沙明洋,张思佳,傅庆财,等.基于动态权重的多模型集成水产动物疾病防治事件抽取方法[J].华中农业大学学报,2023,42(3):80-87.
DOI:10.13300/j.cnki.hnlkxb.2023.03.010

基于动态权重的多模型集成水产动物疾病防治事件抽取方法

沙明洋[1],张思佳[1,2],傅庆财[1],于红[1,2],李枳錡[1],喻文甫[1],刘珈宁[1]

1.大连海洋大学信息工程学院/辽宁省海洋信息技术重点实验室,大连 116023;
2.设施渔业教育部重点实验室(大连海洋大学),大连 116023

摘要 为提高水产动物疾病防治事件抽取的准确性,有效解决抽取过程中出现的专有名词边界模糊和事件实体过长等问题,本研究将动态权重思想引入多模型集成的事件抽取方法中。改进后的方法利用百度自然语言理解开放平台(enhanced representation through knowledge integration,ERNIE)和澎湃BERT(MLM as correction BERT,MacBERT)2个预训练模型来学习文本语义信息;采用动态权重的gate模块融合特征;将学习到的语义信息传入双向长短时记忆网络(bi-directional long shortterm memory,BiLSTM)中,并通过条件随机场(conditional random field,CRF)对输出标签序列进行约束。选取ERNIE⊕MacBERT-CRF模型和ERNIE⊕MacBERT-BiLSTM-CRF模型(⊕代表简单相加求平均的融合方法)作为对照模型对提出的方法进行融合性能对比试验验证,结果显示,该方法F_1值达74.15%,比经典模型BiLSTM-CRF提高了20.02个百分点。结果表明,该方法用于水产动物疾病防治事件抽取具有更好的效果。

关键词 水产动物疾病;事件抽取;ERNIE;MacBERT;动态权重;健康养殖

中图分类号 TP391.41 **文献标识码** A **文章编号** 1000-2421(2023)03-0080-08

随着水产养殖业规模不断扩大,各种疾病的发生频率也越来越高。传统的病害防治技术已无法满足水产养殖业发展的需求。近年来,随着自然语言处理领域的快速发展,构建水产动物疾病防治知识图谱为水产养殖业发展提供了新的途径。知识图谱是一种重要的工具,利用可视化技术来描述知识资源[1]。目前,水产动物疾病防治知识图谱的知识是静态的,主要描述实体及实体之间的关系。但是,这种知识已经无法满足对更广泛学习、推理和理解场景的需求[2-3]。相比“概念”,事件是更加细化、动态和结构化的知识。因此,构建事件知识图谱不仅可以丰富现有的知识图谱,还可以为事件检测、事件脉络分析以及未来事件预测等发展奠定基础[4]。在构建事件知识图谱的过程中,事件抽取是首要任务,可以从描述事件的文本中提取结构化信息。事件抽取的准确性直接影响图谱的构建质量。因此,为了实现水产养殖业的快速稳定发展,亟需一种适用于水产动物疾病防治事件的抽取方法。

早期的事件抽取主要采用基于模式匹配的方法,该方法需要领域专家制定规则并对所有事件元素进行标注[5]。随着标注语料数据量的增加和机器学习方法的兴起,基于模式匹配的方法逐渐被基于机器学习的方法所取代。例如,李浩瑞等[6]提出了一种基于丰富特征和组合不同类型学习器的混合模型。万齐智等[7]提出了一种句法和语义依存分析相结合的中文事件抽取框架。虽然基于机器学习的方法降低了人工成本并解决了特定领域事件成分缺失和事件嵌套等问题,但在系统可移植性和原始文本特征提取方面仍存在一定不足之处。

近年来,神经网络在自然语言处理领域取得了巨大进展,基于预训练模型的深度学习方法逐渐成为主流[8-9]。然而,与其他领域相比,水产动物疾病

收稿日期:2022-09-30
基金项目:设施渔业教育部重点实验室开放课题(2021MOEKLECA-KF-05);计算机体系结构国家重点实验室开放课题(CARCH201921);辽宁省教育厅高等学校基本科研项目面上项目(20220056);辽宁省教育科学“十四五”规划课题(JG21DB076)
沙明洋,E-mail:447412416@qq.com
通信作者:张思佳,E-mail:zhangsijia@dlou.edu.cn

防治事件抽取任务存在大量专有名词、边界模糊以及事件实体过长等问题，这些问题使得已有的方法难以有效解决该领域的事件抽取问题。因此，本研究提出了一种基于动态权重的多模型集成水产动物疾病防治事件抽取方法。该方法通过动态权重融合预训练模型ERNIE和MacBERT学习到的语义信息，并引入BiLSTM-CRF模型进行编码解码，旨在为水产动物疾病防治事件抽取提供一种新的解决方案。

