doi:10.6041/j.issn.1000-1298.2024.06.027

基于连续提示注入与指针网络的农业病害命名实体识别

王春山[1,2]　张宸硕[1]　吴华瑞[2,3]　朱华吉[2,3]　缪祎晟[2,3]　张立杰[4,5]

(1.河北农业大学信息科学与技术学院,保定 071001;2.国家农业信息化工程技术研究中心,北京 100097;
3.农业农村部农业信息化技术重点实验室,北京 100097;4.河北农业大学机电工程学院,保定 071001;
5.河北省农业大数据重点实验室,保定 071001)

摘要:针对农业病害领域命名实体识别过程中存在的预训练语言模型利用不充分、外部知识注入利用率低、嵌套命名实体识别率低的问题,本文提出基于连续提示注入和指针网络的命名实体识别模型 CP-MRC(Continuous prompts for machine reading comprehension)。该模型引入 BERT(Bidirectional encoder representation from transformers)预训练模型,通过冻结 BERT 模型原有参数,保留其在预训练阶段获取到的文本表征能力;为了增强模型对领域数据的适用性,在每层 Transformer 中插入连续可训练提示向量;为提高嵌套命名实体识别的准确性,采用指针网络抽取实体序列。在自建农业病害数据集上开展了对比实验,该数据集包含 2 933 条文本语料,8 个实体类型,共 10 414 个实体。实验结果显示,CP-MRC 模型的精确率、召回率、F1 值达到 83.55%、81.4%、82.4%,优于其他模型;在病原、作物两类嵌套实体的识别率较其他模型 F1 值提升 3 个百分点和 13 个百分点,嵌套实体识别率明显提升。本文提出的模型仅采用少量可训练参数仍然具备良好识别性能,为较大规模预训练模型在信息抽取任务上的应用提供了思路。

关键词:农业病害;命名实体识别;连续提示;指针网络;嵌套实体;预训练语言模型

中图分类号:S126;TP182　　**文献标识码**:A　　**文章编号**:1000-1298(2024)06-0254-08　　**OSID**:

Named Entity Recognition of Agricultural Disease Based on Continuous Prompts Injection and Pointer Network

WANG Chunshan[1,2]　ZHANG Chenshuo[1]　WU Huarui[2,3]　ZHU Huaji[2,3]　MIAO Yisheng[2,3]　ZHANG Lijie[4,5]

(1. *College of Information Science and Technology*, *Hebei Agricultural University*, *Baoding* 071001, *China*
2. *National Engineering Research Center for Information Technology in Agriculture*, *Beijing* 100097, *China*
3. *Key Laboratory of Agricultural Information Technology*, *Ministry of Agriculture and Rural Areas*, *Beijing* 100097, *China*
4. *College of Mechanical and Electrical Engineering*, *Hebei Agricultural University*, *Baoding* 071001, *China*
5. *Hebei Province Key Laboratory of Agricultural Big Data*, *Baoding* 071001, *China*)

Abstract: In response to the problems of insufficient utilization of pretrained language models, low utilization of external knowledge injection, and low recognition rate of nested named entities in the process of named entity recognition in the field of agricultural diseases, a named entity recognition model continuous prompts for machine reading comprehension (CP-MRC) was proposed based on continuous prompt injection and pointer network. This model introduced the bidirectional encoder representation from transformers (BERT) pretraining model, which freezed the original parameters of the BERT model and retained its text representation ability obtained during the pretraining stage. To enhance the applicability of the model to domain data, continuous trainable hint vectors were inserted into each layer of Transformer. To improve the accuracy of nested named entity recognition, a pointer network was used to extract entity sequences. A comparative experiment was conducted on a self built agricultural disease dataset, which included 2 933 text corpora, 8 entity types, and a total of 10 414 entities. The experimental results showed that the accuracy, recall, and F1 values of the CP-MRC model reached

收稿日期:2023-10-31　修回日期:2023-12-27

基金项目:国家科技创新 2030-"新一代人工智能"重大项目(2021ZD0113604)、财政部和农业农村部:国家现代农业产业技术体系项目(CARS-23-D07)和河北省自然科学基金项目(F2022204004)

作者简介:王春山(1978—),男,副教授,博士,主要从事机器学习、知识图谱和智慧农业研究,E-mail:chunshan9701@163.com

通信作者:吴华瑞(1975—),男,研究员,主要从事农业大数据智能服务研究,E-mail:wuhr@nercita.org.cn

83.55%, 81.4%, and 82.4%, which was superior to other models. The recognition rate of nested entities in pathogens and crops was increased by 3 percentage points and 13 percentage points in F1 value compared with that of others, and the recognition rate of nested entities was significantly improved. The model still had good recognition performance with only a small number of trainable parameters, providing ideas for the application of large-scale pretrained models in information extraction tasks.

