doi:10.6041/j.issn.1000-1298.2024.02.028

基于 BERT－BiLSTM－CRF 模型的畜禽疫病文本分词研究

余礼根[1,2]　郭晓利[3]　赵红涛[3]　杨　淦[1,2]　张　俊[1,2]　李奇峰[1,2]

（1.北京市农林科学院信息技术研究中心，北京 100097；2.国家数字畜牧业创新中心，北京 100097；
3.华北电力大学数理学院，北京 102206）

摘要：针对畜禽疫病文本语料匮乏、文本内包含大量疫病名称及短语等未登录词问题，提出了一种结合词典匹配的 BERT－BiLSTM－CRF 畜禽疫病文本分词模型。以羊疫病为研究对象，构建了常见疫病文本数据集，将其与通用语料 PKU 结合，利用 BERT（Bidirectional encoder representation from transformers）预训练语言模型进行文本向量化表示；通过双向长短时记忆网络（Bidirectional long short-term memory network，BiLSTM）获取上下文语义特征；由条件随机场（Conditional random field，CRF）输出全局最优标签序列。基于此，在 CRF 层后加入畜禽疫病领域词典进行分词匹配修正，减少在分词过程中出现的疫病名称及短语等造成的歧义切分，进一步提高了分词准确率。实验结果表明，结合词典匹配的 BERT－BiLSTM－CRF 模型在羊常见疫病文本数据集上的 F1 值为 96.38%，与 jieba 分词器、BiLSTM－Softmax 模型、BiLSTM－CRF 模型、未结合词典匹配的本文模型相比，分别提升 11.01、10.62、8.3、0.72 个百分点，验证了方法的有效性。与单一语料相比，通用语料 PKU 和羊常见疫病文本数据集结合的混合语料，能够同时对畜禽疫病专业术语及疫病文本中常用词进行准确切分，在通用语料及疫病文本数据集上 F1 值都达到 95%以上，具有较好的模型泛化能力。该方法可用于畜禽疫病文本分词。

关键词：畜禽疫病；文本分词；预训练语言模型；双向长短时记忆网络；条件随机场

中图分类号：TP391.1　　**文献标识码**：A　　**文章编号**：1000-1298(2024)02-0287-08　　**OSID**：

Text Word Segmentation of Livestock and Poultry Diseases Based on BERT－BiLSTM－CRF Model

YU Ligen[1,2]　GUO Xiaoli[3]　ZHAO Hongtao[3]　YANG Gan[1,2]　ZHANG Jun[1,2]　LI Qifeng[1,2]

(1. *Research Center of Information Technology*, *Beijing Academy of Agriculture and Forestry Sciences*, *Beijing* 100097, *China*
2. *Innovation Center of National Digital Livestock*, *Beijing* 100097, *China*
3. *School of Mathematics and Physics*, *North China Electric Power University*, *Beijing* 102206, *China*)

Abstract: The diagnosis, prevention and control of livestock and poultry diseases is of great significance to ensure the healthy development of animal husbandry in China. Based on natural language processing, the word segmentation effect of livestock and poultry disease texts was improved to improve the diagnosis level of livestock and poultry diseases. In order to deal with the problems of lacking text corpus in livestock and poultry diseases, and a large number of out of vocabulary words contained in the texts, such as epidemic names and phrases, a word segmentation model was proposed based on BERT－BiLSTM－CRF combined with dictionary matching. Taking sheep diseases as the research object, the text datasets of common diseases were constructed combined with the general corpus PKU, and the text vectorizations were processed by BERT pre-trained language model. Then the context semantic features were obtained through the bidirectional long short-term memory network (BiLSTM), and globally optimal label sequences were outputted by conditional random field (CRF). Based on this, dictionary matching was refined by adding a dictionary in the field of livestock and poultry diseases after the CRF layer, which

收稿日期：2023－11－13　修回日期：2023－12－12

基金项目：云南省重大科技专项计划项目（202102AE090039）、北京市农林科学院创新能力建设专项（KJCX20230204）和北京市数字农业创新团队建设项目（BAIC10－2023）

作者简介：余礼根（1985—），男，副研究员，博士，主要从事畜禽信息感知与智能识别技术研究，E-mail：yulg@nercita.org.cn

通信作者：李奇峰（1981—），男，研究员，博士，主要从事农业信息化技术研究，E-mail：liqf@nercita.org.cn

reduced the ambiguity segmentation caused by the epidemic names and phrases in the process of word segmentation, and the accuracy of word segmentation was further improved. Results showed that the F1 value of the BERT - BiLSTM - CRF model combined with dictionary matching on the text datasets of sheep common diseases was 96.38%, which was increased by 11.01, 10.62, 8.3 and 0.72 percentage points, compared with that of jieba word segmentation, BiLSTM - Softmax model, BiLSTM - CRF model, and BERT - BiLSTM - CRF model that did not combine with dictionary matching, respectively, which verified the effectiveness of BERT - BiLSTM - CRF. Compared with a single corpus, the mixed corpus combined with the general corpus PKU and the text datasets of sheep common diseases could accurately divide the professional terms of livestock and poultry diseases and common words in the texts of diseases at the same time, the F1 values of the general corpus and the text datasets of diseases were more than 95%, which illustrated its better generalization ability. BERT - BiLSTM - CRF model can be effectively used for word segmentation of texts on livestock and poultry diseases.

