·农业信息与电气技术·

基于 BERT 的多特征融合农业命名实体识别

赵鹏飞[1]，赵春江[1,2※]，吴华瑞[2,3,4]，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；4. 北京农业智能装备技术研究中心，北京 100097）

摘 要：命名实体识别是农业文本信息抽取的重要环节，针对实体识别过程中局部上下文特征缺失、字向量表征单一、罕见实体识别率低等问题，提出一种融合 BERT（Bidirectional Encoder Representations from Transformers，转换器的双向编码器表征量）字级特征与外部词典特征的命名实体识别方法。通过 BERT 预训练模型，融合左右两侧语境信息，增强字的语义表示，缓解一词多义的问题；自建农业领域词典，引入双向最大匹配策略，获取分布式词典特征表示，提高模型对罕见或未知实体的识别准确率；利用双向长短时记忆（Bi-directional Long-short Term Memory，BiLSTM）网络获取序列特征矩阵，并通过条件随机场（Conditional Random Field，CRF）模型生成全局最优序列。结合领域专家知识，构建农业语料集，包含 5 295 条标注语料，5 类农业实体。模型在语料集上准确率为 94.84%、召回率为 95.23%、F_1 值为 95.03%。研究结果表明，该方法能够有效识别农业领域命名实体，识别精准度优于其他模型，具有明显的优势。

关键词：农业；命名实体识别；文本；BERT；词典特征；BiLSTM

doi：10.11975/j.issn.1002-6819.2022.03.013

中图分类号：TP391.1 **文献标志码**：A **文章编号**：1002-6819(2022)-03-0112-07

赵鹏飞，赵春江，吴华瑞，等. 基于 BERT 的多特征融合农业命名实体识别[J]. 农业工程学报，2022，38(3)：112-118.
doi：10.11975/j.issn.1002-6819.2022.03.013 http://www.tcsae.org

Zhao Pengfei, Zhao Chunjiang, Wu Huarui, et al. Recognition of the agricultural named entities with multi-feature fusion based on BERT[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2022, 38(3): 112-118. (in Chinese with English abstract) doi：10.11975/j.issn.1002-6819.2022.03.013 http://www.tcsae.org

0 引 言

面对海量的非结构化农业文本数据，农业命名实体识别任务能够快速准确的识别农业实体，获取高质量的语义知识，为农业信息抽取与语义检索提供支撑，最终为基层农业技术员提供专业、个性化的决策信息服务[1]。

命名实体识别（Named Entity Recognition，NER）任务中，基于统计机器学习的方法将实体识别当作序列标注任务来处理，常见的模型有隐马尔可夫模型[2]、最大熵模型[3]和条件随机场等[4]。文献[5-7]基于条件随机场模型，构建不同组合的特征模板，对农业领域实体进行识别。但是，机器学习方法依赖人工制定的特征模板，耗时耗力，不具备领域通用性[8]。近年来，基于深度学习的 NER 研究相继展开。与机器学习方法相比，深度学习通过自动学习特征，以端到端的形式训练模型，在生物化学、医疗文本、军事等领域取得了突破性进展[9-11]。研究者使用 Word2vec[12]工具，预训练获取字向量，作为模型的输入。王欢等[13]提出一种基于 BiLSTM 与具有回路的条件随机场相结合的方法，对机床设备故障领域的实体展开了研究。龚乐君等[14]提出了一种基于领域词典和条件随机场的双层标注模型，从病历文本识别出疾病、症状、药品、操作四类实体。但是 Word2vec 生成的字向量是静态的，表征单一，无法解决一词多义的问题。为更好地提取文本特征信息，BERT[15]被广泛应用于 NER 任务中。尹学振等[16]提出 BERT-BiLSTM-CRF 实体识别模型，基于 BERT 的字向量表达层获取字级别特征，在军事领域进行实体识别研究。李建等[17]将中文特征和句法语义特征相结合，完成对专利文本实体的识别。陈剑等[18]基于 BiLSTM-CRF 模型融入 BERT 层，在司法文书语料库上进行实体识别，解决特征提取效率低的问题。此外，毛明毅等[19]基于 BERT-BiLSTM-CRF 模型，有针对性的减少 BERT 嵌入层数，在中文数据集上验证了模型的有效性。

在农业领域，缺少公开标注的数据集，相关研究仍处于起步阶段。Guo 等[20]基于卷积神经网络和注意力机制搭建 NER 模型，有效识别农业病虫害等实体。目前农业领域的命名实体识别存在以下问题：1）模型无法解决一词多义的现象；2）罕见或未知实体的识别率低；3）农业外部词典利用不充分。

针对上述问题，本文面向农业领域提出一种基于 BERT 和词典特征融合的命名实体识别模型 BERT-Dic-BiLSTM-CRF，该模型引入 BERT 双向编码器，进行预训练，获取丰富的字级语义信息，解决一词多义的问题；针对农业领域外部词典丰富的特点，引入词典特征信息，提升模型对罕见实体的识别率。然后将字级向量与词典特征拼接，作为 BiLSTM-CRF 层的输入，最

收稿日期：2021-09-16 修订日期：2022-01-10
基金项目：国家重点研发计划项目（2019YFD1101105）；国家自然科学基金项目（61871041）；北京市科技计划项目（Z191100004019007）
作者简介：赵鹏飞，博士生，研究方向为农业信息化技术。
Email：zhaopf@nercita.org.cn
※通信作者：赵春江，中国工程院院士，研究员，博士生导师，研究方向为农业信息技术和精准农业技术。Email：zhaocj@nercita.org.cn

