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.
DOI:10.13300/j.cnki.hnlkxb.2023.03.008

基于知识图谱的羊群疾病问答系统的构建与实现

杨喆[1,2],许甜[1],靳哲[1],孔玫[1],李国亮[1],杜小勇[1,2]

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.华中农业大学动物科学技术学院/农业动物遗传育种与繁殖教育部重点实验室,武汉 430070

摘要 为解决羊群疾病检索过程中出现的大量冗余数据及检索后仍需人工挑选准确答案造成的资源浪费,本研究通过以下3个步骤构建基于知识图谱的羊群疾病问答系统:(1)通过爬虫获取数据,人工提取部分信息,再进行自动化信息抽取,在命名实体识别任务中使用双向长短期记忆循环神经网络Bi-LSTM模型,并添加注意力机制提高识别效率,然后使用BIO规则进行实体标注,完成信息抽取,将数据融合后存储在Neo4j图数据库中,构建羊群疾病知识图谱。(2)针对属性映射,构建Bert-softmax模型;根据用户提问,采用Bert模型计算问句和属性的语义相似度,并通过softmax算法进行归一化处理,返回合适答案给用户,实现羊群疾病问答系统算法设计。(3)构建羊群疾病诊断平台,使用Bootstrap、Echarts、Vue组件实现羊群疾病问答系统的可视化,利用Python语言包含的flask框架搭建后台,封装疾病信息,通过web前端呈现给用户,并于后端建立连接,实现数据之间的交互。试验结果显示,基于Bi-LSTM + Attention + CRF模型实体识别的F_1值为83.16%,构建的知识图谱包含实体4 576个,实体关系超13 000条;问答系统添加了预训练模型Bert,对问题识别的F_1值为85.24%。结果表明,该系统实现了对羊群疾病的防治措施等多类问题进行快速检索和精准回答,可以辅助养殖人员在面临羊群疾病时进行生产决策。

关键词 疾病诊断;知识图谱;问答系统;Neo4j;Bert;智慧养殖

中图分类号 TP391 **文献标识码** A **文章编号** 1000-2421(2023)03-0063-08

羊群疾病的诊断和预防一直是养殖人员工作的重点,但专业的畜牧兽医人才的缺乏也是行业面临的问题。因此,帮助从业人员进行疾病诊断和预防的工具及系统受到了广泛的关注。互联网中查询羊群疾病的方式逐渐被用户接受,但由于自然语言(natural language)描述的不准确性和关键词的多样性使得检索结果容易出现偏差,且检索出的答案通常以网页的形式呈现,还需要人工进行挑选才能获得准确的答案。面向专业领域的问答系统能够在很大程度上解决互联网中数据冗余的困境,可以针对目标领域更专业地回答用户提出的问题。早期的问答系统主要以专家诊断系统为主,如陈勇等[1]将290余种不同类型的羊病数据通过产生式规则建成知识库,用作羊病辅助诊断。这种知识库依赖于专家构建规则进行推理问答,需要人工构建海量的规则将用户的问题进行匹配推理。李驰航[2]和聂艳召[3]在已有基于实例汇总的诊断平台基础上进行了知识表示上的改进,但由于不论是基于规则还是基于实例推理的方式构建问答系统,仍需要大量的人工进行规则构建和实例筛选,在大批量的数据面前这种方式便显得略有逊色。现有专家系统的回复慢、检索方式复杂度高、使用方式对一线养殖人员不友好。当羊群遭遇疫病时,为了不耽误动物的病情,使用更加高效的检索方式和系统工具势在必行。

谷歌公司为解决用户的搜索体验欠佳和搜索质量不高的问题,进行了大量的研究,在2012年正式提出了知识图谱(knowledge graph)。知识图谱的出现解决了用户关键词把握不准确导致的检索失败及用户无法进行多轮对话查询等问题[4]。如今,知识图谱的应用领域已经相当广泛,尤其是在医疗疾病问答

收稿日期:2022-12-05
基金项目:国家自然科学基金项目(31872978)
杨喆,E-mail:1148713279@qq.com
通信作者:杜小勇,E-mail:duxiaoyong@mail.hzau.edu.cn

