doi:10.6041/j.issn.1000-1298.2023.06.027

Transformer 优化及其在苹果病虫命名实体识别中的应用

蒲　攀[1]　张　越[1]　刘　勇[1]　聂炎明[1]　黄铝文[1,2]

(1. 西北农林科技大学信息工程学院, 陕西杨凌 712100; 2. 农业农村部农业物联网重点实验室, 陕西杨凌 712100)

摘要: 为提高苹果生产领域实体识别的准确性,提出一种新的 Transformer 优化模型。首先,为解决苹果生产数据集的缺失,基于苹果栽培领域园艺专家的知识经验,创建以苹果病虫害为主的产业数据集。通过字向量与词向量的拼接,提高文本语义表征的准确性;随后,为防止位置信息缺失,引入具有方向和距离感知的注意力机制,平均集成 BiLSTM 的上下文长距离依赖特征;最后,结合条件随机场(Conditional random fields, CRF)约束上下文标注结果,最终得到 Transformer 优化模型。实验结果表明,所提方法在苹果病虫命名实体识别中的 F1 值可达 92.66%,可为农业命名实体的准确智能识别提供技术手段。

关键词: 苹果知识图谱; 病虫害; 自然语言处理; 命名实体识别; Transformer

中图分类号: TP391.1　　**文献标识码**: A　　**文章编号**: 1000-1298(2023)06-0264-08　　**OSID**:

Transformer Optimization and Application in Named Entity Recognition of Apple Diseases and Pests

PU Pan[1]　ZHANG Yue[1]　LIU Yong[1]　NIE Yanming[1]　HUANG Lüwen[1,2]

(1. *College of Information Engineering*, *Northwest A&F University*, *Yangling*, *Shaanxi* 712100, *China*
2. *Key Laboratory of Agricultural Internet of Things*, *Ministry of Agriculture and Rural Affairs*, *Yangling*, *Shaanxi* 712100, *China*)

Abstract: Aiming to improve the accuracy of entity identification in apple production field, a new Transformer optimization model was proposed. Firstly, in order to address the lack of apple production dataset, a corpus focusing on diseases and pests was constructed based on the knowledge and experience of horticultural experts in related field of apple cultivation. The accuracy of semantic representation of text was improved by combining word vector and character vector. Secondly, since the location information was crucial to text semantics, but the traditional Transformer model lacks the directionality of location information, in order to take advantage of the location features of text, an attention mechanism with direction and distance perception was introduced in Transformer encoder. And the contextual long-distance dependence features of BiLSTM was integrated on average to enhance semantic representation. Lastly, with imposing restrictions on labeling results by conditional random fields (CRF), the Transformer optimization model was obtained. The experimental results showed that the F1 score of the proposed method was 92.66% in Chinese named entity recognition of Apple diseases and pests. It indicated that the method proposed could effectively identify the named entities of apple diseases and pest, and provide a technical means for the accurate and intelligent identification of other agricultural named entities.

Key words: apple knowledge map; diseases and pests; natural language processing; named entity recognition; Transformer

0　引言

我国是世界第一大苹果生产国[1],果业病虫害问题对我国苹果产业影响显著,进而直接关系到国家和从业者的经济收益提升。基于知识图谱的问答系统通过对实体间关系描述能够帮助苹果种植人员

收稿日期: 2022-11-22　修回日期: 2023-04-11

基金项目: 陕西省重点研发计划项目(2019ZDLNY07-06-01)

作者简介: 蒲攀(1982—),男,讲师,博士,主要从事农业物联网与传感器网络研究,E-mail: pupan@nwsuaf.edu.cn

通信作者: 黄铝文(1976—),男,副教授,主要从事生物图像处理研究,E-mail: huanglvwen@nwsuaf.edu.cn

快速准确获得病虫管理专业知识，命名实体识别作为一种智能化实体抽取方法，是构建高质量苹果知识图谱的关键环节。因此，如何准确识别出苹果病虫领域相关实体对于苹果种植信息化发展具有重要作用。

近年来，命名实体识别技术被广泛应用在农业[2]、医学[3]等领域，文献[4]通过融合 ALBERT 与规则，针对小麦病虫害 16 类实体进行识别，其 F1 值达到 94.97%；文献[5]提出基于多核的卷积神经网络，对水产医学领域动物名称、发病部位、病原体等 6 项实体进行识别，其 F1 值达到 88.48%。虽然已开展了丰富的相关研究工作，然而在现有的农业命名实体识别的研究中仍存在以下问题：对没有明显边界特征的词汇识别率较低，如文献[6－7]中对于病原的识别 F1 值仅为 88% 和 81.48%；现有研究中常用 BiLSTM 来捕捉文本长距离依赖信息，但当文本距离过长时，其获取长距离依赖信息的能力会有所下降；对苹果病虫害实体识别方面的研究较少，同时缺乏公开权威的数据集，文献[8]构建了苹果病虫相关的 ApdCNER 语料库，提出将字典和相似词汇合并到基于字符的模型中，解决实体类别分布不均、别名和稀有实体识别困难的问题，其 F1 值尽管达到 92.14%，但是该方法需要依靠专家手动构建领域字典，容易对一些实体产生遗漏。此外，不同的人对一些相同实体可能也存在不一样的认识标准，进而字典构建的质量将直接影响模型的识别性能。

