DOI:10.26974/j.cnki.XBGC.2026.01.004

融合多头注意力的中文命名实体识别方法

胡德洲, 李贯峰*, 李 瑞, 王云丽, 高文馨
(宁夏大学 信息工程学院,宁夏 银川 750021)

摘 要:针对现有中文命名实体识别模型特征抽取能力不足,难以捕捉长距离依赖等问题,提出一种融合多头注意力(multi-head attention,MA)和双向门控循环单元(bidirectional gate recurrent unit,BiGRU)的中文命名实体识别方法。首先,使用ALBERT(a lite BERT)预训练模型生成动态表示向量,并将向量序列输入到BiGRU来提取全局语义特征,再利用多头注意力机制捕捉长距离依赖信息来增强语义特征,最后,通过条件随机场(conditional random field,CRF)解码获得最优序列。结果表明,该方法在《人民日报》和MSRA中文数据集上F_1值均超过95%,优于其他模型;同时,该方法相比BERT-BiLSTM-CRF模型,训练时间减少约14.5%,证明了模型的有效性和通用性。

关键词:命名实体识别;多头注意力;双向门控循环单元;条件随机场
中图分类号:TP391.1;TP18 **文献标志码**:A

命名实体识别(named entity recognition,NER)旨在从文本中抽取具有特定意义的实体。命名实体识别在机器翻译、文本理解、问答系统等应用中都发挥着巨大的作用[1]。英文命名实体识别的发展较早,已经取得了不错的成绩,但中文命名实体的识别更加复杂,这主要是因为中文命名实体识别的效果受分词结果的影响较大,并且汉语中一词多义的现象十分普遍,这些都给中文命名实体识别带来了更多的挑战[2]。传统的NER方法主要是基于规则[3]和统计机器学习[4],传统方法具有可解释性,但比较耗费人力,并且可迁移性较差。

得益于深度学习强大的特征抽取能力,基于深度学习的NER方法目前处于主导地位。尤丽珏等[5]使用BiLSTM-CRF(bidirectional long short-term memory-conditional random field)模型对医学影像检查报告进行实体识别,取得了良好的效果,但Word2Vec无法解决中文一词多义的问题。宋佳芮等[6]使用注意力机制对编码层提取的文本信息进行语义补充,在CoNLL-2003数据集上F_1值达到了91.35%,证明了注意力机制在命名实体识别中的有效性,但没有使用BERT(bidirectional encoder representations from transformers)等预训练模型来生成动态语义向量。Zhang等[7]提出了Lattice-LSTM(lattice-long short-term memory)中文NER模型,该模型既可以避免中文分词错误,还可以自动从上下文中匹配词语信息,但是该模型无法并行化训练,且迁移性较差。关斯琪等[8]使用BERT来生成词嵌入向量,融合了左右两侧语境信息,增强了字的语义表示,有效缓解了一词多义的问题。但BERT模型参数量大、训练时间较长,导致基于BERT的模型在专业领域的应用受限[9]。张祺等[10]将BERT-IDCNN-CRF模型应用于军事领域实体识别(IDCNN指iterated dilated convolutional neural network),效果优于Lattice-LSTM模型。Wang等[11]提出了TPlinker联合抽取模型,同时实现了对实体和实体关系的抽取。可见,现有的研究工作并没有将预训练语言模型、深度神经网络和多头注意力三者的优势进行充分结合,BERT参数过大也导致模

收稿日期:2023-12-07
基金项目:国家自然科学基金项目(62066038);宁夏自然科学基金项目(2022AAC03026);宁夏大学研究生创新项目(CXXM202356)
作者简介:胡德洲(2000—),男,硕士研究生,主要从事命名实体识别研究(nxuhdz@163.com)。
* **通信作者**:李贯峰(1979—),男,副教授,博士,主要从事语义计算、知识图谱研究(ligf@nxu.edu.cn)。
引用格式:胡德洲,李贯峰,李瑞,等.融合多头注意力的中文命名实体识别方法[J].西北工程技术学报(中英文),2026,25(1):27-32.