终获得全局最优的标记序列。

1　语料集构建

1.1　数据获取

本文使用轻量级爬虫框架 Scrapy[21]，在中国农药信息网、中国作物种质信息网、百度百科、国家农业科学数据中心等权威机构获取相应的文本数据，通过数据清洗、去噪、去冗等预处理，保证数据可靠性。结合领域专家知识对语料进行类别划分和标注，构建农业语料集，包含 5 295 条标注语料，共 29 483 个实体，涵盖农作物病害、农作物虫害、农药名称、农机名称、农作物品种名称 5 类实体。

1.2　标注体系

本文采用 BIOE 标注体系，其中，B-*代表实体的起始位置、I-*代表实体内部、E-*代表实体的结束位置、O 代表非实体部分、*代表实体类别标签，标注示例如表 1 所示。

表 1　语料库标注示例

Table 1　Corpus labeling example

类别 Type	实体 Entity	标签 Label
农作物病害 Crop diseases	水稻恶苗病、小麦条锈病	Disease
农作物虫害 Crop pests	玉米蚜、小麦叶蝉	Pest
农药名称 Pesticide name	多菌灵可湿性粉剂	Pesticide
农机名称 Machinery name	花生脱壳机	Machinery
农作物品种名称 Crop variety name	中棉所 9711	Crop

农业实体具有很强的领域专业性，通过制定标注策略，更好地确定实体边界，保证实体完整性。标注策略描述如下：

现象 1：同一病害危害不同种类农作物，例如：水稻纹枯病、小麦纹枯病。

策略 1：病害与农作物名称相连，区分不同农作物病害。

现象 2：部分实体包含英文字母、特殊符号。例如：蚕豆萎蔫病毒（Broad Bean Wilt Virus，BBWV）病。

策略 2：英文字母、特殊符号与实体相连作为一个整体。

现象 3：同一虫害可危害不同种类农作物，例如：水稻管蓟马、小麦管蓟马。

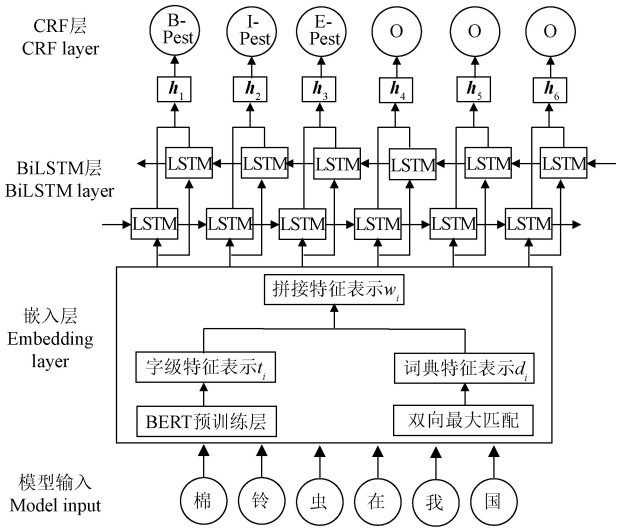
策略 3：虫害与农作物名称相连，区分不同农作物虫害。

2　命名实体识别模型

本文提出的 BERT-Dic-BiLSTM-CRF 模型的整体结构如图 1 所示。模型分为 BERT 层、嵌入层、BiLSTM 层和 CRF 层。对于输入序列 $\boldsymbol{X}=\{x_1,x_2,x_3,\cdots,x_n\}$，$\boldsymbol{x_i}$ 通过 BERT 层生成具有丰富信息的字向量 $\boldsymbol{t_i}$；$\boldsymbol{t_i}$ 与词典特征 $\boldsymbol{d_i}$ 拼接得到向量 $\boldsymbol{w_i}$，$\boldsymbol{w_i}=\boldsymbol{t_i}\oplus\boldsymbol{d_i}$。随后，$\boldsymbol{w_i}$ 输入到 BiLSTM 层进行解码，最后通过 CRF 层输出全局最优序列。

2.1　BERT

自然语言处理领域中，丰富、无监督的预训练是语言理解系统不可或缺的部分[22]，Word2Vec 是使用最广泛的模型。但 Word2Vec 提取的语义信息不足，无法表征字的多义性。农业文本中，实体存在不同语境下不同含义的现象，比如“油葫芦”在不同语境下属于虫害，危害棉花、花生等农作物，也可属于檀香科檀梨属植物—油葫芦。



注：BERT 为转换器的双向编码表征量；BiLSTM 为双向长短时记忆网络；CRF 为条件随机场；$\boldsymbol{h_i}$ 为 BiLSTM 输出的序列向量。

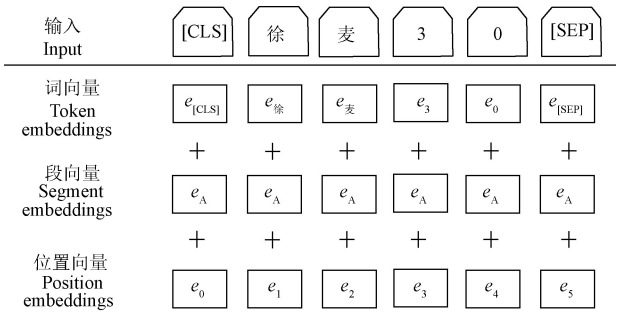
Note: BERT is Bidirectional encoder representations from transformers, BiLSTM is Bi-directional Long Short-Term Memory (LSTM), CRF is conditional random field, $\boldsymbol{h_i}$ is the vector output by the BiLSTM.