Transformer 可以实现并行化计算，同时处理长序列样本，常用于自然语言处理中的机器翻译[9]、文本生成[10]等领域。但是由于文本信息中不同位置的语义信息差别，其内部的绝对位置编码不能很好地表征位置信息，进而对中文语义信息的提取造成了困难。国内外学者对 Transformer 的文本性能优化做出了一系列改进。在较长序列建模方面，文献[11]引入段级递归，将绝对位置编码改为相对位置编码，提出了 Transfortmer－XL 模型。为降低文本分类复杂性，文献[12]使用星型拓扑代替全连通的注意力连接，提出了 Star Transformer 模型。在命名实体识别领域，文献[13]通过改进的相对位置编码，使用非缩放的点乘注意力，提出 TENER 模型。

为提高苹果生产领域实体识别的准确性，本文基于以上研究，通过融合文本的位置特征与语义特征，实现一种新的 Transformer 优化模型。该模型通过结合字向量与词向量以丰富语义信息，平均集成 BiLSTM 和 Transformer，并引入具有方向和距离感知的注意力机制，结合文本上下文依赖特征和位置特征，最后通过条件随机场（Conditional random fields，CRF）得到最优预测序列。

1　材料与方法

1.1　数据来源与标注

实验数据来源于中国农化招商网（http://www.1988.tv/bch/list－4.html）爬取的农业知识，在西北农林科技大学国家级苹果试验示范站的植保和栽培专家团队指导和协助下，人工对所爬取数据进行去空格、空行及特殊符号处理，去除重复数据和无效数据。综合《中国苹果病害病原菌物名录》电子版数据，建立苹果病虫知识数据集，其中包含 3 928 个病虫相关实体。

本文采用“BIO”的标注方式进行实体标注，其中，B 标注实体名称的开始，I 标注实体名称的内部信息，O 标注语料中的非实体部分。实体名称共包括 DIS、PES、NAM、PAR、MED、CAU 共 6 种，DIS 表示苹果病害名称，PES 表示苹果虫害名称，NAM 表示苹果病虫害的别称，PAR 表示为害部位名称，MED 表示防治药剂名称，CAU 表示病原名称。将 B、I、O 三元组与实体名称进行结合，形成标签，可得 B－DIS、B－PES、B－NAM、B－PAR、B－MED、B－CAU、I－DIS、I－PES、I－NAM、I－PAR、I－MED、I－CAU、O 共计 13 种实体标签，并以 8∶1∶1将数据集分为训练集、测试集和验证集，数据集划分如表 1 所示。苹果树腐烂病是我国西北地区苹果树常发病害，以苹果树腐烂病为例，依据此标签进行数据标注，标注示例如图 1 所示。

表 1　苹果知识数据集

Tab.1　Dataset of apple knowledge

数据集	字数	DIS/个	PES/个	NAM/个	PAR/个	MED/个	CAU/个
训练集	99 999	2 609	1 531	411	1 626	7 968	766
测试集	12 786	782	518	92	136	2 822	400
验证集	14 787	375	192	56	260	651	102

图 1　数据集标注示例

Fig.1　Corpus tagging results

1.2　特征分析

与通用语料相比，苹果病虫领域内的实体名称在结构和专业性等方面有明显自身特点，具体体现为以下 4 方面：

(1)构成成分多。苹果病虫领域的实体命名除了单纯的文字外,还常由数字、特殊符号等多种符号构成,如药剂名称“苏脉 1 号”、“多菌灵·异菌脲悬浮剂”等。

(2)生僻字较多。在药剂实体和病原实体中常出现生僻字,如药剂“噻霉酮”,病原“河口槭胶锈菌”,从而造成模型在识别上的困难。

(3)嵌套实体较常见。在药剂名称中常出现由多个子实体构成的实体,如“阿维菌素·哒螨灵乳油”易被拆分为“阿维菌素”“哒螨灵”“乳油”,容易干扰模型判断。

(4)一词多义现象较多存在。与其他农业作物病虫害实体不同的是,在中文文本中,“苹果”具有水果名称和商品品牌(手机、服装)多种含义,在苹果病虫领域,“苹果”一词出现在不同位置,代表着不同标签。如“苹果”在“苹果褐斑病”中的正确标签为“B-DIS I-DIS”,在病原实体“苹果星壳孢”中的正确标签应为“B-CAU I-CAU”,但其单独出现时又不是病虫相关实体,其具体标签由上下文语义共同决定,这给模型提取上下文关系带来难度。

2 模型框架

本文所提模型的整体结构包含嵌入层、Transformer 层、BiLSTM 层、特征融合层和 CRF 层 5 部分,其基本构成如图 2 所示。其中,x1、x2、x3、x4、x5、x6 为嵌入层输出;LSTM 为长短时记忆网络;BiLSTM 为双向长短时记忆网络;Multi-Head Attention 为多头注意力机制;Add&Norm 为残差和标准化;Feed Forward 为前馈神经网络;CRF 为条件随机场。

