第45卷 第3期 华 中 农 业 大 学 学 报 Vol.45 No.3
2026年 5月 Journal of Huazhong Agricultural University May 2026,127~136

文纪威,杨喆,王成,等.预训练模型驱动的猪病知识图谱构建与应用[J].华中农业大学学报,2026,45(3):127-136.
DOI:10.13300/j.cnki.hnlkxb.2026.03.011

预训练模型驱动的猪病知识图谱构建与应用

文纪威[1],杨喆[2,3],王成[1],靳哲[1],邓兵[3],李国亮[1],方亚平[1],刘峰[1],杜小勇[1]

1.华中农业大学信息学院/农业农村部智慧养殖技术重点实验室/农业智能技术教育部工程研究中心/湖北省农业大数据工程技术研究中心,武汉 430070;
2.农业动物遗传育种与繁殖教育部重点实验室,武汉 430070;
3.武汉市农业科学院畜牧兽医研究所,武汉 430200

摘要 为提高猪病防治知识的获取效率,缓解基层畜牧兽医服务资源匮乏的问题,本研究构建了面向猪病防治的垂直领域知识图谱及智能问答系统。针对多源异构数据描述存在冲突的问题,研究依据兽医病理学建立领域本体,引入BERT语义相似度融合策略,构建专业标注语料库。针对兽医术语复合性强且长尾分布的特点,提出ERNIE-Bi-LSTM-CRF实体识别模型。命名实体识别试验及消融实验结果显示,ERNIE独特的实体级掩码机制能有效克服传统字级掩码对长难术语的语义割裂问题,其F_1值达89.63%,显著优于Bi-LSTM+CRF及BERT等基线模型。基于此,集成Neo4j图数据库与意图识别技术开发了微信端问答机器人,测试结果显示该机器人能够准确理解专业性的兽医诊疗问题,并能提供全面、科学的检测指导,有效辅助养殖人员进行疾病确诊。结果表明,所构建系统能够实现从自然语言问题到图谱知识检索的闭环处理,可为猪病辅助诊断与防控知识服务提供技术支持。

关键词 知识图谱;猪病;BERT;XLNet;ERNIE;微信机器人

中图分类号 TP391.4;S858.28 **文献标识码** A **文章编号** 1000-2421(2026)03-0127-10

自2018年以来,非洲猪瘟在中国及多个东亚国家发生和传播,给养猪业及相关产业造成严重冲击。我国生猪养殖行业在遭受非洲猪瘟的影响后也发生变革。为阻断疾病传播路径并强化生物安全屏障,多数猪场实施了严格的封闭式管理策略。然而,当前我国畜牧兽医专业人才缺口显著,导致养殖人员难以获取全面系统的猪病防治知识。猪群健康状况是养殖效益的核心要素,但诊疗资源匮乏与诊断滞后等问题,常导致养殖人员因应对失策而延误病情,或因不当用药引发继发性诊断困难及误诊,最终造成严重的经济损失[1]。

现有搜索引擎虽能提供海量信息,但在面对复杂的诊疗场景时,信息冗余度高且筛选困难,查全率与查准率往往难以兼顾。传统问答系统虽能辅助搜索,但因缺乏语义语境支撑,答案的解释性不足,难以建立用户信任。此外,专家咨询服务虽具备专业性,但响应周期较长,难以满足急性病情的即时诊疗需求。近年来猪病相关数据不断积累,但术语复杂与存储结构异构仍显著增加了有效信息获取难度。知识图谱[2]是一种捕捉和表示复杂关联关系的工具,为上述问题提供了解决方案,通过将知识图谱与智能问答系统相结合,用户能够更准确、高效地检索到所需信息[3]。

尽管知识图谱技术已在各行各业广泛应用,但在畜牧兽医领域仍处于发展阶段[4-6]。已有研究尝试利用ALBERT模型构建猪病知识图谱[7],并实现了面向诊疗的图谱可视化[8],但在长文本语义建模和多意图问答集成方面仍存在不足。随着深度学习技术的快速迭代,该领域的研究重心已逐渐向高精度实体识别与大语言模型(LLM)融合方向转移。近期,基于词典增强BERT的实体识别方法[9]或集成RAG技术的农业大模型[10]的出现,推动了预训练模

收稿日期:2025-09-30
基金项目:湖北省重点研发计划项目(2022BBA0015);中央高校基本科研业务费专项(2662023XXPY005;2662025PY018;2662025DKPY005)
文纪威,E-mail:wenjiwei@webmail.hzau.edu.cn
通信作者:杜小勇,E-mail:duxiaoyong@mail.hzau.edu.cn

型在农业文本理解中的应用。然而，在低资源垂直领域中，如何构建高质量本体基座并实现端到端服务闭环，仍有待进一步研究。预训练语言模型已成为当前文本表示学习的重要方法，并被广泛应用于生物医药、医学图像分析和知识图谱构建等任务。如XLNet-Bi-LSTM模型在智慧医疗领域中被运用到电子病历实体识别[11]，相关研究也发现通过ERNIE预训练语言模型更适合完成中文问答任务[12]。然而，针对猪病领域知识抽取任务，不同预训练语言模型的适用性仍缺乏系统比较，因此需要深入探索适合猪病领域的信息抽取模式。基于上述研究进展与应用需求，本研究构建面向猪病防治的垂直领域知识图谱及智能问答系统，以辅助养殖户进行科学决策。

1 材料与方法

针对猪病知识获取和问答服务中的数据异构、术语识别困难及服务集成不足等问题，本研究从3个方面进行系统性优化：首先，依据兽医病理学确立领域本体，构建核心实体的专业词典，为非结构化数据处理提供高质量本体支撑。其次，针对多源异构数据（如猪易网、猪病通）的描述冲突与属性缺失，提出基于BERT语义相似度的知识融合策略；同时，选定ERNIE-Bi-LSTM-CRF作为核心抽取模型，利用其独特的实体级掩码策略有效攻克了兽医领域中文长难术语的识别瓶颈。

