

家禽诊疗文本多实体关系联合抽取模型研究

胡 滨¹ 汤保虎¹ 姜海燕^{1,2} 霍 傲¹ 韩文笑¹

(1. 南京农业大学人工智能学院, 南京 210095; 2. 南京农业大学国家信息农业技术中心, 南京 210095)

摘要: 针对传统实体关系抽取方法中主体特征与句向量难以有效融合、现有 BIO 标注策略难以有效处理重叠关系的问题,提出一种基于 BERT 和双重指针标注的家禽疾病诊疗文本实体关系联合抽取模型 (Joint extraction of entity relationship of poultry disease diagnosis and treatment text, JEER_PD)。JEER_PD 使用双重指针标注 (Dual-pointer labeling, DPL) 策略,建立头、尾 2 个指针标注器,一次性标注出所有实体的开始和结束位置;引入 CLN (Conditional layer normalization) 网络层,强化主体抽取任务与客体关系联合抽取任务之间的联系;利用概率平衡策略 PBS 对抗正负类标签类别失衡,以加速模型收敛。实验表明,JEER_PD 准确率、召回率和 F1 分别为 97.69%、97.59% 和 97.64%,3 项指标较现有方法均有显著提升,说明 JEER_PD 能够快速、准确地抽取家禽疾病诊疗复杂知识文本中的实体关系三元组。

关键词: 家禽疾病诊疗文本; 实体关系抽取; 关系重叠; BERT 语言模型; 双重指针标注

中图分类号: TP391 文献标识码: A 文章编号: 1000-1298(2021)06-0268-09 **OSID:** 

Joint Extraction Model of Multi-entity Relations for Poultry Diagnosis and Treatment Text

HU Bin¹ TANG Baohu¹ JIANG Haiyan^{1,2} HUO Ao¹ HAN Wenxiao¹

(1. College of Artificial Intelligence, Nanjing Agricultural University, Nanjing 210095, China

2. National Engineering and Technology Center for Information Agriculture,

Nanjing Agricultural University, Nanjing 210095, China)

Abstract: Aiming at the problems that the subject feature and sentence vector in the traditional entity relationship extraction method are difficult to effectively integrate, and the existing BIO annotation strategy is difficult to effectively deal with the overlapping relationships, a joint extraction of entity relationship of poultry disease diagnosis and treatment text (JEER_PD) based on BERT and dual-pointer was proposed. JEER_PD used the dual-pointer labeling (DPL) strategy to establish two pointer labelers at the head and tail, marking the beginning and ending positions of all entities at once; introduced the conditional layer normalization (CLN) network layer to strengthen the connection between the subject extraction task and the object relationship joint extraction task; and used the probability balance strategy (PBS) to combat the imbalance of positive and negative labels to accelerate the model convergence. The experimental results showed that the accuracy, recall and F1 value of JEER_PD were 97.69%, 97.59% and 97.64%, respectively, and the three indicators were significantly improved compared with that of the existing methods, which proved that JEER_PD can quickly and accurately extract the entity relationship triples in the complex knowledge text of the diagnosis and treatment of poultry diseases.

Key words: poultry disease diagnosis and treatment text; entity relationship extraction; relationship overlap; BERT language model; dual-pointer labeling

收稿日期: 2020-09-02 修回日期: 2021-04-02
基金项目: 国家重点研发计划项目 (2016YFD0300607)
作者简介: 胡滨 (1980—), 男, 副教授, 博士, 主要从事网络路径优化和农业农村信息化研究, E-mail: hubin@njau.edu.cn
通信作者: 姜海燕 (1968—), 女, 教授, 博士, 主要从事模式识别和农业人工智能研究, E-mail: jianghy@njau.edu.cn

0 引言

生物医学数据在医学研究、远程会诊和医学诊断等方面具有极高的价值^[1]。现有针对动物医学领域知识抽取的研究较少。为了拓展家禽养殖知识图谱,应从大量非结构化的诊疗文本中自动获取知识^[2-3]。

实体关系三元组抽取作为信息抽取的主要任务之一,旨在从高度碎片化的知识中提取出复杂的实体间关系,并为知识图谱、信息检索等任务提供基础的数据支持^[4]。中文实体关系抽取任务分为有监督、半监督、无监督和开放域抽取4类^[5]。其中有监督的方法可以避免人工特征选择,减小抽取误差,是近年来信息抽取的研究热点。依据实体识别及关系分类的先后顺序,基于深度学习的有监督实体关系抽取方法可以分为流水线法(Pipeline)和联合学习方法(Joint learning)两类^[6]。流水线方法先命名实体识别(Name entity recognition,NER)^[7],再对实体与实体之间存在的关系进行分类(Relation classification,RC)^[8]。流水线方法虽能实现实体关系抽取,但存在以下问题:①实体识别任务中产生的错误会在网络中传播,从而造成误差积累。②先抽取实体再进行关系分类,没有关系的候选实体会产生信息冗余。③忽略2个任务的内在联系,无法捕捉到实体和关系之间的依赖信息^[9]。

联合学习方法^[10]能够利用实体和关系之间的信息,因此具有更好的性能表现。MIWA等^[11]通过在双向LSTM编码层共享实体识别和关系抽取的网络参数实现了联合抽取。郑丽敏等^[12]提出一种基于依存分析面向食品安全事件新闻文本的实体关系抽取方法FSE_ERE。ZHENG等^[13]将输入的句子通过公用的Embedding层和BiLSTM层,使用LSTM和CNN分别进行实体识别命名和关系分类。上述共享参数的方法在顺序上依然是先进行实体抽取再进行关系分类,没有显式刻画2个任务之间的交互。

