病又称火烧瘟，火烧瘟作为病害实体，在文本中不同句子的不同位置多次出现。以句子为处理单元的模型，在脱离上下文语境的情况下，对【火烧瘟】病害实体出现错标或者漏标的现象。为解决实体标注不一致的问题，通常采用基于规则制定的方法，但是特定领域的规则制定较为复杂，需要较强的领域知识，不同的领域规则不具有通用性。

针对农业文本中实体命名方式多样化、实体分布不均匀的特点，在注意力模型基本架构上进行扩展，引入文档级全局信息，并增加余弦距离得分的相似性评估，对处于不同位置的同一实体重点关注。基于注意力的学习模型，能够忽略文本中无关的信息，关注实体关键信息，模型以整篇文本作为训练单元，考虑实体上下文的语境信息，缓解实体标注不一致问题。

本文用 $D = (S_1, S_2, \cdots, S_d)$ 表示文档包含 d 个句子，每个句子 $S = (w_1, w_2, \cdots, w_m)$ 包含 m 个字，文档中包含字的总数是 N。对于文档中的实体，通过注意矩阵 $\boldsymbol{A}$ 处理 BiLSTM 层输出的特征序列，来计算当前目标字与文档中所有字之间的相关性，获取目标字 w_i 基于文档层面的全局特征表示 g_i，计算公式为

$$g_i = \sum_{j=1}^{N} A_{i,j} h_j \tag{7}$$

其中

$$A_{i,j} = \frac{\exp(\mathrm{score}(w_i, w_j))}{\sum_{k=1}^{m} \exp(\mathrm{score}(w_i, w_k))} \tag{8}$$

$$\mathrm{score}(w_i, w_j) = \frac{W_a (w_i w_j)}{|w_i||w_j|} \tag{9}$$

式中 $A_{i,j}$——当前字 w_i 与文档中字 w_j 注意力权重

h_j——BiLSTM 层输出

$\mathrm{score}(w_i, w_j)$——采用余弦距离判定的字 w_i 与字 w_j 相似性得分

W_a——训练过程中学习到的参数

最后，目标字 w_i 在文档级注意力层的输出为 c_i，通过 tanh 函数来获取置信度 e_i，计算公式为

$$c_i = \tanh(\boldsymbol{W}_g [g_i, h_i]) \tag{10}$$

$$e_i = \tanh(\boldsymbol{W}_e c_i) \tag{11}$$

式中 $\boldsymbol{W}_g$、$\boldsymbol{W}_e$——训练时学习到的参数矩阵

2.4 CRF 层

在 CRF 层，采用状态转换矩阵来预测当前标签，获得全局最优的标记序列[24]。设定 $\boldsymbol{P}$ 为 Attention 层的输出矩阵，维度为 $m \times k$，m 表示输入句子包含字的数量，k 表示标签集合的元素数。对

于输入文档 D，对应的输出标签序列 $y=(y_1,y_2,\cdots,y_n)$ 的概率为

$$s(X,y)=\sum_{n}\left(\sum_{i=1}^{m}P_{i,y_i}+\sum_{i=0}^{m}A_{y_i,y_{i+1}}\right) \tag{12}$$

式中　X——输入的文本序列

$A_{y_i,y_{i+1}}$——从标签 y_i 转移到标签 y_{i+1} 的分数，$A_{y_i,y_{i+1}}$ 的值越大表示标签 i 转移到标签 j 的可能性越大

P_{i,y_i}——第 i 个字被预测为第 y_i 个标签的分数

然后，利用 Softmax 函数，得到序列 y 的条件概率。最后，使用 Viterbi[25] 算法将得分最高的序列 y^* 作为模型最终的标注结果。

2.5　模型参数配置及评价

模型的参数配置如表 2 所示，参数通过反复实验确定的，字向量维度设置为 100。模型使用双向的 LSTM 网络，隐藏层维度设置为 128。为减轻模型过拟合问题，引入 Dropout 机制[26]，Dropout 的值直接影响到模型性能，设置为 0.5。选取 ADAM[27] 优化算法，学习率为 0.002。模型训练批处理参数为 16，迭代次数设置为 50。

表 2　参数配置

Tab.2　Parameter setting

参数	数值
字向量维度	100
隐藏单元数	128
Dropout	0.5
学习率	0.002
批处理参数	16
迭代次数	50

与其他实体识别方法相似，采用准确率 P、召回率 R、F 值作为实验的评价指标[28]。

3　实验结果

在不依赖人工设计特征的情况下，通过调整不同的模型参数，在 1.3 节构建的标注数据集上验证模型的识别性能。语料库中训练集、测试集、验证集按 7∶2∶1的比例进行分配，数据集之间无重叠，因此测试数据集的实验结果可作为实体识别效果的评价指标。