领域贡献颇多[5-8]，但有关动物（例如：羊）疾病的知识图谱却寥寥无几。有学者对畜禽疾病领域的问答系统和信息抽取方式进行了探索，如王雅童等[9]完成了以知识图谱为基础的兽药说明书信息问答系统，该系统可以辅助使用者获知兽药信息，防止造成兽药滥用、误用的现象，但关于疾病的防治措施和疾病的相关知识等方面的问答还存在不足。李岩[10]使用问句模板匹配方法构建的畜禽疾病问答系统，在限定领域的问答中可以获得较为准确的答案，但其数据收集方式和问答推理能力还有改进的空间。因此，构建针对羊群疾病的知识图谱对行业发展有积极作用，一方面可以补充羊群疾病知识库的数据，另一方面可以为羊群疾病智能诊断问答系统提供精确的数据来源，辅助养殖人员在羊群疾病发生时进行生产决策。本研究对互联网中羊群疾病数据进行收集，使用Bi-LSTM ＋ Attention ＋ CRF模型将数据中的知识进行信息抽取，将数据存储在Neo4j图数据库中，构建可视化的羊群疾病知识图谱，设计并实现了基于知识图谱的羊群疾病问答系统。

1 材料与方法

1.1 数据收集与预处理

首先采用人工的方式，从互联网的文本信息中手动抽取出疾病实体和实体关系；然后，将这些实体及关系整理为“实体-关系-实体”的三元组形式存放在Excel表格中，最后将Excel文件转换为json格式用于后续的操作。同时还使用网络爬虫的方式从羊群疾病网站（医链）上获取文本信息，将这一部分数据进行实体标注，作为Bi-LSTM-CRF命名实体识别模型的训练数据集。此外，还收集了“中国知网”获取的3 010篇羊病相关文献摘要以及CCL2021智能对话诊疗评测比赛中的数据[11]，将中国知网获取的数据和CCL2021中的数据整理为命名实体识别数据集的一部分。最终，经过数据预处理后获取医链数据4 832条、中国知网数据1 000条及CCL2021数据2 000条以上，作为构建知识图谱及问答系统的基础数据。

在对疾病数据进行命名实体识别之前，需要构建语料库，协助模型进行实体标注[12]。使用已经标注的数据集对添加了Attention机制的Bi-LSTM-CRF模型进行训练，接着用Bi-LSTM ＋ Attention ＋ CRF模型对摘要部分的文本数据进行实体识别，然后通过人工校验的方式对实体进行筛选，确认和人工标注疾病实体，并构建出更多的实体关系，基于Bi-LSTM ＋ Attention ＋ CRF模型的命名实体识别流程如图1所示。

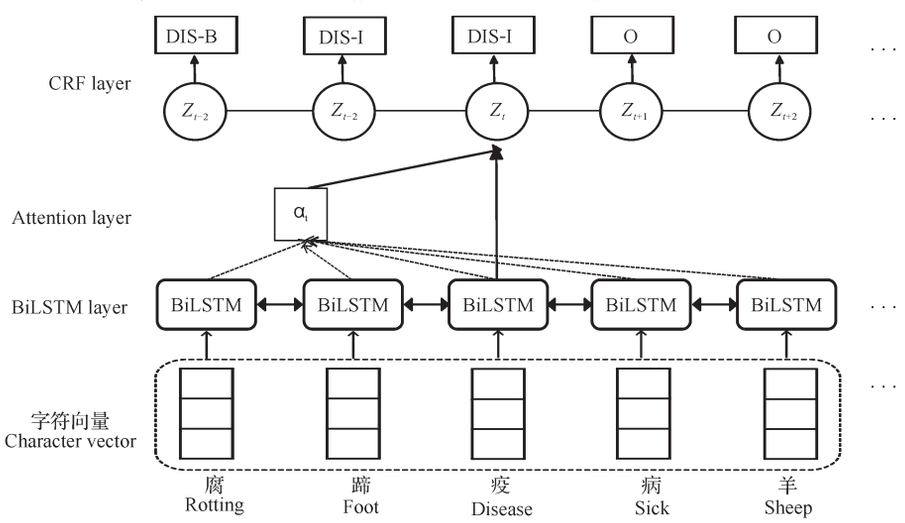

图1 命名实体识别流程图

Fig. 1 Named entity recognition flowchart

1.2 知识图谱的构建

首先收集好数据，将数据进行信息抽取、映射和融合，然后进行知识存储，形成知识图谱，知识图谱构建步骤如图2所示。

1）机器学习及深度学习模型。在命名实体识别任务中用LSTM和CRF等模型进行实体识别，LSTM可以简单理解为更复杂的循环神经网络（RNN）[13]，条件随机场（conditional random field，