Key words: livestock and poultry diseases; text word segmentation; pre-trained language model; bidirectional long short-term memory network; conditional random field

0 引言

据报道,我国确证发生的动物疫病已超过 230 种,其中 90 余种长期流行,平均每年给畜牧业造成的直接和间接经济损失超过 3 400 亿元[1]。因此,畜禽疫病诊断防控对保障我国畜牧业的健康发展具有十分重要的意义。随着新一代信息技术的发展,研究者将人工智能与传统畜禽养殖业融合,极大提高了畜禽疫病诊断防控水平。目前,疫病诊断多基于病例文本诊断报告[2]、文本诊断报告 + 知识图谱[3-4]、病患图像[5]、异常音频[6]等,但还没有很好的模型用于畜禽疫病文本分词,进而形成高质量疫病文本语料库,提升基于病例文本的疫病诊断准确性。

目前,常见的分词模型主要基于机器学习和深度学习。基于机器学习的分词模型主要有最大熵模型(Maximum entropy, ME)、隐马尔科夫模型(Hidden Markov model, HMM)、条件随机场等,但其特征提取依赖人工,对文本的语义信息获取能力不强[7]。

相较于机器学习,基于深度学习的分词模型无需人工进行特征选择,能够挖掘深层语义特征信息,泛化能力更强,逐渐成为文本分词的主流方法。CHEN 等[8]提出长短时记忆网络(Long short-term memory network, LSTM)模型,解决了神经网络利用固定窗口忽略长距离信息问题,但其只能获取单向特征信息。针对这一问题,YAO 等[9]提出采用双向 LSTM 模型进行中文分词,从过去和未来 2 个方向获取上下文信息,提高了文本分词效果。王莉军等[10]构建 BiLSTM - CRF 模型对中医文言文文献进行分词,与中文通用分词器 jieba、ansj 相比,F1 值分别提升 11.1、15.2 个百分点。虽然深度学习方法在分词任务中展现了优越的性能,但其模型训练往往依赖于大规模的训练数据,网络公开的多为通用领域语料,大量畜禽疫病领域数据的获取非常困难。针对领域语料匮乏问题,SHAO 等[11]分别利用通用语料和少量领域语料对 BiLSTM 模型进行训练,得到两个分词模型 metal_BiLSTM 与 msr_BiLSTM 后,在测试阶段将两模型的单词标签概率通过权重比进行组合,使得特定领域文本的分词准确率提升到 95% 左右。成于思等[12]利用迁移学习原理,使用小规模领域文本对已在通用领域语料上训练得到的 BiLSTM - CRF 模型参数进行微调,与通用语料训练的 BiLSTM - CRF 模型相比,F1 值提高 6.42 个百分点。

在畜禽疫病文本分词任务中多含有如疫病名称“传染性胸膜肺炎”、“巴氏杆菌病”,以及短语“胃肠道造影”、“窦性心律”等未登录词,这些词常常会被错误切分使得模型分词准确性降低。目前,在模型训练过程中结合词典或字/词向量是解决未登录词最常用的方法[13]。宫法明等[14]基于自适应隐马尔可夫模型结合领域词典和互信息,实现了对石油领域专业术语和组合词的精确识别,提高了石油领域分词准确率。ZHANG 等[15]基于现有深度学习分词模型缺乏处理生僻词的能力,提出在神经网络的基础上结合领域词典,解决分词的领域自适应性问题。

针对畜禽疫病文本分词过程中面临的领域语料缺乏和领域内专业名词及短语等未登录词难以正确切分问题,本文提出结合词典匹配的 BERT - BiLSTM - CRF 畜禽疫病文本分词模型。基于通用语料 PKU 及构建的羊常见疫病文本数据集,将具有迁移学习能力和更强文本特征表示能力的 BERT 预训练模型作为特征表示层,结合 BiLSTM - CRF 模型,提取文本全局和局部特征进行文本分词。同时,加入构建的领域词典,利用逆向最大匹配算法对首次分词结果进行词典匹配修正,进一步减少领域内未登录词对分词效果的影响。

1　模型原理

本文构建结合词典匹配的 BERT－BiLSTM－CRF 分词模型，如图 1 所示，该模型主要由 3 部分组成：BERT 层、BiLSTM 层、CRF 解码层。BERT 层用于对句子中的字符进行向量化表示，BiLSTM 层用于获取向量的上下文语义特征，CRF 解码层用于输出全局最优标签序列。