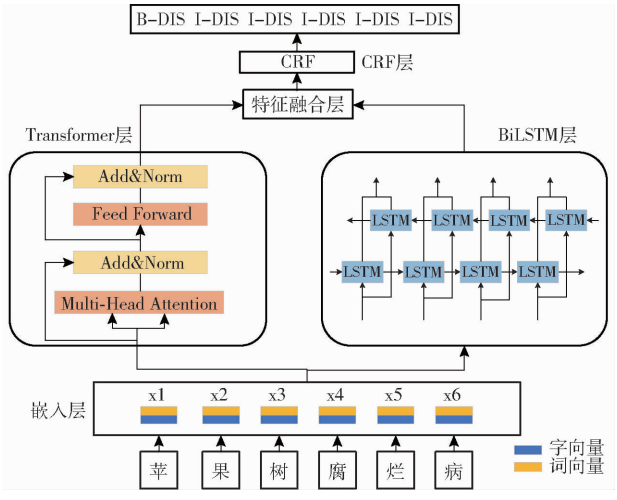

图 2 模型结构

Fig. 2 Model structure

2.1 嵌入层

由于中文分词存在错误拆分的现象[7],如病害名称“斑点落叶病”的分词结果为“斑点/落叶/病”,虫害实体“金纹细蛾”的分词结果为“金/纹/细/蛾”,这些实体的错误拆分会导致模型不能正确获取实体特征[14]。虽有研究表明,在嵌入层中基于字符的模型比基于单词的模型性能要好[15],但在中文里单个字符可表达的语义有限,而通过使用预先训练的词嵌入作为特征可进行改进。本文使用 Lattice LSTM[16]模型提供的预训练向量集,同时采用基于字向量与词向量拼接的嵌入方式来增强文本的语义信息。

2.2 Transformer 层

在中文命名实体识别任务中,文本的位置与语义密切相关。已有的多数研究更偏重文本的语义特征、偏旁特征或拼音特征,而忽略了文本的位置特征。而传统的 Transformer 模型在嵌入层引入了绝对位置编码,其计算方式为

$$P_{t,2i}=\sin(t/10\,000^{2i/d}) \tag{1}$$

$$P_{t,2i+1}=\cos(t/10\,000^{2i/d}) \tag{2}$$

式中 t——文本位置

i——维度位置索引

$P_{t,2i}$——第 t 个 token 在偶数维度的位置编码

$P_{t,2i+1}$——第 t 个 token 在奇数维度的位置编码

d——输入维度

在传统的 Transformer 注意力计算方法中,序列中第 t 个字和第 j 个字的注意力分数计算公式为

$$A_{t,j}=\frac{\boldsymbol{W}_Q(\boldsymbol{X}_t+\boldsymbol{P}_t)(\boldsymbol{W}_K(\boldsymbol{X}_j+\boldsymbol{P}_j))^{\mathrm{T}}}{\sqrt{d}}=\frac{\boldsymbol{W}_Q\boldsymbol{X}_t\boldsymbol{W}_K^{\mathrm{T}}\boldsymbol{X}_j^{\mathrm{T}}+\boldsymbol{W}_Q\boldsymbol{X}_t\boldsymbol{W}_K^{\mathrm{T}}\boldsymbol{P}_j^{\mathrm{T}}+\boldsymbol{W}_Q\boldsymbol{P}_t\boldsymbol{W}_K^{\mathrm{T}}\boldsymbol{X}_j^{\mathrm{T}}+\boldsymbol{W}_Q\boldsymbol{P}_t\boldsymbol{W}_K^{\mathrm{T}}\boldsymbol{P}_j^{\mathrm{T}}}{\sqrt{d}} \tag{3}$$

式中 $\boldsymbol{W}_Q$、$\boldsymbol{W}_K$、$\boldsymbol{W}_V$——输入字符的查询向量 $\boldsymbol{Q}$、键向量 $\boldsymbol{K}$、值向量 $\boldsymbol{V}$ 的权重参数矩阵

$\boldsymbol{X}$——字嵌入向量

$\boldsymbol{P}$——位置编码

j——token 的索引

根据传统的计算方式,$\boldsymbol{P}_t\boldsymbol{P}_j^{\mathrm{T}}$ 包含第 t 个 token 与第 j 个 token 的相对位置信息。令 $\omega_i=1/(10\,000^{2i/d})$,$k=t-j$,对 $\boldsymbol{P}_t\boldsymbol{P}_j^{\mathrm{T}}$ 展开为

$$\boldsymbol{P}_t\boldsymbol{P}_{t+k}=\sum_{i=0}^{\frac{d}{2}-1}(\sin(\omega_i t)\sin(\omega_i(t+k))+\cos(\omega_i t)\cos(\omega_i(t+k)))=\sum_{i=0}^{\frac{d}{2}-1}\cos(\omega_i(t-t-k))=\sum_{i=0}^{\frac{d}{2}-1}\cos(\omega_i k) \tag{4}$$

