文章编号：2095-2163(2023)05-0140-06 中图分类号：TP391 文献标志码：A

基于 BERT-CBG-BiLSTM-CRF 的羊养殖命名实体识别

王 凯，李仁港，王天一
（贵州大学 大数据与信息工程学院，贵阳 550025）

摘 要：羊养殖知识多以文本的形式记录存储，知识量大、碎片化程度严重。为了改善构建羊养殖知识图谱时命名实体识别效果不佳的问题，本文的羊养殖文本命名实体识别模型将双向门控循环单元与卷积神经网络相结合，模型通过 BERT 预处理进行文本向量化处理，处理结果在 CBG 层通过训练字词向量，得到初步提取的上下文语义和词语语义，连接双向长短期记忆网络；条件随机场最终得到最大概率的输出序列。实验对特征、产地、建设、经济价值、品种、产区环境 6 类实体进行识别，最高 *F*1 值为 95.86% 。

关键词：羊养殖；命名实体识别；BERT；神经网络

Sheep breeding named entity identification based on BERT-CBG-BiLSTM-CRF

WANG Kai, LI Rengang, WANG Tianyi
(College of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China)

【Abstract】 Sheep breeding knowledge is mostly recorded and stored in the form of texts, which has the characteristics of large amount of knowledge and serious degree of fragmentation. In order to improve the problem of poor recognition of named entities when constructing sheep breeding knowledge graphs, the named entity recognition model of sheep breeding text in this paper is an optimization model that combines two-way gated circular units with convolutional neural networks. The model performs text vectorization processing through BERT preprocessing, and the processing results are trained in the CBG layer to obtain the contextual semantics and word semantics of the initial extraction, and then connect the two-way long-term short-term memory network; the conditional output sequence with the airport finally obtains the maximum probability. In this paper, six types of entities were identified experimentally for characteristics, origin, construction, economic value, varieties, and production area environment, and the highest *F*1 value was 95.86%.

【Key words】 sheep farming; named entity recognition; BERT; neural networks

0 引 言

随着人工智能的发展，传统手工、非实时的记录方式已经跟不上时代的步伐。中国作为一个羊业大国，养殖智能化能方便农户更精准、高效的管理养殖，促进养殖业的发展，减少人工成本。知识图谱作为养殖智能化中关键一环，其在知识归纳、推理、问答等方面有着举足轻重的地位。

目前，基于深度学习的命名实体识别逐渐受到关注[1]。与需要人工选取特征的基于传统机器学习的方法和耗时长且难以移植的基于规则的方法相比，基于深度学习的命名实体识别得到了广泛的应用[2]。仇增辉等[3]使用条件随机场（Conditional Random Field，CRF）、双向长短期记忆网络（Bi-directional Long Short-Term Memory，BiLSTM）为基础网络对网购评论进行识别，得到不错的识别效果；张帆等[4]用深度学习的方法对医疗文本进行实体识别，得到了比传统方法更高的准确率和召回率；阿依图尔荪·喀迪尔[5]用神经网络强化电子病历识别，为医疗提供了更加精准的服务；王学峰等[6]基于深度学习命名实体识别，在军事语料库识别的准确率、召回率、*F*1 值都得到很大提高；方红等[7]提出一种融合注意力机制的卷积神经网络（Convolutional Neural Network，CNN）和双向门限控循环单元

基金项目：贵州省科学技术基金（ZK[2021]304）；贵州省科技支撑计划（[2021]176）。

作者简介：王 凯（1995-），男，硕士研究生，主要研究方向：知识图谱、自然语言处理；李仁港（1997-），男，硕士研究生，主要研究方向：软件工程；王天一（1989-），男，博士，副教授，主要研究方向：量子通信、图像处理、计算机视觉。

通讯作者：王天一 Email：tywang@gzu.edu.cn

收稿日期：2022-05-22

(Bidirectional Gated Recurrent Unit，BiGRU)结合的网络模型，对产品质量检测进行识别，准确率和 *F*1 值都在 74.7%以上。

