融合 ALBERT 与 BiGRU-Attention-CRF 模型的湖泊命名实体识别

朱道恒 [1]，李志强 [1,*]，刘润 [2]

[1] 广东海洋大学电子与信息工程学院，广东湛江 524088
[2] 广东海洋大学化学与环境学院，广东湛江 524088

摘要： 中国湖泊众多且分布广泛，全面掌握湖泊信息对推动重大水网工程实施、河湖生态环境修复和智慧水利建设等具有重要意义。该文运用深度学习模型对湖泊实体进行识别与提取，为从湖泊文本中挖掘有效湖泊信息提供了参考。首先，自定义了湖泊命名实体识别(Named Entity Recognition, NER)方法和标注规范，并建立湖泊文本语料库。其次，考虑到自建的语料库规模较小，文本信息的稀疏性可能影响模型的性能和识别效果，引入轻量预训练模型(A Little Bidirectional Encoder Representations from Transformers, ALBERT)生成高质量词向量，作为双向门控循环单元-条件随机场（BiGRU-CRF）模型的输入层特征向量，并将注意力机制融入到该模型为实体的语义信息增加特征权重，提升实体特征提取效果。最后，利用自建的语料库完成大量验证试验，并对比 4 种深度学习模型的识别效果。结果表明：ALBERT-BiGRU-Attention-CRF 模型对数据集整体的识别效果良好，准确率、召回率和 F1 分别达到 91.26%、90.38%和 90.81%。该模型对高频出现的 4 类实体识别的性能均优于其他对比模型。

关键词： 命名实体识别；预训练模型；双向门控循环单元；注意力机制；条件随机场

DOI: 10.57237/j.se.2022.01.003

Fusion of ALBERT and BiGRU-Attention-CRF Models for Named Entity Recognition in Lakes

Zhu Daoheng[1], Li Zhiqiang[1,*], Liu Run[2]

[1]School of Electronics and Information Engineering, Guangdong Ocean University, Zhanjiang 524088, China
[2]School of Chemistry and Environment, Guangdong Ocean University, Zhanjiang 524088, China

Abstract: There are many lakes in China and they are widely distributed, so it is important to have comprehensive information on lakes to promote the implementation of major water network projects, ecological and environmental restoration of rivers and lakes, and smart water conservancy construction. This paper uses a deep learning model to identify and extract lake entities, and provides a reference for mining effective lake information from lake text. Firstly, the Named Entity Recognition (NER) method and annotation specification are customized, and a corpus of lake texts is established. Secondly, considering the small size of the self-built corpus and the sparsity of text information may affect the performance of the model and the recognition effect, a lightweight pre-training model (ALBERT) is introduced to generate high-quality word vectors as a two-way gating cycle The unit-conditional random field (BiGRU-CRF) model is

基金项目：本研究得到国家自然科学基金项目 (42176067)和广东海洋大学强校工程项目 (Q18307) 的部分支持.

*通信作者：李志强, qiangzl1974@163.com

收稿日期: 2022-09-29; 接受日期: 2022-10-21; 在线出版日期: 2022-11-01

used to generate the input layer feature vectors, and the attention mechanism is incorporated into the model to add feature weights to the semantic information of the entities to improve the entity feature extraction effect. Finally, a large number of validation experiments are completed, and the recognition effects of the four deep learning models are compared. The results show that the ALBERT-BiGRU-Attention-CRF model has good recognition effect on the dataset as a whole, with accuracy, recall and F1 reaching 91.26%, 90.38% and 90.81%, respectively. In addition, the model outperforms the other comparison models for the recognition of all four types of entities that occur at high frequencies.

Keywords: NER; ALBERT; BiGRU; Attention; CRF

1 引言

我国水利部在 2017 年印发了《关于推进水利大数据发展的指导意见》，旨在水利行业推进数据资源共享开放，促进水利大数据发展与创新应用[1]。湖泊作为我国水资源体系的一个重要分支，在工业、农业、制造业和运输业等方面占有重要地位，而全面准确地掌握湖泊信息是高效合理地利用湖泊、管理和保护湖泊的基础。大数据和人工智能技术的飞速发展使得各行业的数据挖掘变得高效便捷。将大数据和深度学习相关技术方法应用到水利领域，可为水利大数据的深度挖掘及信息整合提供有效途径。

命名实体识别（Named Entity Recognition, NER）作为自然语言处理（Natural Language Processing, NLP）、机器翻译（Machine Translation，MT）和问答系统中的一项基本任务，是指从非结构化文本中识别出规律、认知或定义等内容。它最先由 RAU 等[2]提出，并在 1996 年的国际消息理解会议上被正式确立为信息抽取的主要内容。经过不断发展，它提取的实体范围逐步从狭窄变得广泛，对实体的分类从粗略变得精细，实体标注效率从低能变得高效，同时提取的效率和准确度也不断提升[3-4]。

当前，通用领域和一些特定领域命名实体识别的研究已取得很大进展，如在信息检索方面，Chen 等[5]利用动态多池化层的卷积神经网络对事件进行提取；Voorhees 等[6]提出通过文本检索会议建立数据集以实现更高效的文本检索。在问题回答领域，Fader 等[7]将开放领域的问题映射到 Web 抽取数据库上来做查询，Yao 等[8]结合 Landweber 迭代和在线学习算法提出新的梯度下降算法解决了偏差-方差均衡问题；许宁[9]实现了面向旅游领域的智能问答系统。关系抽取方面，Bunescu 等[10]提出一种最短路径依赖内核方法来提升关系抽取性能；Miwa 等[11]基于序列和数结构的 LSTM 模型实现了实体关系的端到端提取；Singh 等[12]利用联合图形模型减小了实体标记、关系提取和共指三个任务在执行过程中的错误率。事件抽取方面，Han 等[13]提出集成了模式、机器学习模型和词嵌入技术的商业事件提取方法，有效提取了在线中文新闻中的事件；Chen 等[5]提出一种自动提取词汇级和句子级特征的事件提取方法，并取得不错效果；实体链接方面，Upadhyay 等[14]采用联合监督的方式开发了第一个跨语言实体链接（Cross-lingual Entity Linking, XEL）方法；Li 等[15]提出一种从粗粒度到细粒度的集体实体链接算法，有效减少了文本中提及实体的候选数量。