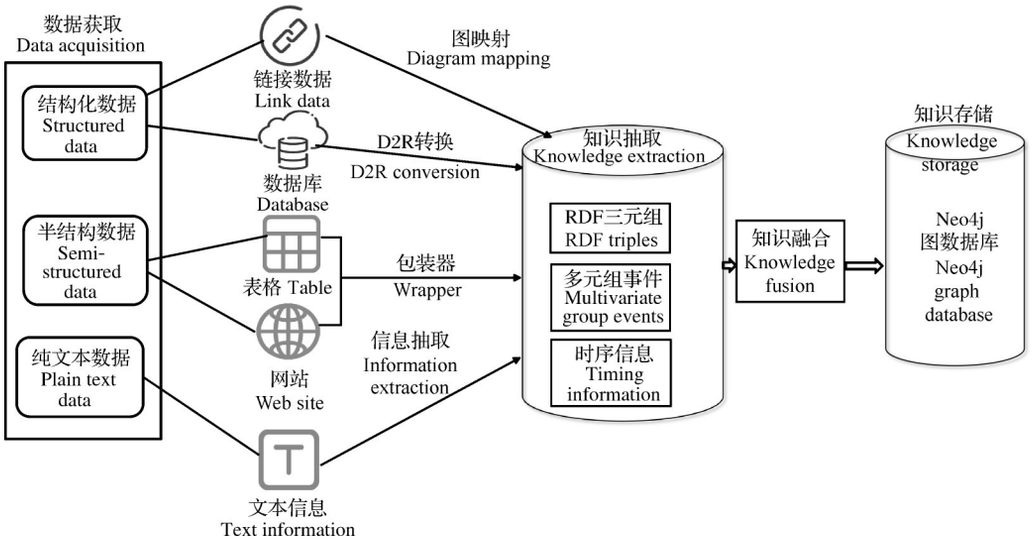

图 2　知识图谱构建步骤

Fig. 2　Knowledge graph construction steps

CRF)可以定义为在同一区域内分布了若干个参数，然后分别对每个参数进行随机赋值。CRF 融合 2 组学习方向的双向长短期记忆网络(Bi-LSTM)被广泛应用于知识图谱的构建中[14]。CRF 结合隐马尔科夫模型和最大熵模型[15]的优点，在序列标记(如：实体标注、命名实体识别)等方面取得了良好的效果。CRF 模型采用了“BIO”的命名规则对输入的句子进行标注。“B”标注的是实体概念的头，“I”标注的是实体概念的中间部分，“O”标注的是实体概念的尾，三者之间的顺序不能颠倒。这也就意味着 CRF 模型能利用这一特性限制序列的前后关系，每条被标注的序列都能当作一条路径，并且能够求出标记序列中的最大路径长度的概率。Attention 机制[16-17]基于对输入权重的关注，添加到 Bi-LSTM-CRF 模型中，更简单地让模型学习到句子语义中的重点。本研究在序列标注的任务中，用双向 LSTM(Bi-LSTM)模型作为支持，完成命名实体识别工作，前向的 LSTM 能够传递上文的信息[18]，后向的 LSTM 能够保存下文的信息，两者相互结合使得模型能够学习文本的上下文信息，从而更好地进行序列标注。Bi-LSTM 模型结构图如图 3 所示。