由以上可知，基于深度学习的命名实体识别相较于基于规则和基于机器学习来说，有着更高的识别率。在羊养殖领域，文本数据来源繁多，没有特定的规则，各种实体定义、关系类别、属性连接都需要人为定义，这导致实体识别难度较大。

为了解决羊养殖知识图谱构建中的命名实体识别问题，本文利用预先训练的语言模型，通过预训练语言模型，充分利用词左右两边的信息，获取词的分布式表示，连接卷积神经网络与双门控循环单元层(Convolutional Neural Network and Bidirectional Gated Recurrent Units，CBG)；在 CBG 层，通过卷积神经网络 CNN 提取羊养殖文本的字向量信息，利用双门控循环单元 BiGRU 网络训练词向量，提取文本语义信息；其次，对两者训练出的词向量结果进行拼接；利用 BiLSTM 网络训练进一步获得文本特征；最后，利用 CRF 层得到最大概率的输出序列，从而识别出实体。

1　数据获取与标注

1.1　数据获取

本试验所采用的资料主要来自于羊养殖相关书籍，以及其他有关羊的资料，结合百度百科，维基百科两大平台。经过整理、归纳得到相关的羊养殖数据，共计 13 859 个句子，532 484 个字符。

1.2　数据标注

本文采用 BIO(Beginning Inside and Outside)标注，数据集共包含实体 17 451 个，各类标注实体数量见表 1。

表 1　标注实体数量

Tab. 1　Number of labeled entities

实体名称	标注数量/条
实体	17 451
特征	5 033
产地	953
建设	2 058
经济价值	3 255
繁殖	2 586
产区环境	3 566

2　模型框架

本文先使用文本数据输入 BERT(Bidirectional Encoder Representations from Transformers，BERT)模型，进行预训练处理，增加文本的泛化能力；其次，将训练好的字、词向量分别送入 CNN 网络和 Bi-GRU 网络，提取字向量信息和上下文信息，并且将两者训练出的词向量结果进行拼接；由于文本较大，为保证长文本语义信息的依赖，将拼接的词向量送入 BiLSTM 网络，通过训练学习到输入向量的双向信息；最后，把 BiLSTM 层学习到的特征输入到 CRF 层中，得到输出序列。BERT-CBG-BiLSTM-CRF 模型整体结构如图 1 所示。

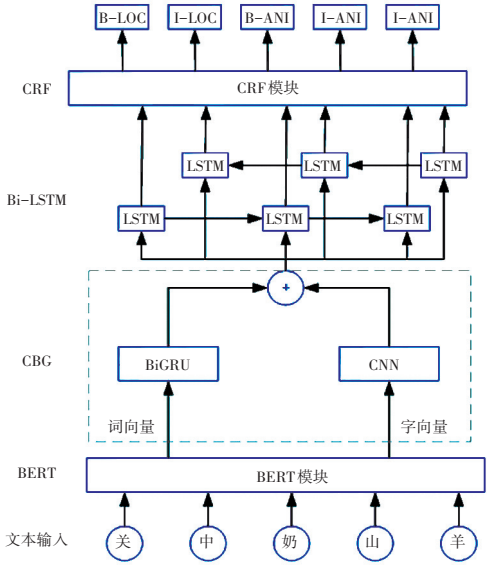

图 1　BERT- CBG-BiLSTM-CRF 模型结构

Fig. 1　BERT-CBG-BiLSTM-CRF model structure

2.1　BERT 层

BERT 模型有别于传统的预训练模型只能单向训练，突破传统语言预训练模型桎梏，通过 MLM(Masked Language Model)及其本身特殊的结构—双向 Transformer 编码，能更深层次获取文本的深层表征。因此，BERT 由于其独特的结构和其预训练任务的创新性，在自然语言处理预训练中取得惊人的效果，其模型结构示意如图 2 所示。