Key words: agricultural diseases; named entity recognition; continuous prompts; pointer network; nested entities; pretrained language model

0　引言

农业病害知识具有多源、多样性、碎片化等特点。通常这些知识分散在各种文献、专家经验、数据库中,形成了一个庞大而复杂的信息网络。由于这些知识片段往往难以获取和利用,限制了病害领域知识的研究与应用。知识图谱是一种用于表示和组织知识的图状结构[1],能够表达不同实体之间的复杂关系,实现分散知识的高效整合,为农业生产与服务提供智能化的数据与决策支撑。

命名实体识别(Named entity recognition, NER)是知识图谱构建过程的重要环节。条件随机场是解决命名实体识别问题最常用的机器学习方法。文献[2]采用条件随机场模型(Conditional random fields, CRF),通过选取词汇、词性和语法作为特征模板来识别病害、品种、农药等命名实体。为了实现特征的自动学习,更多的研究者围绕基于深度学习的 NER 方法展开研究[3-4]。文献[5]提出用于命名实体识别的 BiLSTM-CRF 模型,采用双向长短期记忆(Bi-directional long short-term memory, BiLSTM)结合条件随机场取得了很好的效果。在预处理阶段,研究者使用 Word2vec[6-7]工具获取字向量接入模型的输入端。然而通过这种方法获取到的字向量是静态的,表征往往比较单一,无法解决一词多义的问题[8]。为了更好地对句子进行表征,预训练语言模型 BERT[9-10]被应用到 NER 任务中,并取得了显著效果,如文献[11]提出的 BERT-BiLSTM-CRF 实体识别方法,对法律案件中的案件实体进行识别,取得了良好的识别效果。

为了进一步提高 NER 识别的准确性,研究者围绕在模型中注入先验知识展开了一系列的研究工作。文献[12]通过自建农业领域词典,引入双向最大匹配策略,获取分布式词典特征,提高了模型对罕见或未知实体的识别效果。文献[13]提出实体级遮蔽策略,使模型在预训练阶段学习到实体级的特征信息,从而提高下游 NER 任务的识别效果。考虑到实体通常出现在文档级别的文本中,为了充分利用文档级的上下文信息,文献[14]提出基于机器阅读理解(Machine reading comprehension, MRC)的中文命名实体识别方法,将先验知识融入问题模板,使模型充分利用文档级的上下文语义信息以及先验知识,从而改善 NER 任务的识别效果。为了充分利用标签的知识信息,文献[15]提出了一种标签知识整合方式,让模型高效利用标签先验知识,缓解了 MRC 模型知识利用率低的问题。文献[16]在 MRC 模型的基础上提出先验信息增强的机器阅读理解式命名实体识别方法,通过引入二元字嵌入、词典信息增强的方法,增强了 MRC 模型的特征学习能力。

尽管命名实体识别方面取得了显著进展,但在实体识别过程中仍然存在预训练语言模型利用不充分、外部知识注入利用率低、嵌套命名实体识别率低等问题。

针对以上问题,本文提出基于连续提示注入 BERT 和指针网络的农业病害命名实体识别模型 CP-MRC。首先引入预训练语言模型 BERT,冻结其通过预训练得到的参数以保留其在预训练阶段获取的能力。将连续可训练的提示向量插入到 BERT 中用于适应下游 NER 任务。通过指针网络实现对普通实体、嵌套实体的跨度抽取,最终得到实体序列。

1　数据采集与预处理

1.1　数据采集

在农业病害领域由于缺少公开的中文数据集,本研究使用分布式爬虫框架 Scrapy 从百度百科、惠农网等农技网站上爬取农业病害知识文本。从《蔬菜病虫害防治手册》、《中国蔬菜病虫害原色图谱续集》等专业书籍上通过手工摘录获取知识文本数据作为补充,建立农业病害知识文本数据集。

1.2　数据标注

本研究采用的数据标注格式如图 1 所示。首先对获取到的知识文本数据进行清洗,去除不必要和重复的数据,结合领域专家的意见构建农业病害 NER 数据集。

本文构建的农业病害 NER 数据集共包含 2 932 条文本语料,约 10 414 个实体。实体类型 8 种:病害名称(Disease)、作物名称(Crop)、病害特征(Feature)、病原(Pathogeny)、发病地区(Region)、发

植株轻微发病时或出现轻微根肿病症状时,可用10%氰霜唑悬浮剂1500到2000倍液灌根,移栽成活后用2000倍氟啶胺灌根处理,日本研究表明,用十字花科作物根部内生菌可防治根肿病,根瘤减少率为52到97%

{
"entity_label":"Feature",
"start_position":[42,59,77,89],
"end_position":[42,59,78,89],
"span_position":["42;42","59;59","77;78","89;89"]
},
{
"entity_label":"Disease",
"start_position":[12,85],
"end_position":[14,87],
"span_position":["12;14","85;87"]
}

图 1 语料标注示例

Fig. 1 Annotated sample

病因素(Factor)、防治方法(Method)、发病阶段(Period)。

2 命名实体识别模型

本研究提出的 CP－MRC 模型整体结构如图 2 所示。将随机初始化的伪标签 pseudo tokens 经过嵌入层 Embedding 编码为连续提示向量,插入到冻结参数的预训练 BERT 模型的每层 Transformer 中,输入文本经过 BERT 得到对应的特征表示,然后输入到指针网络中学习标注约束、建立实体首尾位置的匹配关系,最终得到实体标签以及实体边界。