图 1　BERT-Dic-BiLSTM-CRF 模型结构

Fig.1　Main framework of BERT-Dic-BiLSTM-CRF

为充分利用语句上下文信息，获取丰富的字级语义表示，本文引入 BERT 预训练模型，完成对语料集字级特征向量表示。BERT 预训练模型基于双向 Transformer 编码器，通过遮蔽语言模型获取词级特征表示、以及下一句预测模型学习文本序列句子级的语义关系，更好的提取文本特征信息。

BERT 预训练过程中，序列 $\boldsymbol{X}=\{x_1,x_2,x_3,\cdots,x_n\}$ 对应的输入 $\boldsymbol{E}=\{\boldsymbol{E}_1,\boldsymbol{E}_2,\boldsymbol{E}_3,\cdots,\boldsymbol{E}_n\}$ 由三个嵌入特征叠加而成，x_i 为序列 $\boldsymbol{X}$ 的第 i 个字，如图 2 所示。其中，在序列 $\boldsymbol{X}$ 初始位置添加一个特殊标记[CLS]，句子间用[SEP]分隔，三个嵌入特征分别为字符嵌入 e_i^t、句子嵌入 e_i^s、位置嵌入 e_i^p，其中 $\boldsymbol{E}_i=e_i^t+e_i^s+e_i^p$。向量 $\boldsymbol{E}$ 经过多个双向 Transformer 编码器获得含有丰富语义特征的向量 $\boldsymbol{T}=\{t_1,t_2,t_3,\cdots,t_n\}$，向量 $\boldsymbol{T}$ 作为 BiLSTM 层的输入。



注：[CLS]标识序列开始位置；[SEP]标识句子间分割；e_i 表示每个字符的嵌入特征。

Note: [CLS] identifies the starting position of the sequence, and [SEP] identifies the sentence as split, e_i represents the embedding feature of each character.

图 2　BERT 模型输入表示

Fig.2　Input representations of BERT

2.2　词典特征

对于小规模语料库，模型学习到的实体信息有限，

难以识别罕见或未知的实体。农业领域存在丰富的外部词典，词典与语料库文本存在着联系，可完成对语料库的信息补充。本文引入词典特征信息，将词典特征与通过 BERT 获取的字级向量融合，增强序列的语义信息，进一步提升模型性能。选取《农业大词典》中农药、农业机械、农作物等词目，并添加搜狗农业词汇大全词典、百度病害词典进行最新词汇的更新，完成外部词典的构建。词典涵盖 5 类实体，共 9 185 词汇。本文设计了 N-gram 特征模板法和双向最大匹配法两种方式抽取词典特征，用于增强农业实体的外部信息。试验结果表明双向最大匹配法优于 N-gram 特征模板法，适用于农业领域 NER 任务。

2.2.1 N-gram 特征模板法

对于序列 $\boldsymbol{X}=\{x_1,x_2,x_3,\cdots,x_n\}$，基于 N-gram 特征模板法，设计 m 个模板，为字符 x_i 构造二值特征 d_i，最终词典特征表示为 $\boldsymbol{D}=\{d_1,d_2,d_3,\cdots,d_n\}$。本文验证了不同模板数量对模型性能造成的影响，$m$ 取值为 8、10、12，经试验对比分析得出，模板数量为 10 时，实体识别模型性能最佳。特征模板如表 2 所示，其中 2-gram 表示字 x_i 与上一个字 x_{i-1}，以及下一个字 x_{i+1} 组成的字符片段，即 $x_{i-1}x_i$ 和 x_ix_{i+1}。根据表 2 的模板遍历序列 $\boldsymbol{X}$ 中每一个字及它的上下文，如果字符段与词典实体匹配成功，取值为 1；反之，取值为 0。

表 2 N-gram 特征模板

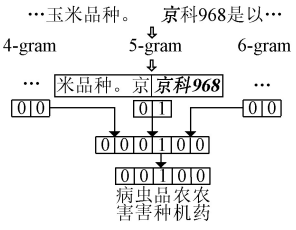
Table 2 N-gram feature templates

类型 Type	模板 Template
2-gram	$x_{i-1}x_i$, x_ix_{i+1}
3-gram	$x_{i-2}x_{i-1}x_i$, $x_ix_{i+1}x_{i+2}$
4-gram	$x_{i-3}x_{i-2}x_{i-1}x_i$, $x_ix_{i+1}x_{i+2}x_{i+3}$
5-gram	$x_{i-4}x_{i-3}x_{i-2}x_{i-1}x_i$, $x_ix_{i+1}x_{i+2}x_{i+3}x_{i+4}$
6-gram	$x_{i-5}x_{i-4}x_{i-3}x_{i-2}x_{i-1}x_i$, $x_ix_{i+1}x_{i+2}x_{i+3}x_{i+4}x_{i+5}$

注：2-gram 表示包含目标字符在内，前后方向选取距离为 2 的字符片段。x_i 为序列 $\boldsymbol{X}$ 的第 i 个字符。

Note: 2-gram represents select character segment with distance of 2 in the front and back direction, including the target character. x_i is the ith character of sequence $\boldsymbol{X}$.