由三角函数的性质 $\cos(x)=\cos(-x)$ 可知，传统计算方式所得出的相对位置信息仅有两个 token 之间的距离关系，而对于两个 token 的位置无法判断，例如在“苹果”与“果苹”中，两个字相对位置虽不同，但其位置编码乘积相同，同时这样的位置信息在经过查询向量和键向量的映射后会消失。

为了充分利用中文文本的位置特征，本文改进使用

$$(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=(\boldsymbol{H}\boldsymbol{W}_Q,\boldsymbol{H}\boldsymbol{W}_K,\boldsymbol{H}\boldsymbol{W}_V) \tag{5}$$

$$\boldsymbol{R}_{t-j}=\left[\cdots\sin\left(\frac{t-j}{10\,000^{2i/d}}\right)\cos\left(\frac{t-j}{10\,000^{2i/d}}\right)\cdots\right]^{\mathrm{T}} \tag{6}$$

$$S_{t,j}=\boldsymbol{Q}_t\boldsymbol{K}_j^{\mathrm{T}}+\boldsymbol{Q}_t\boldsymbol{R}_{t-j}^{\mathrm{T}}+\boldsymbol{u}\boldsymbol{K}_j^{\mathrm{T}}+\boldsymbol{v}\boldsymbol{R}_{t-j}^{\mathrm{T}} \tag{7}$$

$$\boldsymbol{A}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}(S)\boldsymbol{V} \tag{8}$$

式中　H——嵌入层输出

$\boldsymbol{u}$、$\boldsymbol{v}$——可学习的参数

$\boldsymbol{R}_{t-j}$——相对位置编码

$S_{t,j}$——第 t、j 个 token 之间相似度得分

$\boldsymbol{Q}_t$，$\boldsymbol{K}_j$——第 t、j 个 token 的查询向量和键向量

$\boldsymbol{A}$——注意力得分

来计算整体注意力分数，通过式(6)计算位置编码，利用三角函数的性质，通过正弦函数捕捉方向性，余弦函数捕捉字符的绝对位置关系，从而解决传统 Transformer 模型相对位置信息易丢失的问题。通过式(7)计算输入序列中每个单词之间的相关性得分，式(8)对于输入序列中每个单词之间的相关性得分进行了归一化，使每个字与其他字的注意力权重之和为 1。

通过计算每个字与其他字的相关性，即可获得全局特征表示。当实体中的生僻字缺失语义信息时，根据其与前后文本的位置关系，依然可以依据其他文本而获取到。如病原实体中常以“菌”、“壳”、“孢”等字结尾，在实体“河口槭胶锈菌”中，“槭”为生僻字，但由于其后面的“菌”通常是病原中最后一字，且仅相隔两个字，根据其位置信息也可确定其为病原实体的一部分。

2.3　BiLSTM 层

通过分析苹果病虫数据集发现，在长句子中常会出现多类实体，且实体长度不一，为害部位大多常以两个字符出现，而部分药剂名称则多达 9 个字符，如“代森锰锌可湿性粉剂”。使用 LSTM[17] 不仅可以处理长序列问题，同时解决了 RNN 在训练时所产生的梯度爆炸或梯度消失现象[18-19]，而且能够有效利用上一时刻特征来判断下一时刻特征，因此本文使用 LSTM 网络实现对局部语义特征的提取。

在命名实体识别任务中，句子的前向信息和后向信息都很关键，而普通 LSTM 只能捕获前向信息[20]。如病害实体“斑点落叶病”，LSTM 提取到“叶”字时需提取到之前的“斑点落”几个字的特征，而无法考虑到与后面“病”字的关系。针对上述问题，本文选择双向 LSTM(BiLSTM)[21] 结构实现对句子级别的特征提取，以更好地解决苹果领域中一词多义的问题。

2.4　特征融合层

在命名实体识别研究中，许多研究者采用基于 CNN、基于 LSTM 和基于 Transformer 等方法作为上下文编码器，但是采用单一的编码器通常也会引起特征提取不充分问题。LSTM 模型虽然能够在序列信息建模方面凸显优势，却存在冗余信息；Transformer 能够关注重点词汇特征和加速训练速度，但存在上下文信息建模不足的缺陷。基于上述情况，本研究使用平均融合法实现对特征的融合，以降低模型陷入局部极小点的可能，进而达到提高识别率的目的。本文设计平均法、投票法、拼接法 3 种方案，其对应的计算公式分别为