2.1 BERT 模型

BERT 模型是一个在大量无标签的数据集中预训练得到的深层模型,采用多个深层双向 Transformer 编码器(Encoder)堆叠作为主要结构,如图 3 所示。在预训练阶段,对输入文本进行双向编码,充分挖掘文本中的语义信息,因而具备了很强的语义信息提取能力。将输入文本 $X=\{x_1,x_2,\cdots,x_n\}$ 经过嵌入层(Embedding layer)得到的对应文本序列编码表示 $E=\{E_1,E_2,\cdots,E_n\}$ 作为 BERT 的输入。其中每个字符的编码表示 E_i 由词嵌入、句嵌入、位置嵌入叠加得到,如图 4 所示。词嵌入是对句子的每个词进行编码,得到对应词的向量表征,相比于 Word2vec 能够获得更加丰富的语义信息,可以容易解决一词多义问题。句嵌入是对输入句子进行编码,用来区分不同句子。位置嵌入实现了对句子中同一个词所处位置的表征,使模型具备了对时序序列处理的能力。

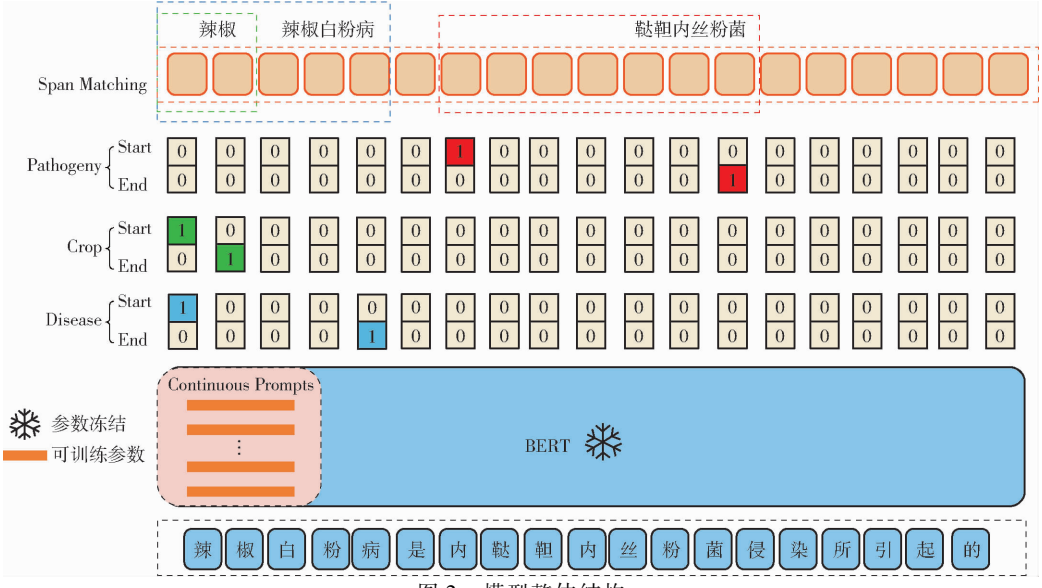


图 2 模型整体结构

Fig. 2 Overall architecture of CP－MRC

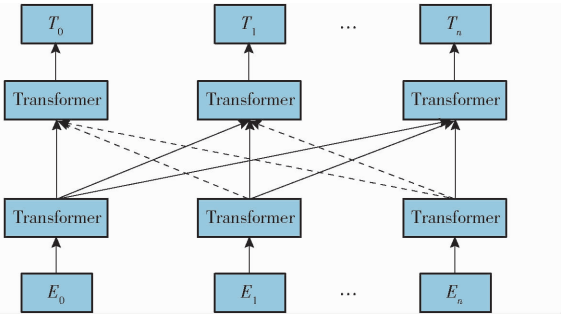


图 3 BERT 模型结构

Fig. 3 Architecture of BERT

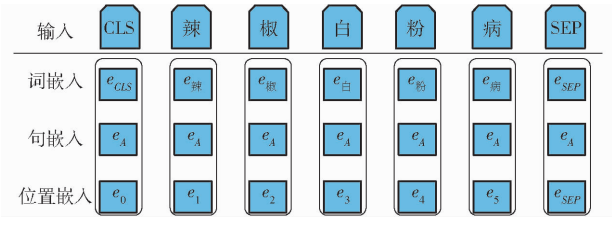


图 4 BERT 模型的输入

Fig. 4 Input of BERT

在预训练过程中,采用掩码语言模型(Masked

language model，MLM）和下一句预测（Next sentence prediction，NSP）2 个任务进行联合训练，使模型具备了良好的捕获句子内以及句子间依赖关系的能力。MLM 是指在输入句子中按照一定的比例随机选中一些单词，将这些单词用[MASK]替换，用分类模型预测[MASK]是什么词。在预训练过程中，随机遮蔽掉 15% 的单词，被遮蔽的单词中有 80% 的单词被替换成[MASK]，10% 的单词被其他单词替换，其余 10% 的单词保持不变。这种遮蔽策略可以在不影响模型语言理解能力的情况下学习到每个输入单词的分布式上下文表示，因而具备很强的文本表征能力。NSP 任务是指训练模型使其具备对句子关系理解的能力。在预训练过程中，每个数据样本为一个句子对“[CLS]A[SEP]B[SEP]”，训练目标为让模型判断句子 B 是否是句子 A 的下一个句子。