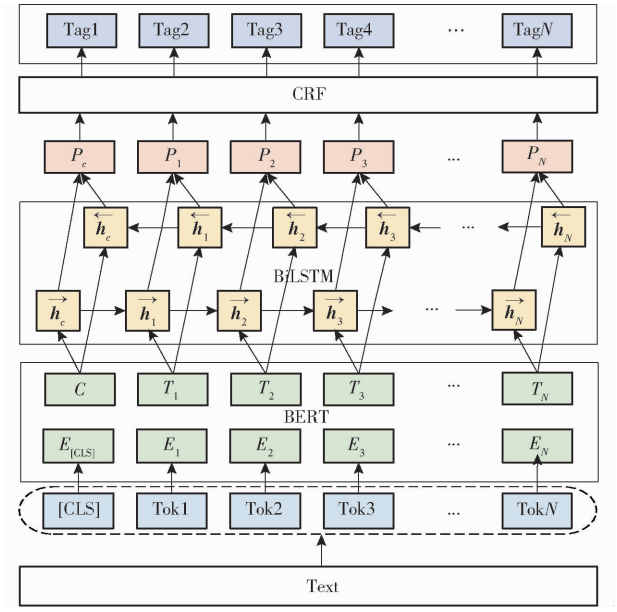

图 1　BERT－BiLSTM－CRF 模型结构图

Fig. 1　Structure diagram of BERT－BiLSTM－CRF model

1.1　BERT 层

BERT[16] 预训练模型属于迁移学习，即利用与目标任务无关的大量文本训练模型，基于该模型，通过目标领域的训练集微调参数，从而完成特定的自然语言处理任务。本文选用基于中文维基百科语料预训练的 BERT 模型 BERT_Base_Chinese，采用通用语料 PKU 和构建的羊常见疫病文本数据集对模型参数进行微调。

BERT 模型输入如图 2 所示，由字向量、句向量、位置向量 3 部分构成，其中含有特殊分类标记（Classification Token，[CLS]）作为起始 Token 与特殊令牌（Special Token，[SEP]）作为结束 Token[17]。[CLS]表示该特征用于分类模型，对于非分类模型可以省略；[SEP]表示输入语料中用于断开 2 个句子的分句符号。若输入由 2 个句子组成，则在第 1 个语句的句首加[CLS]符号，在第 1 个语句和第 2 个语句的末尾分别添加[SEP]符号。

在 BERT 预训练过程中，使用 2 个无监督预测任务对 BERT 进行文本上下文的自监督学习。①掩码语言模型（Mask language model，MLM）：在训练过程中，随机遮盖句中 15% 的词，用模型去预测这些

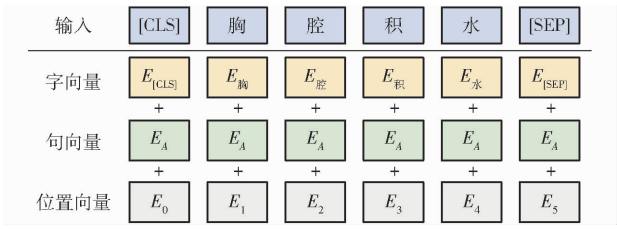

图 2　BERT 模型输入表示

Fig. 2　Input representation of BERT model

被遮盖的词，类似于做完形填空。采用 MLM 来训练语言模型，可解决模型中上下文语境缺失问题。②下一句预测（Next sentence prediction，NSP）：假设给定 A、B 两句话，预测 B 是否为 A 的下一句。训练过程中 50% 为连续正确的句子对，50% 为不正确的句子对，通过预测输入的两句话是否为连续文本，使模型拥有学习相邻句子关系的能力。

1.2　BiLSTM 层

BiLSTM 是一种改进的循环神经网络（Recurrent neural network，RNN）模型，可用于捕获长距离依赖关系和上下文序列信息[18]，通过引入门控机制和存储单元，克服了传统 RNN 由于文本序列过长而导致的梯度消失或梯度弥散问题[19]。BiLSTM 由 2 个独立的 LSTM 组成，LSTM 单元结构如图 3 所示。

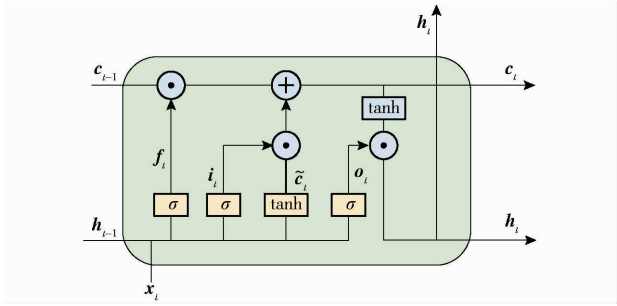

图 3　LSTM 单元结构

Fig. 3　LSTM cell structure

LSTM 主要包括 3 个门控单元：遗忘门、输入门、输出门。遗忘门决定哪些信息需要被丢弃，输入门决定哪些信息被记忆单元所保留，输出门决定哪些信息被输出并进入下一时刻的循环迭代。计算公式为

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_f[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_f) \tag{1}$$

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_i[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_i) \tag{2}$$

$$\tilde{\boldsymbol{c}}_t=\tanh(\boldsymbol{W}_c[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_c) \tag{3}$$

$$\boldsymbol{c}_t=\boldsymbol{f}_t\odot\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\odot\tilde{\boldsymbol{c}}_t \tag{4}$$

$$\boldsymbol{o}_t=\sigma(\boldsymbol{W}_o[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_0) \tag{5}$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t\odot\tanh(\boldsymbol{c}_t) \tag{6}$$