本文构建的语料库中包含 5 种不同类型的实体，每个模板将对应一个 5 维的特征向量，表示其对应的实体类型。基于 N-gram 特征模板法，当 m=10 时，字符 x_i 将产生 50 维的词典特征向量，包含实体边界信息和类型信息，如图 3 所示。



注：字体“京”字为目标字符，“京科 968”为匹配成功的字符片段。

Note: The font “京” is the target character, “京科 968” is character segment that matches successfully.

图 3 N-gram 特征向量

Fig.3 The feature vector of N-gram

2.2.2 双向最大匹配法

基于双向最大匹配算法[23]（Bi-Directional Maximum Matching，BDMM），完成对序列 $\boldsymbol{X}$ 的切分，并与词典实体进行匹配，如果匹配成功则进行标记，保证将词典中存在的最长实体切分出来，如表 3 所示。双向最大匹配算法包含正向最大匹配和逆向最大匹配。其中，正向最大匹配步骤如下：

1）将序列 $\boldsymbol{X}$ 的第一个字符设为当前字符，进行第 2）步；

2）从当前字符开始，按照从左到右的顺序切分，得到字符串 S；在词典中查找字符串 S，如匹配成功，进行标记，进行第 3）步；匹配不成功，进行第 4）步；

3）将字符串 S 的下一个字符设为当前字符，进行第 2）步；

4）去掉字符串 S 的最后一个字，进行第 2）步；

5）重复 2）~4）步，处理完序列 $\boldsymbol{X}$ 为止。

最后，将正向匹配和逆向匹配结果进行对比，选择片段数量少的切分结果，并通过独热编码（One-hot Encoding）和特征嵌入（Feature embedding）两种方式构造特征向量，获得词典特征 $\boldsymbol{D}$。

表 3 基于双向最大匹配法构造词典特征

Table 3 Construction of dictionary features based on bidirectional maximum matching method

项目 Item	序列 Sequence											
	玉	米	品	种	。	京	科	9	6	8	是	以
双向最大匹配结果 BDMM result	玉	米	品	种	。	京科 968					是	以
词典特征 Dictionary features	O	O	O	O	O	B-*	I-*	I-*	I-*	E-*	O	O

注：*为实体类别标签；B-*代表实体的起始位置；I-*代表实体内部；E-*代表实体的结束位置；O 代表非实体。

Note:* indicates the label of entity, the B-* indicates the beginning of an entity, the I-* indicates the inside of an entity, the E-* indicates the end of an entity, the O indicates the character was outside an entity.

2.3 BiLSTM-CRF

NER 任务中，使用循环神经网络处理序列标注，难以学习到长距离的文本信息[24]。针对农业实体上下文关联较强的问题，基于 LSTM（Long Short-Term Memory，长短时记忆）网络，通过门限机制来捕捉序列长距离依赖信息，从两个方向处理输入序列，获取每个字完整的上下文信息。对于拼接向量 $\boldsymbol{W}$，BiLSTM 层输出与之对应的隐式状态序列 $\boldsymbol{H}=\{h_1,h_2,h_3,\cdots,h_n\}$，其中序列 $\boldsymbol{h}_i=[\overrightarrow{h_i};\overleftarrow{h_i}]$。

为了获取全局最优的标签序列，基于 CRF 层考虑相邻标签之间的关系，保证预测标签的合理性[25]。对于序列 $\boldsymbol{X}=\{x_1,x_2,x_3,\cdots,x_n\}$，对应的标签序列为 $\boldsymbol{y}=\{y_1,y_2,y_3,\ldots,y_n\}$，该标签序列的得分 S，如式（1）所示。

$$S(\boldsymbol{X},\boldsymbol{y})=\sum_{i=1}^{n}\boldsymbol{P}_{i,y_i}+\sum_{i=0}^{n}\boldsymbol{A}_{y_i,y_{i+1}} \quad (1)$$

式中 $\boldsymbol{A}_{y_i,y_{i+1}}$ 表示由标签 y_i 转移到标签 y_{i+1} 的概率，$\boldsymbol{P}_{i,y_i}$ 表示第 i 个词预测为第 y_i 个标签的分数。通过 Softmax 函数对 $S(X,y)$进行归一化，得到标签序列 y 的概率分布。最后，利用 Viterbi[26]算法得到序列 $X=\{x_1,x_2,x_3,\cdots,x_n\}$ 的最优标签序列 $\boldsymbol{y}^*$，如式（2）所示。

$$\boldsymbol{y}^*=\arg\max_{\tilde{y}\in Y_X}S(\boldsymbol{X},\tilde{\boldsymbol{y}}) \quad (2)$$

3　案例分析

3.1　试验数据集

为验证 BERT-Dic-BiLSTM-CRF 模型有效性，对自建的语料库按训练集、测试集、验证集为 6∶2∶2 比例进行划分，验证集用于验证模型训练及优化情况，三个数据集无重复交叉，因此测试集的试验结果可作为模型性能的评价指标。