2.2　连续提示

2.2.1　提示模板

预训练语言模型（Pretrained language model，PLM）经过在海量数据预训练蕴含了丰富的知识[17]，然而由于预训练任务与下游任务的不同，针对下游任务，往往需要对预训练语言模型进行微调以适应下游任务的要求。为了充分利用预训练语言模型，缩小预训练任务与下游任务之间的差距，提示学习[18]（Prompt learning）被提出。通过构造提示模板与 PLM 相结合以实现预训练任务与下游任务的统一。提示模板通常有两种形式：由自然语言构成的模板和连续提示向量。

连续提示向量是将模板的构造转换为连续参数的优化问题。保持 PLM 参数冻结，将连续向量插入到 PLM 中[19-20]，通过不断优化连续提示向量，使 PLM 更好地完成下游任务。优化过程为

$$\max_{\sigma} \ln \boldsymbol{P}(y \mid x;\theta;\boldsymbol{\sigma}) = \max_{\sigma} \sum \ln \boldsymbol{P}(y_i \mid h;\theta;\boldsymbol{\sigma}) \tag{1}$$

式中　x——模型预测结果

θ——PLM 的原始参数

$\boldsymbol{\sigma}$——插入的连续向量

h——PLM 中插入连续向量后的模型预测

2.2.2　连续提示注入

本研究中将随机初始化的伪标签（Pseudo token）经过嵌入层得到提示向量矩阵 $\boldsymbol{P} \in \mathbf{R}^{1\times d}$，其中 d 为 PLM 的隐藏层维度。在训练过程中通过训练嵌入层，不断优化连续提示向量，以实现提示向量在嵌入空间中的搜索。在 BERT 中所有的 Transformer 块中，插入连续提示向量，注入过程如图 5 所示。

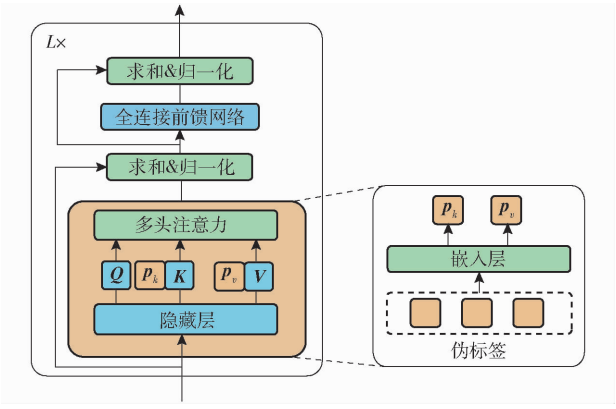


图 5　连续提示注入

Fig. 5　Continuous prompts injection

Transformer 的输入向量经过隐藏层（Hidden states）后通过 3 个不同的线性变换得到索引矩阵 $\boldsymbol{Q}$、目标词矩阵 $\boldsymbol{K}$、词权重矩阵 $\boldsymbol{V}$。

注意力机制[21]（Attention mechanism）是通过将有限的注意力权重分配到不同的特征向量，对贡献度较高的向量分配更高的权重。将提示向量 $\boldsymbol{P}^l = [\boldsymbol{p}_k^l;\boldsymbol{p}_v^l]$ 与 $\boldsymbol{K}^l$、$\boldsymbol{V}^l$ 拼接，计算该层注意力分数方法为

$$Attention^l = \mathrm{softmax}\left(\frac{\boldsymbol{Q}^l[\boldsymbol{p}_k^l;\boldsymbol{K}^l]^{\mathrm{T}}}{\sqrt{d}}\right)[\boldsymbol{p}_v^l;\boldsymbol{V}^l] \tag{2}$$

式中　$\boldsymbol{p}_k^l$、$\boldsymbol{p}_v^l$——第 l 层与 $\boldsymbol{K}$、$\boldsymbol{V}$ 矩阵拼接的连续提示向量

注意力分数捕获了由于添加的提示向量所产生的变化，可训练的提示向量不断引导模型逐渐关注到对实体贡献度较大的关键信息。

注意力分数反映了对输入向量关注的局部信息，为了使模型学习到不同空间的语义特征，采用多头注意力将得到的注意力分数进行拼接和线性变换，最终得到 Multi－Head Attention 层的输出，具体计算方法为

$$\mathrm{MultiHead}(\boldsymbol{Q}^l,\boldsymbol{K}^l,\boldsymbol{V}^l) = \mathrm{Concat}(\boldsymbol{head}_1,\boldsymbol{head}_2,\cdots,\boldsymbol{head}_h)\boldsymbol{W}^O \tag{3}$$

其中　$$\boldsymbol{head}_i = \mathrm{Attention}(\boldsymbol{Q}^l,\boldsymbol{K}^l,\boldsymbol{V}^l) \tag{4}$$

式中　$\boldsymbol{W}^O$——线性变换参数矩阵

为了避免模型退化，对编码后的结果进行残差与归一化处理，得到 Transformer 的输出，用于后续任务。

2.3　指针网络

指针网络由 3 部分构成：实体开始、结束位置分类器、实体跨度匹配分类器。其中实体开始位置与结束位置预测均使用线性二分类器，分别对每个 token 预测其作为实体开始位置概率和作为实体结束位置概率，计算公式为

$$\boldsymbol{P}_{\mathrm{start}} = \mathrm{softmax}(\boldsymbol{E}_{\mathrm{encoder}}\boldsymbol{T}_{\mathrm{start}}) \tag{5}$$

$$\boldsymbol{P}_{\mathrm{end}} = \mathrm{softmax}(\boldsymbol{E}_{\mathrm{encoder}}\boldsymbol{T}_{\mathrm{end}}) \tag{6}$$