有关 NLP 技术在水利工程领域的应用也逐步展开，如冯均等[16]全面调研和深入分析了领域知识图谱的研究现状和发展趋势，并指构建水利领域知识图谱是全面、准确认识水利数据的重要手段。段浩等[17]通过构建水利综合知识图谱，实现了水利知识的跨越查询与检索。刘婷等[18]自建了水利事故数据集，采用 BERT-BiLSTM 模型对事故原因进行分类，并对比了 7 种深度学习模型在该数据集中的性能，验证了所提模型的有高效性。虽然 NLP 技术在水利工程领域的应用有一些成功案例，但是针对湖泊等水资源的文本识别研究却极少。根据第一次全国水利普查公报知，我国湖泊众多，常年水面面积在 1 平方公里及以上的湖泊有 2865 个，全部以汉语命名。汉语词汇边界模糊，实体结构复杂，表现形式多样，缺乏明确的词边界等表征命名实体的线索。例如，我国以“西湖”二字结尾的湖泊共有 36 个，较为著名的有“杭州西湖”、“扬州西湖”、“惠州西湖”、“颍州西湖”、“揭阳西湖”，而“西湖区”又分别是杭州和南昌的一个行政区。在湖泊命名中“一词多义”和“一词多指”的特征使文本在不同场景下包含了丰富的语义信息，这给中文命名实体识别带来巨大挑战[19-21]。

本研究的主要贡献是提出了 ALBERT-BiGRU-Attention-CRF 模型。首先采用预训练模型 ALBERT[22]对文本进行预处理，ALBERT 在自行标注的语料库上学习湖泊文本的语义语法特征，完成词嵌入并

获得动态词向量，解决了一词多义的理解问题。相比 BERT，ALBERT 模型参数量更少，运行速度更快。然后在 BiGRU 层和 CRF 层之间加入注意力机制用于序列标注与解码，能充分考虑与当前实体相关的上下文信息，对文本语义信息的利用率和实体识别率比传统 BiGRU 模型更高，有助于提升模型整体的性能。最后，针对水利领域的湖泊文本信息，制定了一套湖泊实体标注规范。

2 相关工作

2.1 数据来源

目前没有公开的湖泊实体数据集可用，故本研究的数据一部分来自中国湖泊数据库[23]统计的全国湖泊编目数据，共 2928 份。另外一部分来自三大百科网站（百度、搜狗、360）中与湖泊有关的数据，共 24435 条。组成的语料库包括湖泊名称、湖泊代码、位置、面积、年代、所属水资源区等信息，约 42 万余字符。

2.2 实体特征分析

(1) 湖泊文本组成的元素及分布特征。

对 2928 份湖泊编目数据分析发现，存在大量的定量描述，如例句：“洞庭湖位于湖南省，属于长江流域，代码是 F43A001，面积约 2691 平方公里，经度在 111.19°~113.34°之间，纬度在 27.39°~29.51°之间”。分析描述的内容，这些指标是由几个不同类别的实体共同组成，主要包括指标名、指标值、单位和连接词 4 类。例如“代码、面积、经度、纬度”属于指标名，“F43A001，2691，119.19，27.39”是指标值，“平方公里、立方米、‘°’”是单位，“位于、属于、是、约、之间”属于连接词。由于这些指标数目众多且无固定模式，如果直接对整体做识别，那么识别精度和召回率都不理想，所以可将指标进行分类。

通过统计分析语料库中实体出现的频率，发现分为指标名、指标值、单位和连接词这 4 类较合理，并且这几类指标包含的实体字符约占语料文本总量的 17.2%，各类实体数量和分布情况如表 1。

表 1 实体数量和分布

类型	实体数量	字符数量	字符占比(%)
指标名	3294	12066	2.9
指标值	11712	25946	6.2
单位	3620	10478	2.5
连接词	8942	23517	5.6

根据统计结果，将湖泊文本的指标分为 4 类后，每一类都具备了明显的特征，而且特征之间存在具体的关系，这不仅能加快每类指标所包含实体的识别速度，而且能提升实体的识别准确度。

(2) 指标的实体结构特征

通过分析上述 4 种指标的实体构成，发现它们的结构特征存在差异。于是进一步分析预料库中的湖泊文本，将其包含的实体结构特征大致分为 5 类，具体如表 2 所示。

表 2 实体结构特征实例

实体结构特征	实例
指标名+连接词+指标名	洞庭湖+位于+湖南省
指标名+连接词+指标值+单位	面积+约+2691+平方公里
指标名+指标值+单位	总容积+220+亿立方米
指标名+连接词+指标值+指标值+连接词	经度+在+111.19°~113.34°之间
指标名+指标名+连接词+指标值+单位	淡水湖+含盐度+小于+1+g/L

2.3 数据清洗与标注

常用的语料标注方法有 BIO（Begin Inside Other，开始、中间、其他）、BIOES（Begin Inside Other End Single，开始、中间、其他、结束、单字）、BILOU（Begin Inside Last Other Unit，开始、中间、最后、其他、单位）等[24]。由于湖泊文本中的指标分为 4 类，原始 BIO 方法无法完整标注实体的所有字符，故实验采用一种扩展 BIO 的方法，根据文本中的元素和实体的结构特征，对标签类型进行重定义，如表 3 所示。指标名、指标值、单位和连接词分别用 NAM、VAL、UNS 和 CON 表示。