式中　$\boldsymbol{f}_t$、$\boldsymbol{i}_t$、$\boldsymbol{o}_t$、$\boldsymbol{c}_t$——t 时刻的遗忘门、输入门、输出门、记忆细胞

$\boldsymbol{W}_f$、$\boldsymbol{W}_i$、$\boldsymbol{W}_o$、$\boldsymbol{W}_c$——不同控制门对应的权重矩阵

$\boldsymbol{b}_f$、$\boldsymbol{b}_i$、$\boldsymbol{b}_o$、$\boldsymbol{b}_c$——偏置向量

$\tilde{\boldsymbol{c}}_t$——输入的中间状态

$\boldsymbol{x}_t$——t 时刻的输入向量

$\boldsymbol{h}_t$——t 时刻的输出结果

σ——sigmoid 激活函数

tanh——双曲正切激活函数

⊙——点乘运算符

LSTM 只能获取前一时刻信息,例如,文本序列“肺腺瘤病毒”,LSTM 只能获取“腺”前一个字“肺”的特征信息,而 BiLSTM 的输入序列分别以正向和反向输入至 2 个 LSTM 神经网络进行特征提取,可以充分获取“肺”与“瘤”的特征信息。前一时刻与后一时刻信息同等重要,本文采用 BiLSTM 结构模型,假设正向 LSTM 和反向 LSTM 隐藏状态输出序列分别为$\overrightarrow{\boldsymbol{h}_t}$和$\overleftarrow{\boldsymbol{h}_t}$,则将 2 个输出序列进行拼接后的$[\overrightarrow{\boldsymbol{h}_t}:\overleftarrow{\boldsymbol{h}_t}]$作为最终特征表达。

1.3 CRF 层

分词任务通常被看作序列标注任务,前后字的标签结果之间存在很强的依赖性。例如,标签 B 后只能跟标签 M 或标签 E,不能跟标签 B 或 S。而 CRF 是一种序列标记模型,通过考虑相邻标签的关系,对标签序列进行联合概率分析,给出全局最优的标签序列[20],确保了最终预测标签的合理性。

假设 $\boldsymbol{P}$ 为 BiLSTM 层的输出矩阵,n 为输入句子中包含字的数量。若 $X=\{x_1,x_2,\cdots,x_n\}$ 表示输入的文本序列,$Y=\{y_1,y_2,\cdots,y_n\}$ 表示对应的输出标签序列,则输入序列 X 对应预测序列 Y 的得分 $\mathrm{score}(X,Y)$ 计算公式为

$$\mathrm{score}(X,Y)=\sum_{i=1}^{n}A_{y_{i-1},y_i}+\sum_{i=1}^{n}P_{i,y_i} \tag{7}$$

式中 P_{i,y_i}——第 i 个字被预测为标签 y_i 的分数

A_{y_{i-1},y_i}——标签 y_{i-1} 转移到标签 y_i 的分数

A_{y_{i-1},y_i} 越大表示标签 y_{i-1} 转移到标签 y_i 的可能性越大。预测标签序列 Y 的条件概率为 $P(Y/X)$,计算公式为

$$P(Y/X)=\frac{\mathrm{e}^{\mathrm{score}(X,Y)}}{\sum\limits_{\tilde{Y}\in Y_X}\mathrm{e}^{\mathrm{score}(X,\tilde{Y})}} \tag{8}$$

式中 $\tilde{Y}$——真实的标签序列

Y_X——所有可能的标签序列

$\mathrm{score}(X,\tilde{Y})$——所有路径中某一路径的得分

使用 Viterbi 算法将得分最高的标签序列 Y^* 作为模型最终的标注结果,计算公式为

$$Y^*=\underset{\tilde{Y}\in Y_X}{\mathrm{argmax}}\ \mathrm{score}(X,\tilde{Y}) \tag{9}$$

本文采用四位序列标注法,由标签“B”、“M”、“E”分别表示词组的开头、中间、结尾位置,标签“S”表示单字成词。

1.4 逆向最大匹配算法

基于 BERT - BiLSTM - CRF 模型得到首次分词结果后,使用领域词典修正分词结果,将词典作为提高模型分词准确率的辅助工具。一般来说,逆向最大匹配的切分精度高于正向最大匹配,遇到的歧义现象也较少。本文基于构建的领域词典,利用逆向最大匹配算法对模型分词结果进行修正。从右向左将首次分词后待重新切分字符串中几个连续词组组成的字符(通常为词典中的最大词长 N)与领域词典匹配,若匹配成功,则切分出新词,否则去除掉最左边的一个原词组,迭代执行,直到剩下单独一个原词组(原词组指首次分词结果中的词组)。

假设首次分词结果得到的字符串为 S_1,匹配结果为 S_2(初始 S_2 为空字符串),领域词典 D 中最大词长为 N,对 S_1 进行逆向最大匹配的算法步骤如下:

(1)判断字符串 S_1 是否为空,若为空,输出匹配结果 S_2(若 $\mathrm{len}(S_2)>2$,$S_2=S_2[2:]$)。

(2)从 S_1 最右边开始取候选字符串 W,W 所含原词组个数不大于 N。

(3)将 W 与领域词典匹配,若匹配成功转步骤(4),匹配不成功转步骤(5)。

(4)$S_2=$“ ”$+W+S_2$(“ ”为 2 个空格),S_1 最右边去除 W 中所包含原词组,转步骤(1)。

(5)将 W 最左边一个原词组去掉。

(6)判断 W 是否为单个原词组,若是则转步骤(4),否则转步骤(3)。

如文本“上皮细胞浆中出现嗜酸性包涵体”,首次分词结果 S_1 为“上皮 细胞浆 中 出现 嗜酸性 包涵体”,经词典匹配修正后 S_2 为“上皮细胞浆 中 出现 嗜酸性 包涵体”。

逆向最大匹配算法流程如图 4 所示。

2 数据采集与预处理

2.1 数据准备

目前在畜禽疫病领域尚没有公开数据集,为了使模型更好地学习畜禽疫病领域文本分词特点,以羊疫病为研究对象,建立羊常见疫病文本数据集。训练集由通用语料和构建的羊常见疫病文本数据集共同组成。通用语料选择来自 SIGHAN 第二届国际中文分词大赛的 PKU 语料;自建的羊常见疫病文本数据集来源于专业兽医及行业专家共同选定的 3 本具有代表性的书籍:《羊常见疾病诊断图谱与防治技术》、《牛羊病诊治彩色图谱(第二版)》和《羊病

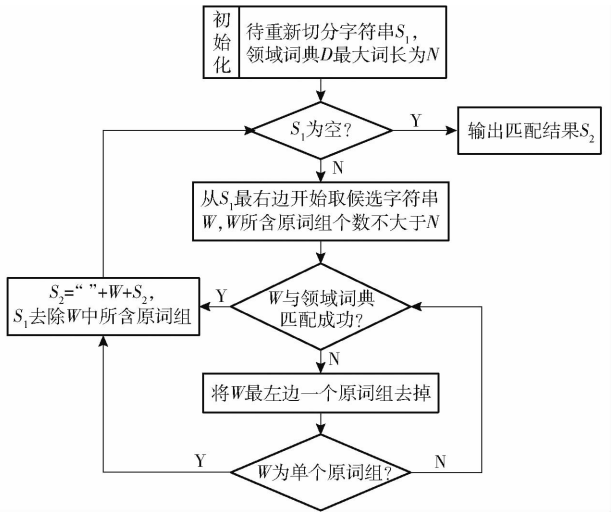

图 4　逆向最大匹配算法流程图

Fig. 4　Flow chart of reverse directional maximum match method

类症鉴别与防控原色图谱》，从每本书里分别提取羊常见疫病（不包含测试集）的文本描述内容，提取内容主要包括流行特点、症状、病理变化、诊断等，共提取文本 1 700 句，作为羊常见疫病文本训练集数据。

针对羊养殖最为典型的 3 类疫病：羊传染性胸膜肺炎、羊口疮、羊疥螨，从上述 3 本书籍中提取出对应疫病文本描述内容作为测试集数据，共 244 句。

在构建羊常见疫病文本数据集的过程中，为节省人力和时间，基于 jieba 分词器加载自定义领域词典对文本进行分词，在此基础上手工修正分词结果，得到精准的数据集。分词标准参照文献[21]。

2.2　领域词典构建

当特定领域文本中出现该领域专有名词或短语等未登录词时，一般情况下，这些词不易被正确切分，而被分割成其他词会失去其特定含义。基于此，建立畜禽疫病领域专有名词及短语词库，将构建的词库作为领域词典，采用逆向最大匹配算法，对首次分词结果进行匹配修正，以提高文本分词准确率。

本文基于搜狗细胞库（https：// pinyin. sogou. com/dict/）、百度百科等途径，搜索畜禽疫病领域关键词，去除重复词并删除其中的非相关词语，构建领域词典 D。词典中共有词语 4 578 条，每个词语单独成行。领域词典 D 示例如表 1 所示。

2.3　数据预处理

对羊常见疫病文本数据集进行规范化处理：

（1）同名替换。将“本病”“此病”“该病”替换为对应羊病病名，防止产生歧义。

（2）去噪。删除类似于“（图 1－54）”、“（图 2－19，图 2－20）”等字样；删除多余空格；删除名词的英文注释，如：金黄色葡萄球菌（*Staphylococcus aureus*）中的（*Staphylococcus aureus*）等；将上述内容视为噪声数据进行清除。

（3）长句拆分。根据标点符号“。”将文本内容由长句分为短句，防止文本句子过长。

表 1　领域词典 D 示例

Tab. 1　Example of field dictionary D

序号	词语
1	巴氏杆菌
2	巴氏杆菌病
3	传染性脓疱
4	非洲马瘟
5	非洲猪瘟
⋮	⋮
4 575	左主支气管
4 576	基底层细胞
4 577	不全角化细胞
4 578	角质蛋白包囊

2.4　实验环境

实验环境采用 Python 编程语言，深度学习框架 Pytorch 的版本为 2. 0. 0 + cu117，处理器为 Intel（R）Core（TM） i5 － 7400 CPU @ 3. 00 GHz，内存为 12. 0 GB。