型在实际应用中受限。

针对现有的中文命名实体识别模型存在着长距离依赖信息捕捉不全、特征抽取能力不足等问题,本文提出了 ALBERT-BiGRU-MA-CRF 命名实体识别模型。BERT 模型可以解决一词多义的问题,但其参数较多,因此,本文选择采用其轻量级版本 ALBERT(a lite BERT)来生成文本序列的动态表示向量;BiLSTM (bidirectional long short-term memory)的网络结构较为复杂、参数较多,采用参数量更少的 BiGRU(bidirectional gate recurrent unit)来获取文本序列的全局特征,可以进一步缩短模型训练时间;再通过引入多头注意力(multi-head attention,MA)机制解决长距离依赖问题,提高模型特征抽取能力,最后通过 CRF(conditional random field)添加规则约束,解码得到最优序列。

1 模型设计

ALBERT-BiGRU-MA-CRF 模型的整体架构如图 1 所示。由 ALBERT 层、BiGRU 层、多头注意力层、CRF 层 4 个模块构成。模型的第一层是 ALBERT 层,在该层中使用 ALBERT 预训练模型获取结合语境信息的动态表示向量,然后通过 BiGRU 层来学习文本序列的全局语义特征,再经过多头注意力机制层获取序列文本的内部联系,更好地考虑每个特征之间的依赖关系,最后通过 CRF 层解码获取全局最优序列。

1.1 ALBERT 层

BERT 模型采用双向 Transformer 进行编码,具有较强的语义表征优势。BERT 模型通过无监督训练从文本数据中学习到丰富的语言表示,只需要在其输出上添加合适的标签分类层就可以处理命名实体识别任务,并通过微调使其适应特定的 NER 任务。ALBERT 和 BERT 一样都采用了 Transformer 作为编码器,针对 BERT 模型参数量过大,ALBERT 主要在 2 个方面进行了优化:

(1)嵌入分解。在 BERT 模型中,嵌入层的维度和隐藏层的维度是相等的,嵌入层的维度伴随隐藏层的维度的增加而增加,这是 BERT 模型参数量较大的原因。ALBERT 模型对嵌入层的维度和隐藏层的维度进行了解绑,引入辅助矩阵对嵌入层的维度进行转化。在完成对嵌入矩阵的分解之后,模型的空间复杂度(参数量)实现了显著优化。尤其当隐藏层维度 H 远大于嵌入层维度 E(即 $H>>E$)时,ALBERT 模型的参数量相较于 BERT 模型大幅减少。

(2)参数共享。BERT 模型中采用了多层 Transformer 编码器,但各层的参数并没有进行关联。为了减小模型参数量,ALBERT 模型将注意力层和全连接层参数全部进行共享,模型参数量在减少的同时,稳定性也得到了一定提高。

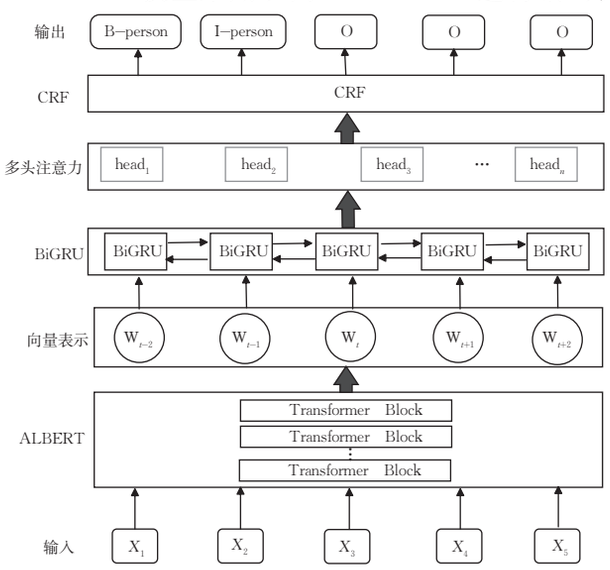

图 1 ALBERT-BiGRU-MA-CRF 模型架构

Fig. 1 Architecture of the ALBERT-BiGRU-MA-CRF model

1.2 BiGRU 层

GRU(gated recurrent unit)是一种特定的循环神经网络,和 LSTM(long short-term memory)一样,也是通过门结构来控制信息的存储和遗忘。LSTM 包括 3 个门:遗忘门、输入门和输出门。GRU 则只有重置门 r_t 和更新门 u_t,其结构如图 2 所示。