1　材料与方法

1.1　试验数据采集与标注

1）数据采集与预处理。水产动物疾病防治知识主要来源于相关书籍、文献以及百度百科等多个渠道。本研究通过爬虫技术爬取这些来源中的文本，并构建了1个近30万字符的水产动物疾病防治事件语料库（DLOU-FZ）。以草鱼出血病为例，该语料库中包含注射草鱼出血病组织浆灭活疫苗、发病季节全池泼洒二氧化氯和表面活性剂等消毒剂以及全池施用大黄或黄芩等抗病毒中草药，用量为1～2.5 g/m^{3}等防治事件。

对DLOU-FZ进行预处理时，首先需要将超长文本切分成短文本，以保持文本的语义一致性。在切分过程中，保留原文本的格式，并使用逗号、顿号、分号、冒号、句号和连接符等符号对短文本进行切分。然后，将DLOU-FZ语料库中包含的一些特殊字符，如表情字符和中英文省略号等无关字符删除。最后，将语料库转换为每行1个字符的格式，并使用UltraEdit文本编辑工具对数据进行编辑，以获得规范的文本。

2）事件触发词及论元划分。通过对DLOU-FZ语料库进行事件分析并咨询水产领域专家后发现，该领域更加关注水产动物疾病相关药物的正确使用方法。经与水产专家讨论后，将水产动物疾病防治事件触发词划分为预防和治疗2类。事件论元包括水产动物疾病、药物、用药频率、用药时间、药物用量和药物用法。其中，药物论元角色包括8类，分别是环境改良剂、消毒剂、抗微生物药、杀虫驱虫药、代谢改善和强壮药、中草药、生物制品和辅助类药物。

3）标注方法。由于水产动物疾病防治事件涉及大量专有名词，且部分疾病和药物名称过长，其分词准确率较低，例如“聚乙烯毗咯烷酮碘”，分词结果可能存在误差，直接影响事件抽取的准确性。为此，本研究采用字符级别的形式处理DLOU-FZ语料库，而非传统的分词方式。整体标注采用BIO标注模式，其中B表示Begin，即一个疾病实体、触发词或论元的开始；I表示Inside，即非起始位置；O表示Outside，即非疾病实体、触发词或论元。这种标注方式可有效处理名称分词不准确的问题，提高事件抽取的精确度。

在DLOU-FZ语料库中，首先对水产动物疾病实体进行标注，标签定义为B-H和I-H，其中H表示Head，即水产动物疾病名称。使用实体名称对应的中文缩写作为实体标签，标注水产动物疾病防治事件的触发词及论元。具体而言，对于预防事件触发词，采用B-TRI-YF和I-TRI-YF标签进行标注；对于治疗事件触发词，采用B-TRI-ZL和I-TRI-ZL标签进行标注。对于只有1个字的事件触发词，采用S-TRI-YF和S-TRI-ZL标签，其中S表示Single，TRI表示事件触发词的英文单词trigger。这种标签定义方式可以准确标注水产动物疾病防治事件中的各种触发词和论元，为后续的事件抽取和分析提供有效的数据支持。

水产动物疾病防治事件的论元和论元角色采用特定的标签来进行定义，具体定义如表1所示。其中，药物论元角色标签采用2个中文首字母作为标签，如“生石灰”对应的标签为B-HJ、I-HJ、I-HJ；“草鱼出血病组织浆灭活疫苗”对应的标签为B-SW、I-SW、I-SW、I-SW、I-SW、I-SW、I-SW、I-SW、I-SW、I-SW、I-SW、I-SW。为降低人工标注成本，采用一个关键字符来区分其他4个论元，例如药物用量的标签用Y表示，药物用法的标签用F表示，用药时间的标签用T表示，用药频率的标签用P表示。

以草鱼出血病为例，采用定义好的标签进行标注。文本为“草鱼出血病【预防方法】注射草鱼出血病组织浆灭活疫苗。”其中，疾病实体为草（B-H）鱼（I-H）出（I-H）血（I-H）病（I-H）；防治事件的触发词为注（B-TRI-YF）射（I-TRI-YF），属于预防类别；论元为草（B-SW）鱼（I-SW）出（I-SW）血（I-SW）病（I-SW）组（I-SW）织（I-SW）浆（I-SW）灭（I-SW）活（I-SW）疫（I-SW）苗（I-SW），属于生物制品类别；未涉及到的字符应被标注为“O”。