3.1　不同嵌入向量性能比较

本文分别以词向量和字向量作为 Att－BiLSTM－CRF 模型的初始输入，验证不同嵌入向量对模型性能的影响，对比结果如表 3 所示。将字向量作为模型的输入，模型识别准确率 P 为 93.48%，相较于词向量作为模型输入，准确率提升了 2.96 个百分点。分析结果得知，基于词向量的输入，实体被错误拆分，导致这些复杂的实体没有被正确识别，例如，水稻品种“广 8 优郁香”被错误地拆分为“广/8/优郁/香/”。接着，验证了不同字向量维度对模型性能的影响。字向量维度设置为 50、100、150、200，模型准确率 P 分别为 91.19%、93.48%、92.15%、91.83%；召回率 R 分别为 89.5%、90.6%、90.08%、90.21%；F 值分别为 90.29%、92.01%、91.04%、91.00%。从实验结果看出，适当增加字向量维度，可以获取质量更好的字级分布式表示，字向量维度为 100 时，模型性能达到最高。随着维度越来越大，训练成本越来越高，模型性能很难得到提升，甚至下降。针对农业实体，字向量维度不是越大越好，在一定范围内存在局部最优值。

表 3　不同嵌入向量实验结果对比

Tab.3　Results of different embedding　%

模型	P	R	F 值
word + Att－BiLSTM－CRF	90.52	90.26	90.38
char + Att－BiLSTM－CRF	93.48	90.60	92.01

3.2　不同注意力机制的性能比较

采用字向量维度 100，并在 BiLSTM－CRF 模型框架上增加句子级和文档级的注意层，并对模型性能进行了评估。结果如表 4 所示。句子级的方法，模型的准确率 P 为 91.23%，召回率 R 为 89.24%，F 值为 90.23%。分析结果发现，同一文本中，部分农药实体“农抗 120/Pesticide”，被错误标记为农作物品种实体“农抗 120/Crop”。这种标记不一致的现象，是由于“农抗 120”与大多数农作物品种实体构词方式相似，都是“词＋数字”的方式，在识别过程中，虽然句子级注意力获取了该实体在句中特征信息，但是并没有考虑全文的语境，从而导致上述错误的判断。

表 4　不同 Attention 机制实验结果对比

Tab.4　Results of different Attention mechanisms　%

模型	P	R	F 值
Att(S)－BiLSTM－CRF	91.23	89.24	90.23
Att(D)－BiLSTM－CRF	93.48	90.60	92.01

与基于句子级的方法相比，文档级方法模型的准确率 P、召回率 R、F 值分别提高了 2.25、1.36、1.78 个百分点。结果表明，文档级方法通过获取文档中字之间的相关信息，通过余弦函数计算文档中目标字与其他字的相似度，调整目标字的权重，在缓解上述讨论的标记不一致问题的同时，有效地提高了模型性能。

3.3　不同模型性能比较

为了验证本文提出的基于 Att－BiLSTM－CRF

在农业语料上的识别性能,在不同的模型上进行对比实验,模型包括:LSTM[29]、LSTM - CRF[30]、BiLSTM - CRF[31]以及本文提出的基于文档级的 Att -BiLSTM - CRF,实验结果如表 5 所示。在准确率 P、F 值两方面,对比了各模型针对 4 类实体的识别性能,结果如图 4 所示。

表 5 不同模型的实验结果对比

Tab. 5 Experimental results of different models %

模型	P	R	F 值
LSTM	80.36	83.80	81.95
LSTM - CRF	83.95	87.67	85.74
BiLSTM - CRF	89.89	88.96	89.41
Att - BiLSTM - CRF	93.48	90.60	92.01