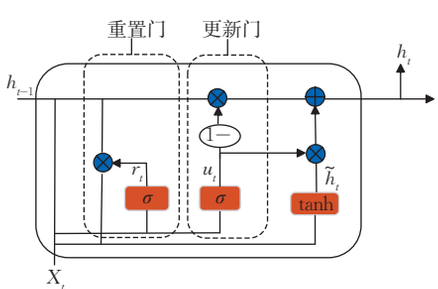

图 2 GRU 网络结构图

Fig. 2 Architecture of the GRU network

由此可知,GRU 较 LSTM 网络结构更加简单,所以参数更少,更容易进行训练。单向 GRU 不能充分考虑上下文信息,因此采用双向 GRU 网络。BiGRU 利用正向和反向 GRU 进行上下文信息特征提取,将输出进行加权求和,最后经过线性层得到最终的输出标签向量列表,计算公式[12]为

$$\overrightarrow{h_t}=\mathrm{GRU}(x_t,\overrightarrow{h_{t-1}}),\tag{1}$$

$$\overleftarrow{h_t}=\mathrm{GRU}\left(x_t,\overleftarrow{h_{t-1}}\right),\tag{2}$$

$$h_t=\left[\overrightarrow{h_t};\overleftarrow{h_t}\right]。\tag{3}$$

式中:GRU 函数表示对输入文本向量进行非线性变换,并将文本向量编码为对应的 GRU 隐藏层状态;$\overrightarrow{h_t}$ 和 $\overleftarrow{h_t}$ 分别代表 t 时刻双向 GRU 中前向隐藏层状态和反向隐藏状态;x_t 表示 t 时刻的文本向量。

1.3　多头注意力层

经过 BiGRU 层编码得到的文本信息向量具有相同的权重,但往往语句中不同词语对于命名实体识别的重要程度不一样。例如,雷军于 2010 年创立了小米公司,“2010 年”对识别“小米”这个公司实体作用较小,应分配较小的权重,“雷军”对识别“小米”这个公司实体作用较大,应分配较大的权重,这有利于提高“小米”这一实体识别的准确率。引入注意力机制可以更加有效地获取与当前信息有关联的上下文语义特征,提高模型的局部抽取能力,并且注意力权重分配不受词间距离的影响,只由词向量本身决定,有助于解决长距离依赖问题。

对于 BiGRU 层所给定的输出 H,注意力机制首先将矩阵 H 投影到矩阵:Q,K,V。本文采用的多头注意力机制通过并行化计算策略实现上下文特征捕获。具体而言,每个注意力头独立执行注意力矩阵的运算过程,随后将所有注意力头的输出结果沿特征维度进行拼接融合,从而有效整合不同语义子空间的信息表征,以增强文本上下文建模能力。Attention 代表单次缩放点积的过程,head_i 代表第 i 个注意力头,Multi(Q,K,V)代表多头注意力机制。具体计算公式[13]为

$$\mathrm{Attention}=\mathrm{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V,\tag{4}$$

$$\mathrm{head}_i=\mathrm{Attention}(QW_i^Q,KW_i^K,VW_i^V),\tag{5}$$

$$\mathrm{Multi}(Q,K,V)=\mathrm{Concat}[\mathrm{head}_1,\cdots,\mathrm{head}_n],\tag{6}$$

式中:Q 为查询向量;K 为键向量;V 为值向量;d_k 为键向量维度;W_i^Q,W_i^K,W_i^V 为权重参数矩阵;n 为自注意力头的个数。

1.4　CRF 层

条件随机场(CRF)是使用机器学习处理 NER 任务的一种常用算法,CRF 可以建模标签之间的依赖关系,也就是说可以充分考虑到相邻词汇之间的约束关系。例如,“B-ORG”标签后面不可以跟“I-LOC”标签,“B-PER”标签后面只可以跟“I-PER”或“O”标签,实体只能以“B-”标签开头等。对于多头注意力层所生成的向量矩阵,首先计算其预测标签序列的得分,其次使用归一化指数函数计算输出标签序列的概率,最后使用维特比算法(Viterbi algorithm)得到文本序列的全局最优标签序列。

2　数据来源和评价标准

2.1　数据来源

实验数据选取的是中文领域较为权威的 1998 年《人民日报》和 MSRA(Microsoft Research Asia)的 NER 语料数据集,对两个数据集中的人名、地名、机构名 3 种实体进行识别,按照 8:2 的比例随机划分训练集和测试集,数据集信息统计如表 1 所示。

表 1　数据集统计信息

Tab. 1　Dataset statistics

数据集	训练集	测试集	总数
《人民日报》	36 257	9 059	45 316
MSRA	64 712	16 172	80 884

数据集采用 BIO(begin,inside,outside)模式标记。其中,B-X 代表命名实体 X 的开头;I-X 代表 X 命名实体的中间或结尾;O 代表不属于任何类型,即非实体部分。