1.2　试验数据与参数

本试验在Ubuntu16.04操作系统和Python3.75编程语言环境下进行，所有的训练数据都是采用相同的DLOU-FZ标注数据，并在不同的模型上进行试

表1 事件论元标签定义

Table 1 Event argument label definition

类别 Category	药物举例 Drug example	标签 Label
环境改良剂 Environment improver	如生石灰、沸石等 Such as quicklime, zeolite, etc	HJ
消毒剂 Disinfectant	如漂白粉、高锰酸钾等 Such as bleaching powder, potassium permanganate, etc	XD
抗微生物药 Antimicrobials	如四环素、复方新诺明等 Such as tetracycline, cotrimoxazole, etc	KW
杀虫驱虫药 Insecticide	如硫酸铜、敌百虫等 Such as copper sulfate, trichlorfon, etc	SC
代谢改善和强壮药 Metabolism improvers and strength pills	如维C、蛋氨酸等 Such as vitamin C, methionine, etc	DX
中草药 Chinese herbal medicine	如大黄、穿心莲等 Such as rhubarb, andrographis, etc	ZC
生物制品 Biological products	包括疫苗、免疫激活剂、某些激素、诊断试剂、生物水质净化剂等 Including vaccines, immune activators, certain hormones, diagnostic reagents, biological water purification agents, etc	SW
辅助类药物 Auxiliary drugs	如山梨酸、叔丁基对羟基茴香醚等 Such as sorbic acid, tert-butyl *p*-hydroxyanisole, etc	FZ

验。已经标注好的数据集随机按照9∶1比例分为训练集和测试集。经过反复试验，确定了最佳的模型参数，包括学习率为5e-5、批次处理大小为32、BiLSTM维度为512、序列最大长度为128和隐藏层维度为768。考虑到Adam[10](adaptive moment estimation)优化算法具有占用资源少、模型收敛快等优点，采用了该算法。

1.3 评价指标

本试验所使用的评价指标为精确率(precision，P)、召回率(recall，R)以及F_1值(F_1-score)。F_1值是P和R的调和平均数，用于综合评估P和R的表现。P、R以及F_1值的计算过程如公式(1)～(3)所示：

$$P=\frac{P_T}{P_T+P_F}\times 100\% \tag{1}$$

$$R=\frac{P_T}{P_T+N_F}\times 100\% \tag{2}$$

$$F_1=\frac{2P\cdot R}{P+R} \tag{3}$$

式(1)～(3)中，P_T是模型预测正确的防治事件实体为真的实体个数，P_F是模型预测错误的防治事件实体为真的实体个数，N_F是模型预测正确的防治事件实体为假的实体个数。

1.4 基于动态权重的多模型集成事件抽取方法

针对水产动物疾病防治事件抽取任务中常见的事件实体过长、边界模糊等问题，提出以下方案。首先，对原始语料进行预处理，并将其输入到预训练模型ERNIE[11]和MacBERT[12]中，利用它们的预训练能力来更好地学习语料的初始特征，从而提高整体模型效果；其次，提出了一种基于动态权重的融合方法，即为ERNIE和MacBERT的输出添加门控(gate)模块，赋予模型动态权重，以使其具备更强的语义信息提取能力，更好地解决水产动物疾病防治事件中的长实体和边界模糊问题；然后，引入BiLSTM[13]模型以获取其输出信息中的上下文语义依赖，进一步提取长距离的语义信息，并输出最初的序列标签；最后，利用条件随机场(CRF)添加约束，去除不合法的标签序列，进而提高抽取效果。通过以上多项优化，模型在水产动物疾病防治事件抽取任务中表现优异，模型框架如图1所示。

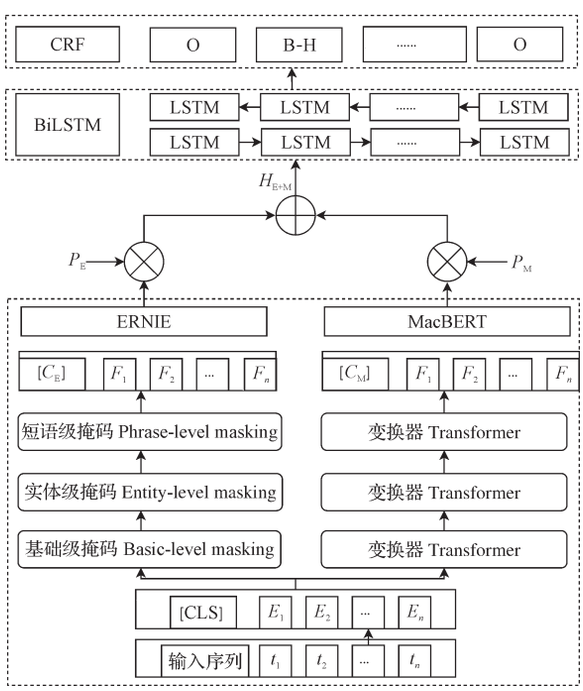

图1 模型框架

Fig.1 Model frame

1)预训练模型。ERNIE通过改进BERT中的MLM(masked language model)任务的实体级掩盖策略和短语级掩盖策略来提高模型性能。MLM任务通过随机掩盖输入中的某些标记(token)，仅根据其上下文预测原始单词。实体级掩盖是将多个单词组成的命名实体进行掩盖，短语级掩盖则是掩盖连续