在此基础上，本研究利用Neo4j图数据库实现了多源数据的关联存储与可视化，并基于微信公众平台开发了智能问答机器人。面向养殖户提问方式的多样性，系统构建了包含12类意图的语义解析算法，实现了从自然语言到Cypher查询语句的精准映射。图1展示了整体架构：底层负责多源数据清洗，中间层利用ERNIE与融合算法完成图谱构建，顶层通过API接口支撑微信端的即时响应，形成了从数据获取到终端服务的完整技术闭环，从而有效提高养殖业的生产效率和猪群健康水平。

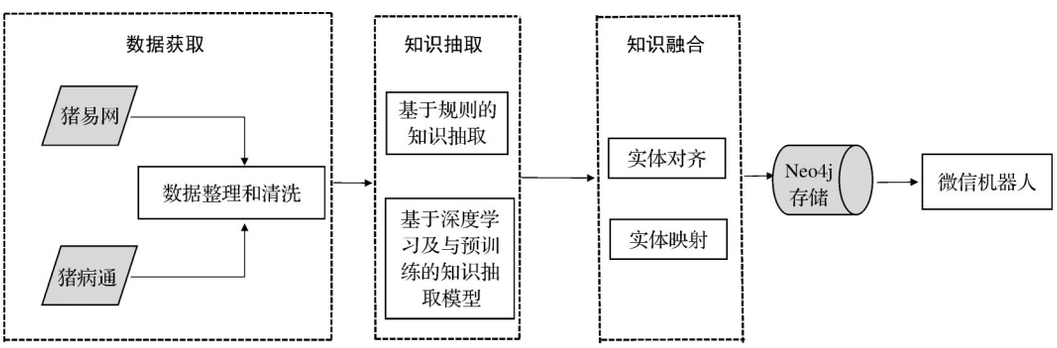


图1 研究框架

Fig.1 Research framework

1.1 数据采集与预处理

为构建高质量的猪病领域知识基座，本研究选取了猪易网和猪病通这2个国内领先的生猪养殖垂直门户网站作为核心数据源。相较于通用网络数据，上述平台不仅汇集了经过兽医专家审核的疾病百科，还拥有活跃的养殖户问答社区，其数据在权威性、专业度及语料覆盖面上具有显著优势。基于此，本研究利用网络爬虫技术定向采集了涵盖47种常见猪病的详细诊疗记录与问答语料。

在获取原始数据后，首先利用正则表达式去除HTML标签及特殊符号，将半结构化数据转化为规范的纯文本。随后，为了构建适用于ERNIE-Bi-LSTM-CRF模型的训练集，研究筛选出2 480条高质量病例描述，依据BIO标注规范利用Doccano工具完成人工标注，并按8∶1∶1的比例随机划分为训练集、验证集和测试集[13]。进一步地，针对多源异构数据中存在的属性缺失与描述冲突挑战，本研究实施了深度的语义处理：一方面利用预训练NER模型从实体描述中自动补全易感动物等缺失字段；另一方面引入基于BERT的语义相似度融合策略，通过计算症状、药物等属性的余弦相似度，实现了同一知识点的自动去重与归一化，从而完成从原始网页到结构化图谱数据的高效转化。

1.2 猪病领域本体构建与三元组抽取策略

为规范化描述猪病领域的复杂知识，本研究首先依据兽医病理学标准，构建包含疾病、症状、药物等在内的实体分类体系（表1），并定义5类核心语义关系（表2）。基于上述定义，本研究针对数据源多为半结构化诊疗记录（如临床症状：高热、呕吐）的特性，采用了基于本体规则映射的策略进行关系抽取

表 1　猪群疾病节点属性

Table 1　Pig disease node properties

节点属性 Node attributes	举例 Example
疾病序列 Disease_id	18
疾病原因 Disease_cause	猪丹毒的病原为红斑丹毒丝菌 The etiologic agent of *Erysipelas suis* is the filamentous bacterium of *Erysipelothrix rhusiopathiae*
治疗周期 Treatment_duration	每日 1～2 次，连用 3～4 d、每 8 h 注射 1 次 1～2 times a day，for 3～4 days，every 8 hours
治疗手段 Cure_way	抗生素治疗，如青霉素类等，以抑制菌体生长并消灭感染 Antibiotic treatments，such as penicillins，are used to inhibit bacterial growth and eliminate infection
疾病描述 Description	猪丹毒是由红斑丹毒丝菌引起猪的一种急性… *Erysipelas suis* is an acute disease in pigs caused by filamentous bacterium of *Erysipelothrix rhusiopathiae*，characterized by…
易感动物 Susceptible_animals	3～6 月龄的猪只 Pigs 3～6 months old
流行病学 Epidemiology	野外环境下，猪对红斑丹毒丝菌的敏感性较高… Under natural conditions，pigs are most susceptible to this bacterium…
疾病名称 Name	猪丹毒 *Erysipelas suis*
病理 Pathological	急性病变，器官受损，组织炎症，神经系统损伤… Acute lesions，damage to organs，inflammation of tissues，damage to the nervous system…
病例预防 Prevent	某些健康猪，猪体内带有猪丹毒丝菌，当… Some healthy pigs，pigs with erysipelas thrix in the body，when…
推荐药物 Recommend_drug	土霉素盐酸盐，青霉素…Oxytetracycline，penicillin…
疾病症状 Symptom	体温正常或稍高，亢进… Normal or slightly elevated，hyperfunction…