2.2　评价标准

采用准确率(P),召回率(R),F_1 值作为评价指标。计算公式[13]分别为

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}\times 100\%,\tag{7}$$

$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\times 100\%,\tag{8}$$

$$F_1=\frac{2\times P\times R}{P+R}\times 100\%。\tag{9}$$

式中:TP 代表正样本预测为正;FP 代表负样本预测为正;FN 代表正样本预测为负。需要注意的是,虽然在数据标注时采用的是字符级的标注策略,但在对模型进行评价时,应该是面向整个实体,只有实体的边界和所属类别均判定为正确才算识别成功。

2.3　实验环境

实验在 Windows 10 操作系统下进行,采用 Pytorch 学习框架进行模型构建,具体各项参数的设置如表 2 所示。

3　结果与分析

3.1　模型对比结果

为了验证本文所提模型的有效性,基于控制变量的思想,设置了以下 4 个对比模型:LSTM-CRF

模型,BiLSTM-CRF 模型,BERT-IDCNN-CRF 模型,BERT-BiLSTM-CRF 模型。不同模型对比实验结果如表 3 和表 4 所示。

表 2 实验参数设置

Tab. 2 Experimental parameter settings

参数名称	参数值
ALBERT-Base层数	12
ALBERT-Base隐藏层维度	768
最大序列长度	128
GRU隐藏层维度	128
Dropout	0.5
Batch size	32
注意力机制头数	8
优化器	Adam

表 3 各对比模型在《人民日报》数据集上的实验结果

Tab. 3 Experimental results on *the People's Daily* dataset

%

模型名称	P	R	F_1
LSTM-CRF	84.56	80.36	82.41
BiLSTM-CRF	87.76	85.47	86.60
BERT-IDCNN-CRF	92.43	91.67	92.05
BERT-BiLSTM-CRF	94.26	94.37	94.31
ALBERT-BiGRU-MA-CRF	95.13	95.32	95.22

表 4 各对比模型在MSRA数据集上的实验结果

Tab. 4 Experimental results on the MSRA dataset

%

模型名称	P	R	F_1
LSTM-CRF	86.79	81.31	83.96
BiLSTM-CRF	89.67	87.26	88.45
BERT-IDCNN-CRF	92.45	91.76	92.11
BERT-BiLSTM-CRF	94.86	93.51	94.18
ALBERT-BiGRU-MA-CRF	94.98	95.07	95.02

从表 3 来看,在 1998 年《人民日报》NER 数据集上,本文模型在 3 项指标上均取得了最好的效果。相比 LSTM-CRF 模型,BiLSTM-CRF 模型的 F_1 值提高了 4.19%,说明 BiLSTM 对文本进行的双向编码比单向编码可以捕捉更多的语义信息。相比 BERT-IDCNN-CRF 模型,BERT-BiLSTM-CRF 模型的 F_1 值提高了 2.26%,说明 BiLSTM 对上下文的语义捕捉能力优于卷积神经网络。相比 BERT-BiLSTM-CRF 模型,本文模型的 P,R,F_1 值分别提高了 0.87%,0.95% 和 0.91%;相比 BiLSTM-CRF 模型,本文模型的 P,R 和 F_1 值分别提高了 7.37%,9.85%,8.62%。BERT-BiLSTM-CRF,BiLSTM-CRF,以及本文所提模型的核心差异在于所采用的预训练模型不同。无论是基于 BERT 模型还是 ALBERT 模型,其生成的动态词向量均显著优于 Word2Vec 生成的静态词向量,因此,相较于 BiLSTM-CRF 模型,其性能均实现了较大提升。本文模型略优于 BERT-BiLSTM-CRF 模型,主要是因为引入了多头注意力机制,让模型更加关注重点内容,有效弥补了神经网络在获取局部特征方面的不足。

从 MSRA 数据集来看(表 4),本文模型的实体识别效果也达到最佳。本文模型的 F_1 值比 BiLSTM-CRF 模型提高 6.57%,比 BERT-IDCNN-CRF 模型提高 2.91%,比 BERT-BiLSTM-CRF 模型提高 0.84%。由此可见,本文提出的命名实体识别模型具有一定的有效性和通用性。