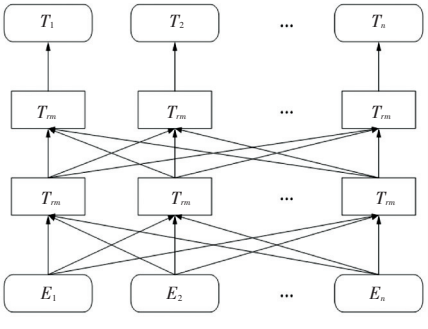

图 2　BERT 结构示意图

Fig. 2　Schematic diagram of the BERT structure

2.2 CBG 层

CBG 层是由 CNN 网络模型和 BiGRU 网络模型拼接而成。通过 CNN 训练 BERT 输入的字符集特征和 BiGRU 网络训练的词语的语义特征，把两者结果进行组合，不仅得到字向量的信息，还得到包含上下文语义信息的词向量。CBG 结构示意图如图 3 所示。

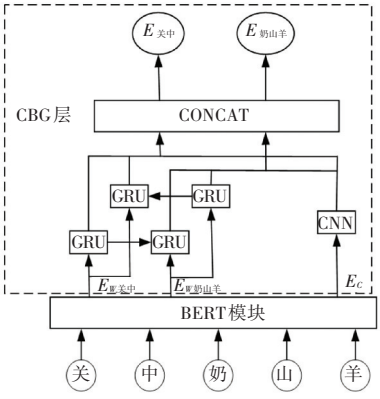

图 3 CBG 结构示意图

Fig. 3 Schematic diagram of CBG structure

2.2.1 CNN

字符级 CNN 用于命名实体识别，利用子词信息消除对形态标记或人工特征的需要并生成新单词，本文基于 CNN 的字符集分布式输入特征表示如图 4 所示。

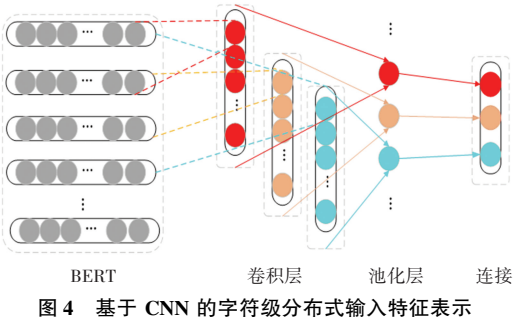

图 4 基于 CNN 的字符级分布式输入特征表示

Fig. 4 CNN-based representation of a character-level distributed input feature

该模型主要包含 4 个结构，即输入层、卷积层、池化层和全连接层。

输入层输入的是文本矩阵，通过 BERT 预训练模型得到字、词向量。

自向量进入神经网络，神经网络的核心是卷积层，通过多层卷积计算，对输入的向量进行特征提取，再经过池化层，最后把提取的特征向量进行拼接。CNN 卷积层的计算如式(1)：

$$s_i = h\left(\sum_{j}^{k-1} \boldsymbol{w}_{j+1}\, \boldsymbol{v}_{j+1} + b\right) \tag{1}$$

其中，$\boldsymbol{v}$ 为输入向量；k 为卷积核大小；$\boldsymbol{w}$ 为权重矩阵；$\boldsymbol{S}$ 为输出值；b 表示偏置。

经过卷积计算后得到的特征向量进入池化层，池化层继续将这些特征进行选择和过滤。全连接层再把这些特征进行分类，最后拼接。

本文采用 CNN 网络来训练字向量，通过卷积、池化、全连接，最后得到新的词级别的特征向量 $\boldsymbol{E}'_c$。

2.2.2 Bi-GRU

羊养殖文本较长，若选用 RNN 网络来进行序列处理，可能因为序列较长引起梯度消失和梯度爆炸，不能保证学习到长距离的依赖特征。本文选用结构跟 LSTM 类似的 GRU 网络，把遗忘门和输入门合二为一，变成新的一个门即更新门，又同时混合细胞状态和隐藏状态。Bi-GRU 能将当前时刻的输入与前一时刻的状态都能与后一时刻的状态产生联系，从而达到很好的学习效果，使羊养殖文本具有连贯性，避免训练空泛。GRU 编码单元如图 5 所示。