表 2　猪群疾病关系类型

Table 2　Pig disease relationship types

节点属性 Node attributes	举例 Example
具有症状 Has_symptom	猪丹毒，具有症状，体温正常或稍高 *Erysipelas suis*，symptomatic，normal or slightly elevated body temperature
疾病类别属于 Belongs_to	猪丹毒，疾病类别属于，传染病 *Erysipelas suis*，the disease category belongs to，infectious diseases
疾病诊断 Need_check	猪丹毒，疾病诊断，细菌学检查… *Erysipelas suis*，disease diagnosis，bacteriological examination…
推荐药物 Recommend_drug	猪丹毒，推荐药物，土霉素盐酸盐 *Erysipelas suis*，recommended drug，oxytetracycline
疾病预防措施 Recommend_method	猪丹毒，疾病预防措施，加强猪的饲养管理 *Erysipelas suis*，disease recommended preventive measures to strengthen the feeding management of pigs

与三元组构建。具体而言，不依赖复杂的监督式关系抽取模型，而是设计了从半结构化字段到图谱边(edge)的映射模板。例如，当NER模型在文本中识别出疾病与症状实体，且它们共现于百科的临床症状字段时，即依据本体约束实例化为(disease，has_symptom，symptom)类型的关系三元组。此外，针对多源数据合并时可能产生的关系冗余，进一步利用BERT语义相似度算法进行冲突消解，最终将清洗后的三元组存储于图数据库中，完成从离散实体到关联图谱的转化。

1.3　模型搭建

LSTM是一种特殊的循环神经网络(recurrent neural network，RNN)，因其在建模和分析序列数据方面的强大能力而受到广泛关注[14]。LSTM在捕捉长序列上的依赖关系和模式方面表现出色，因此特别适用于时间序列和序列数据分析。在医疗健康领域，基于LSTM的模型已成功应用于疾病预测、患者监测和医学信号分析等多个场景[15-16]。研究选取具有双向LSTM网络结构的Bi-LSTM，其中，前向LSTM层旨在捕捉输入序列左侧的上下文特征向量，而反向LSTM层则负责提取右侧上下文信息。接收到预训练模型输出的语义向量序列E后，Bi-LSTM网络通过门控机制(遗忘门、输入门和输出门)进一步提取文本的序列时序特征，其计算逻辑表示为：

$$\overrightarrow{h_t}=f_t\left(e_t,\overrightarrow{h_{t-1}}\right) \tag{1}$$

$$\overleftarrow{h_t}=f_t\left(e_t,\overleftarrow{h_{t+1}}\right) \tag{2}$$

$$H_T=[\overrightarrow{h_t};\overleftarrow{h_t}] \tag{3}$$

式(1)～(3)中，$\overrightarrow{h_t}$表示前向隐藏状态向量；$\overleftarrow{h_t}$表

示后向隐藏状态向量；e_t表示输入的语义特征向量；f_L表示LSTM单元的网络计算函数；H_T表示时刻t拼接后的最终特征向量。该模型可以有效地捕获过去和未来上下文中的模式和依赖关系。此外，CRF条件概率分布模型标签可以约束标签以确保输出结果在合理的范围内[17]，将C词向量的集合输入CRF层，得到每个词最后的实体标注结果。CRF层可以有效解决“BIO”标注中可能出现的顺序错乱问题，例如，CRF层可确保标签遵循“B-I”的合法邻接约束，有效规避了“I”标签前置于“B”标签等非法序列的生成，从而保证了实体边界的准确性。对于给定输入序列特征及其对应的预测标签序列，CRF层引入转移矩阵并计算全局序列得分，最后通过条件概率分布解码，其计算方法为：

$$S(X,y)=\sum_{i=1}^{n}\left(A_{y_{i-1},y_i}+P_{i,y_i}\right) \tag{4}$$

$$P(y|X)=\frac{\exp\left(S(X,y)\right)}{\sum_{\tilde{y}\in Y_X}\exp\left(S(X,\tilde{y})\right)} \tag{5}$$

式(4)～(5)中，S表示序列全局得分计算函数；X表示输入文本的特征向量序列；y表示预测的标签序列向量；n表示输入序列的长度；A_{y_{i-1},y_i}表示标签转移得分；P_{i,y_i}表示发射得分；$\tilde{y}$表示集合中的某一候选标签序列向量。

为使模型在训练阶段能够学习到最优的标签转移规律，本研究采用负对数似然作为损失函数L，其表达式为：

$$L=-\sum_{i=1}^{N}\log\left(P_c\left(y_i|X_i\right)\right)+\frac{\lambda}{2}\|\omega\|^2 \tag{6}$$

式(6)中，L表示训练损失值；N表示训练样本总数；P_c表示标签序列的条件概率分布；λ表示正则化参数；ω表示模型所有可学习参数的集合向量。通过最小化该损失函数，模型能够最大化标注序列的全局得分，从而提升复杂兽医术语的边界识别精度。

Bi-LSTM拼接CRF后在知识图谱的构建中应用广泛[18-21]，它能够利用序列生成过程中已经发生的信息，保证输出值是一组最优解序列。其结构如图2所示。此外，本研究利用预训练模型内部的自注意力机制[22-23]，使得模型能够捕捉句子语义中的重点。

1.4 预训练模型

为比较不同预训练语言模型在猪病实体识别任务中的适用性，本研究分别构建BERT-Bi-LSTM-CRF、XLNet-Bi-LSTM-CRF和ERNIE-Bi-LSTM-CRF 3类模型。模型整体结构保持一致，仅替换前端文本表示层，以排除后续序列建模和标签解码结构差异对试验结果的影响。具体而言，首先将经过BIO标注的猪病文本输入预训练语言模型，获得融合上下文语义信息的字符级或词片级向量表示；随后将该向量序列输入Bi-LSTM层，进一步提取前后向上下文依赖特征；最后通过CRF层对标签序列进行全局约束，输出疾病、症状、药物、检查方法等实体类别标签。