对比模型中,效果最佳的为 BERT-BiLSTM-CRF 模型,表 5 记录了该模型和本文所提模型在《人民日报》数据集训练 1 个 Epoch 所需时间和使用的内存大小。可知,BERT-BiLSTM-CRF 模型花费了 4 287 s,而本文模型只花费 3 667 s,训练时间减少了约 14.5%。这主要得益于使用参数更少的 ALBERT 和 BiGRU 来替换常规的 BERT 和 BiLSTM,并且多头注意力也可以让模型更加关注文本中的重点信息。

表 5 模型运行效率对比

Tab. 5 Comparison of runtime efficiency

模型	使用内存/GB	训练时间/s
BERT-BiLSTM-CRF	3.49	4 287
ALBERT-BiGRU-MA-CRF	3.26	3 667

为了更加直观地展现各模型的性能,图 3 记录了各个模型在《人民日报》数据集上 F_1 值随着每个 Epoch 的更新情况。可以看出,LSTM-CRF 模型和 BiLSTM-CRF 模型的 F_1 值上升比较慢;而另外 3 个模型使用了预训练语言模型,在训练的初始阶段就取得了比较好的效果,收敛速度也更快,证明了预训练模型在实体识别中所发挥的积极作用。BERT-IDCNN-CRF 模型相比 BERT-BiLSTM-CRF 模型更早趋于收敛,这主要是因为卷积神经网络的并行能力优于 BiLSTM。本文模型的 F_1 值始

终高于对比模型,并且在第 8 个 Epoch 时,本文模型率先达到最高 F_1 值(95.22%),并随着训练次数的增加,F_1 值趋于平稳。

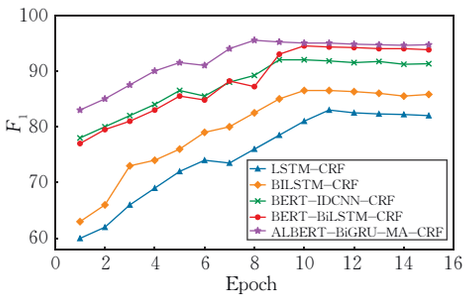

图3　各对比模型在《人民日报》数据集上的F_1值曲线图

Fig. 3　F_1 score curves of the comparison models on *the People's Daily* dataset

3.2　消融实验结果

本文在 ALBERT 预训练模型的基础上引入了 BiGRU,MA,CRF 3 个模块,为研究每个模块对模型的影响,在《人民日报》数据集上进行了对比实验,结果如表 6 所示。

表6　消融实验结果

Tab. 6　Ablation study results

%

模型	P	R	F_1
ALBERT	90.16	90.79	90.48
ALBERT-CRF	91.78	91.37	91.57
ALBERT-BiGRU-CRF	92.96	93.76	93.35
ALBERT-BiGRU-MA-CRF	95.13	95.32	95.22

可以看出,ALBERT-CRF 模型的各项指标都优于 ALBERT 模型。可见,加入 CRF 可以对标签预测提供更多的约束条件,提高模型性能。在 ALBERT-CRF 的基础上添加 BiGRU 模块,P,R,F_1 值分别提高了 1.18%,2.39%,1.78%,其原因是 BiGRU 捕获的全局双向特征对命名实体的识别效果产生了积极的影响。在 ALBERT-BiGRU-CRF 的基础上添加 MA 模块,F_1 值提高了 1.87%,其主要原因是多头注意力机制可以从多个维度捕捉文本特征,从而有效解决了长距离依赖问题。

3　结论

针对目前中文命名实体识别模型特征抽取能力不足,难以捕捉长距离依赖等问题,本文提出了 ALBERT-BiGRU-MA-CRF 中文命名实体识别模型,该模型充分利用了预训练语言模型、多头注意力机制和 BiGRU-CRF 模型的优势,可以更加全面地提取文本中的语义信息。实验表明,本文所提模型在缩短训练时间的同时能够有效识别出各类命名实体,在两个权威中文数据集上 F_1 值均超过了 95%,在一定程度上解决了预训练模型在 NER 研究中应用受限的问题。本文所提模型主要面向非嵌套实体,后续研究工作将进一步针对嵌套命名实体识别任务展开。

参考文献:

[1] Li J, Sun A X, Han J L, et al. A survey on deep learning for named entity recognition[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(1): 50-70.

[2] 彭雪, 赵辉, 郑肇谦, 等. 融合多种嵌入表示的中文命名实体识别[J]. 长春工业大学学报, 2022, 43(1): 81-90.

[3] Shaalan K, Raza H. NERA: Named entity recognition for Arabic[J]. Journal of the American Society for Information Science and Technology, 2009, 60(8): 1652-1663.