$$H(x)=\frac{1}{N}\sum\left(h(x)+f(x)\right) \tag{9}$$

$$H(x)=\mathrm{argmax}\sum_{j=1}^{M}w_j p_{i,j,k} \tag{10}$$

$$H(x)=h(x)\oplus f(x) \tag{11}$$

式中　$h(x)$、$f(x)$——Transformer、BiLSTM 的输出

N——模型个数

w_j——第 j 个模型的权重

$p_{i,j,k}$——第 j 个模型对样本 i 的预测结果为类别 k 的概率

2.5　CRF 层

BiLSTM 与 Transformer 虽然适合处理长距离的文本信息，但都忽略了标签之间的依赖关系[22]，而在命名实体识别任务中，如果不考虑字符标签与相邻标签的相关性则极可能会给出错误标签。CRF[23] 能通过训练数据学习到标签之间的约束性[4-5]，并通过这种约束性获得一个最优的预测序列，具体约束性主要有以下两点：①句子中的第一个字的标签只能是“B-”或者“O”，不能是“I-”。②语句中的标签“B-label I-label I-label”，“label”应该是相同的命名实体标签，如“褐斑病”的标签应为“B-DIS I-DIS I-DIS”，而在苹果病虫数据集中，当“苹果”后为病害实体时，通常当作一个整体，如“苹果褐斑病”的标签为“B-DIS I-DIS I-DIS I-DIS I-DIS I-DIS”；“苹果”后为病原实体时，如“苹果链核盘菌”，其标签应为“B-CAU I-CAU I-

CAU I－CAU I－CAU I－CAU”。

3 实验与分析

3.1 实验环境搭建和模型参数设置

实验运行系统为 Ubuntu 18.04，显卡型号 NVIDIA GeForce RTX 3080 Ti，编程语言使用 Python 3.7 版本，采用 Pytorch 1.7.1 深度框架完成模型构建和训练评估。在实验过程中，所使用的模型参数，是通过前期的实验参数优化调整所得的最优参数组合：即迭代次数为 35，学习率 0.000 6，多头注意力数量为 4，每个头维度为 48，批量大小 4，随机失活率 0.45。由参数优化实验发现，学习率过大，会造成网络不能收敛，而学习率过小容易陷入局部最优解，进而造成识别效果变差；批量大小设置过小会不利于收敛，过大容易陷入局部最小值；为防止模型过拟合，本文添加了随机失活率来减少神经元之间的复杂关系，以增强模型鲁棒性。本文使用的优化实验参数如表 2 所示。对实体抽取模型结果和各项性能进行定量分析，所采用的评价指标为精确率、召回率、F1 值[24]。

表 2 BIO 标注的实验参数

Tab.2 Experimental parameters of BIO

参数	含意	数值
batch_size	批量大小	4
learning rate	学习率	0.000 6
head_num	多头注意力数量	4
head_dims	每个头维度	48
epoch	迭代次数	35
dropout	随机失活	0.45

3.2 实验结果与分析

3.2.1 不同模型实验结果对比

为了对比所提方法的识别效果，本文基于相同数据集，选取命名实体识别领域的 3 种传统常用模型 BiLSTM－CRF、Transformer－CRF 和 TENER 模型分别进行了性能对比实验。各个模型的性能结果如表 3 所示。

表 3 不同模型对比实验结果

Tab.3 Comparative experimental results of different models %

模型	精确率	召回率	F1 值
BiLSTM－CRF	88.58	82.55	85.46
Transformer－CRF	88.47	87.31	89.66
TENER	92.14	89.78	90.95
本文方法	94.08	91.28	92.66

从表 3 可以看出，使用 BiLSTM 对序列提取特征，F1 值仅为 85.46%，其可能原因是当文本序列长度过长时 BiLSTM 对上下文特征的提取能力会下降；而在传统的 Transformer 模型中，其位置编码信息没有得到充分利用，较于 BiLSTM 模型其仅在召回率和 F1 值上有所提高；TENER 通过改进位置编码函数，采用非缩放点积的注意力机制，与 Transformer 相比精确率、召回率和 F1 值分别提高 3.67、2.47、1.29 个百分点。本文提出的模型利用了 TENER 的优势，使用了具有方向和距离感知的注意力机制，充分结合了文本的位置特征，同时通过引入 BiLSTM 增强了上下文语义特征，使模型对上下文相关性的提取能力有所提高，其精确率、召回率和 F1 值分别达到 94.08%、91.28% 和 92.66%，在所比较模型中均达到最高。对比结果表明所提方法对于句子语义特征的学习是有效的。

3.2.2 不同实体类别实验结果对比

为了验证模型在各类实体上的提取能力，表 4 列出了所提方法在苹果病虫数据集中对各类实体的识别结果。从表中可知，本文方法对病害名称、虫害名称、为害部位、药品名称、病原 5 类实体的 F1 值均在 90% 以上，但对别名的 F1 值仅为 83.33%，这可能是由于在数据集中别名实体数量较少，同时病害与虫害的别名通常具有相同的边界特征，如病害名称与其别名都常以“病”字结尾，导致模型在判断中容易混淆，造成 F1 值较低。

表 4 Transformer 优化模型在各类实体上的表现

Tab.4 Optimized Transformer's performance on various entities %

实体名称	精确率	召回率	F1 值
病害名称	97.37	87.06	91.93
虫害名称	91.84	93.75	92.78
别名	90.91	76.92	83.33
为害部位	95.56	93.48	94.51
药品名称	93.24	92.00	92.62
病原	89.19	91.67	90.41