表 3 标签类型及定义

标签类型	标签含义
B-NAM	指标名首字符
I-NAM	指标名非首字符
B-VAL	指标值首字符
I-VAL	指标值非首字符
B-UNS	单位首字符
I-UNS	单位非首字符
B-CON	连接词首字符
I-CON	连接词非首字符
O	其他字符（包括标点等）

由于原始文本包含 42 万余字符，其中大量无用的噪声数据，因此在模型训练之前必须经过清洗和标注。数据清洗与标注过程如下：（1）通过正则表达式去除无用的网址、特殊的标点符号和一些符号化的字等信息，同时保留逗号、句号等重要标点；（2）将上一步得到的

标准数据用 Jieba 分词对语料进行切分，然后采用扩展 BIO 的标注方法对实体数据进行自动化标注。（3）完成数据集的标注工作，得到湖泊文本数据集(Lake Text Dataset, LTD)，共计 32037 条。数据清洗与标注流程如图 1 所示。将 LTD 按 8:2 的比例分为训练集和测试集，训练集用于模型训练，测试集用于验证模型性能。

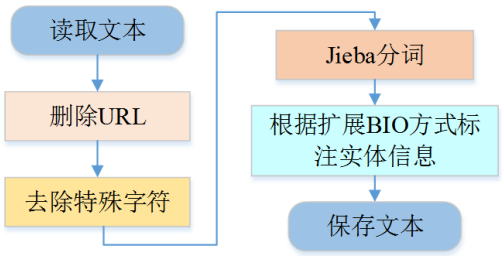

图 1 数据清洗与标注过程

3 ALBERT 与 BiGRU-Attention-CRF 模型融合

通过研究常见的深度学习模型，发现 BERT 预训练模型具有预测推断位置向量和结构信息的能力，而 ALBERT 较 BERT 的参数量更少、速度更快[22]。BiGRU 在保证短序列语义学习效果的基础上增加了学习长序列语义的能力；注意力机制增强特征词的权重，可以解决特征词语义稀释问题；CRF 具备对标签进行约束的能力。因此设计了融合注意力机制与 ALBERT-BiGRU-CRF 的模型对湖泊文本做实体识别。

图 2 展示了模型的总体框架。将 ALBERT 层作为第一层，先对语料进行初始化，然后将 ALBERT 模型输出位置和结构信息向量作为 BiGRU 层中各个时间点的输入，送入 BiGRU 模型；为了强化位置信息作用，将位置向量和字向量同时输入 BiGRU 模型，经过双向处理后，输出结果同时具备长时记忆和短时记忆。另外，考虑到长时记忆和短时记忆权重的不同可能引起长序列语义稀释问题，引入注意力层来处理 BiGRU 模型的输出结果，提升重点词语在句子中的权重，使模型将注意力集中在目标实体上，降低其他无关词作用。CRF 层可计算注意力层的输出得到最优结果，并转化成序列标签得到最终预测结果。

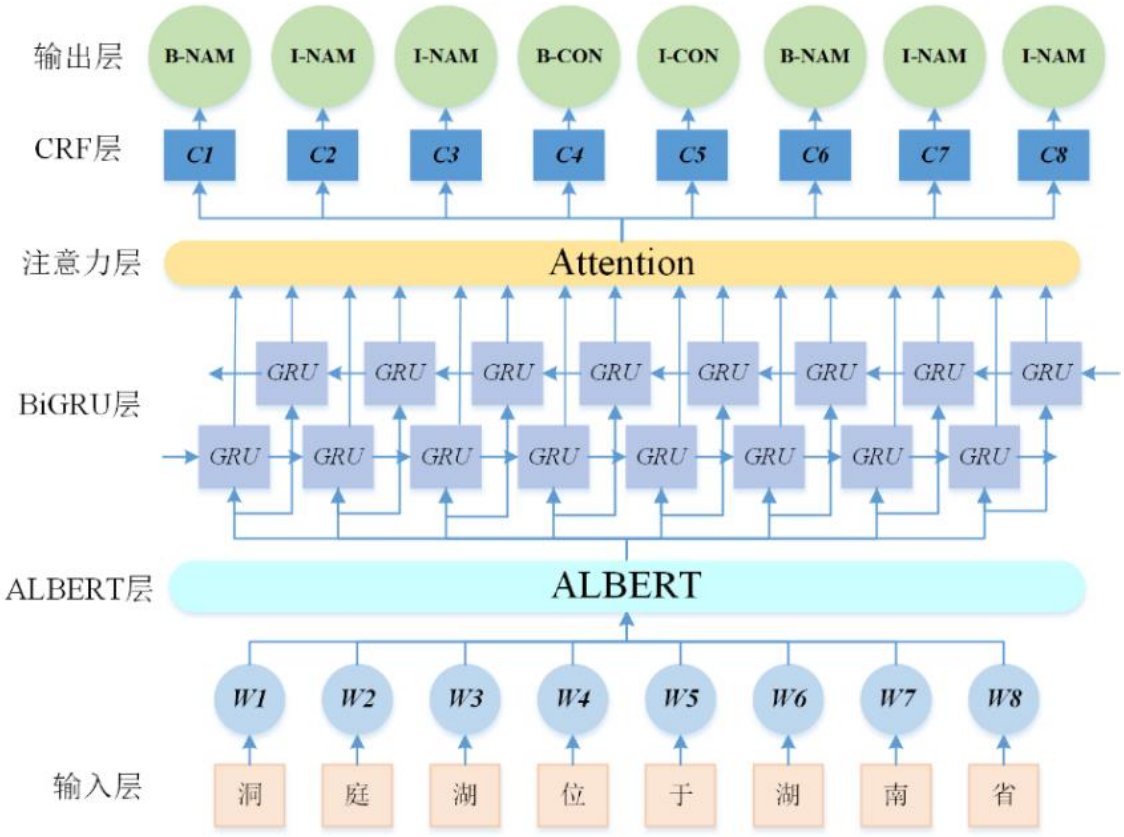

图 2 模型总体框架

3.1 ALBert 预训练模型

ALBERT[22]是一个轻量级的 BERT 预训练模型，使用 Transformer 编码器和 GELU 非线性激活函数。定义 V 为词汇表大小，E 为词典大小，L 为编码器层数，H 为隐藏层大小。为加快 BERT 的训练速度和取得更好的训练效果，谷歌工作人员从三个方面改进了 BERT。