文献[14]提出端到端神经网络,以序列标注的方式实现实体关系的联合抽取,但该模型采取就近原则分配关系,即一个实体在上下文中与多个实体存在关系时,模型只能将关系分配给距离最近的实体对,不能解决实体对的关系重叠问题。关系重叠文本信息丰富、结构复杂、抽取难度较大,是限制三元组抽取准确率的主要原因之一。为处理此类文本,文献[15-16]通过将Token编码传入Sigmoid网络层构建多头选择任务来同时进行实体头和关系的决策。但是,将主客体隐藏层表示简单相加作为关系向量的做法并不能有效表达实体间的信息关

联。文献[17]提出一种基于CRF解码的总体标注框架,针对关系重叠问题,对文本序列的每一个Token都做一次BIES标注。但大量的编码工作导致算法复杂度高、计算成本巨大,难以在实际中应用。文献[18]将关系建模为主体映射到客体的函数,按照SPO的顺序对文本进行联合抽取,通过多指针网络解码加强模型对关系重叠问题的处理能力。但对于文本中的长实体,多个Softmax激活的指针网络造成的类别失衡会使模型倾向于学习数量较多的样本(可能是无关信息),而对数量较少的正类样本则需要更多次的学习。此外,将主体特征与句子特征直接相加作为客体抽取任务的输入,不能在抽取客体时充分融合主体特征。

家禽疾病诊疗知识文本关系重叠现象普遍,实体关系抽取难度大。本文在现有联合抽取方案的基础上提出基于BERT(Bidirectional encoder representations from transformers)和双重指针标注(Dual-pointer labeling,DPL)策略的家禽诊疗文本实体关系联合抽取模型(Joint extraction of entity relationship of poultry disease diagnosis and treatment text,JEER_PD)。

1 JEER_PD 模型

1.1 家禽诊疗文本语料特征分析

家禽疾病诊疗文本包括家禽常见疾病的病症、致病因素、疾病类型、可用药物等实体信息。与农业其他领域的知识文本不同,家禽疾病诊疗知识文本因实体种类繁多而具有十分复杂的实体间对应关系。常见情况,如一种疾病对应多种病症,各个病症又存在不同的治疗方案等。

现有研究并未在实体关系级别对该领域文本做出统一分类。如图1所示,本文从主客体之间的联系出发,将诊疗知识文本划分为4种类型:

- (1)普通类型(Normal),此类文本结构简单,抽取难度最低。单个文本中只含有一对实体关系,文本中不存在关系重叠现象。
- (2)单主体,多关系及客体类型(Ss-Mpo),此类文本单个主体与多个客体之间存在关系。如图1所示,主体“青霉素”分别与客体“肌注”和“20~40 mg/kg”存在“用法”和“用量”关系。
- (3)多主体,单关系及客体类型(Ms-Spo),此类文本多个主体与单个客体之间存在关系。图1中,2个主体“土霉素”和“四环素”与客体“鸡白痢”都存在“治疗”关系。
- (4)多主体,多关系及客体类型(Ms-Mpo),此类文本结构最为复杂。文本中的多个主体和多个客

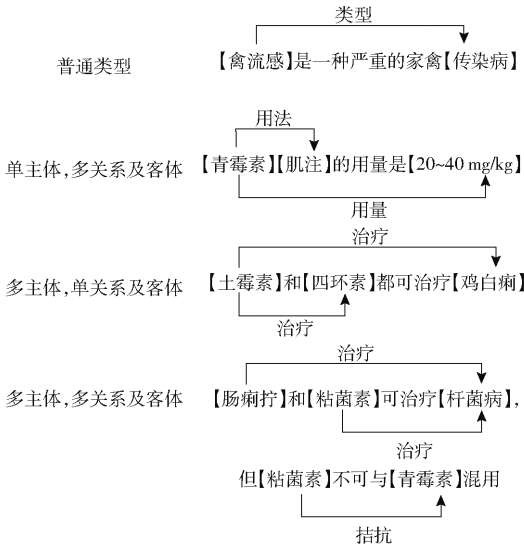


图 1 实体关系类型及示例

Fig. 1 Types and examples of entity relationships

体之间存在一种或多种对应关系。图 1 中,主体“肠痢宁”、“粘菌素”与客体“杆菌病”存在“治疗”关系,同时“粘菌素”与客体“青霉素”存在“拮抗”关系。

依本文定义, $S_s - M_{po}$ 、 $M_s - S_{po}$ 和 $M_s - M_{po}$ 3 类均属于关系重叠类型。处理此类文本时,现有的

深度学习方法在抽取精度或处理速度上难以达到令人满意的效果。

1.2 模型框架

JEER_PD 主要分为 3 部分:

(1) 编码部分。先对文本序列进行 Token、Segment 和 Position 3 级别的向量化操作。再将 3 个嵌入向量相加作为 BERT 语言模型的输入,使用多层 Transformer 进行特征提取得到句编码 h_N 。

(2) 主体抽取部分。建立 2 个主体指针标注器,分别对句编码 h_N 中所有主体的开始位置 (head) 和结束位置 (tail) 进行标注来完成主体抽取任务。

(3) 客体关系联合抽取部分。将主体向量参数化与句编码 h_N 经过 CLN 网络层得到特征序列 h'_N 。对于每一种关系分别建立 2 个指针标注器,抽取 h'_N 包含的所有客体及关系。最后将抽取出的主体、客体和关系解码,以三元组的形式输出。