式中　$\boldsymbol{E}_{\mathrm{encoder}}$——上层模型的编码结果

$\boldsymbol{T}_{\mathrm{start}}$、$\boldsymbol{T}_{\mathrm{end}}$——线性变换参数矩阵

$\boldsymbol{P}_{\text{start}}$、$\boldsymbol{P}_{\text{end}}$——实体开始、结束位置概率

由于输入文本中可能存在多个相同类别的实体,因此在上述预测模块中可能预测出了多个开始、结束位置,如果仅仅按照先后顺序匹配可能会导致错误传播。这里将得到的 $\boldsymbol{P}_{\text{start}}$、$\boldsymbol{P}_{\text{end}}$ 的每行使用 argmax 得到实体开始、结束位置的预测结果矩阵 $\boldsymbol{E}_{\text{start}}$、$\boldsymbol{E}_{\text{end}}$,计算公式为

$$\boldsymbol{E}_{\text{start}} = \text{argmax}(\boldsymbol{P}_{\text{start}}) \tag{7}$$

$$\boldsymbol{E}_{\text{end}} = \text{argmax}(\boldsymbol{P}_{\text{end}}) \tag{8}$$

$$\boldsymbol{P}_{\text{start,end}} = \text{sigmoid}(\text{concat}(\boldsymbol{E}_{\text{start}}, \boldsymbol{E}_{\text{end}})\boldsymbol{T}_{\text{span}}) \tag{9}$$

式中 $\boldsymbol{T}_{\text{span}}$——线性变换参数矩阵

$\boldsymbol{P}_{\text{start,end}}$——实体跨度匹配概率

sigmoid( )——激活函数

concat( )——矩阵拼接

$\boldsymbol{E}_{\text{start}}$、$\boldsymbol{E}_{\text{end}}$均为长度为 n 的 0/1 矩阵。在预测结果矩阵中,采用实体开始位置的索引和结束位置的索引确定实体位置。使用线性二分类器预测实体跨度匹配概率。

2.4 损失函数

在训练阶段,使用交叉熵(Cross entropy)损失计算模型预测值与真实值之间的损失。实体开始、结束位置的损失函数定义为

$$L_{\text{start}} = \text{CE}(\boldsymbol{P}_{\text{start}}, Y_{\text{start}}) \tag{10}$$

$$L_{\text{end}} = \text{CE}(\boldsymbol{P}_{\text{end}}, Y_{\text{end}}) \tag{11}$$

式中 Y_{start}、Y_{end}——开始、结束位置标签

CE( )——交叉熵损失函数

实体开始位置与结束位置的匹配预测,使用的损失函数定义为

$$L_{\text{start, end}} = \text{CE}(\boldsymbol{P}_{\text{start,end}}, Y_{\text{start,end}}) \tag{12}$$

式中 $Y_{\text{start,end}}$——实体跨度标签

总体损失定义为

$$L_{\text{total}} = \alpha(L_{\text{start}} + L_{\text{end}}) + \beta L_{\text{start,end}} \tag{13}$$

式中 α、β——超参数

2.5 训练参数配置与评价指标

本研究与对照组试验均在 Ubuntu 20.04 环境下进行。处理器为 Intel core i9 - 10920X,内存容量为 64 GB,显卡为 GeForce RTX 3090,采用 Python 编程语言,深度学习框架 Pytorch 1.10.1。PLM 选用 RoBERTa - base 和 RoBERTa - large 中文模型[22],具体参数如表 1、2 所示。

采用精确率 P、召回率 R 和 F1 值作为不同模型性能的评判指标。

3 实验结果与分析

3.1 不同模型性能比较

为验证本文提出的模型对中文命名实体识别的有效性,本模型与其他基准模型的实验结果对比如表 3 所示。BiLSTM - CRF 模型不采用预训练模型,其余均采用 RoBERTa - base 预训练模型,参数设置如表 1 所示。

表 1 预训练模型参数设置

Tab.1 Pretrained model parameters setting

预训练模型	RoBERTa - base	RoBERTa - large
Transformer 层数	12	24
隐藏层维度	768	1 024
可接收最大序列长度	256	512
参数量	1.1×10^8	3.4×10^8

表 2 训练参数设置

Tab.2 Training parameters setting

参数	数值
批尺寸	8
输入最大长度	256
提示向量长度	16
迭代次数	50
学习率	0.007
权重衰减量	0.001
失活率	0.1

表 3 命名实体识别总体结果

Tab.3 Overall results of NER %

模型	精确率	召回率	F1 值
BilSTM - CRF	71.93	73.66	72.79
BERT	84.17	79.01	81.51
BERT - BiLSTM	83.92	79.01	81.39
BERT - CRF	82.63	79.28	80.92
BERT - BiLSTM - CRF	84.36	78.74	81.53
CP - MRC	83.55	81.40	82.40