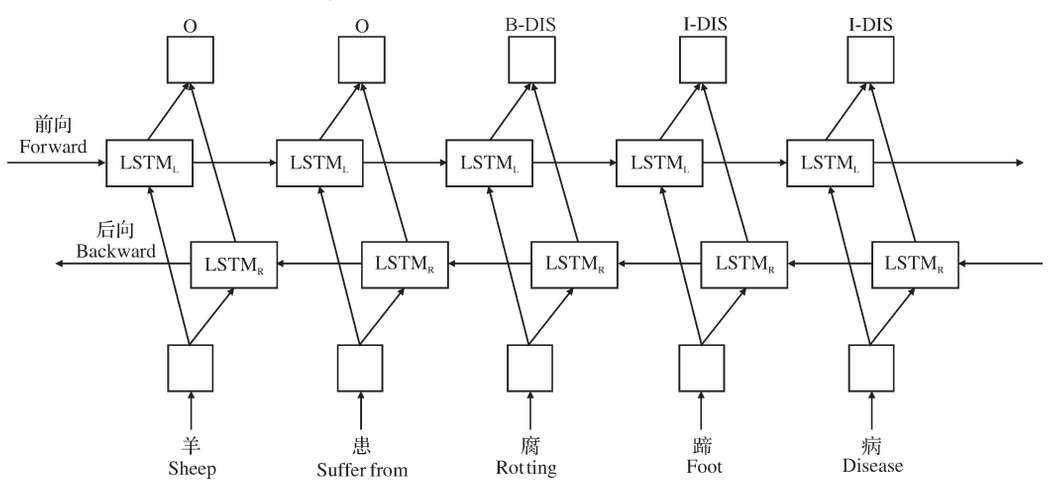

图 3　Bi-LSTM 模型结构图

Fig. 3　Bi-LSTM model structure diagram

实体命名识别模型的效果通过精确率(P)、召回率(R)以及 F 值(也被称为 F_1 值，即 P 值和 R 值的调和平均数)的大小来进行评价，其公式如(1)、(2)、(3)所示。本研究基于 Bi-LSTM 模型＋Attention 机制＋CRF 模型对“中国知网”获取以“羊病”为关键词的中文文献摘要部分进行命名实体识别，且进行了对照试

验来分析Bi-LSTM-CRF与CRF模型存在的差异。

$$P=\frac{\text{正确识别的实体数目}}{\text{识别出的实体总数}} \quad (1)$$

$$R=\frac{\text{正确识别出的实体数目}}{\text{人工标注的总实体数}} \quad (2)$$

$$F_1=(2\times P\times R)/(P+R) \quad (3)$$

2)Neo4j图数据库存储。本研究使用Neo4j图数据库进行数据存储，Neo4j是NoSQL数据库中最具代表性的高质量图数据库[19]。使用图数据库Neo4j存储羊病的数据不仅可以更直观地了解疾病和诊断方式及预防措施等关联信息，在问答系统的实现中也能够起到简化查询语句的作用。

1.3 基于知识图谱的羊群疾病问答系统算法设计

智能问答系统的设计可以将知识图谱中丰富的结构化语义信息合理地应用，使得人机之间的交互更加高效[20]。用户可以根据自己的意图来针对系统进行提问。系统理解用户的问句后，在语义库中进行查询，最终将最佳答案反馈给用户。本系统的设计和实现后端采用Python语言和flask框架编写，前端采用JavaScript语言和Vue框架。由于Python具备更加高级的数据结构，开发应用速度较快，并且具备很强大的扩展性，因此，被广泛地应用到NLP任务中。后端服务器选择用flask框架搭建，一方面是由于扩展性较好，另一方面是相比较于Django框架灵活度更高且更容易上手。JavaScript是一种解释型语言，以跨平台、简单且容易上手的优势被广泛地应用到网页开发中，从而提高网页的交互能力。Vue框架提供了双向数据绑定原理，可以更好地维护系统的数据，并且使用虚拟DOM。当数据发生改变时，能够减少页面重新渲染的次数，从而降低资源上的浪费。

智能问答系统基于命名实体识别和属性映射两部分实现，整体设计流程如图4所示。

针对用户在系统中的提问，如："羊发烧感冒能吃什么药？"。首先确定疾病实体和属性，然后在知识图谱中进行匹配。"感冒"这个实体通过Bi-LSTM+CRF模型命名实体识别得到，属性通过Bert模型得到。Bert模型可以同时输入A、B两个句子，得到它们的相似度在0～1之间，并根据对应实体取出相似度最高的属性作为结果。