3.2　试验设置

模型试验参数设置如下：利用 BERT-Base 模型，含有 12 个 Transformer 层，768 维隐藏层和 12 头多头注意力机制。最大序列长度采用 256，BiLSTM 隐藏层维度为 128，dropout 设置为 0.5，使用 Adam 优化算法[27]，训练学习率 0.001，批处理参数 32，迭代次数 100。通过准确率（Precision，P）、召回率（Recall，R）、F_1 值三个指标对模型进行评估[28]。

3.3　结果与分析

3.3.1　不同字级嵌入的性能对比

本文以 BiLSTM-CRF[29]为基准模型，采用 Word2Vec 和 BERT 两种字级嵌入进行对比试验，结果如表 4 所示。基于 BERT 的字级嵌入与 Word2Vec 方式相比，模型准确率提高了 5.5 个百分点，F_1 值提高了 5.25 个百分点。试验发现，Word2Vec 方式无法处理一词多义的问题，错误地把陕北民歌《东方红》识别为农机设备实体“东方红”，如图 4 所示。“东方红”为一词多义实体，在不同语境下，可表达为陕北民歌《东方红》，也可表达为农机实体“东方红”。BERT 的嵌入方式，通过多层 Transformer 编码器，能够学习更多的语义特征，获取丰富的字级特征信息，正确识别“东方红”这类实体，有效缓解一词多义的问题。

表 4　不同嵌入向量模型性能对比

Table 4　Performance comparison of model with different embedded vector

%

模型 Model	准确率 Precision	召回率 Recall	F_1 值 F_1 score
Word2Vec-BiLSTM-CRF	88.01	88.76	88.38
BERT-BiLSTM-CRF	93.51	93.76	93.63

…碰巧有一天，厂区内有人高唱陕北民歌《***东方红***》：“***东方红***，太阳升…，多好的名字呀，既歌颂共产党，又歌颂新中国，这个名字一经提出，马上获得一拖职工的认可，很快也得到了上级批准。…***东方红***全系列25至188马力大中功率轮式拖拉机…

a. Word2Vec识别结果

a. Recognition result of Word2Vec

…碰巧有一天，厂区内有人高唱陕北民歌《东方红》：“东方红，太阳升…，多好的名字呀，既歌颂共产党，又歌颂新中国，这个名字一经提出，马上获得一拖职工的认可，很快也得到了上级批准。…***东方红***全系列25至188马力大中功率轮式拖拉机…

b. BERT识别结果

b. Recognition result of BERT

注：字体“***东方红***”为不同嵌入方式的模型识别结果。

Note: The font “***东方红***” is the model recognition results of different embedding methods.

图 4　多义词识别结果

Fig.4　Result of polysemy recognition

3.3.2　不同词典特征的性能对比

基于 BERT-BiLSTM-CRF 模型，融入不同词典特征，在农业领域数据集上进行对比试验，结果如表 5 所示。融入词典特征的模型性能优于基准模型，其准确率分别提高了 0.24、0.44、0.3、0.84、1.33 个百分点。结果表明，词典特征的融入相较于单一字向量作为模型输入，能够有效补充序列语义信息，提升模型的识别准确度。

基于 N-gram 特征模板抽取词典特征，模板数量 m 为 8、10、12 时，模型的准确率 P 分别为 93.75%、93.95%、93.81%；召回率 R 分别为 92.86%、94.12%、92.95%。从试验结果看出，适当增加模板数量，模型抽取的词典特征信息越丰富，当模板数量为 10 时，模型性能达到最优。随着模板数量的增加，词典特征信息维度越大，训练周期越来越长，模型识别准确率降低。

相较于 N-gram 特征模板法，基于双向最大匹配法模型性能得到进一步提升，而将词典匹配结果进行特征嵌入的方式优于独热编码方式。分析得出，N-gram 方法忽略了实体内部结构，知识表示能力有限，对模型性能的提升低于双向最大匹配法。而采用特征嵌入的方式将匹配结果映射为低维的向量表示，能够获取更多的潜在信息，优于独热编码，模型准确率提高了 0.49 个百分点，F_1 值提高了 0.91 个百分点。

表 5　不同词典特征模型性能对比

Table 5　Performance comparison of model with different dictionary feature

%

模型 Model	准确率 Precision	召回率 Recall	F_1 值 F_1 score
BERT-BiLSTM-CRF	93.51	93.76	93.63
+N-gram feature(m=8)	93.75	92.86	93.30
+N-gram feature(m=10)	93.95	94.12	94.03
+N-gram feature(m=12)	93.81	92.95	93.38
+BDMM feature(one-hot)	94.35	93.89	94.12
+BDMM feature(embedding)	94.84	95.23	95.03

注：“+”以 BERT-BiLSTM-CRF 为基准模型，融入不同方式的词典特征；m 为 N-gram 模板数量。

Note: “+” represents integrate different dictionary features taking BERT-BiLSTM-CRF as the benchmark model; m is the number of N-gram templates.