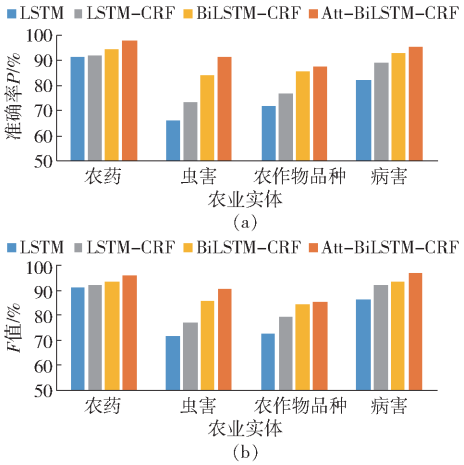

图 4 农业实体实验结果对比

Fig. 4 Experimental results of agricultural entities

由表 5 可知,LSTM 模型通过隐藏层获取过去的序列信息,结构比较单一,模型准确率为 80.36%。LSTM - CRF 模型相比于 LSTM 模型,通过添加 CRF 层,利用实体间相邻的标签动态规划最优的序列标注,模型准确率为 83.95%。为了获得输入序列丰富的上下文信息,基于 BiLSTM - CRF 模型框架,模型准确率为 89.89%,与 LSTM - CRF 模型相比,提升了 5.94 个百分点。基于文档级注意力的 Att - BiLSTM - CRF 模型,通过添加注意力层,获取文本中实体间的相似系数,与其他 3 个模型相比,准确率 P 和 F 值最高,分别为 93.48% 和 92.01%。

图 4 展示了 4 种模型对于农药、虫害、农作物品种以及病害 4 类实体的识别率 P 和 F 值,4 种模型对病害和农药实体识别准确率较高,虫害和农作物品种较低。LSTM 模型结构单一,对于复杂的虫害和农作物品种实体,模型不能获取丰富的特征信息,识别率为 65.92% 和 71.64%,F 值为 71.33% 和 72.48%。LSTM - CRF 模型对虫害实体识别率为 73.18%,农作物品种实体识别率为 76.59%,相较于 LSTM 模型分别提高了 7.26、4.95 个百分点。

分析得出,病害和农药具有较规则的后缀组成词,例如,病害的“病”、农药的“乳油”等,这些明显的字特征信息提高了这类实体识别的准确率。而虫害和农作物的构词比较复杂,例如“数字 + 词”、“数字 + 字母”等方式,因此这类实体需要提升模型的复杂性,来获取更丰富的特征信息。

BiLSTM - CRF 模型对农药和病害实体识别率相对较高,为 94.35%、92.70%,对虫害和农作物两类实体识别率为 83.66%、85.47%,相较于 LSTM - CRF 模型,分别提升了 10.48、8.88 个百分点。模型通过双向 LSTM 隐藏层提取过去和未来的序列信息,对复杂、长度较大的实体识别率有较大提升。但是,模型依然存在实体标签不一致的现象。

本文 Att - BiLSTM - CRF 模型对农药实体识别率达到 97.58%,虫害实体识别率为 91.15%,对于构词更复杂的农作物品种实体识别率达到最高的 87.26%,F 值为 84.92%。进一步验证了添加文档级的注意力机制,结合实体所在文本的语境信息,获取实体关注度能够提高农业实体的识别效果。

实验结果表明,本文提出的 Att - BiLSTM - CRF 模型不使用任何字典或外部注解资源,在训练过程中动态地获取实体间的相似关系,能够有效地识别农业复杂实体,F 值达到 92.01%。

3.4 不同模型识别效率比较

为了验证语料集的规模对模型性能的影响,本文新增了 3 个语料库,包含实体数量分别为 9 906、15 020、20 618,新增的语料库同样按照 7∶2∶1 的比例进行分配,数据集之间无重叠,实验结果如下:LSTM 模型由于结构比较单一,在 4 种规模语料库准确率较低,分别为 64.52%、72.85%、83.93%、85.11%。LSTM - CRF 模型通过添加 CRF 层,获取标签转移的最优概率,与 LSTM 相比,模型准确率分别提高了 1.40、2.03、0.43、1.57 个百分点。BiLSTM - CRF 和 Att - BiLSTM - CRF 在语料集较小的情况下,模型达到较好的识别效果。随着语料集规模的扩大,融入注意力机制的 Att - BiLSTM - CRF 模型,在 4 种规模语料库识别准确率均达到最高,分别为 85.11%、86.68%、90.29%、93.48%。

最后,本文通过中国农技推广信息平台,在农技问答板块,抽取了相应的农户问答文本数据,应用 Att - BiLSTM - CRF 模型对文本数据进行了实体识别,结果如表 6 所示。

表 6　问答数据识别结果示例

Tab. 6　Examples of Q&A data recognition results

数据序号	数据内容	识别结果
1	小麦赤霉病什么时间用药防治？	病害：小麦赤霉病
2	治疗大豆根腐病的药物有哪些？	病害：大豆根腐病
3	小麦锈病俗称“黄疸病”，分条锈病、秆锈病、叶锈病 3 种，是中国小麦生产上分布广、传播快、危害面积大的重要病害。	病害：小麦锈病；黄疸病；条锈病；秆锈病；叶锈病
4	这是桃蚜，用吡虫啉兑水喷雾防治。	虫害：桃蚜；农药：吡虫啉
5	这是蛴螬虫，危害植物的根部。	虫害：蛴螬