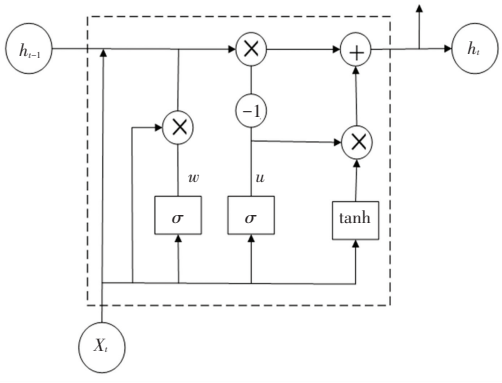

图 5 GRU 结构

Fig. 5 GRU structure

GRU 的计算方式：在 t 时刻，Z_t 为更新门，用来控制当前状态中前序记忆与候选记忆所占的比例，如式(2)：

$$Z_t = \sigma(w_u x_t + U_Z h_{t-1}) \tag{2}$$

r_t 为重置门，用于控制当前内容是否被记忆，计算如式(3)：

$$r_t = \sigma(w_r x_t + U_r h_{t-1}) \tag{3}$$

h'_t 代表候选隐藏层，计算如式(4)：

$$h'_t = \tanh(W x_t + r_t * U h_{t-1}) \tag{4}$$

h_t 代表隐藏层，计算如式(5)：

$$h_t = z_t * h_{t-1} + (1 - z_t) * h'_t \tag{5}$$

其中，W_r、W_z、W、U_r、U_Z、U 都是 GRU 的权重值；σ 代表 *sigmoid* 激活函数；h_{t-1} 为 $t-1$ 时刻隐含状态的输入。

将输入词向量 $\boldsymbol{E}_w$ 通过 BiGRU 网络训练，即可

得到初步提取过语义信息的词向量 $\boldsymbol{h}_{cbg}$，将其与 CNN 的输出 $\boldsymbol{E}'_C$ 拼接；在 CBG 层获得了拼接后的词向量 $\boldsymbol{E}_{cbg}$，融合了初步提取的上下文语义和词语语义；将其输到 BiLSTM 网络训练，提取深层特征，由前向后的拼接所得的输出将会产生 BiLSTM 层的输出 h'_{cbg}，将其引入 CRF 层，经过 CRF 得到最大概率输出序列。

2.3　BiLSTM 模型

经过 CBG 网络的训练，从 CNN 网络得到训练好的词向量，又从 BiGRU 网络得到深层特征的词向量，但对于羊养殖文本而言，经过这两个网络并没有考虑到词语在文本中的前后顺序，也没有考虑词语之间的依赖关系。如“关中奶山羊”，经过训练只知道“奶山羊”，而不知“关中”这个限定。因此，本文加了 BiLSTM 网络对文本进行训练。

BiLSTM 网络主要有两个作用：一是可以考虑前后句子之间的相互关系，对文本向前和向后两个方向进行训练，在训练过程中学到保存哪些信息，遗弃哪些信息；二是对更微小的分类进行限定，更好地捕获句子之间的语义信息。门机制中各个门和记忆细胞的表达式介绍如下：

在 t 时刻遗忘门 F_t 的表达式(6)：

$$F_t = \sigma(\boldsymbol{W}_i[\boldsymbol{H}_{t-1}, \boldsymbol{X}_t + \boldsymbol{b}_f]) \tag{6}$$

在 t 时刻输入门 I_t 的表达式(7)：

$$I_t = \sigma(\boldsymbol{W}_{\mathrm{i}}[\boldsymbol{H}_{t-1}, \boldsymbol{X}_t + \boldsymbol{b}_{\mathrm{i}}]) \tag{7}$$

在 t 时刻记忆门 C_t 的表达式(8)：

$$C_t = F_t\, C_{t-1} + I_t\, \tilde{\boldsymbol{C}}_t \tag{8}$$

在 t 时刻输出门 O_t 的表达式(9)：

$$O_t = \sigma(\boldsymbol{W}_{\mathrm{o}}[\boldsymbol{H}_{t-1}, \boldsymbol{X}_t + \boldsymbol{b}_o]) \tag{9}$$