[4] Zhou G D, Su J. Named entity recognition using an HMM-based chunk tagger [C]//Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, Pennsylvania, USA Association for Computational Linguistics, 2002: 473-480.

[5] 尤丽珏, 尹远芳. 基于BiLSTM-CRF模型的医学影像检查报告信息实体识别[J]. 微型电脑应用, 2023, 39(10): 134-137.

[6] 宋佳芮, 陈艳平, 王凯, 等. 基于Affix-Attention的命名实体识别语义补充方法[J]. 山东大学学报(工学版), 2023, 53(2): 70-76.

[7] Zhang Y, Yang J. Chinese NER using lattice LSTM [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia Association for Computational Linguistics, 2018: 1554-1564.

[8] 关斯琪, 董婷婷, 万子敬, 等. 基于BERT-CRF模型的火灾事故案例实体识别研究[J]. 消防科学与技术, 2023, 42(11): 1529-1534.

[9] Sun S Q, Cheng Y, Gan Z, et al. Patient knowledge distillation for BERT model compression[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China Association for Computational Linguistics, 2019: 4323-4332.

[10] 张祺, 李成军, 刘敬蜀. 基于BERT-IDCNN-CRF的

军事领域命名实体识别研究[J]. 航天电子对抗, 2021, 37(5): 56-60.
[11] Wang Y C, Yu B W, Zhang Y Y, et al. TPLinker: Single-stage joint extraction of entities and relations through token pair linking [C]//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona, Spain (Online) Association for Computational Linguistics, 2020: 1572-1582.
[12] Zhang J R, Liu F A, Xu W Z, et al. Feature fusion text classification model combining CNN and BiGRU with multi-attention mechanism [J]. Future Internet, 2019, 11(11): 237.
[13] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]// Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, Long Beach, 2017: 5998-6008.

A Chinese Named Entity Recognition Method Integrating Multi-Head Attention

HU Dezhou, *LI Guanfeng**, *LI Rui*, *WANG Yunli*, *GAO Wenxin*
(School of Information Engineering, Ningxia University, Yinchuan 750021, China)

Abstract: To address the limited feature extraction capacity and the difficulty in modeling long-range dependencies in existing Chinese named entity recognition(NER) models, this study proposed a method that integrates multi-head attention(MA) with a bidirectional gated recurrent unit(BiGRU). Specifically, it first used the ALBERT(a lite BERT) pre-trained language model to generate contextualized representations, which are then fed into a BiGRU to extract global semantic features. A multi-head attention mechanism is subsequently applied to capture long-range dependencies and further enhance the semantic representation then the multi-head attention mechanism is used to capture long-distance dependent information to enhance semantic representations. Finally, a conditional random field (CRF) layer decodes the optimal label sequence. Experimental results show that the proposed method achieves F_1 scored above 95% on both *the People's Daily* and MSRA datasets, outperforming competing models. In addition, compared with the BERT-BiLSTM-CRF model, this approach reduces training time by approximately 14.5%, demonstrating its effectiveness and generalizability.
Keywords: named entity recognition; multi-head attention; bidirectional gated recurrent unit; conditional random field

(责任编辑 李 琼)

(上接第26页)

Video-Based Discharge Measurement Method for Main Canals in the Yellow River Irrigation District Using Intelligent Image Recognition

YIN Ting
(Ningxia Hui Autonomous Region Hanyan Canal Management Office, Yinchuan 750001, China)

Abstract: To evaluate the measurement accuracy and adaptability of video-based discharge gauging, this study targets the main canals of the Ningxia Yellow River Diversion Irrigation District. Using video data from two representative cross-sections, it proposed a real-time dynamic discharge measurement method for irrigation canals based on intelligent video-image recognition and validated it against synchronous discharge observations from a vertical acoustic Doppler current profiler (V-ADCP). The results indicate that in wide and shallow cross-sections, the video-based method achieves higher monitoring and recognition accuracy, with relative deviations ranging from −21.5% to 15.2%. Moreover, when the water surface exhibits abundant texture and floating tracers, measurement accuracy improves and adaptability is further enhanced. These findings provide a technical reference for video-based discharge gauging in the main canals of the Ningxia Yellow River Diversion Irrigation District.
Keywords: video-based discharge measurement; intelligent image recognition; comparative analysis; relative deviation; Yellow River Diversion Irrigation District

(责任编辑 李 琼)