使用 12 层的谷歌 BERT_Base_Chinese 预训练模型，隐藏层数为 768，multi－head 个数为 12。其他超参数设置如表 2 所示。

表 2　超参数设置

Tab. 2　Hyperparameter settings

参数	数值
最大长度（max_len）	64
批量大小（batch size）	32
LSTM 隐藏层维度（LSTM hidden_dim）	384
学习率（learning rate）	1×10^{-4}
优化器（optimizer）	AdamW
训练轮数（epoch）	5

2.5　实验流程

结合领域词典 D，基于预训练模型 BERT 构建 BERT－BiLSTM－CRF 模型应用于羊常见疫病文本数据集的分词任务。实验流程如图 5 所示。

2.6　评价指标

评价模型实验结果需要将模型分词序列标注结果与手工标注矫正的分词结果进行对比，将精确率（Precision，P）、召回率（Recall，R）和 F1 值（F-measure）作为其评价指标。

3　实验结果与分析

3.1　分词性能比较

为验证结合词典匹配的 BERT－BiLSTM－CRF

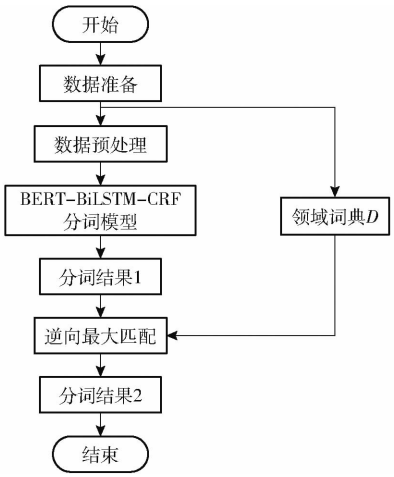

图 5 实验流程图

Fig. 5 Flow chart of experiment

模型的有效性,将其与其他模型在羊常见疫病文本数据集上的分词效果进行对比分析。其中,BiLSTM - Softmax 与 BiLSTM - CRF 模型的字嵌入维度为 300,隐藏层维度为 256。实验结果如表 3 所示。

表 3 分词实验结果

Tab. 3 Experimental results of word segmentation

%

模型	*P*	*R*	F1 值
jieba	82.63	88.29	85.37
BiLSTM - Softmax	83.94	87.66	85.76
BiLSTM - CRF	86.76	89.44	88.08
BERT - BiLSTM - CRF	95.17	96.15	95.66
BERT - BiLSTM - CRF(带词典)	96.37	96.39	96.38

由表 3 可以看出:

(1)BERT - BiLSTM - CRF 模型的精确率、召回率和 F1 值分别为 95.17%、96.15%、95.66%,各项评价指标均优于其他分词模型,与 BiLSTM - Softmax 模型和 BiLSTM - CRF 模型的 F1 值相比分别高出 9.9、7.58 个百分点。这是因为 BERT 预训练模型通过双向 transformer 结构动态生成了字符的上下文语义表示,与传统的词嵌入方法相比,能更加精准地理解语义信息,特征提取能力更强[19]。

(2)在 BiLSTM - Softmax 模型中,Softmax 层的输出是相互独立的,仅基于 BiLSTM 层的输出直接将概率最大的标签作为最后序列标注结果。而 CRF 中有转移特征,利用标签间的依赖关系,对标签序列进行联合概率分析,获取全局最优的标签序列[22]。与 BiLSTM - Softmax 模型相比,BiLSTM - CRF 模型分词性能更佳,本文模型也将 CRF 作为输出层。

(3)jieba 分词各评价指标值最低,与之相比,BERT - BiLSTM - CRF 模型的 F1 值提升 10.29 个百分点。表明仅靠公开的常用分词软件对特定领域文本分词,不能达到最优效果,也不能满足对分词性能的需求。

(4)基于逆向最大匹配算法进行词典匹配的 BERT - BiLSTM - CRF 模型,与未进行词典匹配的 BERT - BiLSTM - CRF 模型相比,F1 值提升 0.72 个百分点,这是因为领域词典的引入有效利用了词典信息,说明通过词典匹配的辅助分词能较好解决特定领域的未登录词问题,提高文本分词效果。

上述结果表明,结合词典匹配的 BERT - BiLSTM - CRF 模型在具有少量畜禽疫病文本数据集的分词任务上具有更高的准确率,与 jieba 分词器、BiLSTM - Softmax 模型、BiLSTM - CRF 模型相比,F1 值分别提升 11.01、10.62、8.3 个百分点,验证了本文方法的有效性。

3.2 BERT - BiLSTM - CRF 模型在不同数据集上的实验对比

将 BERT - BiLSTM - CRF 模型分别在通用语料 PKU、构建的羊常见疫病文本数据集、通用语料 PKU + 羊常见疫病文本数据集下进行实验并对比分析结果。其中,仅基于羊常见疫病文本数据集进行模型训练时,数据量较少,将 batch size 设置为 2。实验结果如表 4 所示。