由表 3 可知,本文提出的 CP - MRC 模型精确率、召回率、F1 值分别达到 83.55%、81.40%、82.40%,对比无预训练的模型 BiLSTM - CRF,精确率提升 11.62 个百分点,召回率提升 7.74 个百分点,F1 值提升 9.61 个百分点,各方面指标均有较大提升。对比性能较好的 BERT - BiLSTM - CRF 模型,召回率、F1 值分别提升 2.66、0.87 个百分点,模型综合性能表现最优。

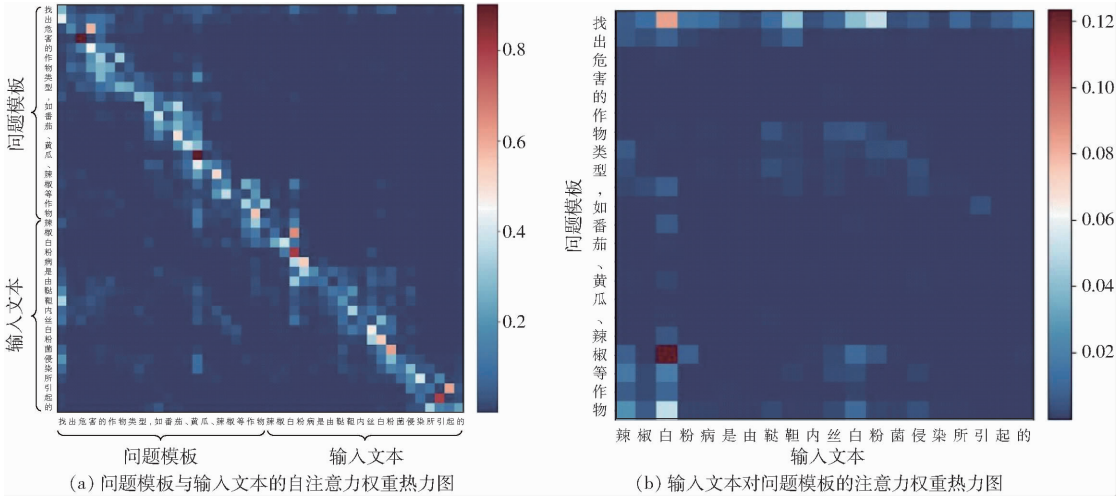
3.2 不同提示模板注入性能比较

为了验证连续提示注入的有效性,本文与采用自然语言构造提示模板的机器阅读理解模型 MRC 做了对比实验。MRC 模型是通过为每类实体构造问题模板,将先验知识融入模板中,然后将问题模板与输入文本共同作为 BERT 的输入。本质上是将 NER 任务建模成 BERT 的 NSP 任务,通过建立问题模板与输出文本之间的联系,同时融入先验知识以完成下游 NER 任务。然而这种方式存在两方面问题:

(1)问题模板由人工定义,模板构造的质量会影响最终性能。

(2)问题模板仅在输入层注入,随着模型层数不断加深,问题模板对模型的影响逐渐降低,模型对问题模板中的先验知识利用率低。

图 6 为使用 MRC 模型对农业病害 NER 数据实例的注意力权重热力图,问题模板为“找出危害的作物类型,如番茄、黄瓜、辣椒等作物”,输入文本为“辣椒白粉病是由鞑靼内丝白粉菌侵染所引起的”,从问题模板的内容上看,在作物类型实体构造的问题模板中,人工加入了部分先验知识作为提示:番茄、黄瓜、辣椒等均可作为作物类型实体。图 6a 为问题模板与输入文本自注意力的对齐情况,热力图展示出明显的对角矩阵的形式,表明每个位置对自身的注意力最高,而与其他位置之间的注意力较低。图 6b 为输入文本对问题模板的注意力权重可视化结果,从热力图中可以看出输入文本与问题模板之间的交互关系较弱。由于问题模板内容上相对固定,针对不同的输入文本,问题模板对输入文本的贡献往往比较有限。



(a) 问题模板与输入文本的自注意力权重热力图　　(b) 输入文本对问题模板的注意力权重热力图

图 6　注意力权重可视化图

Fig. 6　Attention matrix visualization

与全参数微调的 MRC 模型相比,使用 RoBERTa - base 预训练模型,分别采用 2 种方式构建问题模板,其中 MRC - label 模型在构建问题模板时只添加实体标签,不额外注入其他信息,构建方式如表 4 中引入标签信息所示;MRC - label - desc 模型是在构建问题模板时不仅注入实体标签信息,而且还添加了对部分实体类别本身的表述,帮助模型理解实体类型的含义,从而更有利于在给定的文本中找出对应的实体,比如为作物 Crop 实体构造的问题模板:找出危害的作物类型,如番茄、黄瓜、辣椒等作物;为防治方法 Method 实体构造问题模板:找出防治方法,包括播种前的准备工作、栽培方式、护理工作、防控传播策略等。完整问题模板构建方式如表 4 中引入标签及描述所示。从表 5 可以看出,添加了先验知识后,MRC 模型的性能的确有所提升,可见原有的 MRC 模型最终的性能仍然受问题模板构造的影响。相比之下,连续提示注入的方式可以在训练的过程中自动搜索并插入连续提示模板,在不需外部知识注入的情况下仍具备良好的性能。