2 结果与分析

2.1 命名实体识别效果

本研究对比了Bi-LSTM-CRF模型和CRF模型

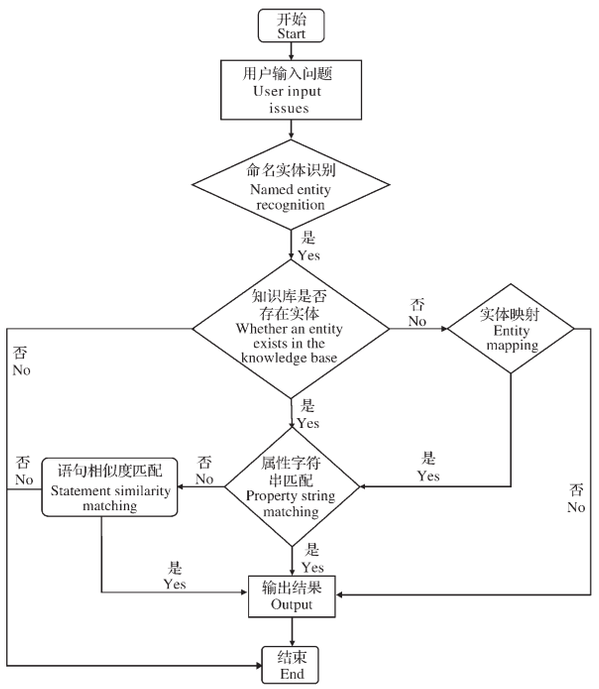

图4 羊群疾病问答算法整体设计流程图

Fig. 4 Flow chart of overall design of question answering algorithm for herd disease

对同一数据集的命名实体识别效果，试验结果如表1所示，Bi-LSTM-CRF模型对本数据集的命名实体识别的效果更佳，相比于CRF模型，其准确率、召回率和F_1值都有了明显提高。

表1 2种不同命名实体识别的模型效果评价

Table 1 Evaluation of two different named entity recognition %

方法 Method	P	R	F_1
CRF	76.16	86.64	81.06
Bi-LSTM-CRF	78.75	88.06	83.16

2.2 知识图谱数据结果

将数据用训练好的Bi-LSTM-CRF模型对采集的数据进行自动标注，通过该模型，最终得到疾病名称实体352个，疾病症状实体3 206个以及治疗方案实体1 429个。通过信息抽取和知识融合等步骤，最终搜集到的数据见表2～3。

将抽取好的数据使用Python逻辑代码转化为json格式，通过读取json文件路径，建立Neo4j数据库连接。对json文件中保存的数据进行判断，将判断后的数据加入到对应的列表中；创建知识图谱的节点和边，并将实体、关系、属性和属性值等信息存储到图数据库Neo4j中。可视化的羊病知识图谱的部分实体和边如图5所示。

表 2　羊群疾病知识图谱实体类型

Table 2　Sheep disease knowledge graph entity type

实体类型 Entity types	实体数量 The number of entities	举例 Example
诊断结果 Check	914	口腔、蹄部皮肤出现水疱和溃烂 Blisters and ulcers on the skin of the mouth and hooves
疾病类别 Department	8	传染病 Infectious disease
疾病名称 Disease	266	口蹄疫 Foot-and-mouth disease
药物 Drug	750	青霉素软膏，磺胺软膏 Penicillin ointment，sulfonamide ointment
预防措施 Method	698	保护粘膜、皮肤，做好检疫消毒 Protect mucous membrane，skin，do quarantine disinfection
疾病症状 Symptom	1 940	唇部出现小红斑，口角出现水疱 Small erythema appeared on the lips and blisters appeared at the corner of the mouth
总计 Total	4 576	——

表 3　羊群疾病实体关系类型

Table 3　Sheep disease entity relationship types

实体关系类型 Entity relationship types	举例 Example
属于 Belongs_to	＜口疮，属于，传染病＞ ＜Oral sore，belongs_to，infectious disease＞
所需诊断 Need_check	＜口疮，所需诊断，口腔皮肤出现水疱＞ ＜Oral sore，need_check，blisters appear on the oral skin＞
推荐药物 Recommend_drug	＜口疮，推荐药物，碘甘油＞ ＜Oral sore，recommend_drug，iodine glycerin＞
预防措施 Recommend_method	＜口疮，预防措施，做好检疫消毒＞ ＜Oral sore，recommend_method，do quarantine disinfection＞
疾病症状 Has_symptom	＜口疮，疾病症状，唇部出现小红斑＞ ＜Oral sore，has_symptom，small red spots appear on the lipsn＞
疾病症状并发疾病 Accompany_with	＜跛行，发生感染，并发疾病，腐蹄病＞ ＜Lameness，infection，accompany_with，foot rot＞