3.3.3　不同模型的性能对比

为验证 BERT-Dic-BiLSTM-CRF 模型在农业语料的识别性能，分别与 BiLSTM-CRF 、CNN-BiLSTM-CRF[30]、BERT-BiLSTM-CRF 等主流模型进行了对比试验，试验结果如表 6 所示。BiLSTM-CRF 模型准确率为 88.01%、F_1 值为 88.38%。相较于 BiLSTM-CRF 模型，CNN-BiLSTM-CRF 模型通过 CNN 层抽取文本局部特征信息，与 Word2Vec 训练得到的字向量拼接作为 BiLSTM 层的输入，模型的准确率提高了 1.71 个百分点、F_1 值提高了 0.14 个百分点。但 CNN-BiLSTM-CRF 无法聚焦实体上下文信息，不能解决一词多义的问题。

引入 BERT 层的 BiLSTM-CRF 模型，通过 BERT 预训练模型充分提取字符级和序列上下文特征信息，更好地表征农业实体在不同语境下的语义表示，提升模型识别性能。相较于 CNN-BiLSTM-CRF 模型，识别准确率提高了 3.79 个百分点。本文提出的 BERT-Dic-BiLSTM-CRF 模型识别效果优于其他 3 种模型，识别准确率最高达到 94.84%、F_1 值为 95.03%。

表 6 不同模型性能对比

Table 6 Performance comparison of different models %

模型 Model	准确率 Precision	召回率 Recall	F_1 值 F_1 score
BiLSTM-CRF	88.01	88.76	88.38
CNN-BiLSTM-CRF	89.72	87.35	88.52
BERT-BiLSTM-CRF	93.51	93.76	93.63
BERT-Dic-BiLSTM-CRF	94.84	95.23	95.03

3.3.4 词典特征对模型性能影响

为验证模型引入词典特征可提高对罕见或未知实体识别准确率，统计测试集中实体在训练集出现的次数，将实体类型分为未知实体、罕见实体、高频实体，进行对比试验，试验结果如表 7 所示。其中，未知实体：测试集中实体从未出现在训练集；罕见实体：测试集中实体在训练集出现次数少于 5 次。高频实体：测试集中实体在训练集出现次数高于 5 次。

表 7 不同实体性能对比

Table 7 Performance comparison of different entity

实体类型 Entity type	模型 Model	准确率 Precision/%	召回率 Recall/%	F_1 值 F_1 score/%
未知实体	BERT-Δ	73.85	74.19	74.02
Unknown entity	BERT-Dic-Δ	80.29	80.05	80.17
罕见实体	BERT-Δ	85.61	85.75	85.68
Rare entity	BERT-Dic-Δ	91.54	90.18	90.85
高频实体	BERT-Δ	94.17	94.05	94.11
High-frequency entity	BERT-Dic-Δ	96.82	96.05	96.43

注：Δ 为 BiLSTM-CRF 模型，BERT-Δ 为 BERT-BiLSTM-CRF 模型。

Note: Δ represents BiLSTM-CRF model, BERT-Δrepresents BERT-BiLSTM-CRF model.

由表 7 可知，因高频实体在训练集出现的次数较多，模型可学习到较丰富的特征信息，BERT-BiLSTM-CRF 模型和 BERT-Dic-BiLSTM-CRF 模型识别准确率分别为 94.17%、96.82%。对于未知实体和罕见实体，实体在训练集出现的频率较低，BERT-BiLSTM-CRF 模型自身学习能力有限，识别准确率分别为 73.85%、85.61%。引入词典特征信息的 BERT-Dic-BiLSTM-CRF 模型，通过外部领域知识信息的辅助，模型的泛化能力得到提升，相较于 BERT-BiLSTM-CRF 模型，识别准确率分别提高了 6.44 个百分点、5.93 个百分点。进一步验证融入词典特征，能够提升模型对罕见或未知实体的识别准确率。

为验证词典规模对模型的影响，从自建的外部词典随机抽取 60%、70%、80%、90%的实体构造 4 个大小不同的词典进行对比试验，试验结果如图 5 所示。

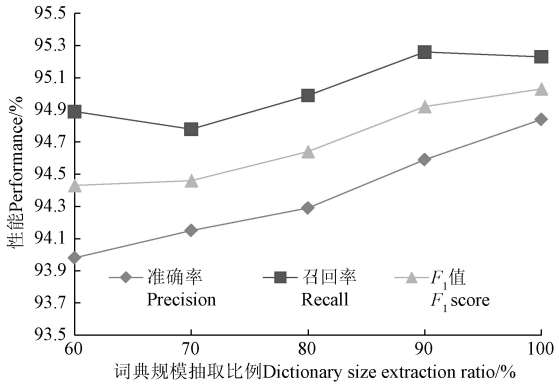


图 5 词典规模对模型性能的影响

Fig.5 The impact of the different dictionary size on model performance

由图 5 可知，随着词典规模的增大，模型学习到的特征信息更丰富，模型性能也随之得到提升，准确率达到 94.84%。

4 结 论

1）针对农业领域命名实体识别任务，提出 BERT-Dic-BiLSTM-CRF 模型，该模型结合字向量和词典特征能够处理一词多义问题，提升模型对罕见或者未知实体的识别准确率，模型准确率为 94.84%、F_1 值为 95.03%。

2）基于 BERT 预训练模型获得字级别的特征表示，能够聚焦实体上下文的语境，丰富农业文本的语义表示，缓解一词多义的问题，提升模型识别性能。

3）本文构建了农业领域外部词典，并设计了 2 种构建词典特征信息的方法。经试验验证，基于特征嵌入方式的双向最大匹配法能够获取丰富的词典特征，适用于农业 NER 任务。