(1)嵌入向量参数化的因式分解。

研究者将词嵌入参数分解为 $V\times E$ 和 $E\times H$ 两个矩阵，不再将 one-hot 向量直接映射到大小为 H 的隐藏空间，而是先将它们映射到一个低维词嵌入空间 E，然后再映射到隐藏空间。通过这种分解，可以将词嵌入参数从 $O(V\times H)$降低到 $O(V\times E+E\times H)$，在 H 远远大于 E 的时候，参数量会明显减少。

(2)跨层参数共享。

研究者提出了一种跨层参数共享机制来提升参数效率，ALBERT 采用共享所有层的所有参数，极大地减少了参数量，使模型参数更加稳定。

(3)句间连贯性损失。

研究者使用了一个句子顺序预测（Sentence-order prediction，SOP）函数代替 BERT 的下一句预测（Next-sentence prediction，NSP）任务，它会避免预测主题，而只关注建模句子之间的连贯性。

ALBERT 模型结构如图 3 所示，$[E_1, E_2, \ldots, E_{n-1}, E_n]$是输入文本向量，trm 是 Transformer 模型，$[T_1, T_2, \ldots, T_{n-1}, T_n]$是输出文本向量，包含全部序列的文本信息。在语料规模较大的场景下，模型也能表达丰富的语义信息，同时极大地减少参数量，达到加快速度的目的。

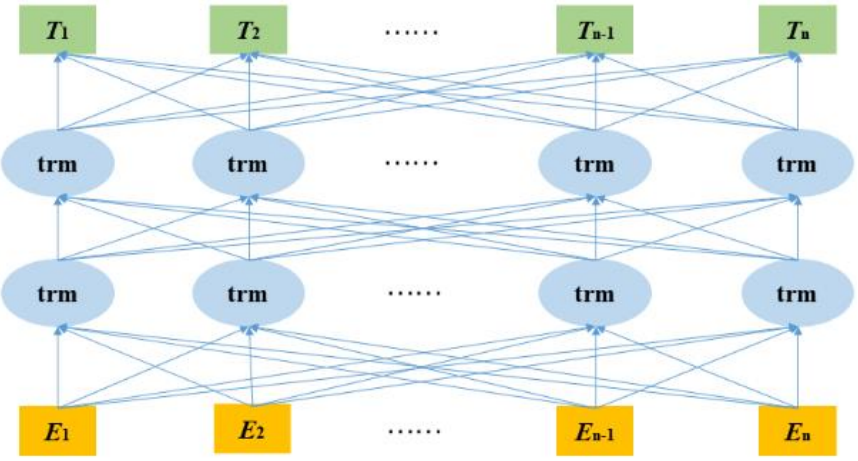

图 3 ALBERT 模型结构

3.2 BiGRU 层

门控循环单元[25](Gated Recurrent Unit, GRU)是长短时记忆神经网络(Long Short-Term Memory, LSTM)的变体，是循环神经网络 (RNN, Recurrent Neural Network) 的改进模型。由于 RNN 在处理序列时具有严重的梯度消失问题，即越靠后的节点对于前面的节点感知能力越低。为解决梯度消失问题，Hinton 等人[28]提出了 LSTM 模型。GRU 作为 LSTM 的变体，对序列数据处理同样高效，也是通过“门机制”来记忆前面节点的信息，可以顺利解决梯度消失的问题。

GRU 模型只有两个门，分别是更新门和重置门，如图 4 中的z_t和r_t。更新门用来控制上一时刻的状态信息被传入到当前状态中的程度，更新门的值越大说明上一时刻的状态信息传入越多。重置门用来控制遗忘前一时刻的状态信息的程度，重置门的值越小说明遗忘得越多。

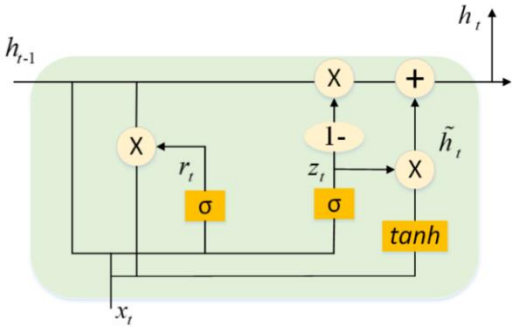

图 4 GRU 模型结构

x_t是输入数据，h_t是 GRU 单元的输出。从h_{t-1} 隐藏状态到h_t隐藏状态的计算由z_t和r_t共同控制，更新门同时控制当前输入数据和先前记忆信息h_{t-1}，输出一个在 0~1 之间的数值z_t，z_t决定以多大程度将h_{t-1}向下一个状态传递。门单元的计算如式(1)-(4)所示：