采用较大规模的预训练语言模型 RoBERTa - large 模型,由表 5 可以看出,模型的性能进一步提升,这也验证了连续提示向量能够实现将预训练模型对下游任务的适配,与全参数微调的 MRC 模型相比,本文提出的模型在大幅降低可训练参数的情况下仍具备良好性能。

表 4　MRC 问题模板

Tab. 4　Question template for MRC

实体类型	引入标签信息(query_label)	引入标签及描述(query_label_desc)
Pathogeny	找出病原	找出病原,包括病毒、真菌、病菌等
Crop	找出危害的作物类型	找出危害的作物类型,如番茄、黄瓜、白菜等作物
Factor	找出致病因素	找出致病因素,包括温度、湿度、天气、气候条件等
Period	找出发病阶段	找出发病阶段,如育苗期、幼苗期、开花期、结果期等
Disease	找出病害名称	找出病害名称
Feature	找出病害特征和危害植物部位	找出病害特征和危害植物部位
Method	找出防治方法	找出防治方法,包括播种前的准备工作、栽培方式、护理工作、防控传播策略等
Region	找出发病地区名称	找出发病地区名称,如南方、长江流域等

表 5 MRC 模型命名实体识别总体结果

Tab. 5 Overall results of NER for MRC %

预训练模型	模型	精确率	召回率	F1 值
RoBERTa - base	MRC - label	80.22	81.69	80.95
	MRC - label - desc	82.08	80.34	81.20
	本文模型	83.55	81.40	82.40
RoBERTa - large	MRC - label	82.69	82.85	82.77
	MRC - label - desc	83.09	83.81	83.45
	本文模型	83.07	83.30	83.19

3.3 嵌套命名实体识别

从模型预测标签的方式上看,在实体预测模块中,本文提出的模型是对每一个 token 分别预测实体的开始标签和结束标签,而基准模型是对每个 token 预测一个具体的标签,由于数据集中存在大量嵌套实体,有的 token 可能会对应多个标签,因此不能较好地解决实体嵌套问题。由于本文提出的 CP - MRC模型采用指针网络对每个 token 分别预测其作为某类实体是否为开始、结束位置,可以很好地解决实体嵌套问题。如图 7 所示,病原、作物两类嵌套实体的识别率较其他模型 F1 值提升 3 个百分点和 13 个百分点,本文提出的模型在解决嵌套命名实体问题上具备良好的性能表现。

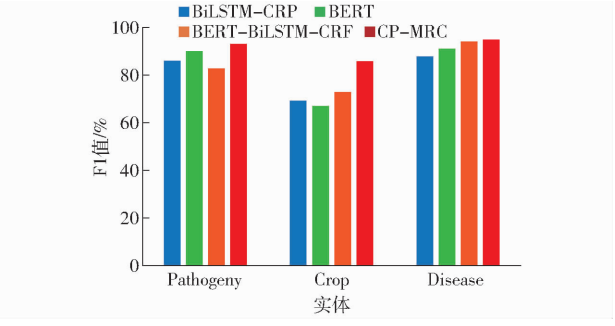


图 7 嵌套实体识别结果

Fig. 7 Nested entity recognition results

3.4 可训练参数量比较

从模型训练方式上看,MRC 模型采用预训练模型全参数微调的方式,而本文提出的 CP - MRC 模型在训练过程中冻结了预训练模型原有参数,模型采用不同规模预训练语言模型性能如图 8 所示。由图 8 可以看出,相较于全参数微调的 MRC 模型,本文提出的模型仍具备良好的性能。另外从图 9 可以看出,与 MRC 模型相比,本文提出的 CP - MRC 模型仅采用 1% ~2% 的可训练参数,实现了模型性能与可训练参数量之间的平衡。

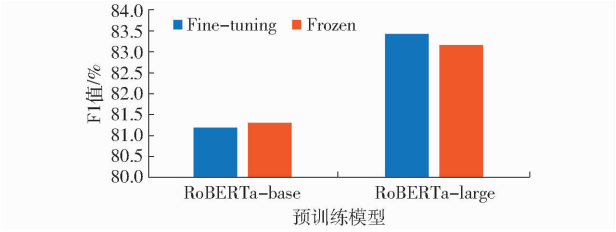


图 8 MRC 与 CP - MRC 对比

Fig. 8 Comparison of MRC and CP - MRC

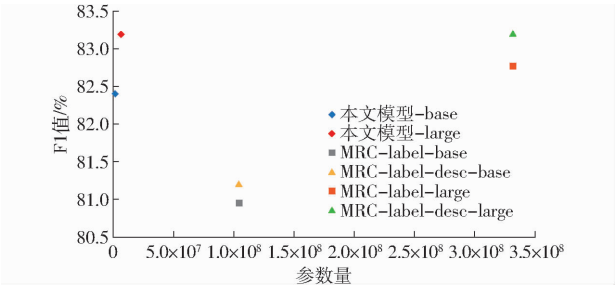


图 9 F1 值与可训练参数量关系

Fig. 9 Relationship between model trainable parameters and F1-score

4 结论

(1)针对农业病害命名实体识别过程中存在的预训练语言模型利用不充分、外部知识注入利用率低、嵌套实体识别率低的问题,提出基于连续提示和指针网络的命名实体识别模型 CP - MRC,提升了模型识别性能,同时模型可训练参数量大幅下降。

(2)采用连续提示注入 BERT 的方法实现了下游任务的适配,通过与采用自然语言模板的提示学习模型对比,证明将连续提示注入的方法适用于农业病害 NER 任务,模型精确率、召回率、F1 值分别达到 83.55%、81.4%、82.4%。

(3)采用指针网络实现对嵌套实体的抽取,通过与其他模型实验对比,证明模型对嵌套实体的识别具备一定优势。

参 考 文 献

[1] PUJARA J,MIAO H,GETOOR L,et al. Knowledge graph identification[C]//International Semantic Web Conference,2013:542 - 557.