整体框架如图 2 所示。

1.3 BERT 语义编码

BERT 是谷歌公司提出的预训练语言模型^[19]。

该模型融合 Mask language model 和 Next sentence

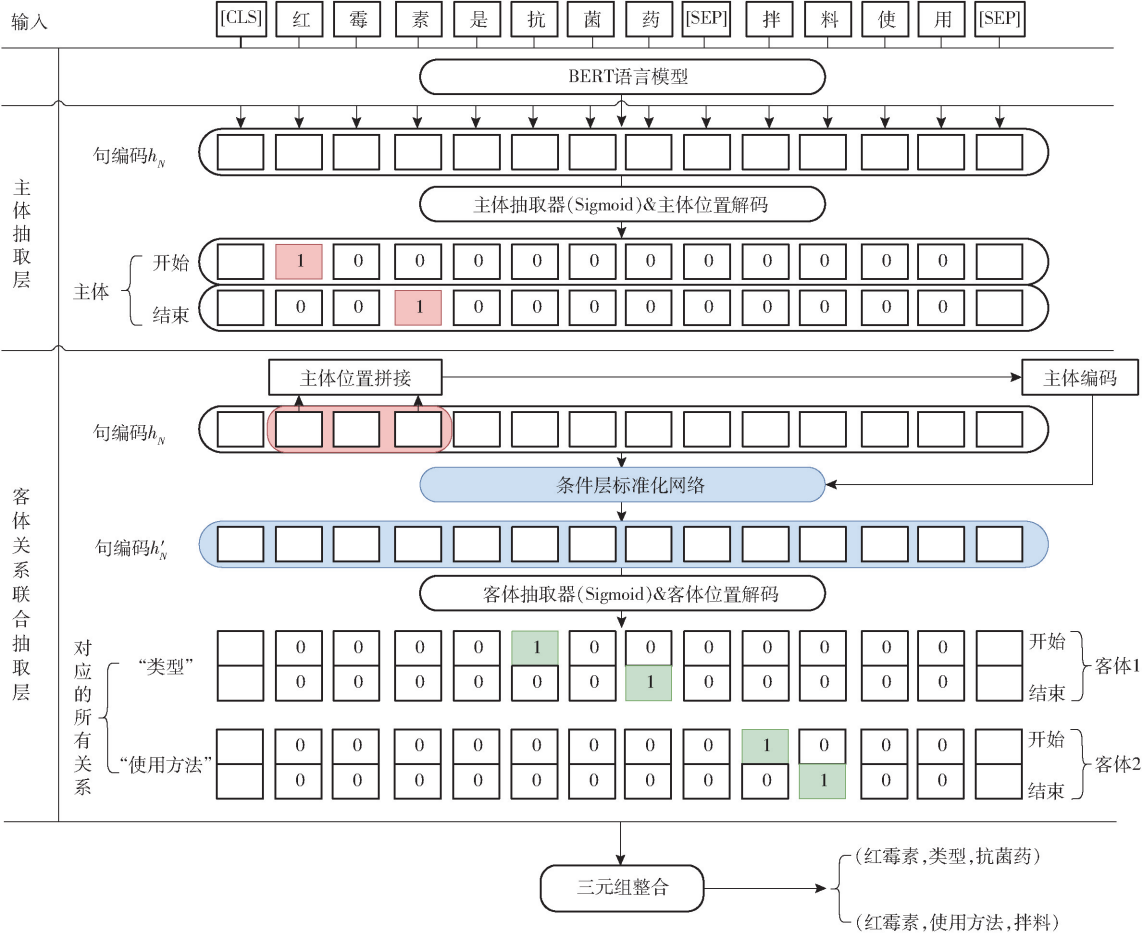


图 2 JEER_PD 模型框架

Fig. 2 JEER_PD model framework

prediction 2 种训练策略,极大提升了模型捕捉句子和词语级别语义的能力。与 Word2vec 不同,BERT 的语义编码可以随着模型训练而更新,因此可以在实体关系联合抽取任务中通过微调获取更好的语义表达。

本文采用 BERT 发布的预训练语言模型 BERT-base,网络层数为 12 层,隐藏层维度为 768,总参数量为 1.1×10^5 。BERT 对输入句子的编码过程如图 3 所示。

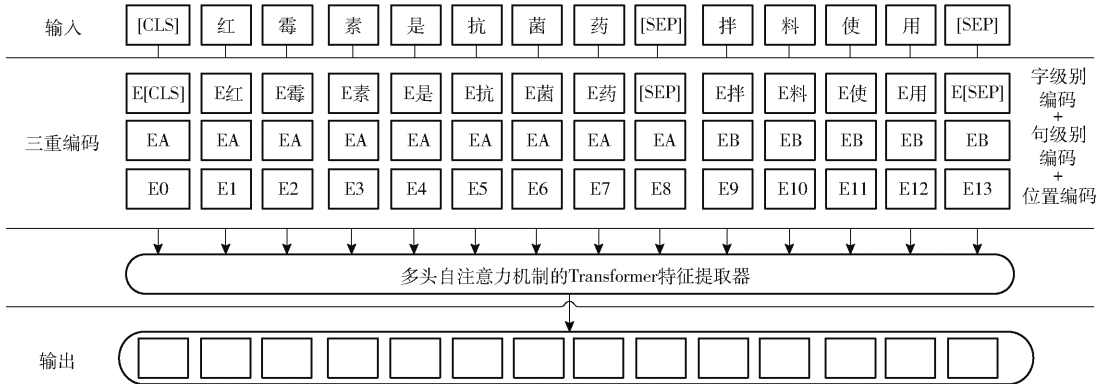


图 3 BERT Embedding 结构
Fig. 3 Construction of BERT Embedding

1.4 双重指针标注方法

文本输入经过 BERT 编码后得到融合文本信息和位置向量的句编码 h_N ,模型的关键任务就是对 h_N 进行实体关系解码。

序列标注和指针网络方法是解码任务常用的方法。序列标注任务本质上是 将解码问题转换为多分类问题。序列标注常采用 BIO 标注方法,使用 CNNs、RNNs 或是 BERT 进行特征编码,连接 CRF 条件随机场或 Softmax 函数辅助实体识别^[20]。标注过程中每一个 Token 只能被标识为一个实体类型。当实体发生重叠时,此方法无法对序列进行有效标注^[21]。图 4 为 BIO 标注方法对重叠实体的标注实例。