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \tag{1}$$

$$h_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \tag{2}$$

$$\widetilde{h_t} = \tanh(W \cdot [r_t \times h_{t-1}, x_t]) \tag{3}$$

$$h_t = (1 - z_t) \times h_{t-1} + z_t \times \widetilde{h_t} \tag{4}$$

式中 σ 是 Sigmoid 函数，W_z，W_r，W 分别为更新门，重置门以及候选隐含状态的权重矩阵．重置门控制h_{t-1}对结果h_t的重要程度，当先前记忆h_{t-1}和新的记忆完全相关性较大时，可以通过重置门发挥作用，提升先前记忆的影响。根据重置门、更新门和隐含状态的计算结果通过公式计算得当前时刻的输出h_t。

GRU 只在某一个方向上提取信息，而 BiGRU 是双向输入的 GRU，可以同时在两个方向上提取信息，即上下文信息。若文本序列正向输入 GRU，则获取的序列为“过去信息”，若文本序列反向输入 GRU，则获取的序列为“未来信息”，再通过连接合并两者就实现了提取上下文信息的目标，与传统的单向 GRU 相比，有更强大的上下文记忆能力。

3.3 注意力机制层

由于 GRU 可以在一定程度上解决长期记忆的问题，提取全局特征。但难以解决湖泊文本中长距离依赖问题，在长文本中难以保留局部细节信息。为弥补 BiGRU 提取局部特征所存在的缺陷，本文引入 Attention 机制[26]提取句子中不同的字符与上下文的关联程度，有利于解决湖泊命名实体字符长度大导致的长距离依赖问题。Attention 机制对与湖泊命名实体相关的语义增加特征权重，提升局部特征提取效果。注意力机制对挖掘自然语言问句词语和属性词语之间的语义关系有帮助，使注意力始终集中在最重要的信息上。在实验中通过计算问句词向量输入得到注意力矩阵，获得问句和属性的注意力表示之后，与原词向量矩阵进行拼接操作，再将拼接后的向量矩阵作为卷积神经网络模型的输入。注意力机制层对 BiGRU 层输出的特征向量进行权重分配，计算得到时刻 BiGRU 层和注意力层的共同输出特征向量。

$$C_t = \sum_{i=1}^{n} a_{t,i} h_i \tag{5}$$

$$a_{t,i} = \frac{exp(S(s_{t-1}, h_i))}{\sum_{i'}^{n} exp(S(s_{t-1}, h_{i'}))} \tag{6}$$

$$S(s_t, h_i) = vtanh(w[s_t, h_i]) \tag{7}$$

其中$a_{t,j}$为注意力函数。函数 S 为对齐模型，它是基于 i 时刻的输入和输出的匹配程度来分配分数，定义每个输出给每个输入隐藏状态多大的权重。

3.4 CRF 层

CRF 是序列标注任务中的一种常见算法[27]，通常用作序列的标注和分割。它综合考虑了输入的状态特征和序列中各标签的转移矩阵的关系，逐步在命名实体识别中得到广泛应用。定义一组输入序列$X = (x_i, \dots, x_n)$，设它的对应输出标签$Y = (y_i, \dots, y_i)$，如果两者满足式（8）：

$$P(y \mid X, y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n) = P(y \mid X, y_{i-1}, y_{i+1}) \tag{8}$$

设$P(N, K)$是经解码后输出的权重矩阵，则可计算出评估分数$S(x, y)$和输入序列与输出标签序列的对应概率$P(y \mid x)$，如式(9)和(10)所示：

$$S(x, y) = \sum_{i=1}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i} \tag{9}$$

$$P(y \mid x) = \frac{e^{S(x,y)}}{\sum_{y \in W_s} e^{S(x,y)}} \tag{10}$$

式中 $\boldsymbol{A}$ 是转移矩阵，k 是标签个数，n 是序列长度，W_s是所有可能的标签序列。

BiGRU 层输出的分数矩阵传入 CRF 层后，在 CRF 层结合学习得到的转移矩阵，计算出最终评估分数和输入序列与对应标签序列概率，最后通过维特比解码获得最优序列标注。

4 实验与分析

4.1 实验环境与参数

本研究的实验环境和配置信息如表 4 所示。训练过程中使用 Adam 优化器来计算和更新模型训练中的网络参数。最大序列长度设为 128，batch 大小设为 32，为防止过拟合在模型中引入 dropout，并设为 0.2。GRU 隐层维数为 128，优化函数学习率为5×10^{-5}，具体的参数设置如下表 5。

表 4 实验环境和配置信息

环境参数	版本
操作系统	Ubuntu 18.04 LTS x86_64
CPU	Intel Core i7 1068G7
内存	64G
GPU	Geforce RTX 3060Ti
开发语言	Python 3.7.6
Tensorflow	1.12.0

表 5 训练参数设置

实体识别训练参数	参数值
max_seq_length	128
batch_size	32
epochs	10
GRU_units	128
learning_rate	5×10^{-5}
dropout	0.1

4.2 评估指标

采用的评估指标主要有准确率(*Precision, P*)、召回率(*Recall, R*)和 *F1* 分数，它们的定义分别如下：

$$P=\frac{TP}{TP+FP}\times100\% \tag{11}$$

$$R=\frac{TP}{TP+FN}\times100\% \tag{12}$$

$$F1=\frac{2PR}{P+R}\times100\% \tag{13}$$

其中，*TP* 表示正确地把正样本预测为正，*FN* 表示错误地把正样本预测为负，*FP* 表示错误地把负样本预测为正。

4.3 结果与分析

为验证本研究所设计的 ALBERT-BiGRU-Attention-CRF 模型的实体识别效果，选取相同规模的自建数据集，分别进行 3 组实验：（1）比较 BiGRU-CRF、BiLSTM-CRF、BiGRU-Attention-CRF、BERT-BiGRU-CRF 模型和本研究所提模型在 LTD 的全部实体识别结果；（2）比较 BERT-BiGRU-CRF 模型和本研究所提模型识别全部实体消耗的平均预测时间和模型的参数量情况；（3）比较 BiGRU-CRF、BiLSTM-CRF、BiGRU-Attention-CRF、BERT-BiGRU-CRF 模型和本研究所提模型分别对指标名、指标值、单位和连接词 4 类实体的识别结果。3 组实验的结果分别如图 5、图 6 和表 6 所示。