的单词。通过这种掩盖策略，ERNIE可以从DLOU-FZ语料库中隐式地学习到更多的语法和语义信息，提高总体模型效果。

MacBERT采用了全词掩盖和N-gram掩盖策略，并使用了基于Word2Vec相似度计算的同义词工具包来获取相似的单词。当选择N-gram进行掩盖时，MacBERT将找到对应的相似单词进行替换。如果没有相似单词可用，MacBERT将随机选择1个单词进行替换。在预测下一个句子方面，MacBERT使用了ALBERT[14]引入的句子顺序预测任务，取代了BERT中的NSP（next sentence prediction）任务。这一改变在下游任务中取得了不错的效果。

ERNIE和MacBERT采用了特殊的学习策略，可以隐式地学习比BERT更长的语义信息。这使得ERNIE和MacBERT在解决水产动物疾病防治事件文本中存在的事件实体边界模糊导致实体识别效果不佳的问题方面表现优异。MacBERT使用同义词工具包来训练文本，并在训练过程中逐渐增加掩码遮住的词语数量，以降低模型对周边词的依赖，使得模型学习得更加充分，特别是在语料非常充分的情况下，模型可以学习到更远距离的特征表示。这一策略在一定程度上解决了水产动物疾病防治事件实体过长的问题，并提高了事件抽取效果。因此，本研究引入ERNIE和MacBERT模型同时训练DLOU-FZ语料库。

2）基于动态权重的融合方法。在ERNIE和MacBERT模型学习语义信息的过程中，添加了gate模块来有侧重地学习每个特征，使得模型能够自适应地调整权重学习最为适合的超参，并采用加权求和的方法进行特征融合。通过动态权重融合，模型可以具备更强大的特征提取能力，从而有效融合更准确的语义信息，进一步提高水产动物疾病防治事件抽取的效果。例如对于给定句子$X_i=\{x_1,x_2,\ldots,x_n\}$，其中$X_i$表示输入句子的第$i$个字，$n$为句子所包含字的个数。如公式（4）～（8）所示，$L$表示Logits，是预训练模型训练文本输出的概率。E表示ERNIE，M表示MacBERT。W表示线性层，b表示偏置，P表示gate模块输出的结果，即权重。H表示将ERNIE和MacBERT模型通过gate提供动态权重并进行融合后的输出。

$$L_{\mathrm{E}}=(X) \tag{4}$$

$$P_{\mathrm{E}}=\frac{1}{1+\mathrm{e}^{-x}}(W_{\mathrm{E}}\cdot L_{\mathrm{E}}+b_{\mathrm{E}}) \tag{5}$$

$$L_{\mathrm{M}}=\mathrm{MacBERT}(X) \tag{6}$$

$$P_{\mathrm{M}}=\frac{1}{1+\mathrm{e}^{-x}}(W_{\mathrm{M}}\cdot L_{\mathrm{M}}+b_{\mathrm{M}}) \tag{7}$$

$$H=P_{\mathrm{E}}\cdot L_{\mathrm{E}}+P_{\mathrm{M}}\cdot L_{\mathrm{M}} \tag{8}$$

3）BiLSTM层。采用双向递归神经网络BiLSTM模型对输入文本进行处理，使文本序列中的每个单词都包含完整的前后双向特征，从而提供更全面的语义信息。分析DLOU-FZ语料库发现，水产动物疾病防治事件大多为长文本，因此采用BiLSTM模型获取文本中远距离的防治事件实体之间的联系，进一步解决文本中存在的长距离依赖问题。

4）CRF层。条件随机场（conditional random field，CRF）是目前解决序列标注问题的主流方法。在水产动物疾病防治事件抽取中，BiLSTM模型能够提取文本的双向语义信息，但未能考虑实体之间的依存关系。因此，为消除不合法的标签序列，采用CRF对预测输出的标签进行约束。例如，在句子中标注触发词时，标签的首字符应该是以“B-”或“S-”开头，而不是“I-”。如果事件实体或者句子以“I-，I-”开头，则不符合规则。水产动物疾病实体标签以“B-H”开始时，后续标签只能是若干个“I-H”标签。引入CRF后，可以使预测的标签序列更加规范和合理、从而提高预测准确率[15]。表2展示了水产动物疾病防治事件抽取输出结果。

表2　草鱼出血病防治事件抽取输出结果

Table 2　Grass carp haemorrhagic disease prevention event extraction output result

抽取框架 Extraction framework	抽取结果 Extract result
原始内容 Original content	草鱼出血病【预防方法】注射草鱼出血病组织浆灭活疫苗 Grass carp hemorrhagic disease【prevention method】inject grass carp hemorrhagic disease tissue plasma inactivated vaccine
事件触发词 Event trigger	“注射”：预防事件 “Injections”: preventing events
事件论元 Event argument	“草鱼出血病”：疾病名称 “Grass carp hemorrhagic disease”: the name of the disease “草鱼出血病组织浆灭活疫苗”：论元生物制品 “Grass carp hemorrhagic disease histological plasma inactivated vaccine”: lunyuan biological products