文本	“	庆	大	霉	素	注	射	液	”
BIO 标签	O	B-成分	I-成分	I-成分	I-成分	I-药物	I-药物	I-药物	O
标注结果	✓	✓	✓	✓	✓	×	×	×	✓

图 4 BIO 标注重叠实体
Fig. 4 BIO labels overlapping entities

指针网络方法最早应用于机器阅读理解任务 (Machine reading comprehension, MRC) 中^[22],该方法通常先对整个序列做 2 次 Softmax 预测答案的头指针和尾指针,然后将 2 个指针的位置拼接输出预测的答案。实体关系抽取任务的句子序列中可能存在一个或者多个实体。这时通过 Softmax 函数只能识别出概率最大的一个实体,无法解决多实体问题。

Sigmoid 函数常用于处理多标签二分类任务。在实体识别任务中,使用 Sigmoid 函数识别特征序列的实体,可以输出所有满足激活阈值的结果。Sigmoid 激活函数表达式为

$$\text{Sigmoid}(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

本文综合序列标注和指针网络 2 种方法,使用双重指针标注方式 DPL,将实体解码拆分成 2 个相互独立的指针标注任务。DPL 以“0/1”的标注方式,分别标注实体的开始位置和结束位置,然后将实体首尾位置拼接输出实体位置。这种方法可以对重叠实体进行有效标注,图 5 为 DPL 标注示意图。

文本	“	庆	大	霉	素	注	射	液	”
成分	头	0	1	0	0	0	0	0	0
	尾	0	0	0	0	1	0	0	0
药品名	头	0	1	0	0	0	0	0	0
	尾	0	0	0	0	0	0	1	0

图 5 DPL 方法
Fig. 5 DPL strategy

1.4.1 主体标注

在主体抽取层中通过 2 个 Sigmoid 指针标注器为每个 Token 分配一个 0/1 标签以表示该 Token 是否为主体的开始或结束位置。具体操作为

$$p_i^{s_head} = f(\mathbf{W}_{s_head} \mathbf{x}_i^{sub} + \mathbf{b}_{s_head}) \tag{2}$$

$$p_i^{s_tail} = f(\mathbf{W}_{s_tail} \mathbf{x}_i^{sub} + \mathbf{b}_{s_tail}) \tag{3}$$

式中 $\mathbf{x}_i^{sub}$ ——句编码 h_N 第 i 个 Token 的编码表示
 $\mathbf{W}_{s_head}$ ——头指针参数矩阵
 $\mathbf{W}_{s_tail}$ ——尾指针参数矩阵
 $\mathbf{b}_{s_head}$ ——头指针偏置矩阵
 $\mathbf{b}_{s_tail}$ ——尾指针偏置矩阵
 f ——Sigmoid 激活函数
 $p_i^{s_head}$ ——主体开始位置的输出序列
 $p_i^{s_tail}$ ——主体结束位置的输出序列

1.4.2 客体标注

与主体识别层类似,客体关系联合抽取层对每一种关系都分别建立 2 个用以识别客体头和尾的指针标注器。

$$p_i^{o_head} = f(W_{o_head}x_i^{obj} + b_{o_head}) \tag{4}$$

$$p_i^{o_tail} = f(W_{o_tail}x_i^{obj} + b_{o_tail}) \tag{5}$$

式中 x_i^{obj} ——融合主体向量信息的新的句编码 h_N 第 i 个 Token 的编码表示

W_{o_head} 、 W_{o_tail} ——客体头、尾指针的参数矩阵

b_{o_head} 、 b_{o_tail} ——客体头、尾指针偏置矩阵

$p_i^{o_head}$ 、 $p_i^{o_tail}$ ——客体开始和结束位置的输出序列

1.5 条件层标准化

网络层数的增加可以提高神经网络的表达能力,同时也会引发模型训练过程中梯度消失或梯度爆炸的现象^[23]。为了解决这种情况,通常会在网络层之间引入标准化(Normalization)操作加速训练并在一定程度上对抗过拟合。在 BERT 等基于 Transformer 的模型中,常用的标准化方法是 Layer normalization^[24]。相比于另一个常用的标准化方法 Batch normalization^[25],前者在文本序列处理上有着更好的表现。一般 Layer normalization 的计算公式为

$$LN(x_i) = \alpha \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta \tag{6}$$

式中 x_i ——句编码 h_N 第 i 个输入特征

$LN(x_i)$ ——网络层输出

μ ——均值 σ ——方差

ϵ ——很小的正数

α 、 β ——模型参数

α 和 β 是模型训练过程中由梯度下降得到的参数, ϵ 用以防止除 0 错误。

抽取实体关系三元组时应充分考虑主体与客体之间的联系。参照文献[26]提出的条件批标准化(Conditional batch normalization, CBN),本文在客体关系联合抽取层中使用 CLN(Conditional layer normalization)网络层实现主体向量的特征融合。实现过程如下:

(1)将识别出的主体向量 s 通过 2 个不同的变换矩阵转换为 2 个与 α 和 β 维度相同的向量。

(2)将这 2 个向量加到 α 和 β 中,使 α 和 β 成为主体向量 s 的函数并参与网络层标准化。

CLN 使主体向量能够以输入条件的形式参与客体及关系的抽取任务,表达式为

$$\alpha(s) = \alpha + W_\alpha s \tag{7}$$

$$\beta(s) = \beta + W_\beta s \tag{8}$$

$$CLN(x_i) = \alpha(s) \frac{x_i - \mu}{\sqrt{(\sigma')^2 + \epsilon}} + \beta(s) \tag{9}$$

式中 $CLN(x_i)$ ——条件层标准化网络(CLN)输出
为防止扰乱初始预训练权重,变换矩阵 W_α 和 W_β 全零初始化。两矩阵随着训练过程更新参数。

1.6 损失函数

分析家禽疾病诊疗知识文本后发现,句子序列中需要抽取的目标实体数目(正类标签)比非目标实体(负类标签)少。以图 5 为例,主体的正类标签“1”与负类标签“0”的比值为 1/9,这种类别失衡现象会使模型倾向于学习数量较多的负类标签的特征,影响模型的分类性能^[27]。

欠采样或过采样这类数据层面的处理办法虽能缓解类别失衡,但同时也会对样本数量多的类别造成消极影响^[28]。为缓解标签类别失衡问题,本文使用了结合概率平衡策略(Probabilistic balancing strategy, PBS)的损失函数,计算方法为:

初始损失函数为二元交叉熵损失(Binary crossentropy)

$$Loss_{PBS} = -\frac{1}{m} \sum_{i=1}^m [y_i \lg p_i + (1 - y_i) \lg(1 - p_i)] \tag{10}$$

式中 y_i ——标签类别,取 0 或 1

p_i ——取正类样本的概率

m ——训练批次大小

设平衡后的正类样本概率为 p'_i ;正负类样本比值为 K 。对于平衡后的正负类样本有

$$\frac{\lg(1 - p'_i)}{\lg p'_i} = K \tag{11}$$

p'_i 与初始概率 p_i 存在数量关系

$$p'_i = p_i^n \tag{12}$$

$$n = \log_{p_i} p'_i \tag{13}$$

求得的 n 为指数平衡因子。

平衡后的二元交叉熵损失函数为

$$Loss_{PBS} = -\frac{1}{m} \sum_{i=1}^m [ny_i \lg p_i + (1 - y_i) \lg(1 - p_i^n)] \tag{14}$$

由于 $0 \leq p_i \leq 1$, p_i^n 整体会更接近于 0,使得初始状态符合目标分布。从式(14)中发现,当标签 y_i 为 1 时,指数平衡因子 n 会加大正类标签的 loss 权重;而标签 y_i 为 0 时,负类标签的 loss 权重会被削减。PBS 策略时模型在训练过程中能够自适应地调节梯度权重来对抗标签类别失衡。

JEER_PD 的整体 Loss 由主体抽取任务的损失 $Loss_{sub}$ 与客体关系抽取任务的损失 $Loss_{obj}$ 两部分相

加得到,公式为

$$Loss_{JEER_PD} = Loss_{sub} + Loss_{obj} \tag{15}$$

2 实验设计与结果分析

2.1 文本知识获取

使用 OCR 软件对电子书籍进行文本识别,然后人工标记知识文本中的实体关系三元组。为提高标记质量,数据标注采用多人交叉标记验证策略,最后得到包含家禽常见疾病、病症、诱因、防控方法、可用药等实体与关系的文本数据 2 555 条。

通过网络爬虫实现数据集拓展,从兽药信息网上获取常见兽医药的使用方法和用药剂量等文本数据。经过数据清洗、数据蒸馏得到药物使用方法和剂量文本数据 1 124 条,总计 3 704 条文本数据,其中含实体 12 422 个,关系 6 211 个。按 7:3 划分训练和测试集,得到训练数据 2 591 条,测试数据 1 113 条。文本数据分布情况如表 1 所示。

表 1 文本数据信息
Tab.1 Text data information

文本类别	训练集	测试集
Normal	1 324	568
Ss - Mpo	674	289
Ms - Spo	348	150
Ms - Mpo	245	106
总计	2 591	1 113

2.2 实验环境及模型参数设置

本实验的硬件环境:处理器为 Intel Xeon E5 - 2698,运行内存 16 GB;运行环境:Tensorflow - 1.14.0, Keras - 2.2.4, Python 3.2。编译软件为 PyCharm Community Edition 2019.3.3 版 64 位;操作系统为 Windows 7;文中插图绘制工具为 diagrams。

通过对样本数据进行分析,主体和客体的正负类标签数量比约为 0.06 和 0.23,代入式(13),得到平衡因子 $n_{sub} \approx 3$ 、 $n_{obj} \approx 2$;最长数据文本有 115 字,因此将 Max-len 设置为 115,其余短文本向量化时在末尾补 0。指针标注器的激活阈值 K 默认值为 0.5;使用权重滑动平均的自适应优化器 Adam-EMA,初始学习率为 5.0×10^{-5} ,最低学习率为 1.0×10^{-5} ,训练批次大小为 64,迭代次数为 200。

2.3 评价指标

实验模型采用准确率 P (Precision)、召回率 R (Recall) 和 F_1 (F1-score) 3 项综合评价指标评估模型抽取实体关系三元组的性能表现。其中 F_1 是准确率和召回率的加权平均,能够更全面地衡量模型效果。