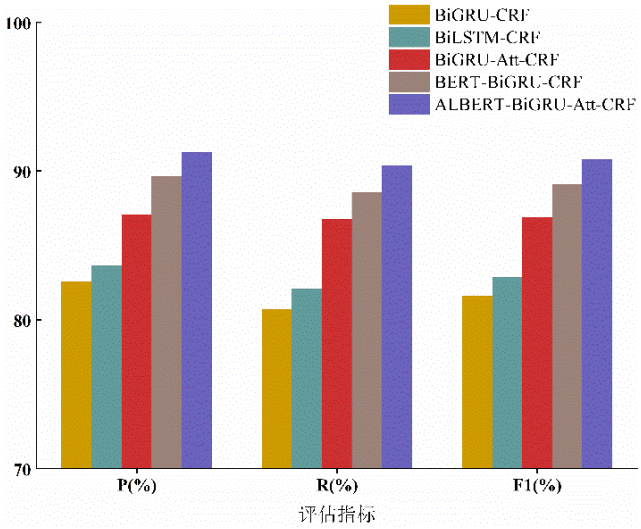

图 5 5 种不同的模型在 LTD 上的全部实体识别结果

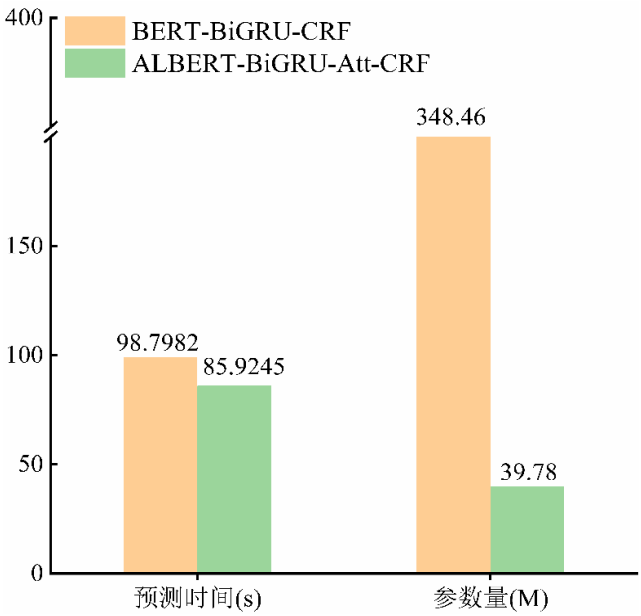

图 6 两种模型在 LTD 中的预测时间和参数量对比情况

图 5 中的实验结果表明，与 BiGRU-CRF 模型对比，BERT-BiGRU-CRF 模型的准确率、召回率和 *F1* 值分别提高 7.07%、7.86% 和 7.48%。ALBERT-BiGRU-Attention-CRF 模型的准确率、召回率和 *F1* 值比 BiGRU-CRF 模型分别提高 8.7%、9.67%和 9.19%，说明经过词向量模型 BERT 和 ALBERT 训练之后样本的整体识别效果明显提升，实体识别准确率更高。同样与 BiGRU-CRF 模型对比，BiGRU-Attention-CRF 模型的准确率、召回率和 *F1* 值分别提高 4.49%、6.05%和 5.28%；与 BERT-BiGRU-CRF

模型对比，ALBERT-BiGRU-Attention-CRF 模型的准确率、召回率和 *F1* 值分别提高 1.63%、1.80%和 1.71%，说明通过注意力机制增加各类指标权重的方式对识别效果的提升有正向影响。

表 6 不同模型对不同类别实体的识别效果

实体类别	模型	P(%)	R(%)	F1(%)
指标名	BiGRU-CRF	78.44	77.56	77.79
	BiLSTM-CRF	80.52	78.38	79.59
	BiGRU-Attention-CRF	83.25	81.19	82.21
	BERT-BiGRU-CRF	87.31	85.21	86.25
	ALBERT-BiGRU-Attention-CRF	88.42	87.36	87.89
指标值	BiGRU-CRF	82.37	81.95	82.16
	BiLSTM-CRF	84.64	83.47	84.05
	BiGRU-Attention-CRF	87.05	85.32	86.18
	BERT-BiGRU-CRF	90.61	88.59	89.59
	ALBERT-BiGRU-Attention-CRF	92.02	90.54	91.27
单位	BiGRU-CRF	85.18	84.46	84.82
	BiLSTM-CRF	86.32	85.17	85.74
	BiGRU-Attention-CRF	89.45	88.76	89.10
	BERT-BiGRU-CRF	91.36	90.53	90.94
	ALBERT-BiGRU-Attention-CRF	92.27	91.83	92.05
连接词	BiGRU-CRF	84.24	83.67	83.95
	BiLSTM-CRF	86.31	85.82	86.06
	BiGRU-Attention-CRF	88.45	87.76	88.10
	BERT-BiGRU-CRF	89.27	88.64	88.95
	ALBERT-BiGRU-Attention-CRF	90.32	89.81	90.06

第 2 组实验对比了 BERT-BiGRU-CRF 和 ALBERT-BiGRU-Attention-CRF 模型的性能。根据图 6 中两种模型的平均预测时间和参数量大小，可得出 ALBERT-BiGRU-Attention-CRF 模型对 LTD 数据集的预测时间更短，并且产生的参数量比 BERT-BiGRU-CRF 模型小很多。由于 ALBERT 模型采用跨层参数共享机制，几大地减少了参数量，促进模型整体性能的提升。另外，引入注意力机制使模型中资源的分配效率更高，从而表现为更优的识别准确率和更短的预测时间。