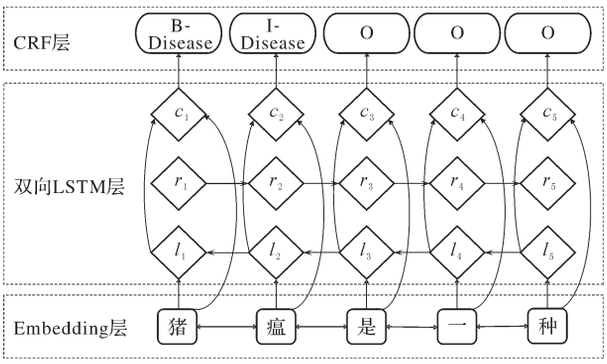


图2 Bi-LSTM+CRF模型

Fig.2 Bi-LSTM+CRF model

BERT[24]采用基于Transformer的双向编码结构，通过掩码语言模型和下一句预测任务学习上下文语义表示(图3)；XLNet[25]在自回归语言模型基础上引入排列语言建模机制，以增强对词序关系和长距离依赖的建模能力；ERNIE[26]则在预训练阶段引入实体级和短语级掩码机制，使模型能够学习完整实体单元内部的语义关联(图4)。考虑到猪病文本中存在大量“猪流行性感冒”“土霉素盐酸盐”等复合型专业术语，本研究重点考察不同预训练表示方式对长难兽医实体边界识别和类别判定的影响。

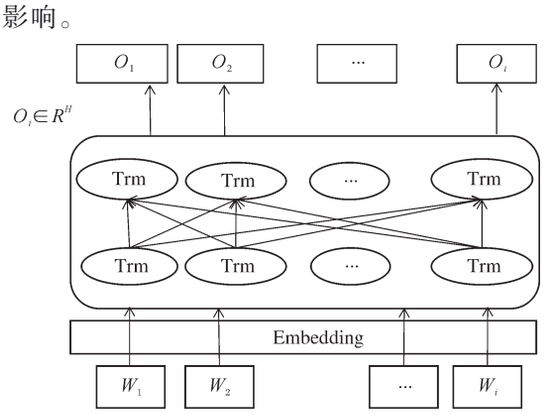


图3 BERT模型架构

Fig.3 BERT model architecture

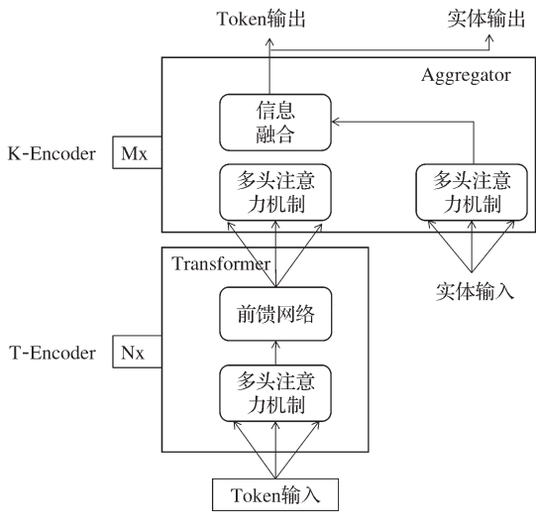


图4　ERNIE模型架构

Fig.4　ERNIE model architecture

预训练语言模型的上下文表示主要由多头自注意力机制完成。对于任意一个注意力头，其计算公式为：

$$Z=f_s\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{7}$$

式(7)中，Z表示自注意力输出矩阵；f_s表示Softmax激活函数；Q表示查询矩阵；K表示键矩阵；V表示值矩阵；d_k表示键向量的维度。

ERNIE针对猪病领域长难术语多的特点，在预训练阶段采用了3层掩码策略。对于输入文本序列T，其掩码学习目标J可表述为对实体级单元m的最大似然估计：

$$J(\theta)=\sum_{m\in M}\log P(m|T_{\setminus M};\ \theta) \tag{8}$$

式(8)中，$J(\theta)$表示预训练目标函数；M表示被掩码处理的实体或短语单元集合；m表示集合中具体的被遮蔽单元；$T_{\setminus M}$表示序列中未被遮蔽的上下文部分；θ表示模型参数。相比于字级遮蔽，该机制强制模型学习实体内部字词间的语义关联，显著改善了生僻猪病术语的特征表示。这一机制使模型能够更准确地结合上下文信息，因此在多种评估中ERNIE模型相对于其他模型(如BERT)表现更为优异，能够更准确地理解和处理文本信息，提升任务性能。

1.5　模型评价指标

选择精确率(precision)、召回率(recall)、F_1值(F_1-score)3个指标进行模型效果评估，评价指标的值越大说明模型实体识别的效果越好，其计算方法如公式(9)～(11)所示。

$$P=\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{P}}} \tag{9}$$

$$R=\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{N}}} \tag{10}$$

$$F_1=\frac{2\times P\times R}{P+R} \tag{11}$$

式(10)～(11)中，P表示识别出的所有实体中，正确实体的比例；R表示在所有真实存在的实体中，被正确识别出来的比例；F_1表示精确率和召回率的调和平均数，用于综合评价模型效果；T_{P}表示识别正确的实体数量；F_{P}表示错误识别为实体的数量；F_{N}表示未被识别出的真实实体数量。

1.6　图数据库存储

本研究使用Neo4j图数据库进行数据存储，Neo4j数据库通过向节点和边添加键值属性来有效压缩信息，以提供高度的灵活性，从而使建模更易于阅读和逐步改进[27]。使用图数据库Neo4j存储猪病的数据使得模型更容易被人理解，相较于传统SQL查询，图数据库查询语言(cypher)在处理复杂关联检索时具有更优的便捷性与执行效率。