4　结论

（1）针对农业领域命名实体识别中实体识别类别众多、实体类型组成复杂，造成分词不准确等问题，提出基于注意力机制的 Att－BiLSTM－CRF 神经网络模型方法，提升了识别性能，F 值为 92.01%。

（2）通过预训练的方法获取农业实体字级的分布式表示，缓解分词错误造成的性能影响。通过多种向量维度的实验，证明基于字向量的识别方法适用于农业领域 NER 任务，字向量维度设置为 100，模型准确率 P 达到 93.48%，召回率为 90.60%。

（3）基于文档级的注意力机制获取实体间的相似度，可确保农业实体标签的一致性，避免错标或者漏标的情况，提高了模型识别性能。

参　考　文　献

[1] 赖英旭，李亚娟，刘静. 基于本体的水稻育种方法应用知识库构建[J]. 北京工业大学学报，2019，45(12)：1181－1191.
LAI Yingxu，LI Yajuan，LIU Jing. Construction of ontology-based rice breeding method knowledge base[J]. Journal of Beijing University of Technology，2019，45(12)：1181－1191.（in Chinese）

[2] 李想，魏小红，贾璐，等. 基于条件随机场的农作物病虫害及农药命名实体识别[J/OL]. 农业机械学报，2017，48(增刊)：178－185.
LI Xiang，WEI Xiaohong，JIA Lu，et al. Recognition of crops，diseases and pesticides named entities in Chinese based on conditional random fields[J/OL]. Transactions of the Chinese Society for Agricultural Machinery，2017，48(Supp.)：178－185. http：//www. j-csam. org/jcsam/ch/reader/view_abstract. aspx？ flag = 1&file_no = 2017s029&journal_id = jcsam. DOI：10.6041/j. issn. 1000-1298. 2017. S0. 029.（in Chinese）

[3] 王春雨，王芳. 基于条件随机场的农业命名实体识别研究[J]. 河北农业大学学报，2014，37(1)：132－135.
WANG Chunyu，WANG Fang. Study on recognition of chinese agricultural named entity with conditional random fields[J]. Journal of Agricultural University of Hebei，2014，37(1)：132－135.（in Chinese）

[4] 黄念娥，黄河，王儒敬. 本体与条件随机场结合的涉农商品名称抽取与类别标注[J]. 计算机应用，2017，37(1)：233－238.
HUANG Niane，HUANG He，WANG Rujing. Agriculture-related product name extraction and category labeling based on ontology and conditional random field[J]. Journal of Computer Applications，2017，37(1)：233－238.（in Chinese）

[5] HABIBI M，WEBER L，NEVES M，et al. Deep learning with word embeddings improves biomedical named entity recognition[J]. Bioinformatics，2017，33(14)：i37－i48.

[6] TONG F，LUO Z H，ZHAO D S. A deep network based integrated model for disease named entity recognition[C]//2017 IEEE International Conference on Bioinformatics and Biomedicine，2017：618－621.

[7] ZHAO Z H，YANG Z H，LUO L，et al. Disease named entity recognition from biomedical literature using a novel convolutional neural network[J]. BMC Medical Genomics，2017，10(Supp. 5)：75－83.

[8] XU J J，HE H F，XU S，et al. Cross-domain and semi-supervised named entity recognition in chinese social media：a unified model[J]. ACM Transactions on Audio，Speech，and Language Processing，2018，26(11)：2142－2152.

[9] 李丽双，郭元凯. 基于 CNN－BLSTM－CRF 模型的生物医学命名实体识别[J]. 中文信息学报，2018，32(1)：116－122.
LI Lishuang，GUO Yuankai. Biomedical named entity recognition with CNN－BLSTM－CRF[J]. Journal of Chinese Information Processing，2018，32(1)：116－122.（in Chinese）

[10] 李健龙，王盼卿，韩琪羽. 基于双向 LSTM 的军事命名实体识别[J]. 计算机工程与科学，2019，41(4)：713－718.
LI Jianlong，WANG Panqing，HAN Qiyu. Military named entity recognition based on bidirectional LSTM[J]. Computer Engineering & Science，2019，41(4)：713－718.（in Chinese）