1.5 水产动物疾病防治事件抽取模型对比设计

目前，事件抽取的主流方法是基于预训练模型（pre-trained model，PTM）-BiLSTM-CRF的方法。本研究在此基础上进一步将ERNIE和MacBERT预训练模型融合。为了验证这2个预训练模型的有效性，选取在中文抽取任务上表现较为突出的预训练模型BERT[16]、RoBERTa[17]和ELECTRA[18]作为对照模型；为了验证引入预训练模型与BiLSTM对水产动物疾病防治事件抽取的有效性，选取BiLSTM-CRF[19]、PTM-CRF和PTM-BiLSTM-CRF作为对照模型；另外，为了验证本研究采用的动态权重gate模块的有效性，选取不同的融合方法进行对比试验。

2 结果与分析

2.1 消融实验

为测试本研究模型的性能，将近年来基于不同预训练模型的深度学习方法与本研究模型进行性能对比试验。为保证试验客观和公平，所有模型在相同的训练集和测试集下进行训练和测试。结果如表3所示。

表3 消融实验结果

Table 3 Ablation experiment result

模型 Models	精确率 Precision	召回率 Recall	F_1值 F_1-score
BERT-CRF	0.809 6	0.558 6	0.661 1
ERNIE-CRF	0.819 7	0.576 0	0.676 6
RoBERTa-CRF	0.830 6	0.556 0	0.666 1
ELECTRA-CRF	0.817 0	0.540 8	0.650 8
MacBERT-CRF	0.835 3	0.575 2	0.681 3
BiLSTM-CRF	0.731 4	0.429 6	0.541 3
BERT-BiLSTM-CRF	0.857 4	0.567 3	0.682 8
ERNIE-BiLSTM-CRF	0.830 9	0.591 5	0.691 1
RoBERTa-BiLSTM-CRF	0.845 9	0.558 3	0.672 6
ELECTRA-BiLSTM-CRF	0.825 5	0.568 1	0.673 0
MacBERT-BiLSTM-CRF	0.846 0	0.582 8	0.690 2
本研究模型 Proposed method	0.852 2	0.656 2	0.741 5

从表3中的F_1值可以看出，预训练模型中以ERNIE和MacBERT为基础的模型表现最为出色，相比其他预训练模型能更好地解决水产动物疾病防治事件抽取任务所面临的问题。相反，以BERT、RoBERTa和ELECTRA为基础的预训练模型在此任务中表现不佳，说明这些预训练模型的训练策略并不适用于水产动物疾病防治事件抽取任务。因此，本研究需要根据抽取任务和文本特点选择适合的预训练模型，以达到更好的抽取效果。

从表3的数据可以看出，PTM-CRF在水产动物疾病防治事件抽取任务中的表现比BiLSTM-CRF模型提升了14个百分点，ERNIE-BiLSTM-CRF模型的F_1值比其他模型提升了14.98个百分点。本研究模型的F_1值比BiLSTM-CRF模型提高了20.02个百分点。这充分说明了预训练模型在该任务中的有效性。因此，在本研究中采用预训练模型能够更精准地捕捉文本的语义信息，提高水产动物疾病防治事件抽取效果。

此外，PTM-BiLSTM-CRF的试验结果优于单一的PTM-CRF的结果。这表明BiLSTM通过获取文本前后双向的语义特征，进一步解决了水产动物疾病防治事件文本中存在的中长距离依赖问题，证明了BiLSTM模型在水产动物疾病防治事件抽取中的有效性。因此，在本研究抽取任务中，采用预训练模型PTM与BiLSTM结合的模型，以获得更好的性能表现。

2.2 融合方法性能对比

为验证本研究提出的基于动态权重的多模型集成水产动物疾病防治事件抽取方法的有效性，选取ERNIE ⊕ MacBERT-CRF模型和ERNIE ⊕ MacBERT-BiLSTM-CRF模型（⊕代表简单相加求平均的融合方法）作为对照模型，与本研究提出的基于动态权重的方法进行性能对比试验，结果如表4所示。

表4 不同融合方法性能对比

Table 4 Comparison of performance of different fusion methods

模型 Models	精确率 Precision	召回率 Recall	F_1值 F_1-score
ERNIE⊕MacBERT-CRF	0.825 8	0.613 5	0.704 0
ERNIE⊕MacBERT-BiLSTM-CRF	0.849 6	0.630 0	0.723 5
本研究模型 Proposed method	0.852 2	0.656 2	0.741 5