[2] 张剑,吴青,羊昕旖, 等. 基于条件随机场的农业命名实体识别[J]. 计算机与现代化,2018(1):123 - 126.
ZHANG Jian,WU Qing,YANG Xinyi,et al. Chinese agricultural named entity recognition based on conditional random fields [J]. Computer and Modernization,2018(1):123 - 126. (in Chinese)

[3] 蒲攀,张越,刘勇,等. Transformer 优化及其在苹果病虫命名实体识别中的应用[J]. 农业机械学报,2023,54(6):264 - 271.
PU Pan,ZHANG Yue,LIU Yong,et al. Transformer optimization and application in named entity recognition of apple diseases and pests[J]. Transactions of the Chinese Society for Agricultural Machinery,2023,54(6):264 - 271. (in Chinese)

[4] 刘新亮,张梦琪,谷情,等. 基于 BERT - CRF 模型的生鲜蛋供应链命名实体识别[J]. 农业机械学报,2021,52(增刊):

519 – 525.
LIU Xinliang, ZHANG Mengqi, GU Qing, et al. Named entity recognition of fresh egg supply chain based on BERT – CRF architecture[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(Supp.): 519 – 525. (in Chinese)
[5] HUANG Z, XU W, YU K. Bidirectional LSTM – CRF models for sequence tagging[J]. arXiv Preprint, arXiv: 1508.01991, 2015.
[6] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]// Advances in Neural Information Processing Systems, 2013: 3111 – 3119.
[7] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv Preprint, arXiv: 1301.3781, 2013.
[8] 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.
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 CSAE, 2022, 38(3): 112 – 118. (in Chinese)
[9] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [J]. arXiv Preprint, arXiv: 1810.04805, 2018.
[10] CUI Y, CHE W, LIU T, et al. Pre-training with whole word masking for Chinese BERT[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504 – 3514.
[11] 郭知鑫,邓小龙. 基于 BERT – BiLSTM – CRF 的法律案件实体智能识别方法[J]. 北京邮电大学学报,2021,44(4):129 – 134.
GUO Zhixin, DENG Xiaolong. Intelligent identification method of legal case entity based on BERT – BiLSTM – CRF[J]. Journal of Beijing University of Posts and Telecommunications, 2021, 44(4): 129 – 134. (in Chinese)
[12] 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)
[13] 韦紫君,宋玲,胡小春,等. 基于实体级遮蔽 BERT 与 BiLSTM – CRF 的农业命名实体识别[J]. 农业工程学报,2022, 38(15):195 – 203.
WEI Zijun, SONG Ling, HU Xiaochun, et al. Named entity recognition of agricultural based entity-level masking BERT and BiLSTM – CRF[J]. Transactions of the CSAE, 2022, 38(15): 195 – 203. (in Chinese)
[14] 刘奕洋,余正涛,高盛祥,等. 基于机器阅读理解的中文命名实体识别方法[J]. 模式识别与人工智能,2020,33(7): 653 – 659.
LIU Yiyang, YU Zhengtao, GAO Shengxiang, et al. Chinese named entity recognition method based on machine reading comprehension[J]. Pattern Recognition and Artificial Intelligence, 2020, 33(7): 653 – 659. (in Chinese)
[15] YANG P, CONG X, SUN Z, et al. Enhanced language representation with label knowledge for span extraction[J]. arXiv Preprint, arXiv: 2111.00884, 2021.
[16] LI Ren, MO Tianjin, YANG Jianxi, et al. Bridge inspection named entity recognition via BERT and lexicon augmented machine reading comprehension neural model[J]. Advanced Engineering Informatics, 2021, 50: 101416.
[17] PETRONI F, ROCKTASCHEL T, LEWIS P, et al. Language models as knowledge bases? [J]. arXiv Preprint, arXiv: 1909.01066, 2019.
[18] LIU P, YUAN W, FU J, et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing[J]. arXiv Preprint, arXiv: 2107.13586, 2021.
[19] LI X L, LIANG P. Prefix-tuning: optimizing continuous prompts for generation[J]. arXiv Preprint, arXiv: 2101.00190, 2021.
[20] LIU X, JI K, FU Y, et al. P-tuning v2: prompt tuning can be comparable to fine-tuning universally across scales and tasks [J]. arXiv Preprint, arXiv: 2110.07602, 2021.
[21] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems, 2017: 5998 – 6008.
[22] LIU Y, OTT M, GOYAL N, et al. Roberta: a robustly optimized bert pretraining approach[J]. arXiv Preprint, arXiv: 1907.11692, 2019.