2.4 实验结果及分析

2.4.1 综合表现分析

本文设置多组对比实验来验证不同语言模型和标注方法的组合在实体关系抽取任务中的表现。对比实验均采用统一评价指标,分别在同一数据集中对不同模型进行 5 次实验并计算模型性能表现的均值。各模型的性能数据如表 2 所示。

表 2 不同抽取模型的性能
Tab.2 Performance of different extraction models

模型	标注方法	$P/\%$	$R/\%$	$F_1/\%$
BERT	BIO	85.33	75.29	80.03
BERT + BiLSTM + CRF	BIO	89.27	75.88	82.03
Word2vec + BiLSTM + CNN	DPL	89.58	85.96	87.74
BERT	DPL	93.28	94.05	93.66
JEER_PD	DPL	97.69	97.59	97.64

由表 2 可见,JEER_PD 在 3 个不同指标上均取得了最优表现。值得注意的是,结合 DPL 方法的 BERT 模型在不引入任何优化策略的情况下 F_1 较 BIO 标注的同模型提升了 13.63 个百分点,这一点从效果上体现出 DPL 方法的优越性。

相较于 BERT 语言模型,Word2vec 模型表现不佳。主要原因是:①预训练的 BERT 语言模型比静态编码的 Word2vec 在表义能力方面更优。②Transformer 特征提取器较 BiLSTM + CNN 的组合性能更优。

2.4.2 不同标注方法实验及分析

本文设计两组实验以研究标注方法对抽取结果的影响。

实验 1:按关系重叠类型划分数据集,旨在纵向比较不同标注方法对模型处理关系重叠文本的影响。

实验 2:按关系三元组数量重新划分数据集,以进一步探索不同标注方法下的各模型抽取多个三元组的能力。

两实验的对比模型和实验数据一致。

实验 1 的结果如图 6 所示。由图 6 可知,4 个模型在抽取 Normal 类型的单个三元组时性能表现较为接近,但处理 Ss - Mpo、Ms - Spo、Ms - Mpo 这 3 类关系重叠类型的文本时,基于 DPL 方法的模型 F_1 仍能够维持 93% 以上,而基于 BIO 标注方法的模型随着文本复杂度的提升, F_1 大幅下降。造成这种差距的主要原因是,基于 DPL 方法的模型按关系类别建立客体抽取任务,倾向于从给定的句子中抽取尽可能多的实体关系。在处理关系重叠文本时,这种做法不仅简化了抽取过程还可以增强模型抽取关系重叠文本的性能。

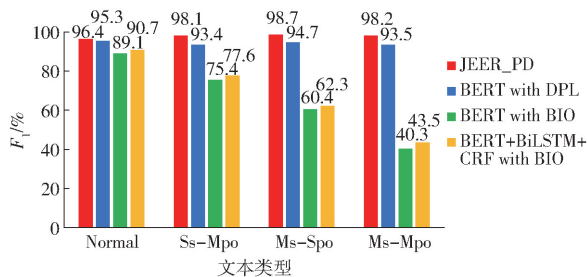


图 6 从不同重叠类型文本中抽取三元组的 F_1

Fig. 6 F_1 - score of extracting triples from texts with different overlap types

因数据样本中拥有 4 个三元组以上的句子数量较少,故将此类文本合并为一类。文本数量分布如表 3 所示,图 7 为实验 2 中各模型的抽取效果。

表 3 拥有不同数量三元组的文本数量

Tab.3 Number of texts with different numbers of triples

三元组数目/个	训练集	测试集
1	1 324	568
2	355	153
3	402	173
4	304	131
>4	206	88
总计	2 591	1 113

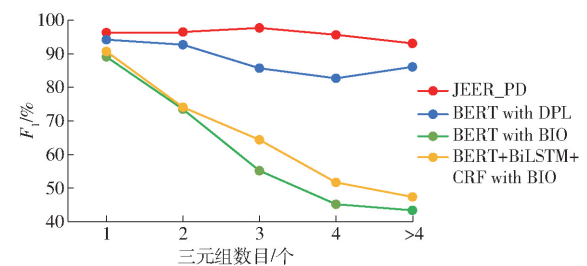


图 7 从不同三元组数目的文本中抽取三元组的 F_1

Fig. 7 F_1 - score of extracting triples from sentences containing different numbers of triples

由图 7 可见,参与实验的 4 个模型中 JEER_DP 表现最优,基于 DPL 方法的 BERT(BERT with DPL)次之。其余 2 个基于 BIO 标注方法的模型在抽取 2 个以上三元组数目的文本时与基于 DPL 方法的模型性能差距显著。

实验 1 和实验 2 充分验证了 DPL 方法从复杂文本中抽取实体关系三元组(一个或多个)的优越

表 5 消除 CLN 网络层导致识别错误的案例

Tab.5 Extraction errors caused by eliminating CLN layer

文本	基模型	消去 CLN 的模型
马立克氏病可经呼吸道、消化道及羽毛而传播。	马立克氏病,传染途径,呼吸道 马立克氏病,传染途径,消化道 马立克氏病,传染途径,羽毛	马立克氏病,传染方式,呼吸道 马立克氏病,传染方式,消化道
泰农用药可能导致家禽注射部位坏死或采食量下降。	泰农,不良反应,注射部位坏死 泰农,不良反应,采食量下降	泰农,病症,注射部位坏死 泰农,病症,采食量下降

性。DPL 方法不仅能使模型有效抽取关系重叠文本,还能提升对多三元组文本的抽取能力。

2.4.3 CLN 网络层消融实验及分析

从图 6 和图 7 中可以观察到 JEER_PD 在 2 个实验中性能均优于 BERT with DPL。原因是 JEER_PD 拥有用于融合主体特征的 CLN 网络层,而 BERT with DPL 在关系客体联合抽取任务时仅将主体特征与句子特征相加作为输入,这种做法不能充分体现主体抽取任务和关系客体联合任务之间的联系。