图 3～5 展示了不同模型在各实体上的性能表现对比，本文方法在药品名称（MED）、病害名称（DIS）、病原（CAU）、为害部位（PAR）、别名（NAM）实体上的 F1 值均优于其他模型。对实体虫害名称（PES）的 F1 值为 92.78%，低于其他模型，原因可能是在虫害实体中，大多以“虫”、“蛾”、“蚜”等特征词作为结尾，但存在部分虫害名称中不存在这些特征词，如“金龟子”，从而导致模型易将这些虫害实体识别错误。参考苹果病虫领域的实体特征，一词多义、生僻字、构成成分多等情况通常多出现在药品名称、病原、别名的实体中，因此，可以说明本文方法能够有效解决苹果病虫领域的命名实体识别问题。

以上结果表明，本文利用了 Transformer 中注意力机制，使模型具有更加关注重点词、抑制无用词的特点。采用具有方向和距离感知的注意力机制来充分利用文本位置特征，通过引入 BiLSTM 来增强上下文信息，在本文所构建的数据集中综合识别能力优于所比较的其他传统模型。

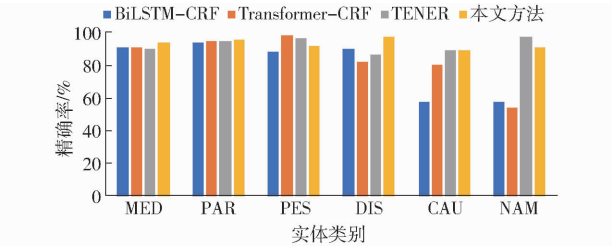

图 3　各实体精确率对比

Fig. 3　Comparison of precision of each entity

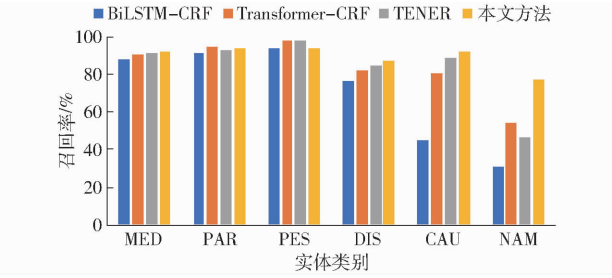

图 4　各实体召回率对比

Fig. 4　Comparison of recall rate of each entity

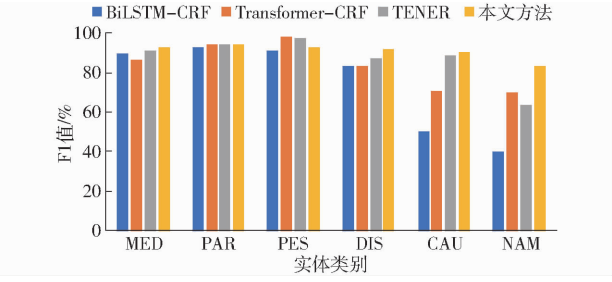

图 5　各实体 F1 值对比

Fig. 5　Comparison of F1 score of each entities

3.2.3　Transformer 与不同模型融合结果对比

学习任务假设空间往往很大，会有多个假设在训练集上达到同等性能，使用单一学习器可能会出现泛化能力不佳的情况，通过结合多个学习器可以有效降低这一风险。为了得到最佳的模型，本文设计了 Transformer 分别与 RNN、LSTM、BiLSTM 和 BiGRU 进行融合，实验结果如表 5 所示。

表 5　Transformer 与不同模型融合结果

Tab. 5　Transformer integrated with different models　%

模型	精确率	召回率	F1 值
Transformer－RNN－CRF	92.87	91.49	92.18
Transformer－LSTM－CRF	93.06	89.36	91.12
Transformer－BiGRU－CRF	94.65	90.43	92.49
本文方法	94.08	91.28	92.66

由表 5 可以看到，与传统 RNN 进行融合，模型精确率为 92.87%。而 LSTM 与 GRU 都属于门控 RNN，更适合处理长序列数据，与 LSTM 融合后的精确率达到 93.06%。为了可以更好地捕捉双向的语义依赖，选择分别与 BiGRU 和 BiLSTM 模型进行融合，在精确率和 F1 值上均有提高，结果表明，BiLSTM 模型可以更大范围地补充 Transformer 提取不到的依赖关系和语义特征，因此与 BiLSTM 融合的效果更好。

3.2.4　不同融合方法对模型的影响

为了选择最佳的特征融合方法，本文使用 3 种不同的融合方法（平均法、投票法、拼接融合法），比较不同融合方法对于模型性能的影响。在不同融合方法下，模型的性能参数结果如表 6 所示。

表 6　不同融合方法对比

Tab. 6　Comparison of different fusion methods　%

方法	精确率	召回率	F1 值
拼接融合	91.56	90.00	90.77
投票法	91.81	90.64	91.22
平均法	94.08	91.28	92.66

实验表明，使用简单拼接融合效果最差，平均法效果最好，其可能原因是拼接融合将多个特征向量拼接在一起，其中一些特征可能存在多个特征重复出现，导致模型在进行训练时过度依赖重复特征，从而降低模型性能。而由于本文苹果数据集中实体占比较小，投票法可能会受到噪声数据的干扰，从而导致错误的预测结果。平均融合法通过平均多个模型的预测结果，可以减少单个模型的偏差和方差，从而提高整体的准确性。