最后的输出为 H_t，表达式(10)：

$$H_t = O_{\mathrm{t}}\tanh(C_t) \tag{10}$$

其中，$\tilde{\boldsymbol{C}}_t$ 为输入的中间状态向量，式(11)：

$$\tilde{\boldsymbol{C}}_t = \tanh(\boldsymbol{W}_c[\boldsymbol{H}_{t-1}, \boldsymbol{X}_t + \boldsymbol{b}_c]) \tag{11}$$

其中，σ 代表 *sigmoid* 激活函数；tanh 为双曲正切激活函数；$\boldsymbol{W}_f$、$\boldsymbol{W}_i$、$\boldsymbol{W}_c$、$\boldsymbol{W}_o$、分别代表遗忘门权重矩阵、输入门权重矩阵、当前输入单元权重矩阵和输出门权重矩阵；$\boldsymbol{X}_t$ 为 t 时刻的输入向量；$\boldsymbol{H}_{t-1}$ 为 $t-1$ 时刻的输出向量；$\boldsymbol{b}_f$、$\boldsymbol{b}_i$、$\boldsymbol{b}_c$、$\boldsymbol{b}_o$ 分别为遗忘门偏置向量、输入门偏置向量、当前输入单元偏置向量和输出门偏置向量。

2.4　CRF 模块

通过 BiLSTM 网络输出的是经过标注标签的预测值，但这些预测值杂乱无序，为了知道输出的标签对应实体，需要将这些预测值输入 CRF 层。

CRF 模块主要作用就是考虑相邻数据的标记信息，自动对 BiLSTM 网络输出的预测分值进行约束，确保尽量输出的是合法序列，降低非法序列输出概率。

对于输入序列 $X = (x_1, x_2, \cdots, x_n)$ 预测输出序列 $Y = (y_1, y_2, \cdots, y_n)$ 的得分可以用式(12)表示，即转移概率和状态概率之和。

$$S(X, y) = \sum_{i=0}^{n} (\boldsymbol{A}_{yi,yi+1} + \boldsymbol{P}_{i,yi}) \tag{12}$$

其中，$\boldsymbol{A}$ 表示转移矩阵，$\boldsymbol{P}$ 表示 BiLSTM 的输出得分矩阵。

再利用 softmax 求得标签序列 Y 的概率值，式(13)：

$$\mathrm{p}(\mathrm{y} \mid \mathrm{X}) = \frac{\mathrm{e}^{\mathrm{S}}(X, y)}{\sum_{\tilde{Y} \in Y_X} S(X, \tilde{y})} \tag{13}$$

CRF 网络中的每个节点都代表一个预测值，在 BiLSTM 输出的预测序列的基础上，该方法在网络中找到最有可能的路径，以确定所输出的指定实体的标签，以实现标识实体的标识。因而训练的目标就是最大化概率 $P(y \mid X)$，可通过对数似然的方式实现，式(14)：

$$\log p(y \mid X) = S(X, y) - \sum_{i=0}^{n} S(x, yi) \tag{14}$$

最后利用维比特算法预测解码，得到求解的最优路径，式(15)：

$$y^* = \arg\max_{y \in Yx} S(x, y) \tag{15}$$

3　实验结果分析

实验采用 Pytorch1.7.1 框架，实验环境设置为：Intel（R）Core（TM）i7-9700K CPU 6 核处理器；GPU 为 RTX 2080，运行内存 32 G。

3.1　实验设置

本文实验参数具体设置见表 2。

表 2　参数设置

Tab. 2　Parameter settings

参数	数值
Transformer 层	12
Dropout	0.5
学习率	0.000 01
字向量维度	100
隐藏单元数	128
批量大小	32
迭代周期	20