从表4可以看出，ERNIE与MacBERT模型进行特征融合后不使用BiLSTM模型，比表3中所有试验效果都好。在采用简单相加求平均的融合方法的同时，引入BiLSTM-CRF模型，相比于ERNIE ⊕ MacBERT-CRF模型的F_1值提升了1.95个百分点。本研究提出的基于动态权重的多模型集成方法，通过运用动态权重思想，在学习语义信息的过程中进行自适应调整2个预训练的权重，并采取加权和的融合方法，进一步提高了模型提取特征能力；同时，针

对DLOU-FZ语料的结构特点，采用BiLSTM-CRF模型解决了文本中存在的长距离依赖问题，并添加约束以克服标签偏差问题。试验结果表明，本研究提出的模型精确率、召回率和F_1值分别达到85.22%、65.62% 和 74.15%。与 ERNIE ⊕ MacBERT-BiLSTM-CRF 模型相比，F_1值提高了1.8个百分点。由此可见，本研究提出的模型在水产动物疾病防治事件抽取具有更好的效果。

2.3　模型对比结果

针对水产动物疾病防治事件存在的事件实体过长和边界模糊问题，选取2个输入句子进行抽取结果对比，试验结果如表5和表6所示。

表5　长事件实体抽取结果对比

Table 5　Comparison of long event entity extraction results

输入句子 Input sentence	事件实体 Event entity	抽取方法 Extraction method	抽取结果 Extract result
草鱼出血病【预防方法】注射草鱼出血病组织浆灭活疫苗	“草鱼出血病”:(论元,疾病名称),“注射”:(触发词,预防),“草鱼出血病组织浆灭活疫苗”:(论元,生物制品)	BiLSTM-CRF	**“注射”**:(触发词,治疗),“出血病”:(论元,疾病名称),“疫苗”:(论元,生物制品)
		ERNIE-BiLSTM-CRF	**“注射”**:(触发词,预防),**“草鱼出血病”**:(论元,疾病名称),“灭活疫苗”:(论元,生物制品)
		MacBERT-BiLSTM-CRF	**“注射”**:(触发词,预防),**“草鱼出血病”**:(论元,疾病名称),“组织浆灭活疫苗”:(论元,生物制品)
		本研究模型 Proposed method	**“注射”**:(触发词,预防),**“草鱼出血病”**:(论元,疾病名称),**“草鱼出血病组织浆灭活疫苗”**:(论元,生物制品)

从表5可以看出，相比于其他模型，仅使用BiLSTM-CRF模型只能识别短事件实体，如“注射”。而对于较长的事件论元，如“草鱼出血病组织浆灭活疫苗”，该模型的识别效果较差。这是因为该模型在学习过程中缺乏预训练模型的输入，导致对实体的语义信息掌握不足，从而对“注射”这个触发词预测错误。其他模型使用了预训练模型，并通过不同的掩码策略学习到实体级和短语级等语义信息，因此可以更准确地预测“注射”触发词为预防类别。在这些模型中，MacBERT-BiLSTM-CRF 模型相比 BiLSTM-CRF 和 ERNIE-BiLSTM-CRF 模型能够找到并识别出更长的论元，如“组织浆灭活疫苗”，但仍不能识别完整的长事件论元。相比之下，本研究提出的模型能够准确找到并识别长事件论元，如“草鱼出血病组织浆灭活疫苗”，并学习到更加丰富的语义信息，从而提高了事件抽取的性能。

表6　事件实体边界模糊抽取结果对比

Table 6　Comparison of event entity boundary fuzzy extraction results

输入句子 Input sentence	事件实体 Event entity	抽取方法 Extraction method	抽取结果 Extract result
细锚三代虫病【治疗方法】2%～3%的食盐溶液浸洗 10 min，可达到杀死三代虫的目的	“细锚三代虫病”:(论元,疾病名称),“浸洗”:(触发词,治疗),“食盐溶液”:(论元,消毒剂),“2%～3%”:(药物用量),“10 min”:(论元,用药时间)	BiLSTM-CRF	“虫病”:(论元,疾病名称),**“浸洗”**:(触发词,治疗),“食盐”:(论元,辅助类药物),“3%”:(药物用量),**“10 min”**:(论元,用药时间)
		ERNIE-BiLSTM-CRF	“三代虫病”:(论元,疾病名称),**“浸洗”**:(触发词,治疗),“食盐”:(论元,消毒剂),“2%～3%”:(药物用量),**“10 min”**:(论元,用药时间)
		MacBERT-BiLSTM-CRF	**“细锚三代虫病”**:(论元,疾病名称),**“浸洗”**:(触发词,治疗),“食盐”:(论元,辅助类药物),**“2%~3%”**:(药物用量),**“10 min”**:(论元,用药时间)
		本研究模型 Proposed method	**“细锚三代虫病”**:(论元,疾病名称),**“浸洗”**:(触发词,治疗),**“食盐溶液”**:(论元,消毒剂),**“2%~3%”**:(论元,药物用量),**“10 min”**:(论元,用药时间)