随着智慧农业的不断发展，农业信息化决策服务更具体、更快捷。因此，下一步工作是增加农业病虫害病原、病害部位实体丰富语料库，并制定更规范、更完善的标注策略，在保证模型性能的基础上，对模型结构进行优化。

[参 考 文 献]

[1] 张善文，王振，王祖良. 结合知识图谱与双向长短时记忆网络的小麦条锈病预测[J]. 农业工程学报，2020，36(12)：172-178.
Zhang Shanwen, Wang Zhen, Wang Zuliang. Prediction of wheat stripe rust disease by combining knowledge graph and bidirectional long short term memory network[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2020, 36(12): 172-178. (in Chinese with English abstract)

[2] Zhang J, Shen D, Zhou G D, et al. Enhancing HMM-based biomedical named entity recognition by studying special phenomena[J]. Journal of Biomedical Informatics, 2004, 37(6): 411-422.

[3] Saha S K, Sarkar S, Mitra P. Feature selection techniques for maximum entropy based biomedical named entity recognition[J]. Journal of Biomedical Informatics, 2009, 42(5): 905-911.

[4] Sun C J, Guan Y, Wang X L, et al. Rich features based conditional random fields for biological named entities recognition[J]. Computers in Biology and Medicine, 2007, 37(9): 1327-1333.

[5] 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]. 农业机械学报，2017，48(增刊 1)：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]. Transactions of the Chinese Society for Agricultural Machinery, 2017, 48(Supp.1): 178-185. (in Chinese with English abstract)

[6] 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 Nian'e, 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 with English abstract)

[7] 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 with English abstract)

[8] Xu K, Yang Z G, Kang P P, et al. Document-level attention-based BiLSTM-CRF incorporating disease dictionary for disease named entity recognition[J]. Computers in Biology and Medicine, 2019, 108(22): 122-132.

[9] Maryam H, Leon W, Mariana N, et al. Deep learning with word embeddings improves biomedical named entity recognition[J]. Bioinformatics, 2017, 33(14): 37-48.

[10] Wang Q, Zhou Y M, Ruan T, et al. Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition[J]. Journal of Biomedical Informatics, 2019, 92: 103133.

[11] Wu Y H, Jiang M, Lei J B, et al. Named entity recognition in Chinese clinical text using deep neural network[J]. Studies in Health Technology and Informatics, 2015, 216: 624-628.

[12] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]// Advances in neural information processing systems. Lake Tahoe, US: MIT Press, 2013, 26: 3111-3119.

[13] 王欢，朱文球，吴岳忠，等. 基于数控机床设备故障领域的命名实体识别[J]. 工程科学学报，2020，42(4)：476-482.
Wang Huan, Zhu Wenqiu, Wu Yuezhong,et al. Named entity recognition based on equipment and fault field of CNC machine tools[J]. Chinese Journal of Engineering, 2020, 42(4): 476-482. (in Chinese with English abstract)

[14] 龚乐君，张知菲. 基于领域词典与 CRF 双层标注的中文电子病历实体识别[J]. 工程科学学报，2020，42(4)：469-475.
Gong Lejun, Zhang Zhifei. Clinical named entity recognition from Chinese electronic medical records using a double-layer annotation model combining a domain dictionary with CRF[J]. Chinese Journal of Engineering, 2020, 42(4): 469-475. (in Chinese with English abstract)

[15] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the Association for Computational Linguistics, Minneapolis, Minnesota, 2019: 4171-4186.

[16] 尹学振，赵慧，赵俊保，等. 多神经网络协作的军事领域命名实体识别[J]. 清华大学学报：自然科学版，2020，60(8)：648-655.
Yin Xuezhen, Zhao Hui, Zhao Junbao, et al. Multi-neural network collaboration for Chinese military named entity recognition[J]. Journal of Tsinghua University: Science and Technology, 2020, 60(8): 648-655. (in Chinese with English abstract)

[17] 李建，靖富营，刘军. 基于改进 BERT 算法的专利实体抽取研究-以石墨烯为例[J]. 电子科技大学学报，2020，49(6)：883-890.
Li Jian, Jing Fuying, Liu Jun. Study on patent entity extraction based on improved BERT algorithms-a case study of graphene[J]. Journal of University of Electronic Science and Technology of China, 2020, 49(6): 883-890. (in Chinese with English abstract)

[18] 陈剑，何涛，闻英友，等. 基于 BERT 模型的司法文书实体识别方法[J]. 东北大学学报：自然科学版，2020，41(10)：1382-1387
Chen Jian, He Tao, Wen Yingyou, et al. Entity recognition method for judicial documents based on BERT model[J]. Journal of Northeastern University Natural Science, 2020, 41(10): 1382-1387. (in Chinese with English abstract)

[19] 毛明毅，吴晨，钟义信，等. 加入自注意力机制的 BERT 命名实体识别模型[J]. 智能系统学报，2020，15(4)：772-779.
Mao Mingyi, Wu Chen, Zhong Yixin, et al. BERT named entity recognition model with self-attention mechanism[J]. CAAI Transactions on Intelligent Systems, 2020, 15(4): 772-779. (in Chinese with English abstract)

[20] Guo X C, Zhou H, Su J, et al. Chinese agricultural diseases and pests named entity recognition with multi-scale local context features and self-attention mechanism[J]. Computers and Electronics in Agriculture, 2020, 179(5): 105830.