表 4 BERT - BiLSTM - CRF 模型在不同数据集上的实验结果

Tab. 4 Experimental results of BERT - BiLSTM - CRF model on different datasets

%

训练集	评价指标	测试集	
		通用语料 PKU	羊常见疫病文本数据集
通用语料 PKU	*P*	95.83	82.49
	R	95.78	90.17
	F1 值	95.80	86.16
羊常见疫病文本数据集	*P*	83.57	95.13
	R	76.92	95.75
	F1 值	80.11	95.44
通用语料 PKU + 羊常见疫病文本数据集	*P*	95.83	95.17
	R	95.61	96.15
	F1 值	95.72	95.66

由表 4 可知:

(1)基于通用语料 PKU 与构建的羊常见疫病文本数据集结合的混合语料训练 BERT - BiLSTM - CRF 模型,在通用语料 PKU 和羊常见疫病文本数据集上的分词效果最好,F1 值分别达到 95.72%、95.66%,这是因为混合语料的使用同时考虑了通用领域和畜禽疫病领域的文本特征[23],能够对畜禽疫病专业术语及疫病文本中常用词准确

切分。

（2）仅使用通用语料 PKU 进行模型训练时，在羊常见疫病文本数据集上分词性能不佳，与混合语料的 F1 值相比低 9.5 个百分点，这是因为仅使用通用语料训练模型，未充分学习到疫病文本分词特点，在对疫病文本中专业术语进行切分时准确率较低。

（3）仅基于羊常见疫病文本数据集进行模型训练，在通用语料 PKU 上分词效果不佳，与混合语料训练得到的 F1 值相比低 15.61 个百分点，这是因为模型未充分学习到常用文本分词特点，对常用词的分词准确率较低[24]。

畜禽疫病文本分词任务中主要是疫病专业术语的切分，其次是文本中常用词的切分。本文选取通用语料 PKU 与羊常见疫病文本数据集结合的混合语料训练模型，进一步提高了模型泛化能力。

4　结论

（1）面向畜禽疫病领域，提出了结合词典匹配的 BERT－BiLSTM－CRF 畜禽疫病文本分词模型。基于 BERT 预训练模型，充分考虑了输入序列的全局特征，使嵌入层的输出包含更丰富的上下文语义信息。在首次分词结果基础之上，引入构建的畜禽疫病领域词典，利用逆向最大匹配算法进行匹配修正，减少了在疫病文本分词过程中出现的疫病名称及短语等错误切分问题，提升了分词性能，F1 值为 96.38%。

（2）通过将通用语料 PKU 与构建的羊常见疫病文本数据集结合的混合语料训练模型，能够同时对畜禽疫病专业术语及疫病文本中常用词进行准确切分，在通用语料 PKU 和羊常见疫病文本数据集上 F1 值都达到 95% 以上，验证了模型具有较好的泛化能力。

参　考　文　献

[1] 常帅，刘嘉，叶静，等. 新发展理念视阈下的我国畜禽疫病防控[J]. 中国科学院院刊，2019，34(2)：145－151.
CHANG Shuai，LIU Jia，YE Jing，et al. Study on prevention and control of animal infectious diseases in China from perspective of new development vision[J]. Bulletin of Chinese Academy of Sciences，2019，34(2)：145－151.（in Chinese）

[2] TERRADA O，CHERRADI B，RAIHANI A，et al. A novel medical diagnosis support system for predicting patients with atherosclerosis diseases[J]. Informatics in Medicine Unlocked，2020，21：100483.

[3] DONG L L，CHENG J，ZHANG X，et al. Research on disease diagnosis method combining knowledge graph and deep learning[J]. Journal of Frontiers of Computer Science & Technology，2020，14(5)：815.

[4] WANG H，SHEN W，ZHANG Y，et al. Diagnosis of dairy cow diseases by knowledge-driven deep learning based on the text reports of illness state[J]. Computers and Electronics in Agriculture，2023，205：107564.

[5] DAI Y，BAI W H，TANG Z，et al. Computer-aided diagnosis of Alzheimer's disease via deep learning models and radiomics method[J]. Applied Sciences，2021，11(17)：8104.

[6] DESPOTOVIC V，ISMAEL M，CORNIL M，et al. Detection of COVID－19 from voice，cough and breathing patterns：dataset and preliminary results[J]. Computers in Biology and Medicine，2021，138：104944.

[7] 周晓进，徐陈铭，阮彤. 面向中文电子病历的多粒度医疗实体识别[J]. 计算机科学，2021，48(4)：237－242.
ZHOU Xiaojin，XU Chenming，RUAN Tong. Multi-granularity medical entity recognition for Chinese electronic medical records[J]. Computer Science，2021，48(4)：237－242.（in Chinese）

[8] CHEN X，QIU X，ZHU C，et al. Long short-term memory neural networks for Chinese word segmentation[C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing，2015：1197－1206.

[9] YAO Y，HUANG Z. Bi-directional LSTM recurrent neural network for Chinese word segmentation[C]// International Conference on Neural Information Processing，2016：345－353.

[10] 王莉军，周越，桂婕，等. 基于 BiLSTM－CRF 的中医文言文文献分词模型研究[J]. 计算机应用研究，2020，37(11)：3359－3362，3367.
WANG Lijun，ZHOU Yue，GUI Jie，et al. Chinese medicine literature word segmentation based on BiLSTM－CRF[J]. Application Research of Computers，2020，37(11)：3359－3362，3367.（in Chinese）

[11] SHAO D，ZHENG N，YANG Z，et al. Domain-specific Chinese word segmentation based on bi-directional long-short term memory model[J]. IEEE Access，2019，7：12993－13002.