1.7　智能问答系统的逻辑架构与实现

研究基于微信公众平台开发了猪病智能问答机器人，其核心交互逻辑包含意图识别、实体链接与知识检索3个阶段。首先，在系统搭建方面，利用Flask框架构建后端服务器，通过微信API接口实现与前端用户的实时消息交互。其次，在问句解析方面，当后端接收到用户输入的自然语言(如猪得了口蹄疫怎么治?)时，系统并行调用训练好的BERT意图分类模型与ERNIE实体识别模型：前者识别出用户意图为查询治疗方案(treatment)，后者提取出核心实体口蹄疫(disease)。具体分类逻辑如下：系统接收到问句特征向量u后，通过线性映射及Softmax运算判定用户意图类别k：

$$p_k=\frac{\exp(\omega_k u+b_k)}{\sum_{j=1}^{12}\exp(\omega_j u+b_j)} \tag{12}$$

式(12)中，p_k表示问句属于第k类意图的概率；u表示问句提取后的全局特征向量；ω_k，b_k分别表示第k类意图对应的权重系数与偏置项；j表示意图类别索引，本研究中取值为1～12。最后，在知识库调用方面，系统预置了与12类意图对应的Cypher查询模板。程序将识别到的实体填入模板，通过Py2neo驱动连接Neo4j数据库执行查询，并将返回的结构化结果(如药物列表)封装为文本消息反馈给用户，从而实现从自然语言到知识图谱的精准检索。

2 结果与分析

2.1 命名实体识别试验

本研究在自建的猪病数据集(self-built swine disease dataset,SSDD)及2个公开通用数据集上,分别对Bi-LSTM+CRF、BERT+Bi-LSTM+CRF以及本研究提出的ERNIE+Bi-LSTM+CRF模型进行全方位的对比试验。

作为基线模型的Bi-LSTM+CRF试验结果如表3所示。从表3中数据可以看出,虽然该模型在通用数据集上的表现尚可,但在猪病垂直领域数据集中却出现了明显的性能滑坡,F_1值仅为46.00%。进一步观察精确率和召回率指标发现,其召回率极低(仅46.64%),这表明在缺乏预训练知识库支持的情况下,仅依赖字向量的模型难以有效识别出未登录词(OOV)和生僻的兽医专业术语,导致大量关键实体被漏检。引入BERT预训练模型后,BERT架构在猪病数据集上的F_1值相较于基线模型实现了38.36百分点的巨大跃升,达到了84.36%。值得注意的是,BERT的引入不仅大幅提升了召回率,还使得精确率稳定在81.87%,这一结果有力地证明了预训练模型所提供的动态词向量技术能够有效平衡模型的查全率与查准率,显著改善了小样本场景下的实体抽取效果。

表3 Bi-LSTM+CRF和BERT+Bi-LSTM+CRF模型效果评价

Table 3 Experimental results of Bi-LSTM+CRF and BERT+Bi-LSTM+CRF models %

模型 Model	名称 Name	精确率 Precision	召回率 Recall	F_1值 F_1-score
Bi-LSTM+CRF	cEHRNER	61.83	65.59	63.65
	Yidu-S4K	60.97	69.16	64.81
	SSDD	45.48	46.64	46.00
BERT+Bi-LSTM+CRF	cEHRNER	75.29	77.73	76.49
	Yidu-S4K	74.30	82.97	78.40
	SSDD	81.87	87.01	84.36

与Bi-LSTM+CRF和BERT+Bi-LSTM+CRF模型对比,XLNet架构在各数据集上的表现具有较强的竞争力(表4),在SSDD数据集上,XLNet+Bi-LSTM+CRF模型取得了F_1值为85.63%的优异成绩。在Yidu-S4K和cEHRNER这2个数据集上,XLNet模型分别获得79.23%和77.81%的F_1值,显示出XLNet在处理更广泛语料时的稳健性。通过引入XLNet预训练模型,进一步提升了在猪病数据集上的性能,尤其在召回率方面取得良好的表现,表明该模型在识别关键实体方面有着较好的能力,尤其是在数据稀缺的领域中,XLNet具有良好的泛化能力。

表4 XLNet+Bi-LSTM+CRF模型试验结果

Table 4 Experimental results of XLNet+Bi-LSTM+CRF model %

名称 Name	精确率 Precision	召回率 Recall	F_1值 F_1-score
Yidu-S4K	75.21	83.53	79.23
cEHRNER	77.17	78.02	77.81
SSDD	84.17	87.13	85.63

表5展示了本研究提出的ERNIE+Bi-LSTM+CRF模型的试验结果。通过横向对比表3至表5的数据可知,ERNIE模型展现出了最优的综合性能。一方面,在猪病数据集上,ERNIE将F_1值刷新至89.63%,相比BERT模型,其精确率和召回率均有稳步提升,表明模型对猪病语境的理解更加细腻;另一方面,在2个公开数据集上,ERNIE模型同样取得了最优或极具竞争力的结果。这就更有力地证明了ERNIE模型的优势不仅仅局限于特定数据,而是具有强大的跨领域泛化能力,既能处理规范的语料,也能很好地适配包含大量非规范描述的垂直领域文本。

表5 ERNIE+Bi-LSTM+CRF模型试验结果

Table 5 Experimental results of ERNIE+Bi-LSTM+CRF model %

名称 Name	精确率 Precision	召回率 Recall	F_1值 F_1-score
cEHRNER	84.22	79.78	81.94
Yidu-S4K	86.72	85.40	86.05
SSDD	91.68	87.66	89.63

2.2 消融实验

为了进一步探究不同预训练机制对猪病实体识别的具体贡献,并验证本研究选择ERNIE的理论合理性,本研究在命名实体识别试验的基础上,以猪病数据集为核心对象设计了消融对比实验。实验以Bi-LSTM+CRF为基线,分别引入BERT、XLNet和ERNIE 3种主流预训练模型作为增强模块,消融实验结果如表6所示。