[21] Wang J, Guo Y. Scrapy-based crawling and user-behavior characteristics analysis on taobao[C]// Proceedings of the 2012 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery. Sanya, China: IEEE Conference Publishing Services, 2012: 44-52.

[22] 吴俊，程垚，郝瀚，等. 基于 BERT 嵌入 BiLSTM-CRF 模型的中文专业术语抽取研究[J]. 情报学报，2020，39(4)：409-418.
Wu Jun, Cheng Yao, Hao Han, et al. Automatic extraction of Chinese terminology based on BERT embedding and BiLSTM-CRF model[J]. Journal of The China Society for Scientific and Technical Information, 2020, 39(4): 409-418. (in Chinese with English abstract)

[23] Gai R L, Gao F, Duan L M, et al. Bidirectional maximal matching word segmentation algorithm with rules[J]. Advanced Materials Research, 2014, 926-930: 3368-3372.

[24] 李明扬，孔芳. 融入自注意力机制的社交媒体命名实体识别[J]. 清华大学学报：自然科学版，2019，59(6)：461-467.
Li Mingyang, Kong Fang. Combined self-attention mechanism for named entity recognition in social media[J]. Journal of Tsinghua University: Science and Technology, 2019, 59(6): 461-467. (in Chinese with English abstract)

[25] 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 with English abstract)

[26] Strubell E, Verga P, Belanger D, et al. Fast and accurate

entity recognition with iterated dilated convolutions[C]// Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark: Association for Computational Linguistics, 2017: 2670-2680.
[27] Kingma D, Ba J. Adam: A method for stochastic optimization[C]// Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, 2015: 1-15.
[28] Li X, Zhang H, Zhou X H. Chinese clinical named entity recognition with variant neural structures based on BERT methods[J]. Journal of Biomedical Informatics, 2020, 107(5): 103422.
[29] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. Computer Science, 2015, 4(1): 1508-1519.
[30] Chiu J P C, Nichols E. Named entity recognition with bidirectional LSTM-CNNs[J]. Transactions of the Association for Computational Linguistics, 2016, 4: 357-370.

Recognition of the agricultural named entities with multi-feature fusion based on BERT

Zhao Pengfei[1], Zhao Chunjiang[1,2※], Wu Huarui[2,3,4], Wang Wei[2,3]

(1. *School 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*; 4. *Beijing Research Center of Intelligent Equipment for Agriculture*, *Beijing* 100097, *China*)

Abstract: Agricultural named entity recognition is a fundamental task for information extraction in the agricultural domain. Aiming at the problems of local context features、unable to solve the polysemy of the word、low recognition rate of rare entities in the process of entity recognition, the model combined with character level features and dictionary feature was proposed to automatically identify entities in the text，the character level features were obtained from the BERT(Bidirectional Encoder Representations from Transformers)model. Firstly, the BERT pre-trained language model was used to integrate the left and right contextual information to obtain the character level features, enhance the semantic representation of words, in order to alleviate the problem of polysemy; Secondly, we built an agricultural dictionary and introduced external dictionary information through the feature extraction strategy to improve the recognition accuracy of the model for rare or unknown entities. Among them, two feature extraction strategies were designed to capture the dictionary features, included N-gram feature template algorithm and bi-direction maximum matching algorithm. Then, the character level features and dictionary features were fused as the input of the next neural network layer. Finally, the fused feature information were encoded by the BiLSTM (Bi-directional Long-short Term Memory) neural network layer, obtained the sequence feature matrix, and the optimal text label sequence was obtained by CRF (Conditional Random Field). Based on the knowledge of domain experts, a labeling strategy of named entities in the agricultural field was proposed to solve the problem of fuzzy boundaries of agricultural named entities, in order to ensure the integrity of the entities. The experiments were carried out on the corpus of agricultural, which contained 5 295 labeled corpora and 5 categories of agricultural entities. The results showed that better overall performance was achieved in the corpus, where the recognition precision, recall, and F_1-score were 94.84%, 95.23%, and 95.03%, respectively. In terms of specific categories, due to the obvious boundary characteristics of crop diseases and pesticide, the model achieved higher recognition precision than the remaining three entities of agricultural, such as machinery, pests, and crop variety. Experimental comparison showed that for the effectiveness of the dictionary feature extraction strategy, the performance of the model based on the bi-direction maximum matching algorithm was better than the N-gram feature template algorithm. When the number of templates was 10, the performance of the model based on N-gram feature template was the best with the recognition precision of93.95%and F_1-score of 94.03%. The bi-directional maximum matching algorithm using feature embedding can obtain more potential information, which was better than one-hot encoding. The precision and F_1-score of the model were improved by 0.49 and 0.91 percentage points, respectively. Compared with the models based on BiLSTM-CRF, BERT-BiLSTM-CRF, the precision of the BERT-Dic-BiLSTM-CRF model proposed in this paper had obvious performance advantages with the highest recognition precision of 94.84%. Compared with the BERT-BiLSTM-CRF model, for the recognition performance of rare or unknown entities, the recognition precision of the BERT-Dic-BiLSTM-CRF model was improved by 5.93 and 6.44 percentage points, respectively. Further verifying that the integration of dictionary features into the model can improve the recognition accuracy of the model for such entities.

Keywords: agriculture; named entity recognition; text; BERT; dictionary feature; BiLSTM