第 3 组实验对比了 5 种模型分别对指标名、指标值、单位和连接词 4 类实体的识别结果。在这 4 类实体中，单位和指标值的识别效果最好，连接词次之，指标名较差。由于指标值和单位绝大部分都是数字、字母和特殊字符，不同于上下文的汉字，它们有更明显的实体特征，所以容易被识别出。虽然文本中的连接词实体类型较多，但它们的上下文多与指标值或指标值相连，这使它们的特征也很明显，而且连接词的词向量长度短，不容易与指标名实体混淆或产生歧义，因此识别的准确率、召回率和 *F1* 值都很高。

表 6 中的识别结果表明，ALBERT-BiGRU-Attention-CRF 模型在上述 4 类实体识别任务中的表现均优于其他 3 种模型，特别是对后 3 类实体的识别准确率、召回率和 *F1* 值已超过 90%，识别效果提升明显，进一步验证了将 ALBERT 预训练模型和注意力机制融合到 BiLSTM-CRF 基础识别模型中是可行且高效的。

湖泊文本中的指标名全以汉字命名，与上下文的汉字相似度很高。例如：“洞庭湖在湖南省与湖北省之间，是中国第二大淡水湖。”短短的几字却包含了 4 个“湖”字，如果不与上下文明确划分边界就可能产生误判，生成错误的标签序列并传递给下一层，从而降低识别的准确率。另外一些以地名命名的实体具有独特的地域特点，它们比一般的汉字实体相比更长，而且在语料库中出现的次数也很少，如“恰尔嘎木错”、“艾里西苏阿木巴”等。这些生僻的地名词可能不会被准确识别，也会降低识别的精度。深度学习模型在训练过程中需要大量的数据来学习其中的特征，才能保证对输入的文本做出准确识别。而从数据集的规模来看，本文的语料库有待扩充。

本文部分实体的识别效果如表 7 所示，不难发现，地理指标名实体“云梦”、“九江”、“崇湖”，指标值实体和

单位实体“平方公里”等基础命名实体都能被有效识别。但是仍然存在一些识别失败的情况，主要包括三种：（1）连接词的首字符识别失败；如“俗称”的首字符“俗”被识别为其他实体，而次字符“称”被有效识别为连接词实体。（2）部分包含小数点的指标值实体识别失败；将小数点后一位的指标值判断为另一指标值的首字符。（3）部分实体指标名冗长，模型仅识别出部分信息；如“戈吉力特古尔班诺尔湖”是内蒙地区以当地地名命名的湖，模型将“戈吉力”和“班诺尔湖”识别为指标名，而将“特古尔”识别为其他字符。针对这些问题，一方面可通过扩大湖泊文本语料库规模，完善指标名实体的标注信息，以此提升模型训练标注数据时的覆盖度；另一方面可尝试结合各类实体的数量和分布情况，对模型损失函数进行加权优化，以增强稀疏实体的识别效果。

表 7 部分实体识别结果示例

语料	标注信息	识别信息
洞庭湖原名云梦、九江、崇湖，位于长江中游荆江南岸	洞 B-NAM 庭 I-NAM 湖 I-NAM 原 O 名 O 云 B-NAM 梦 I-NAM、O 九 B-NAM 江 I-NAM 、O 崇 B-NAM 湖 I-NAM，O 位 B-CON 于 I-CON 长 B-NAM 江 I-NAM 中 O 游 O 荆 B-NAM 江 I-NAM 南 O 岸 O	洞 B-NAM 庭 I-NAM 湖 I-NAM 原 O 名 O 云 B-NAM 梦 I-NAM、O 九 B-NAM 江 I-NAM、O 崇 B-NAM 湖 I-NAM，O 位 B-CON 于 I-CON 长 B-NAM 江 I-NAM 中 O 游 O 荆 B-NAM 江 I-NAM 南 O 岸 O
20 世纪 90 年代末，据水利部门计算，有面积 2579.2 平方公里，俗称中国第二大淡水湖	2 B-VAL 0 I-VAL 世 O 纪 O 9 B-VAL 0 I-VAL 年 O 代 O 末 O，O 据 B-CON 水 B-NAM 利 I-NAM 部 I-NAM 门 I-NAM 计 O 算 O，O 有 B-CON 面 B-NAM 积 I-NAM 2 B-VAL 5 I-VAL 7 I-VAL 9 I-VAL.O 2 I-VAL 平 B-UNS 方 B-UNS 公 B-UNS 里 B-UNS，O 俗 B-CON 称 I-CON 中 B-NAM 国 I-NAM 第 O 二 O 大 O 淡 B-NAM 水 I-NAM 湖 I-NAM	2 B-VAL 0 I-VAL 世 O 纪 O 9 B-VAL 0 I-VAL 年 O 代 O 末 O，O 据 B-CON 水 B-NAM 利 I-NAM 部 I-NAM 门 I-NAM 计 O 算 O，O 有 B-CON 面 B-NAM 积 I-NAM 2 B-VAL 5 I-VAL 7 I-VAL 9 I-VAL.O 2 B-VAL 平 B-UNS 方 B-UNS 公 B-UNS 里 B-UNS，O 俗 O 称 B-CON 中 B-NAM 国 I-NAM 第 O 二 O 大 O 淡 B-NAM 水 I-NAM 湖 I-NAM