首先,预训练模型的引入解决了垂直领域语义稀疏的难题。相比于Baseline模型(F_1值46.00%)仅

表 6　不同预训练模型的消融实验结果对比

Table 6　Ablation study of different pre-trained models　%

模型架构 Model	预训练模块 Pre-trained module	精确率 Precision	召回率 Recall	F_1值 F_1-score
Bi-LSTM+CRF	None	45.48	46.54	46.00
BERT+Bi-LSTM+CRF	BERT	81.87	87.01	84.36
XLNet+Bi-LSTM+CRF	XLNet	84.17	87.13	85.63
ERNIE+Bi-LSTM+CRF	ERNIE	91.68	87.66	89.63

依赖静态字向量，BERT 模型的引入将 F_1 值大幅提升至 84.36%。这表明在大规模语料上训练得到的动态词向量，能够为猪病文本中的生僻术语提供丰富的上下文语义表征，这是提升模型性能的第一关键要素。

其次，在预训练模型的横向对比中，尽管 XLNet 凭借其特有的排列语言建模机制，在处理长距离上下文依赖时优于 BERT（F_1 值提升至 85.63%），但其提升幅度依然有限。究其原因，无论是 BERT 还是 XLNet，其核心掩码任务主要基于字粒度或子词粒度进行预测。在面对猪病领域中大量存在的复合专有名词时，这种细粒度的掩码方式容易割裂实体的完整语义，导致模型只见字，不见义。

反观本研究提出的 ERNIE+Bi-LSTM+CRF 模型，其 F_1 值高达 89.63%，明显优于 BERT 和 XLNet。这种优势主要归功于 ERNIE 独有的实体级和短语级掩码策略。ERNIE 在预训练阶段强制模型对整个实体进行掩码预测，迫使模型学习实体内部字与字之间的强关联性以及实体整体的语义信息。这种机制恰好契合了兽医病理描述中长难术语多、实体边界模糊的语言学特征，从而在召回率和精确率上均实现了对通用预训练模型的全面超越。

2.3　知识图谱构建及存储

在进行知识融合和处理后，所有猪病领域的三元组数据被存储在 Neo4j 本地图数据库平台中。通过使用 py2neo 库创建节点和实体关系图，成功实现了猪病领域知识的可视化图谱。这一步骤不仅提供了高质量的数据用于后续微信问答机器人的应用，同时也通过图谱的可视化展示，使猪病领域的知识结构更加清晰和直观。部分可视化图谱如图 5 所示。

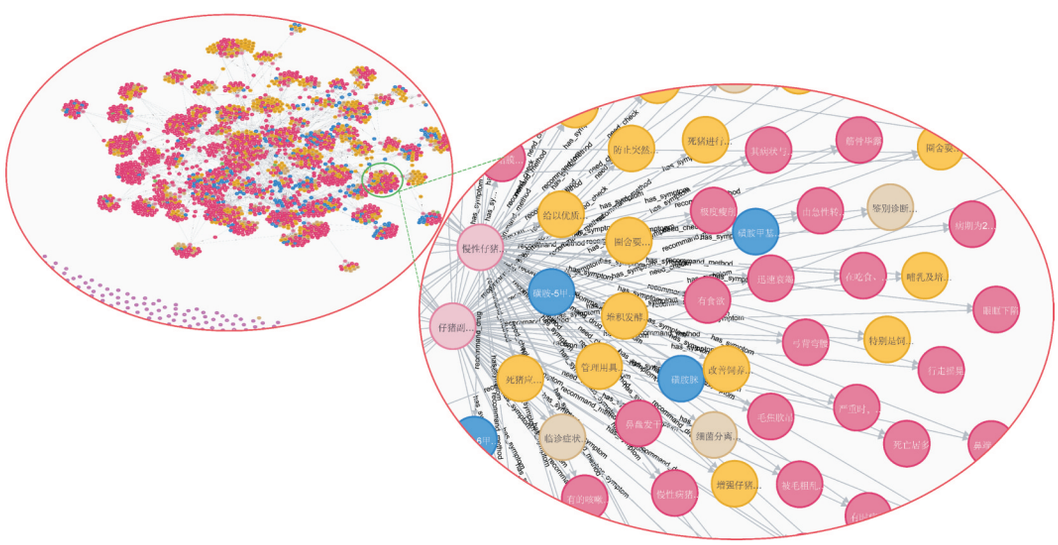


图 5　猪病知识图谱

Fig.5　Knowledge graph of pig disease

2.4　微信猪病问答机器人的应用

基于前文所述的系统架构，本研究构建了面向终端用户的微信猪病智能问答机器人。为了验证系统的实际应用效果，选取生猪养殖中常见的疾病诊断/检查场景进行测试，具体交互实例如图 6 所示。

在图 6 的测试案例中，用户输入自然语言问题："猪流行性感冒怎么检查"。系统后台首先通过意图识别模块准确判断出用户意图为"查询检查项目（check）"，并提取出核心实体"猪流行性感冒（swine influenza）"。随后，系统利用图谱查询接口从 Neo4j 数据库中检索该疾病关联的诊断属性值，并反馈了包含"血清学试验""血凝抑制（HI）检测""酶联免疫吸附试验（ELISA）"及"PCR 检测"等在内的详细诊断方案。



图6 微信问答机器人使用实例

Fig.6 Examples of using WeChat Q&A robot

3 讨 论

本研究旨在探索不同模型在猪病领域实体识别方面的效果，并成功构建了常见猪病的知识图谱、语料库与猪病相关信息服务系统。为生猪养殖户等相关从业人员提供更好的常见猪病预防和治疗支持，帮助养殖户实现科学管理猪群。本研究明确界定了5种关系属性与12种实体类别，提出的ERNIE-Bi-LSTM-CRF模型表现优异，其精确率达到91.68%，F_1值达到89.63%。