在水产领域，由于专有名词众多以及边界模糊的问题，事件抽取变得更加困难。从表6可以看出，句子中的“食盐”是一个典型的例子。这个词既可以作为辅助类药物拌入药饵中投喂给水产动物，也可以用作消毒剂和杀虫药。与其他模型相比，BiLSTM-CRF 和 MacBERT-BiLSTM-CRF 模型只能识别出“食盐”是中草药，而 ERNIE-BiLSTM-CRF 模型可以识别出“食盐”是消毒剂。本研究提出的模

型能够准确识别出存在边界模糊的事件论元，如“食盐溶液”。这表明，本研究模型能够更好地处理水产动物疾病防治事件存在的复杂语境，从而提高事件抽取的效率。

3 讨 论

本研究提出了一种基于动态权重的多模型集成水产动物疾病防治事件抽取方法。该方法使用ERNIE和MacBERT预训练模型获取DLOU-FZ语料更全面的语义信息；通过gate模块赋予模型动态权重，并采取加权求和的方式将2种预训练模型的输出进行融合，充分考虑了语料的原始语义信息并提高了语义的准确性；同时，利用BiLSTM模型提取融合后的语义信息，并解决文本中长距离语义依赖问题；最后，使用CRF添加约束去除非法标签，有效提高了模型性能。该方法能够解决水产动物疾病防治事件抽取中存在的事件实体过长、边界模糊等问题，并获得更加准确的抽取结果。与其他模型对比试验表明，本研究模型具有更好的事件抽取性能，有效提升了水产动物疾病防治事件抽取的效果。然而，在水产动物疾病防治事件中仍存在一些有抽取意义的事件论元，例如“pH值”，由于样本较少且句式较为复杂，本研究中抽取效果并不理想。因此，下一步的研究重点是探索事件抽取任务中的少样本、零样本学习。此外，由于预训练模型较大，对设备要求较高，如何减少试验成本且不降低模型性能也是未来的研究方向。综上所述，本研究提出的基于动态权重的多模型集成方法，能够应用于水产动物疾病防治事件抽取任务，促进水产健康养殖，并可为后续深入研究提供参考。

参考文献 References

[1] 张善文，王振，王祖良.结合知识图谱与双向长短时记忆网络的小麦条锈病预测[J].农业工程学报，2020，36(12)：172-178.ZHANG S W，WANG Z，WANG Z L.Prediction of wheat stripe rust disease by combining knowledge graph and bi-directional long short term memory network[J].Transactions of the CSAE，2020，36(12)：172-178 (in Chinese with English abstract).

[2] 杨鹤，于红，孙哲涛，等. 基于双重注意力机制的渔业标准实体关系抽取[J].农业工程学报，2021，37(14)：204-212.YANG H，YU H，SUN Z T，et al.Fishery standard entity relation extraction using dual attention mechanism[J].Transactions of the CSAE，2021，37(14)：204-212 (in Chinese with English abstract).

[3] 刘巨升，杨惠宁，孙哲涛，等.面向知识图谱构建的水产动物疾病诊治命名实体识别[J].农业工程学报，2022，38(7)：210-217.LIU J S，YANG H N，SUN Z T，et al.Named-entity recognition for the diagnosis and treatment of aquatic animal diseases using knowledge graph construction[J].Transactions of the CSAE，2022，38(7)：210-217 (in Chinese with English abstract).

[4] 项威.事件知识图谱构建技术与应用综述[J].计算机与现代化，2020(1)：10-16.XIANG W.Reviews on event knowledge graph construction techniques and application[J].Computer and modernization，2020(1)：10-16(in Chinese with English abstract).

[5] 贾美英，杨炳儒，郑德权，等.基于模式匹配的军事演习情报信息抽取[J].现代图书情报技术，2009(9)：70-75.JIA M Y，YANG B R，ZHENG D Q，et al.Sham battle information extraction based on pattern matching[J].New technology of library and information service，2009(9)：70-75(in Chinese with English abstract).

[6] 李浩瑞，王健，林鸿飞，等.基于混合模型的生物事件触发词检测[J].中文信息学报，2016，30(1)：36-42.LI H R，WANG J，LIN H F，et al.A hybrid approach to trigger detection in biological event extraction[J].Journal of Chinese information processing，2016，30(1)：36-42(in Chinese with English abstract).

[7] 万齐智，万常选，胡蓉，等.基于句法语义依存分析的中文金融事件抽取[J].计算机学报，2021，44(3)：508-530.WAN Q Z，WAN C X，HU R，et al.Chinese financial event extraction based on syntactic and semantic dependency parsing[J].Chinese journal of computers，2021，44(3)：508-530(in Chinese with English abstract).

[8] YANG S，FENG D W，QIAO L B，et al.Exploring pre-trained language models for event extraction and generation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.Stroudsburg，PA，USA：Association for Computational Linguistics，2019：5284-5294.