图 8 展示了 CLN 网络层整合主体向量的过程。主体向量 s 通过可更新参数的矩阵转换为 CLN 网络层的输入,与句编码 h_N 各个 Token 的特征(x_i)进行标准化融合,最终得到一个新的句编码 h'_N 。由于 h'_N 的每一个 Token 都融合了主体特征,模型在对 h'_N 解码时更容易抽取正确的客体及关系。

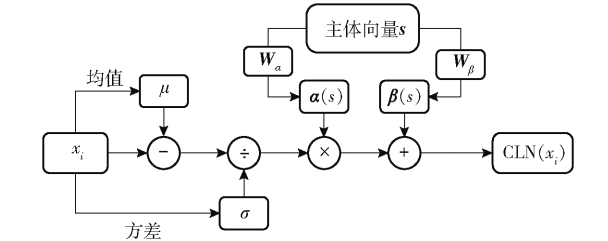


图 8 CLN 融合主体向量

Fig. 8 CLN fusion subject vector

为验证 CLN 网络层对模型整体性能的影响,本文设计了网络层消融实验。结果如表 4 所示。

表 4 CLN 消融实验结果

Tab.4 CLN ablation experiment

模型	P	R	F_1
基模型	97.57	97.61	97.59
消去 CLN 模型	96.40	93.88	94.94

对比实验结果可以发现,CLN 网络层在改善模型精度的同时也能进一步提升模型对三元组的整体抽取性能(Recall 指标)。

表 5 列举了因删除 CLN 网络层导致模型实体关系三元组抽取错误的样例。对比样例可以发现,删除 CLN 网络层的模型在抽取关系重叠文本时,可能因主体特征融合不充分而造成客体缺失或关系错误识别。

上述实验证明了 CLN 网络层融合主体特征的有效性。CLN 网络层使模型注重于抽取与主体特征联系更为紧密客体及关系,能在一定程度上提升模型的准确率和召回率并增强模型鲁棒性。

3 结论

(1)使用 BERT 预训练语言模型进行文本向量化,借助 Transformer 实现句子特征编码。相较于 Word2vec 编码的同类模型,JEER_PD 取得了 9.9 个百分点的 F1 提升。

(2)使用 DPL 方法提升模型对 3 类关系重叠文

本的抽取能力,较 BIO 标注策略的 BERT 模型最高取得 17.61 个百分点的 F1 提升。

(3)在客体关系联合抽取任务中使用 CLN 网络层融合主体特征,强化主体抽取和关系客体联合抽取任务之间的联系,进一步提升了模型的整体抽取性能。

(4)实验证明,JEER_PD 能够对家禽诊疗知识文本中复杂的实体关系进行有效抽取(实验 F1 达到 96.74%)。JEER_PD 抽取三元组的准确率高、性能稳定,可以为构建家禽养殖知识图谱和信息化养殖体系赋能。

参 考 文 献

- [1] 管红英,关同峰,张坤丽,等. 面向医学文本的实体关系抽取研究综述[J]. 郑州大学学报(理学版),2020,52(4):1-15.
ZAN Hongying, GUAN Tongfeng, ZHANG Kunli, et al. A review of research on entity relationship extraction for medical texts [J]. Journal of Zhengzhou University (Natural Science Edition), 2020, 52(4): 1-15. (in Chinese)
- [2] 刘峤,李杨,段宏,等. 知识图谱构建技术综述[J]. 计算机研究与发展,2016,53(3):582-600.
LIU Qiao, LI Yang, DUAN Hong, et al. Knowledge graph construction techniques [J]. Journal of Computer Research and Development, 2016, 53(3): 582-600. (in Chinese)
- [3] 李涓子,侯磊. 知识图谱研究综述[J]. 山西大学学报(自然科学版),2017,40(3):454-459.
LI Juanzi, HOU Lei. Reviews on knowledge graph research [J]. Journal of Shanxi University (Natural Science Edition), 2017, 40(3):454-459. (in Chinese)
- [4] 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 Guilin, GAO Huan, WU Tianxing. The research advances of knowledge graph [J]. Technology Intelligence Engineering, 2017,3(1):4-25. (in Chinese)
- [5] 武文雅,陈钰枫,徐金安,等. 中文实体关系抽取研究综述[J]. 计算机与现代化,2018(8):21-27.
WU Wenya, CHEN Yufeng, XU Jin'an, et al. Review of Chinese entity relation extraction [J]. Computer and Modernization, 2018(8):21-27. (in Chinese)
- [6] 鄂海红,张文静,肖思琪,等. 深度学习实体关系抽取研究综述[J]. 软件学报,2019,30(6):1793-1818.
E Haihong, ZHANG Wenjing, XIAO Siqi, et al. Survey of entity relationship extraction based on deep learning [J]. Journal of Software, 2019, 30(6): 1793-1818. (in Chinese)
- [7] MARRERO M, URBANO J, SÁNCHEZ-CUADRADO S, et al. Named entity recognition: fallacies, challenges and opportunities [J]. Computer Standards & Interfaces, 2013, 35(5):482-489.
- [8] KUMAR S. A survey of deep learning methods for relation extraction [J]. arXiv preprint arXiv:1705.03645, 2017.
- [9] 李冬梅,张扬,李东远,等. 实体关系抽取方法研究综述[J]. 计算机研究与发展,2020,57(7):1424-1448.
LI Dongmei, ZHANG Yang, LI Dongyuan, et al. Review of entity relation extraction methods [J]. Journal of Computer Research and Development, 2020, 57(7): 1424-1448. (in Chinese)
- [10] ZHENG S, XU J, BAO H, et al. Joint learning of entity semantics and relation pattern for relation extraction [C] // Joint European Conference on Machine Learning and Knowledge Discovery in Database, Springer, Cham, 2016:443-458.
- [11] MIWA M, BANSAL M. End-to-end relation extraction using LSTMs on sequences and tree structures [C] // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016: 1105-1116.
- [12] 郑丽敏,齐珊珊,田立军,等. 面向食品安全事件新闻文本的实体关系抽取研究[J/OL]. 农业机械学报,2020,51(7):244-253.
ZHENG Limin, QI Shanshan, TIAN Lijun, et al. Entity relation extraction of news texts for food safety events [J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(7): 244-253. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?file_no=20200728&flag=1. DOI:10.6041/j.issn.1000-1298.2020.07.028. (in Chinese)
- [13] ZHENG S, HAO Y, LU D, et al. Joint entity and relation extraction based on a hybrid neural network [J]. Neurocomputing, 2017, 257:59-66.
- [14] ZHENG Suncong, WANG Feng, BAO Hongyun, et al. Joint extraction of entities and relations based on a novel tagging scheme [C] // ACL (1), 2017.
- [15] GIANNIS B, JOHANNES D, THOMAS D, et al. Joint entity recognition and relation extraction as a multi-head selection problem [J]. Expert Systems with Application, 2018, 114(12):34-45.
- [16] BEKOULIS G, DELEU J, DEMEESTER T, et al. Adversarial training for multi-context joint entity and relation extraction [C] //