3.2.5　与同领域相似研究比较

文献[9]通过整理苹果病虫领域相关书籍，构建了包含 130 448 个汉字的苹果病虫害库 ApdCNER，共标注相关实体 11 876 个；本文先通过爬虫技术爬取网页中对苹果病虫害的描述，随后在专家指导下进行修正，构建了苹果病虫语料库，共标注 127 572 个汉字，包含 3 928 个实体，与文献[9]相比，本文构建的数据集中各实体数量相对更少，分布更稀疏；在实体的标注方面，文献[9]将苹果相关各实体分为 21 个类别，更考验模型的提取能力。在本文构建的数据集中根据日常使用情况将实体分为 6 个类别，更适用于普通问答系统的构建；在模型提取能力方面，文献[9]将苹果数据集中的字典和类似的单词纳入 BiLSTM CRF 模型，其精确率、召回率和 F1 值分别为 92.29%、91.99% 和 92.14%，本文所提方法的精确率、召回率和 F1 值分别为 94.08%、

91.28%、92.66%,F1 值相较提高 0.52 个百分点,表明了本文方法与当前同领域较先进的模型达到同等性能水平。考虑到所需的样本数量,本方法对小样本量的苹果领域命名实体识别任务具有较高的特征提取能力。

另外,由于与文献[9]采用了不同的实体标注方式,为了排除标注方法的可能影响,本文又使用文献[9]的 BMES 标注方式对所建数据集进行了实体标注,B 表示该汉字是一个词语的开头,E 表示该汉字是一个词语的结尾,M 表示该汉字是一个词语的中间部分,S 则表示该汉字单独构成一个词语,该标注方法对所标注实体增加了对应标签,而对于实体数量较 BIO 标注并未发生改变。BMES 标注后重新训练该模型并对识别性能评估,模型的优化实验参数如表 7 所示,重新标注后所建模型的精确率为 94.40%、召回率为 92.21%、F1 值达到 93.29%。对比 3.2.1 节使用 BIO 标注方法的模型运行结果,可以发现两种标注方法的运行效果基本达到相同水平,使用 BMES 方法更加准确地标注出每个汉字的位置和类型,提供了比 BIO 方法更多的信息,方便后续进行分词处理,因而是其 F1 值略有提升的可能原因。对比表 7 与表 2 可以看到,模型的优化参数会受标注方式影响。

表 7 BMES 标注的实验参数

Tab.7 Experimental parameters of BMES

参数	含意	数值
batch_size	批量大小	8
learning rate	学习率	0.001 8
head_num	多头注意力数量	4
head_dims	每个头维度	48
epoch	迭代次数	40
dropout	随机失活	0.30

最后,为了测试所建模型在实际使用过程中的有效性,通过百度贴吧平台抽取了苹果种植相关的问题,并使用本文模型对文本数据进行了实体识别,结果如表 8 所示,本模型对实际病虫提问均准确提取出了问题中的实体。

表 8 识别结果示例

Tab.8 Examples of recognition results

问题	识别结果
苹果根腐烂病怎么治疗?	病害:苹果根腐烂病
治疗红蜘蛛用什么?	虫害:红蜘蛛
苹果灰霉病的防治方法	病害:苹果灰霉病
这个是斑点落叶病吗?	病害:斑点落叶病

4 结论

(1)针对苹果生产领域存在病虫害相关数据集缺失的问题,本文基于西北农林科技大学在陕西省渭南市白水县的苹果试验示范站所收集的苹果病虫知识,以及通过爬取中国农化招商网,综合《中国苹果病害病原菌物名录》电子版数据,建立了苹果病虫知识数据集。

(2)为提高苹果病虫害实体识别的准确性,本文通过在 Transformer 中引入具有方向和距离感知的注意力机制,融合 BiLSTM 提取到的语义特征来提高 Transformer 在苹果病虫害实体识别领域的识别效果,通过对比 BiLSTM - CRF、Transformer - CRF、TENER 模型,验证了传统 Transformer 模型在命名实体识别领域的不足。实验结果表明,本文所提方法在苹果命名实体识别中的 F1 值可达 92.66%。相较于传统的识别方法,其性能进一步提升,对小样本量数据集优势明显。

参 考 文 献

[1] 张山清,普宗朝,李新建,等. 气候变化对新疆苹果种植气候适宜性的影响[J]. 中国农业资源与区划, 2018, 39(8): 255 - 264.
ZHANG Shanqing, PU Zongchao, LI Xinjian, et al. Impact of climate change on apple-planting climatic suitability in Xinjiang [J]. Chinese Journal of Agricultural Resources and Regional Planning, 2018, 39(8): 255 - 264. (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]. 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]. Transactions of the Chinese Society for Agricultural Machinery, 2017, 48(Supp.): 178 - 185. (in Chinese)

[3] 黄健格,贾真,张凡,等. 基于多特征嵌入的中文医学命名实体识别[J/OL]. 计算机科学, 2023: 1 - 12. http://kns.cnki.net/kcms/detail/50.1075.TP.20230308.1155.008.html.
HUANG Jian'ge, JIA Zhen, ZHANG Fan, et al. Chinese medical named entity recognition based on multi-feature embedding [J/OL]. Computer Science, 2023: 1 - 12. http://kns.cnki.net/kcms/detail/50.1075.TP.20230308.1155.008.html. (in Chinese)

[4] 刘合兵,张德梦,熊蜀峰,等. 融合 ALBERT 与规则的小麦病虫害命名实体识别[J/OL]. 计算机科学与探索, 2022: 1 - 12. http://kns.cnki.net/kcms/detail/11.5602.TP.20220704.1056.002.html.