[12] 成于思，施云涛. 基于深度学习和迁移学习的领域自适应中文分词[J]. 中文信息学报，2019，33(9)：9－16，23.
CHENG Yusi，SHI Yuntao. Domain adaption of Chinese word segmentation based on deep learning and transfer learning[J]. Journal of Chinese Information Processing，2019，33(9)：9－16，23.（in Chinese）

[13] WANG Q，ZHOU Y，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.

[14] 宫法明，朱朋海. 基于自适应隐马尔可夫模型的石油领域文档分词[J]. 计算机科学，2018，45(6A)：97－100.
GONG Faming，ZHU Penghai. Word segmentation based on adaptive hidden Markov model in oil field[J]. Computer Science，2018，45(6A)：97－100.（in Chinese）

[15] ZHANG Q, LIU X Y, FU J L. Neural networks incorporating dictionaries for Chinese word segmentation[C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2018.

[16] 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 for Chapter of the Association for Computational Linguistics: Human Language Technoligies, 2019:4171 - 4186.

[17] 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)

[18] DU J L, MI W, DU X L. Chinese word segmentation in electronic medical record text via graph neural network-bidirectional LSTM - CRF model[C]// 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2020: 985 - 989.

[19] GAO W, ZHENG X, ZHAO S. Named entity recognition method of Chinese EMR based on BERT - BiLSTM - CRF[J]. Journal of Physics: Conference Series, 2021, 1848(1):012083.

[20] WANG S, PANG M, PAN C, et al. Information extraction for intestinal cancer electronic medical records[J]. IEEE Access, 2020, 8:125923 - 125934.

[21] 俞士汶,段慧明,朱学锋,等. 北京大学现代汉语语料库基本加工规范[J]. 中文信息学报, 2002,16(5):49 - 64.
YU Shiwen, DUAN Huiming, ZHU Xuefeng, et al. The basic processing of contemporary Chinese corpus of Peking University specification[J]. Journal of Chinese Information Processing, 2002, 16(5): 49 - 64. (in Chinese)

[22] DAI Z, WANG X, NI P, et al. Named entity recognition using BERT BiLSTM CRF for Chinese electronic health records[C]//2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP - BMEI). IEEE, 2019:1 - 5.

[23] CHEN J W, CHEN J G, WANG C B, et al. Research on segmentation of geological mineral text using conditional random fields[J]. China Mining Magazine, 2018, 27(9):69 - 74, 101.

[24] WEI D, LIU Z, XU D, et al. GeoBERTSegmenter: word segmentation of Chinese texts in the geoscience domain using the improved BERT model[J]. Earth and Space Science, 2022, 9(10):e2022EA002511.

(上接第 248 页)

[13] 刘阳. 面向星箭对接环的视觉位姿测量方法研究[D]. 哈尔滨:哈尔滨工业大学, 2022.
LIU Yang. Research on visual pose measurement for adapter ring[D]. Harbin: Harbin Institute of Technology, 2022. (in Chinese)

[14] 温卓漫. 复杂场景下合作靶标的准确快速识别与定位[D]. 长春:中国科学院长春光学精密机械与物理研究所, 2017.
WEN Zhuoman. Accurate and fast identification and localization of a cooperative target in complex background[D]. Changchun: Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Science, 2017. (in Chinese)

[15] SADIQ H A, ABD R R, BASHEERA M M, et al. Image edge detection operators based on orthogonal polynomials[J]. International Journal of Image and Data Fusion, 2017, 8(1/4):293 - 308.

[16] 任栋宇,李晓娟,林涛,等. 基于 Kinect v2 传感器的果树枝干三维重建方法[J]. 农业机械学报, 2022,53(增刊2):197 - 203.
REN Dongyu, LI Xiaojuan, LIN Tao, et al. 3D reconstruction method for fruit tree branches based on Kinect v2 sensor[J]. Transactions of the Chinese Society for Agricultural Machinery, 2022, 53(Supp.2):197 - 203. (in Chinese)

[17] 叶应辉. 基于深度学习的卫星遥感图像边缘检测方法[J]. 计算机测量与控制, 2022, 30(10): 39 - 44.
YE Yinghui. An edge detection method of satellite remote sensing image based on deep learning[J]. Computer Measurement and Control, 2022, 30(10): 39 - 44. (in Chinese)

[18] 周小军,谭薇. 一种改进 Canny 算子芯片图像边缘检测方法[J]. 工业仪表与自动化装置, 2022(4): 91 - 94.
ZHOU Xiaojun, TAN Wei. An improved algorithm for chip image edge detection based on Canny[J]. Industrial Instrumentation & Automation, 2022(4): 91 - 94. (in Chinese)

[19] SUN Zhaohui, YANG Feng, GUO Peng, et al. Flash-based in-memory computing for stochastic computing in image edge detection[J]. Journal of Semiconductors, 2023, 44(5): 159 - 164.