与现有同类代表性研究进行横向对比，可以发现本研究成果在核心性能指标上实现了优化。在底层算法维度，陈海辉[7]采用的ALBERT-BiGRU-CRF模型虽然通过参数因式分解降低了计算量，但其在猪病数据集上的F_1值为88.79%，精确率为88.95%。相比之下，本研究的F_1值较其提升了0.84百分点，精确率更是显著提升了2.73百分点。这种性能提升主要归因于ERNIE模型独特的实体级掩码机制，它比ALBERT架构采用的字级掩码策略更能有效捕捉兽医领域长难术语的完整语义，从机理上克服了由于术语复合性强而导致的语义割裂瓶颈。

进一步对比朱思强[8]提出的BERT-CNN-MH-SA-BiLSTM-CRF模型，该模型通过叠加多头自注意力与卷积神经网络等复杂模块获得了91.26%的精确率，仍然低于本研究的精确率(91.68%)。更重要的是，本研究采用的架构更为精简高效，在保证高识别精度的同时降低了计算开销，展现出更强的工程落地价值。此外，在应用层面的集成深度上，它侧重于实现单一任务的图谱展示，且问答功能多依赖于初级的关键词模板映射。而本研究开发的微信端问答机器人深度集成了包含12类诊疗意图的语义解析算法，能够精准理解并转化养殖户多样化的自然语言提问，真正实现了从知识挖掘到实时交互的高效闭环。

但是，本研究的数据规模仍有待进一步扩充，以持续优化模型泛化能力，从而构建更具行业应用价值的猪病知识图谱。未来工作将致力于扩展微信机器人的功能模块，例如引入多模态技术以支持语音识别与图像诊断，从而进一步提升人机交互的便捷性与系统的智能化水平。同时，针对意图不明显或极其复杂的长尾提问，机器人还需增加多轮对话澄清机制，以确保为用户提供更高准确度的决策支持。

参考文献 References

[1] 佟丹丹，曲艳鹏，管鹏宇，等. 我国畜牧兽医类企业人才需求现状调研分析[J]. 黑龙江畜牧兽医，2021(14)：15-22. TONG D D, QU Y P, GUAN P Y, et al. Investigation and analysis on the current situation of talent demand in animal husbandry and veterinary industry enterprises in China[J]. Heilongjiang animal science and veterinary medicine, 2021(14): 15-22 (in Chinese with English abstract).

[2] PUJARA J, MIAO H, GETOOR L, et al. Knowledge graph identification[C]//The Semantic Web-ISWC 2013. Berlin, Heidelberg: Springer, 2013: 542-557.

[3] 漆桂林，高桓，吴天星. 知识图谱研究进展[J]. 情报工程，2017，3(1)：4-25. QI G L, GAO H, WU T X. The research advances of knowledge graph[J]. Technology intelligence engineering, 2017, 3(1): 4-25 (in Chinese with English abstract).

[4] 李岩. 基于知识图谱的禽畜疾病问答系统分析与设计[D]. 石家庄：河北经贸大学，2022. LI Y. Analysis and design of livestock disease question answering system based on knowledge graph[D]. Shijiazhuang: Hebei University of Economics and Business, 2022 (in Chinese with English abstract).

[5] 王雅童. 基于知识图谱的兽药知识问答系统研究与实现[D]. 泰安：山东农业大学，2022. WANG Y T. Research and implementation of veterinary drug knowledge question answering system based on knowledge graph[D]. Tai'an: Shandong Agricultural University, 2022 (in Chinese with English abstract).

[6] 杨喆，许甜，靳哲，等. 基于知识图谱的羊群疾病问答系统的构建与实现[J]. 华中农业大学学报，2023，42(3)：63-70. YANG Z, XU T, JIN Z, et al. Construction and application of knowledge graph of sheep & goat disease[J]. Journal of Huazhong Agricultural University, 2023, 42(3): 63-70 (in Chinese with English abstract).

[7] 陈海辉. 基于Albert的猪病知识图谱构建研究与实现[D]. 合肥：安徽农业大学，2022. CHEN H H. Research and imple-

mentation of swine disease knowledge graph construction based on Albert[D]. Hefei: Anhui Agricultural University, 2022 (in Chinese with English abstract).

[8] 朱思强.面向猪病的命名实体识别及知识图谱构建研究与应用[D].合肥:安徽农业大学,2023. ZHU S Q.Research and application of named entity recognition and knowledge graph construction for pig disease[D].Hefei:Anhui Agricultural University,2023(in Chinese with English abstract).

[9] PENG C,WANG X,LI Q,et al.A new Chinese named entity recognition method for pig disease domain based on lexicon-enhanced BERT and contrastive learning [J/OL].Applied sciences, 2024, 14 (16): 6944 [2025-09-30]. https://doi. org/10.3390/app14166944.

[10] YANG S T,LIU Z H,MAYER W,et al.ShizishanGPT:an agricultural large language model integrating tools andResources[C]//Web Information Systems Engineering-WISE 2024. Singapore:Springer,2025:284-298.

[11] 沈宙锋,苏前敏,郭晶磊.基于XLNet-BiLSTM的中文电子病历命名实体识别方法[J].智能计算机与应用,2021,11(8):97-102.SHEN Z F,SU Q M,GUO J L.Named entity recognition model of Chinese clinical electronic medical record based on XLNet-BiLSTM[J].Intelligent computer and applications, 2021, 11(8):97-102 (in Chinese with English abstract).

[12] 王鑫雷,李帅驰,杨志豪,等.基于预训练语言模型的中文知识图谱问答系统[J].山西大学学报(自然科学版),2020,43(4):955-962.WANG X L,LI S C,YANG Z H,et al.Chinese knowledge base question answering system based on pre-trained language model[J].Journal of Shanxi University (natural science edition), 2020, 43(4):955-962 (in Chinese with English abstract).