LIU Hebing, ZHANG Demeng, XIONG Shufeng, et al. Named entity recognition of wheat diseases and pests fusing ALBERT and rules[J/OL]. Journal of Frontiers of Computer Science and Technology, 2022: 1 - 12. http://kns.cnki.net/kcms/detail/11.5602.TP.20220704.1056.002.html. (in Chinese)

[5] 刘巨升,于红,杨惠宁,等. 基于多核卷积神经网络(BERT + Multi - CNN + CRF)的水产医学嵌套命名实体识别[J]. 大连海洋大学学报, 2022, 37(3): 524 - 530.
LIU Jusheng, YU Hong, YANG Huining, et al. Recognition of nested named entities in aquature medicine based on multi-kernel convolution(BERT + Multi - CNN + CRF)[J]. Journal of Dalian Ocean University, 2022, 37(3): 524 - 530. (in Chinese)

[6] 郭旭超,唐詹,刁磊,等. 基于部首嵌入和注意力机制的病虫害命名实体识别[J]. 农业机械学报, 2020, 51(增刊 2): 335 - 343.
GUO Xuchao, TANG Zhan, DIAO Lei, et al. Recognition of Chinese agricultural diseases and pests named entity with joint radical-embedding and self-attention mechanism[J]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(Supp.2): 335 - 343. (in Chinese)

[7] 李林,周晗,郭旭超,等. 基于多源信息融合的中文农作物病虫害命名实体识别[J]. 农业机械学报, 2021, 52(12): 253 - 263.
LI Lin, ZHOU Han, GUO Xuchao, et al. Named entity recognition of diseases and insect pests based on multi source information fusion[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(12): 253 - 263. (in Chinese)

[8] ZHANG J, GUO M, GENG Y, et al. Chinese named entity recognition for apple diseases and pests based on character augmentation[J]. Computers and Electronics in Agriculture, 2021, 190: 106464.

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

[10] 李想,王卫兵,尚学达. 指针生成网络和覆盖损失优化的 Transformer 在生成式文本摘要领域的应用[J]. 计算机应用, 2021, 41(6): 1647 - 1651.
LI Xiang, WANG Weibing, SHANG Xueda. Application of Transformer optimized by pointer generator network and coverage loss in field of abstractive text summarization[J]. Journal of Computer Applications, 2021, 41(6): 1647 - 1651. (in Chinese)

[11] DAI Z, YANG Z, YANG Y, et al. Transformer - XL: attentive language models beyond a fixed-length context[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019:2978 - 2988.

[12] GUO Q, QIU X, LIU P, et al. Star-Transformer[C]// Proceedings of NAACL-HLT, 2019:1315 - 1325.

[13] YAN H, DENG B, LI X, et al. TENER: adapting transformer encoder for named entity recognition[J]. arXiv preprint. arXiv:1911.04474, 2019.

[14] 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)

[15] LI X, MENG Y, SUN X, et al. Is word segmentation necessary for deep learning of Chinese representations? [C]// The Association for Computational Linguistics, 2019:3242 - 3252.

[16] ZHANG Y, YANG J. Chinese NER using lattice LSTM[J]. arXiv preprint. arXiv:1805.02023,2018.

[17] SAK H, SENIOR A W, BEAUFAYS F. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition[J]. arXiv preprint. arXiv.1402.1128,2014.

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

[19] 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)

[20] WANG Y, SUN Y, MA Z, et al. A hybrid model for named entity recognition on Chinese electronic medical records[J]. ACM Transactions on Asian and Low-Resource Language Information Processing, 2021, 20(2): 1 - 12.

[21] GRAVES A, SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM networks[C]// Proceedings of the IEEE International Joint Conference on Neural Networks, 2005: 2047 - 2052.

[22] 张云秋,汪洋,李博诚. 基于 RoBERTa - wwm 动态融合模型的中文电子病历命名实体识别[J]. 数据分析与知识发现, 2022, 6(2/3): 242 - 250.
ZHANG Yunqiu, WANG Yang, LI Bocheng. Identifying named entities of Chinese electronic medical records based on RoBERTa - wwm dynamic fusion model[J]. Data Analysis and Knowledge Discovery, 2022, 6(2/3): 242 - 250. (in Chinese)

[23] LAFFERTY J, MCCALLUM A, PEREIRA F. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]// Proceedings of the Eighteenth International Conference on Machine Learning, 2001: 282 - 289.

[24] JOHNSON S, SHEN S, LIU Y. CWPC_BiAtt: character-word-position combined BiLSTM-attention for Chinese named entity recognition[J]. Information, 2020, 11(1): 45.