[11] 秦娅，申国伟，赵文波，等. 基于深度神经网络的网络安全实体识别方法[J]. 南京大学学报(自然科学版)，2019，55(1)：29－40.
QIN Ya，SHEN Guowei，ZHAO Wenbo，et al. Research on the method of network security entity recognition based on deep neural network[J]. Journal of Nanjing University(Natural Science)，2019，55(1)：29－40.（in Chinese）

[12] 孙娟娟，于红，冯艳红，等. 基于深度学习的渔业领域命名实体识别[J]. 大连海洋大学学报，2018，33(2)：265－269.
SUN Juanjuan，YU Hong，FENG Yanhong，et al. Recognition of nominated fishery domain entity based on deep learning

architectures[J]. Journal of Dalian Fisheries University, 2018, 33(2):265 - 269. (in Chinese)

[13] 张晗, 郭渊博, 李涛. 结合 GAN 与 BiLSTM - Attention - CRF 的领域命名实体识别[J]. 计算机研究与发展, 2019, 56(9):1851 - 1858.
ZHANG Han, GUO Yuanbo, LI Tao. Domain named entity recognition combining GAN and BiLSTM - Attention - CRF[J]. Journal of Computer Research and Development, 2019, 56(9):1851 - 1858. (in Chinese)

[14] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems, 2017: 5998 - 6008.

[15] 田萱, 王子亚, 王建新. 基于语义分割的食品标签文本检测[J/OL]. 农业机械学报, 2020, 51(8):336 - 343.
TIAN Xuan, WANG Ziya, WANG Jianxin. Text detection of food labels based on segmentation[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(8): 336 - 343. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? file_no=20200837&flag=1. DOI:10.6041/j.issn.1000-1298.2020.08.037. (in Chinese)

[16] BHARADWAJ A, MORTENSEN D, DYER C, et al. Phonologically aware neural model for named entity recognition in low resource transfer settings[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016:1462 - 1472.

[17] PANDEY C, IBRAHIM Z, WU H 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.

[18] ALI M N A, TAN G Z, HUSSAIN A. Boosting arabic named-entity recognition with multi-attention layer[J]. IEEE Access, 2019, 7:46575 - 46582.

[19] YIN M W, MOU C J, XIONG K N, et al. Chinese clinical named entity recognition with radical-level feature and self-attention mechanism[J]. Journal of Biomedical Informatics, 2019, 98:103289.

[20] 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.
LI Ni, GUAN Huanmei, YANG Piao, et al. BERT - IDCNN - CRF for named entity recognition in Chinese[J]. Journal of Shandong University(Natural Science), 2020, 55(1): 102 - 109. (in Chinese)

[21] WANG Y S, LIU S J, AFZAL N, et al. A comparison of word embeddings for the biomedical natural language processing[J]. Journal of Biomedical Informatics, 2018, 87:12 - 20.

[22] REI M, CRICHTON G, PYYSALO S. Attending to characters in neural sequence labeling models[J]. arXiv:1611.04361, 2016.

[23] WANG Q, ZHOU Y M, TONG R, et al. Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition[J]. Journal of Biomedical Informatics, 2019, 92:103133.

[24] 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)

[25] XU G H, WANG C Y, HE X F. Improving clinical named entity recognition with global neural attention[C]// Asia-Pacific Web and Web-Age Information Management Joint International Conference on Web and Big Data, 2018:264 - 279.

[26] 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.

[27] 买买提·阿依甫, 吾守尔·斯拉木, 帕丽旦·木合塔尔, 等. 基于 BiLSTM - CNN - CRF 模型的维吾尔文命名实体识别[J]. 计算机工程, 2018, 44(8):230 - 236.
MAIMAITI Ayifu, WUSHOUER Silamu, PALIDAN Muhetaer, et al. Uyghur named entity recognition based on BiLSTM - CNN - CRF model[J]. Computer Engineering, 2018, 44(8):230 - 236. (in Chinese)

[28] GRIDACH M. Character-level neural network for biomedical named entity recognition[J]. Journal of Biomedical Informatics, 2017, 70:85 - 91.

[29] SAK H, SENIOR A, BEAUFAYS F. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition[C]// Proceedings of the 15th Annual Conference of the International Speech Communication Association, 2014:338 - 342.

[30] 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]// International Conference on Computer Processing of Oriental Languages, 2016:239 - 250.

[31] HUANG Z H, XU W, YU K. Bidirectional LSTM - CRF models for sequence tagging[J]. Computer Science, 2015, 4(1): 1508 - 1519.