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018: 2830 – 2836.

[17] DAI D, XIAO X, LYU Y, et al. Joint extraction of entities and overlapping relations using position-attentive sequence labeling [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(1): 6300 – 6308.

[18] YU B, ZHANG Z, SHU X, et al. Joint extraction of entities and relations based on a novel decomposition strategy[J]. arXiv preprint arXiv:1909.04273, 2019.

[19] 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, 2019: 4171 – 4186.

[20] ZHANG Yue, YANG Jie. Chinese NER using lattice LSTM[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018: 1554 – 1564.

[21] 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/OL]. 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/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(1): 185 – 192. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?file_no=20210121&flag=1. DOI: 10.6041/j.issn.1000-1298.2021.01.021. (in Chinese)

[22] LI Xiaoya, FENG Jingrong, MENG Yuxian, et al. A unified MRC framework for named entity recognition[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 5849 – 5859.

[23] 刘建伟,赵会丹,罗雄麟,等. 深度学习批归一化及其相关算法研究进展[J]. 自动化学报,2020,46(6):1090 – 1120. LIU Jianwei, ZHAO Huidan, LUO Xionglin, et al. Research progress on batch normalization of deep learning and its related algorithms[J]. Acta Automatica Sinica, 2020, 46(6): 1090 – 1120. (in Chinese)

[24] BA J L, KIROUS J R, HINTON G E. Layer normalization[J]. arXiv preprint arXiv:1607.06450, 2016.

[25] IOFFE S, SZEGEDY C. Batch normalization: accelerating deep network training by reducing internal covariate shift[C]// International Conference on Machine Learning. PMLR, 2015: 448 – 456.

[26] DE V H, STRUB F, MARY J, et al. Modulating early visual processing by language[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 6597 – 6607.

[27] AKKASI A, VAROLU E, DIMILILER N. Balanced undersampling: a novel sentence-based undersampling method to improve recognition of named entities in chemical and biomedical text[J]. Applied Intelligence, 2017, 48(4): 1 – 14.

[28] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 318 – 327.

(上接第 241 页)

[19] LEE M, HUANG Y, YAO H, et al. Effects of sample storage on spectral reflectance changes in corn leaves excised from the field[J]. Journal of Agricultural Science, 2014, 6(8): 214.

[20] XIE C, SHAO Y, LI X, et al. Detection of early blight and late blight diseases on tomato leaves using hyperspectral imaging [J]. Scientific Reports, 2015, 5(1): 1 – 11.

[21] ASAARI M S M, MISHRA P, MERTENS S, et al. Close-range hyperspectral image analysis for the early detection of stress responses in individual plants in a high-throughput phenotyping platform[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2018, 138(4): 121 – 138.

[22] THOMAS S, KUSKA M T, BOHNENKAMP D, et al. Benefits of hyperspectral imaging for plant disease detection and plant protection: a technical perspective[J]. Journal of Plant Diseases and Protection, 2018, 125(1): 5 – 20.

[23] LIU Z, QI J, WANG N, et al. Hyperspectral discrimination of foliar biotic damages in rice using principal component analysis and probabilistic neural network[J]. Precision Agriculture, 2018, 19(6): 973 – 991.

[24] SENA F, SOTELO-SILVEIRA M, ASTRADA S, et al. Spectral phasor analysis reveals altered membrane order and function of root hair cells in Arabidopsis dry2/sqe1-5 drought hypersensitive mutant[J]. Plant Physiology and Biochemistry, 2017, 119(4): 224 – 231.

[25] PIAZZOLLA F, AMODIO M L, COLELLI G. Spectra evolution over on-vine holding of Italia table grapes: prediction of maturity and discrimination for harvest times using a Vis – NIR hyperspectral device[J]. Journal of Agricultural Engineering, 2017, 48(2): 109 – 116.

[26] SUN C, LI C, ZHANG C, et al. Reflectance and biochemical responses of maize plants to drought and re-watering cycles[J]. Annals of Applied Biology, 2018, 172(3): 332 – 345.