5 结论

本研究自定义了湖泊命名实体识别(Named Entity Recognition, NER)方法和标注规范，并建立湖泊文本语料库。对于小规模语料库，文本信息的稀疏性可能影响模型的性能和识别效果，因此设计了 ALBERT-BiGRU-Attention-CRF 模型。通过引入轻量预训练模型 ALBERT 生成高质量词向量，并作为 BiGRU-CRF 模型的输入层特征向量，再将注意力机制融入到该模型为实体的语义信息增加特征权重，提升实体特征提取效果。在自建的湖泊文本语料库中进行大量试验，结果表明 ALBERT-BiGRU-Attention-CRF 模型对数据集整体的识别效果良好，准确率、召回率和 F1 分别达到 91.26%、90.38%和 90.81%。另外对比了几种主流深度学习模型，发现该模型对高频出现的 4 类实体识别的性能均优于其他模型，验证了 ALBERT-BiGRU-Attention-CRF 模型在小规模语料场景下的命名实体识别的高效性。

在下一步的工作中，考虑融入更多水利领域的专业实体特征，在大规模数据集中进一步验证该模型的性能，同时与主流深度学习模型作对比，不断提升模型的性能并应用到更多的领域。

参考文献

[1] 第一次全国水利普查公报 [J]. 中华人民共和国水利部公报, 2013 (02): 53-57.

[2] Rau L F. Extracting Company Names from Text [C] // The Seventh IEEE Conference on Artificial Intelligence Application. IEEE, 1991: 29-32.

[3] 李军怀，陈苗苗，王怀军，崔颖安，张爱华. 基于 ALBERT-BGRU-CRF 的中文命名实体识别方法 [J]. 计算机工程, 2022, 48 (06): 89-94+106.

[4] 巩敦卫，张永凯，郭一楠，王斌，樊宽鲁，火焱. 融合多特征嵌入与注意力机制的中文电子病历命名实体识别 [J]. 工程科学学报, 2021, 43 (09): 1190-1196.

[5] Chen Y, Xu L, Liu K, et al. Event extraction via dynamic multi-pooling convolutional neural networks [C] // Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2015: 167-176.

[6] Voorhees E, Harman DK (2006) Trec: experiment and evaluation in information retrieval. J Am Soc Inf Sci Technol 32 (4): 563–567.

[7] Fader A, Zettlemoyer L, Etzioni O. Paraphrase-driven learning for open question answering [C] // Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2013: 1608-1618.

[8] Yao X, Van Durme B. Information extraction over structured data: Question answering with freebase [C] // Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014: 956-966.

[9] 许宁. 面向旅游领域的智能问答系统设计与实现 [D]. 内蒙古大学, 2021.

[10] Bunescu R, Mooney R. A shortest path dependency kernel for relation extraction [C] // Proceedings of human language technology conference and conference on empirical methods in natural language processing. 2005: 724-731.

[11] Miwa M, Bansal M. End-to-end relation extraction using lstms on sequences and tree structures. In: ACL. 2016: 1105–1116.

[12] Singh S, Riedel S, Martin B, Zheng J, McCallum A. Joint inference of entities, relations, and coreference. In: AKBC. 2013: 1–6.

[13] Han S, Hao X, Huang H. An event-extraction approach for business analysis from online Chinese news [J]. Electronic Commerce Research and Applications, 2018, 28: 244-260.

[14] Upadhyay S, Gupta N, Roth D. Joint multilingual supervision for cross-lingual entity linking [J]. arXiv preprint arXiv: 1809. 07657, 2018.

[15] Li J, Bu C, Li P, et al. A coarse-to-fine collective entity linking method for heterogeneous information networks [J]. Knowledge-Based Systems, 2021, 228: 107286.

[16] 冯钧, 杭婷婷, 陈菊, 王云峰, 王秉发, 张涛. 领域知识图谱研究进展及其在水利领域的应用 [J]. 河海大学学报(自然科学版), 2021, 49 (01): 26-34.

[17] 段浩, 韩昆, 赵红莉, 蒋云钟, 李豪, 毛文山. 水利综合知识图谱构建研究 [J]. 水利学报, 2021, 52 (08): 948-958.

[18] 刘婷, 张社荣, 王超, 李志竑, 关炜, 王枭华. 水利施工事故文本智能分析的 BERT-BiLSTM 混合模型 [J]. 水力发电学报, 2022, 1 (07): 1-12.

[19] Cao P, Chen Y, Liu K, et al. Adversarial transfer learning for Chinese named entity recognition with self-attention mechanism [C] // Proceedings of the 2018 conference on empirical methods in natural language processing. 2018: 182-192.

[20] Xu C, Wang F, Han J, Li C. Exploiting multiple embeddings for Chinese named entity recognition. In: Proceedings of the 28th ACM international conference on information and knowledge management. 2019: 2269-2272.

[21] Greenberg N, Bansal T, V erga P, McCallum A. Marginal likelihood training of bilstm-crf for biomedical named entity recognition from disjoint label sets. In: Proceedings of the 2018 conference on empirical methods in natural language processing. 2018: 2824–2829.

[22] Lan Z, Chen M, Goodman S, et al. Albert: A lite bert for self-supervised learning of language representations [J]. arXiv preprint arXiv:1909.11942, 2019.

[23] http://www.lakesci.csdb.cn

[24] Ju M, Miwa M, Ananiadou S. A neural layered model for nested named entity recognition. In: Proceedings of the 2018 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies, vol 1 (Long Papers). 2018: 1446–1459.

[25] Chung J, Gulcehre C, Cho K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling [J]. arXiv preprint arXiv: 1412. 3555, 2014.

[26] Zhu Q L, Li X L, Conesa A, et al. GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text. Bioinformatics. 2018, 34 (9): 1547.

[27] WOJEK C SCHIELE B. A dynamic conditional random field model for joint labeling of object and scene classes [C] // Proceedings of the 10th European Conference on Computer Vision, LNCS 5305. Berlin: Springer. 2008: 733-747.