2.3　羊群疾病问答系统

本研究通过语料库生成 4 000 个问答语句（如图 6 所示），作为 Bert 模型训练的数据集。首先，将数据集按照 9∶1 的比例分为训练集和测试集（测试集由“问题＋属性＋标记”组成），针对问句将正确的答案属性标记为 1；然后，从属性列表中抽取其中的 5 个作为负极属性（即错误的属性）；最后，放入 Bert 模型中进行训练。根据输入的问句，采用 Bi-LSTM＋CRF 模型进行命名实体识别，并在知识库中进行实体检索；然后，使用 Bert 模型进行属性映射；最终，将属性相似度最高的答案返回给用户。通过试验数据分析可知，在问答模块中添加 Bert 模型，可以显著提高羊群疾病系统的问答能力，其 F_1 值达到 85.24%，未增加前为 83.16%。

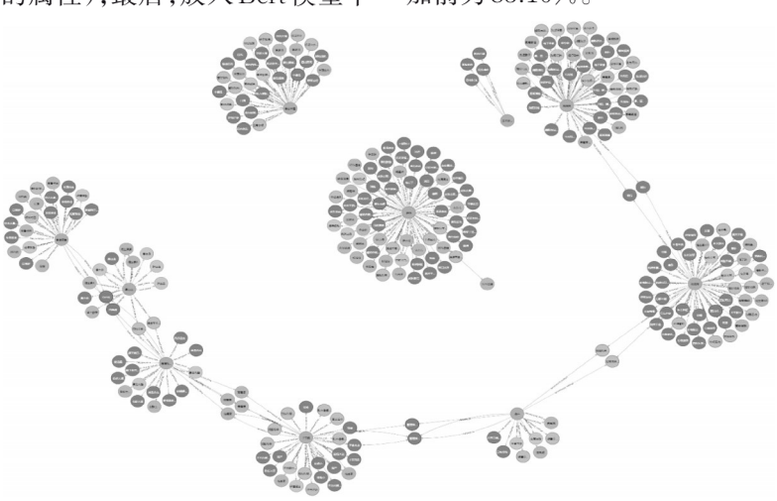

图 5　羊群疾病知识图谱中 Neo4j 实体和关系可视化图

Fig. 5　Neo4j entity and relationship visualization diagram of sheep disease knowledge graph

羊群疾病问答系统可以实现以下 4 个方面的问答：根据症状诊断病情、根据疾病回答疾病症状、根据疾病描述给出治疗方案及根据疾病查询预防措施。例如，用户输入“羊痘的症状有哪些？”，系统则会返回结果，如图 7 所示。

<question id=1>羊痘属于哪一类疾病呢? What kind of disease does sheeppox belong to?
<triple id=1>羊痘 Sheeppox|||疾病 Disease|||传染病 Contagious disease
<answer id=1>传染病 Contagious disease

<question id=2>羊破伤风治疗周期为多久呢 How long is the treatment cycle for sheep tetanus?
<triple id=2>羊破伤风 Sheep tetanus|||治疗周期 Treatment cycle|||4~6 d 4–6 days
<answer id=2> 4~6 d 4–6 days

<question id=3>羊副结核病易感羊群有哪些呢? What are the sheep susceptible to tuberculosis?
<triple id=3>羊副结核病 Sheep paratuberculosis|||易感羊群 Susceptible flocks|||幼龄羊 Young sheep
<answer id=3>幼龄羊 Young sheep