[9] 陈星月，倪丽萍，倪志伟.基于ELECTRA模型与词性特征的金融事件抽取方法研究[J].数据分析与知识发现，2021，5(7)：36-47.CHEN X Y，NI L P，NI Z W.Extracting financial events with ELECTRA and part-of-speech[J].Data analysis and knowledge discovery，2021，5(7)：36-47(in Chinese with English abstract).

[10] KINGMA D P，ADAM B A J. A method for stochastic optimization [C]//International conference on learning representations. Ithaca：NYarXiv.org，2014.

[11] 李舟军，范宇，吴贤杰.面向自然语言处理的预训练技术研究综述[J].计算机科学，2020，47(3)：162-173.LI Z J，FAN Y，WU X J.Survey of natural language processing pre-training techniques [J].Computer science，2020，47(3)：162-173 (in Chinese with English abstract).

[12] CUI Y，CHE W，LIU T，et al. Revisiting pre-trained models for chinese natural language processing[C]//Findings of the association for computational linguistics：EMNLP 2020. [S.l.]：[s.n.]，2020：657-668.

[13] 王子牛，姜猛，高建瓴，等.基于BERT的中文命名实体识别方法[J].计算机科学，2019，46(S11)：138-142.WANG Z N，JIANG M，GAO J L，et al.Chinese named entity recognition method based on BERT[J].Computer science，2019，46(S11)：138-142 (in Chinese with English abstract).

[14] 李军怀,陈苗苗,王怀军,等.基于ALBERT-BGRU-CRF的中文命名实体识别方法[J].计算机工程,2022,48(6):89-94,106. LI J H, CHEN M M, WANG H J, et al. Chinese named entity recognition method based on ALBERT-BGRU-CRF[J]. Computer engineering, 2022, 48(6): 89-94, 106(in Chinese with English abstract).

[15] 余本功,范招娣.面向自然语言处理的条件随机场模型研究综述[J].信息资源管理学报,2020,10(5):96-111.YU B G,FAN Z D.A review of conditional random field models for natural language processing[J]. Journal of information resources management,2020,10(5):96-111(in Chinese with English abstract).

[16] DEVLIN J,CHANG M W,LEE K,et al.BERT:pre-training of deep bidirectional transformers for language understanding[DB/OL]. arXiv, 2018: 1810.04805. https://doi. org/10.48550/arXiv.1810.04805.

[17] 喻雪寒,何琳,徐健.基于RoBERTa-CRF的古文历史事件抽取方法研究[J].数据分析与知识发现,2021(7):26-35.YU X H, HE L, XU J.Extracting events from ancient books based on RoBERTa-CRF[J]. Data analysis and knowledge discovery, 2021(7):26-35(in Chinese with English abstract).

[18] CLARK K, LUONG M T, LE Q V, et al. ELECTRA: pre-training text encoders as discriminators rather than generators[C]//International conference on learning representations. arXiv:computation and language.[S.l.]:[s.n.],2020.

[19] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition[C]//Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016: 260-270.

Multi-model integrated event extraction for aquatic animal disease prevention and control based on dynamic weight

SHA Mingyang[1], ZHANG Sijia[1,2], FU Qingcai[1], YU Hong[1,2], LI Zhiqi[1], YU Wenfu[1], LIU Jianing[1]

1.*College of Information Engineering/Liaoning Provincial Key Laboratory of Marine Information Technology, Dalian Ocean University, Dalian* 116023, *China*;
2.*Key Laboratory of Environment Controlled Aquaculture(Dalian Ocean University), Ministry of Education, Dalian* 116023, *China*

Abstract In order to enhance the accuracy of event extraction for aquatic animal disease prevention and control, and effectively address issues such as ambiguous boundaries of proprietary terms and excessively lengthy event entities during the extraction process, the research introduces the idea of dynamic weight into the event extraction method of multi-model integration. Two pre-training models, ERNIE (enhanced representation through knowledge integration) and MacBERT (MLM as correction BERT), are used to learn the text semantic information. A gate module with dynamic weights is used to fuse features to enhance the semantic information of the original text. Pass the learned semantic information into BiLSTM (bi-directional long shortterm memory), and constrain the output label sequence through CRF (conditional random field). Select the ERNIE⊕MacBERT-CRF model and the ERNIE⊕MacBERT-BiLSTM-CRF model (⊕ represents the fusion method of simple addition and averaging) as the control model to conduct a comparative test of the fusion performance of the proposed method. The results show that the *F*1-score of this method reaches 74.15%, which is 20.02 percentage points higher than the classic model BiLSTM-CRF. The results show that this method has a better effect in the extraction of aquatic animal disease prevention and control events.

Keywords aquatic animal diseases; event extraction; ERNIE; MACBERT; dynamic weight; healthy aquaculture

(责任编辑:边书京)