3.2　评价指标

本文采用精确率、召回率和 $F1$ 值作为评价指标，如式(16)~式(18)：

$$P = \frac{T_P}{T_P + F_P} \times 100\% \quad (16)$$

$$R = \frac{T_P}{T_P + F_N} \times 100\% \quad (17)$$

$$F1 = \frac{2PR}{P + R} \times 100\% \quad (18)$$

其中，T_P 表示正确把正样本预测为正；F_P 表示错误把负样本预测为正；F_N 表示错误把正样本预测为负。

3.3 实验结果

本文把数据集分为训练集和测试集，比率为 7∶3。各种实体相互独立，确保实验的独立性。各种实体信息见表 3。

表 3 实体信息

Tab. 3 Entity information

类别	训练集实体数量/条	测试集实体数量/条
特征	3 523	1 510
产地	667	286
建设	1 440	618
经济价值	2 279	976
品种	1 810	776
产区环境	2 496	1 070

3.4 不同模型的性能比较

为了验证不同模型对于羊养殖数据集识别效果，本文做了 4 组实验，用当前比较热门的模型和本文提出的模型作对比，实验结果见表 4。

表 4 4 种模型实验

Tab. 4 Four model experiments

模型	*P*	*R*	*F*1
BiLSTM-CRF	93.21	92.85	93.03
BERT-LSTM-CRF	94.53	94.66	94.60
BERT-BiLSTM-CRF	96.32	93.26	94.79
BERT-CBG-BiLSTM-CRF	95.98	95.74	95.86

通过表 4 可知，BiLSTM-CRF 模型 *F*1 值为 93.03%，识别效果最差；本文提出的 BERT-CBG-BiLSTM-CRF 模型的 *F*1 值为 95.86%，识别效果最好；BiLSTM-CRF 模型没有对数据进行预训练，导致识别效果不佳；BERT-LSTM-CRF 模型虽然对数据进行了预训练，只使用单向长短期记忆网络，训练只能从一个方向训练，丢失了部分句子之间的语义信息。本文提出的 BERT-CBG-BiLSTM-CRF 模型在 CBG 层通过 CNN 网络进行字向量的训练，又通过 Bi-GRU 网络训练词向量，充分学习到文本数据的上下文信息特征，从而达到很好的学习效果，使羊养殖文本具有连贯性，较 BERT-BiLSTM-CRF 模型提高了 1.07%。

3.5 不同实体实验结果比较

对不同的网络模型进行了识别实验后，本文又对数据集进行了不同的实体分类，并将其送入本文模型进行命名实体识别，实验结果见表 5。

表 5 BERT-CBG-BiLSTM-CRF 模型下不同实体识别

Tab. 5 Identification of different entities under the BERT-CBG-BiLSTM-CRF model

类别	*P*	*R*	*F*1
特征	96.78	97.42	97.10
产地	97.26	97.85	97.56
建设	97.18	98.41	97.80
经济价值	95.25	93.22	94.24
繁殖	96.59	97.87	97.23
产区环境	93.12	92.48	92.80

通过表 5 可以看出，相较于特征、产地、建设和繁殖，经济价值和产区环境的准确率、召回率和 *F*1 值都较低。原因有两点：一是由于某些不成功的实体是未登录的，如："关中奶山羊的皮毛和骨等为毛纺、制革、化工提供原料"中，"制革"在实体识别中就属于未登录词，因此实体识别有很大概率识别不出来；二是不同来源的知识说法不一致，语料新旧不同，导致未能识别出来。比如"奶质优良"，有的说法是"奶中含有多种营养物质"。

3.6 非数据集实体识别验证

本文还对非数据集内容进行实体识别验证，结果见表 6，可以看出，对于非数据集内容，本文模型仍然可以将其识别出来。

表 6 实体识别结果

Tab. 6 Entity Recognition Results

编号	内容	识别结果
1	屠宰率 42.2%，净肉率 38.4%	经济价值
2	遗传性能稳定	特征
3	一般 5~7 月龄配种	繁殖

4 结束语

在构建羊养殖知识图谱过程中，针对羊养殖实体识别效果不佳的问题，本文提出了改进的命名实体识别模型 BERT-CBG-BiLSTM-CRF，该模型在已有模型的基础上增加了 CBG 层，通过对字词向量的训练，且将训练结果进行拼接，最终的识别结果 *F*1 值为 95.86%。

（下转第 150 页）