图6 数据集样例图

Fig. 6 A case diagram of sample dataset

图7 羊群疾病问答系统中的问答示例

Fig. 7 Q & A example of sheep disease knowledge graph

3 讨 论

本研究构建了羊群疾病的知识图谱，并进行了问答系统的应用。重点采用Bi-LSTM-CRF模型进行命名实体识别，并添加Attention机制，增加关键词的权重，从而实现了高质量的自动抽取过程。羊群疾病问答系统通过Bert模型进行属性映射或者语句相似度计算，实现了比较复杂的问答场景应用。羊群疾病诊断平台通过Web端的形式进行可视化，并通过发送ajax请求，实现了问答系统的交互。用户可以通过该平台查询羊群疾病、治疗方法、预防措施等信息。本研究通过不同来源的数据构建羊群疾病的知识图谱，一方面，拓宽了中文知识图谱的领域范围，将知识图谱应用于畜牧行业垂直领域，为畜牧行业数字化、智能化提供思路。另一方面，填补了羊病知识图谱的空缺，不仅可以为羊群疾病在线诊断平台提供数据支撑，还可以为动物疾病库提供结构化的数据来源，有助于养殖人员科学地管理羊群。此外，在对“中国知网”文献摘要部分进行知识抽取过程中构建了Bi-LSTM-CRF模型，并与模型CRF对比进行效果评价，发现Bi-LSTM-CRF进行实体标注的准确率较高。

在问答系统方面，先分析用户需求及系统需求，然后进行系统总体架构设计，结合图谱构建技术和疾病问答算法，搭建了一个羊群疾病智能诊断平台。其中系统需求被细分为系统功能设计、数据库设计、系统部署和系统功能展示几个部分。平台具有羊群疾病知识构建、羊群疾病检索功能、羊群疾病总体内容和羊群疾病关系问答4个模块，数据层绑定在flask框架中，使用py2neo工具包对Neo4j数据库进行访问，前端页面的设计使用HTML标签，并且结合CSS进行调整。使用JavaScript语言结合Vue框架实现与用户的交互。与传统基于规则匹配的问答系统不同，本系统使用了Bert等模型，通过计算语义和属性之间的相似度，在复杂的问答情景下提高问答的泛化能力，将相似度最高的结果作为答案反馈给用户，从而实现了基于神经网络的智能问答。虽然本研究构建了近2万个实体及关系，但要想成为更加专业的问答系统仍有扩充数据集的必要。此外，问答系统的应用效果也还有进步的空间，未来增加用户及专家反馈模块将更便于系统的优化。

参考文献 References

[1] 陈勇，李书琴，张平. 羊病诊断与防治专家系统的研制与应用[J]. 动物医学进展，2003，24(5)：61-64. CHEN Y，LI S Q，

ZHANG P. Development & application of computer diagnosis, prevention and treatment expert system on sheep and goat disease [J]. Progress in veterinary medicine, 2003, 24(5): 61-64 (in Chinese with English abstract).

[2] 李驰航. 基于Web的羊病诊断专家系统关键技术的研究[D]. 杨凌：西北农林科技大学, 2009. LI C H. The research on key technologies of sheep & goat diagnosis expert system based on Web[D]. Yangling: Northwest A & F University, 2009 (in Chinese with English abstract).

[3] 聂艳召. 基于案例推理的羊病诊断专家系统研究与实现[D]. 杨凌：西北农林科技大学, 2007. NIE Y Z. The study and implementation of sheep & goat diagosis expert system based on CBR[D]. Yangling: Northwest A & F University, 2007 (in Chinese with English abstract).

[4] SINGHAL A. Oficial google blog: introducing the knowledge graph: things, not strings[EB/OL]. [2012-05-16]. https://blog.google/products/search/introducing-knowledge-graph-things-not/.

[5] ABACHA A B, ZWEIGENBAUM P. MEANS: a medical question-answering system combining NLP techniques and semantic Web technologies [J]. Information processing & management, 2015, 51(5): 570-594.

[6] GOODWIN T R, HARABAGIU S M. Medical question answering for clinical decision support [C]//Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. New York: ACM, 2016: 297-306.

[7] SHI L X, LI S J, YANG X R, et al. Semantic health knowledge graph: semantic integration of heterogeneous medical knowledge and services[J/OL]. BioMed research international, 2017, 2017: 2858423[2022-12-05]. https://doi.org/10.1155/2017/2858423.

[8] KUHN M, LETUNIC I, JENSEN L J, et al. The SIDER database of drugs and side effects[J]. Nucleic acids research, 2016, 44(1): 1075-1079.

[9] 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).

[10] 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).

[11] CHEN W, LI Z W, FANG H Y, et al. A benchmark for automatic medical consultation system: frameworks, tasks and datasets [J/OL]. Bioinformatics, 2023, 39(1): btac817 [2022-12-05]. https://doi.org/10.1093/bioinformatics/btac817.