[13] JAFFE A,KLUGER Y,LINDENBAUM O,et al.The spectral underpinning of word2vec [J/OL]. Frontiers in applied mathematics and statistics, 2020, 6[2025-09-30].https://doi.org/10.3389/fams.2020.593406.

[14] KAUSHIK S,CHOUDHURY A,SHERON P K,et al.AI in healthcare:time-series forecasting using statistical,neural,and ensemble architectures[J/OL].Frontiers in big data,2020,3:4 [2025-09-30].https:// doi: 10.3389/fdata.2020.00004

[15] ABSAR N,UDDIN N,KHANDAKER M U,et al.The efficacy of deep learning based LSTM model in forecasting the outbreak of contagious diseases[J].Infectious disease modelling, 2022,7(1):170-183.

[16] SKARAMAGKAS V,PENTARI A,KEFALOPOULOU Z, et al.Multi-modal deep learning diagnosis of Parkinson's disease:a systematic review[J].IEEE transactions on neural systems and rehabilitation engineering,2023,31:2399-2423.

[17] CATELLI R,CASOLA V,DE PIETRO G,et al.Combining contextualized word representation and sub-document level analysis through Bi-LSTM+CRF architecture for clinical de-identification[J/OL].Knowledge-based systems, 2021, 213: 106649 [2025-09-30]. https://doi. org/10.1016/j. knosys.2020.106649.

[18] DERNONCOURT F,LEE J Y,UZUNER O,et al.De-identification of patient notes with recurrent neural networks [J]. Journal of the American medical informatics association,2017, 24(3):596-606.

[19] HUANG Z H,XU W,YU K.Bidirectional LSTM-CRF models for sequence tagging [PP/OL]. arXiv, 2015: 1508.01991 [2025-09-30].https://arxiv.org/abs/1508.01991.

[20] LAMPLE G,BALLESTEROS M,SUBRAMANIAN S,et al.Neural Architectures for Named Entity Recognition [C]// KNIGHT K,NENKOVA A,RAMBOW O.Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.San Diego: Association for Computational Linguistics,2016: 260-270.

[21] ZHAO Y S,ZHANG K L,MA H C,et al.Leveraging text skeleton for de-identification of electronic medical records[J/OL].BMC medical informatics and decision making,2018,18(1):18[2025-09-30]. https://doi.org/10.1186/s12911-018-0598-6.

[22] BAHDANAU D,CHO K,BENGIO Y.Neural machine translation by jointly learning to align and translate[PP/OL]. arXiv, 2014: 1409.0473 [2025-09-30]. https://arxiv. org/abs/1409.0473.

[23] MNIH V,HEESS N,GRAVES A,et al.Recurrent models of visual attention [C]// Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2. Red Hook, NY, USA: Curran Associates, Inc., 2014: 2204-2212.

[24] 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 North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics,2019:4171-4186.

[25] YANG Z,DAI Z,YANG Y,et al.XLNet: generalized autoregressive pretraining for language understanding [C]// Proceedings of the 33rd International Conference on Neural Information Processing Systems.Vancouver,Canada: Curran Associates,Inc.,2019: 5753-5763.

[26] SUN Y,WANG S H,LI Y K,et al.ERNIE:enhanced representation through knowledge integration [PP/OL]. arXiv, 2019: 1904.09223 [2025-09-30]. https://arxiv. org/abs/1904.09223.

[27] NEO4J INC. Neo4j graph data platform [CP/OL]. San Mateo, California: Neo4j, Inc., 2022 [2025-09-30]. https://neo4j.com.

Pre-trained model-driven construction and application of a swine disease knowledge graph

WEN Jiwei[1], YANG Zhe[2,3], WANG Cheng[1], JIN Zhe[1], DENG Bing[3],
LI Guoliang[1], FANG Yaping[1], LIU Feng[1], DU Xiaoyong[1]

1.*College of Informatics, Huazhong Agricultural University/ Key Laboratory of Smart Farming for Agricultural Animals, Ministry of Agriculture and Rural Affairs/ Engineering Research Center of Agricultural Intelligent Technology, the Ministry of Education/Hubei Engineering Technology Research Center of Agricultural Big Data, Wuhan 430070, China*;
2.*Key Laboratory of Agricultural Animal Genetics, Breeding and Reproduction of Ministry of Education, Wuhan 430070, China*;
3.*Institute of Animal Husbandry and Veterinary Sciences, Wuhan Academy of Agricultural Sciences, Wuhan 430200, China*

Abstract To enhance the efficiency of acquiring knowledge on swine disease prevention and control and to alleviate the shortage of veterinary service resources at the grassroots level, this study develops a vertical domain knowledge graph and an intelligent question-answering system specifically for swine disease prevention and control. To resolve descriptive conflicts in multi-source heterogeneous data, the study establishes a domain ontology based on veterinary pathology and introduces a BERT-based semantic similarity fusion strategy to construct a professionally annotated corpus. Given the complex and long-tail distribution of veterinary terms, we propose an ERNIE-Bi-LSTM-CRF entity recognition model. Named entity recognition experiments and ablation tests show that ERNIE's unique entity-level masking mechanism effectively overcomes the semantic fragmentation problem caused by traditional character-level masking of complex terms. The F_1 score reaches 89.63%, significantly outperforming baseline models such as Bi-LSTM+CRF and BERT. Based on this, the study integrates the Neo4j graph database and intent recognition technology to develop a WeChat-based question-answering chatbot. Tests show that the chatbot can accurately understand professional veterinary diagnosis and treatment issues, providing comprehensive, and scientifically sound diagnostic guidance. This effectively assists farmers in disease diagnosis. The results indicate that the developed system can achieve a closed-loop process, from natural language queries to knowledge retrieval within the knowledge graph, thereby offering technical support for swine disease diagnosis and prevention knowledge services.

Keywords knowledge graph; pig diseases; BERT; XLNet; ERNIE; Wechat robot

(责任编辑:边书京)