[12] JAFFE A, KLUGER Y, LINDENBAUM O, et al. The spectral underpinning of word2vec[J/OL]. Frontiers in applied mathematics and statistics, 2020, 6: 593406[2022-12-05]. https://doi.org/10.3389/FAMS.2020.593406.

[13] SHERSTINSKY A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network [J/OL]. Physica D: nonlinear phenomena, 2020, 404: 132306 [2022-12-05]. https://doi.org/10.1016/j.physd.2019.132306.

[14] SIAMI-NAMINI S, TAVAKOLI N, NAMIN A S. The performance of LSTM and BiLSTM in forecasting time series [C]//2019 IEEE International Conference on Big Data (Big Data). Los Angeles, CA, USA: IEEE, 2020: 3285-3292.

[15] 李素建, 王厚峰, 俞士汶, 等. 关键词自动标引的最大熵模型应用研究[J]. 计算机学报, 2004, 27(9): 1192-1197. LI S J, WANG H F, YU S W, et al. Research on maximum entropy model for keyword indexing [J]. Chinese journal of computers, 2004, 27(9): 1192-1197 (in Chinese with English abstract).

[16] MNIH V, HEESS N, GRAVES A, et al. Recurrent models of visual attention [DB/OL]. arXiv, 2014: 1406.6247. https://doi.org/10.48550/arXiv.1406.6247.

[17] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate [DB/OL]. arXiv, 2014: 1409.0473. https://doi.org/10.48550/arXiv.1409.0473.

[18] GREFF K, SRIVASTAVA R K, KOUTNÍK J, et al. LSTM: a search space odyssey [J]. IEEE transactions on neural networks and learning systems, 2017, 28(10): 2222-2232.

[19] VUKOTIC A, WATT N, ABEDRABBO T, et al. Neo4j in action[M]. Manning: Manning Publications Co., 2015: 14-17.

[20] TURE F, JOJIC O. Ask your TV: real-time question answering with recurrent neural networks [C]//Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. New York: ACM, 2016: 457-458.

Construction and application of knowledge graph of sheep & goat disease

YANG Zhe[1,2], XU Tian[1], JIN Zhe[1], KONG Mei[1], LI Guoliang[1], DU Xiaoyong[1,2]

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.*College of Animal Sciences & Technology, Huazhong Agricultural University/Key Laboratory of Agricultural Animal Genetics, Breeding and Reproduction of Ministry of Education, Wuhan* 430070, *China*

Abstract In order to solve the problem of a large amount of redundant data in the retrieval process of sheep disease and the waste of resources caused by manual selection of accurate answers after retrieval, this study constructed a question-and-answer system based on the knowledge graph of sheep & goat disease through the following three steps: (1) The data was obtained through web crawlers and some is manually extracted, automated information extraction was carried out using the bidirectional long short-term memory recurrent neural network (Bi-LSTM) model with an attention mechanism for improved recognition efficiency in the named entity recognition task. The entity annotation was performed using the BIO rule to complete the information extraction. The data was then integrated and stored in the Neo4j graph database . (2) For the attribute mapping, we constructed the Bert-softmax model; according to the user's question, the Bert model was used to calculate the semantic similarity between the question and the attribute to determine the user's intention, then the softmax algorithm was used to for normalization, finally, the most suitable answer was found and fed back to the system. (3) We built a sheep & goat disease diagnosis platform using Bootstrap, Echarts, and Vue components to visualize the sheep disease question-and-answer system. We used flask framework included in the Python language to build a backend, encapsulate disease information, present it to users through the web frontend, and establish a connection on the backend to enable data interaction. The results in the study show that the F_1 value of entity recognition based on Bi-LSTM + Attention + CRF model is 83.16%, and the constructed knowledge graph contains 4 576 entities and more than 13 000 entity relationships. The pre-trained model Bert was added to the question answering system, and the F_1 value of problem recognition was 85.24%. The results indicated that the system can quickly retrieve and accurately answer various types of questions such as the prevention and control measures of sheep diseases, and assist the farmers to make production decisions when faced with sheep diseases.

Keywords disease diagnosis; knowledge graph; question answering system; Neo4j; Bert; smart farming

(责任编辑:边书京)