引文格式：郑旭野,陈涛,周婧娟. 利用扩展词嵌入BERT的地表水系地理命名实体抽取模型[J]. 地理空间信息, 2025, 23(2): 1-6. doi:10. 3969/j. issn. 1672-4623.2025.02.001

利用扩展词嵌入BERT的地表水系地理命名实体抽取模型

郑旭野[1]，陈　涛[2]，周婧娟[2]
（1. 武汉大学 测绘学院，湖北 武汉 430079；2. 湖北省测绘成果档案馆，湖北 武汉 430014）

摘　要：地理知识图谱构建的重要任务之一是地理命名实体的识别。中文文本中词汇结构灵活，词汇边界不明显，地理领域的中文标注数据集稀缺，因此中文文本的地理命名实体识别一直是研究难点之一。针对蕴含地理信息的海量网络文本中的地理命名实体识别任务，建立了基于维基百科数据的地表水系数据集以及领域词典，并提出了一种基于扩展词嵌入的词汇增强方法，对于BERT预训练模型进行词汇增强，并结合了BiGRU与CRF网络进行上下文特征识别与学习，构建了EXP-BERT-BiGRU-CRF的命名实体识别模型，实验表明，该模型在地表水系数据集上达到了95.94%的*F*1值，比无词汇增强Bert模型提高了4.94%，相较于其他模型的精度也有大幅度提升，能更加准确地识别地理命名实体。

关键词：地理知识图谱；BERT；命名实体识别；词汇增强
中图分类号：P208　　**文献标志码：**B　　**文章编号：**1672-4623（2025）02-0001-06

Geographical Named Entity Extraction Model of Surface Water System Based on Expanded Word Embedding BERT

ZHENG Xuye[1], CHEN Tao[2], ZHOU Jingjuan[2]
(1. School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China;
2. Province Surveying Mapping Production Archives of Hubei, Wuhan 430014, China)

Abstract: One of the important tasks in constructing a geographical knowledge graph is the recognition of geographical named entities. Chinese text has flexible vocabulary structures and unclear word boundaries, making the recognition of geographical named entities in Chinese text a challenging research area, especially due to the scarcity of annotated datasets in the geographical domain. To address the task of geographical named entities recognition in massive network texts containing geographical information, we established a dataset of surface water system based on Wikipedia data and a domain dictionary, and proposed a vocabulary enhancement method based on expanded word embedding to enhance the vocabulary of BERT pre-training model. We constructed EXPBERT-BiGRU-CRF named entity recognition model by combining BiGRU and CRF networks for context feature recognition and learning. Experimental results show that this model achieves F1_score of 95.94% on the surface water system dataset, which is a 4.94% improvement compared to the BERT model without vocabulary enhancement, along with significant accuracy improvements compared to other models, and can accurately identify geographical named entities.

Key words: geographical knowledge graph, BERT, named entity recognition, vocabulary enhancement

知识图谱作为一种高效的知识表示和表达方式，是自然语言识别领域中备受关注的研究方向。它是一种图结构的知识表示网络[1]，以实体和实体之间的关系为基础，将丰富多样的知识组织成一个个节点和边的形式。每个节点代表一个实体，而边则表示不同实体之间的关系。通过这种方式，知识能够以更加形式化和结构化的方式进行表达，并且可以被计算机高效地处理和推理。

地理知识图谱是知识图谱在地理领域的扩展，其特点在于地理知识图谱的关系不仅包含语义关系，还包含实体之间的时空关系[2]。地理知识图谱可以用于实现基于地理数据库的智能知识问答[3]、辅助交通预测和决策等[4]。

尽管知识图谱在地理应用方面已经取得了一些重要成果，但仍面临着一些挑战和问题。如何从海量的网络文本中抽取地理知识，挖掘与分析地理信息，是地理科学在新地理信息时代面临的新挑战[5]。因此，进一步研究和探索网络文本中地理知识抽取以及知识图谱在地理领域的应用是具有重要意义的。

针对以上问题，本文将聚焦于地理命名实体识别，并探索深度学习在该领域的应用。我们提出了一种基于扩展BERT+BiGRU+CRF模型的地理命名实

项目来源：湖北省自然科学基金资助项目（2022CFB194）。
第一作者简介：郑旭野（1999—），硕士研究生，研究方向为空间数据建模与分析，E-mail：zireael@whu.edu.cn。

体识别方法，以提高地理信息处理的准确性和效率。通过实证研究和实验评估，验证了该方法在地理命名实体识别任务中的有效性，为地理知识图谱的构建提供了技术支持。

1 相关技术

早期命名实体抽取的方法主要是基于规则和特定模板，需要相关领域的专家撰写大量的识别规则，费时费力且识别准确率不算高，迁移性和泛化能力都很低。随着机器学习的发展，机器学习的方法也应用于命名实体的抽取中[6]，HMM（隐马尔科夫模型）和CRF（条件随机场）是常用于文本序列标注预测的机器学习方法[7-9]，但由于机器学习的方法需要构造特征，精度也不是很理想，命名实体识别的技术进一步延伸到深度学习领域。

基于深度学习进行命名实体抽取的方法以循环神经网络（RNN）为主。RNN能有效学习到句子中的上下文特征。双向长短期记忆网络BiLSTM被广泛应用于命名实体抽取（NER）任务中。在此基础上，有学者提出了将条件随机场（CRF）与BiLSTM结合的BiLSTM-CRF模型[10]，其中，条件随机场能有效减少标签的错误判别概率。曾青霞等[14]将Transformer的自注意力机制与BiLSTM-CRF相结合提出了Self-Att-BiLSTM-CRF模型。Jin等[12]采用将BILSTM-CRF神经网络与门控自注意力相结合的方法，用于中文命名实体识别。在Google提出BERT（bidirectional encoder representations from transformers）模型[13]之后，将BERT作为文本嵌入的工具并与BiGRU-CRF模型结合的BERT-BiGRU-CRF模型[14-16]，大幅度提高了命名实体识别的精度。

中文命名实体抽取任务的文本嵌入方法通常分为词嵌入和字符嵌入的方法，由于基于单个中文字符的字符嵌入方法忽略了中文词汇的边界信息，研究人员开始探索将词汇边界信息加入到模型中的方法。Li等[17]在Lattice LSTM[18]的基础上，提出平面的晶格结构——FLAT方法使用Transformer模型和位置编码来处理字符-词的晶格结构，将其转换为由跨度组成的扁平化结构，相较于Lattice LSTM拥有更高的效率。Peng等[19]在此基础上进一步简化了模型，提出了Soft Lexicon的词汇嵌入方法，将词汇信息简单有效的组合到字符嵌入中，相较于晶格结构的模型具有更快的速度，其与BERT结合的模型在Resume NER数据集中得到了96.11%的$F1$值。Chen等[20]提出的CRW-NER模型，不仅将词汇的信息汇入了嵌入层，还将汉字的字形结构作为特征对字嵌入进行增强。上述的研究都从不同角度探索了词汇增强模型在命名实体识别中的应用，并获取了有效的成果。

由上述可见，中文NER任务由于中文字符的原因，在采用分词之后嵌入的方法会受到分词误差的影响，而采用字嵌入的方法又会忽略词汇的边界信息。有许多学者致力于将词汇信息融入到字符嵌入中。李发东等[21]的研究证实了在各种改进模型的比较中，词汇增强模型对地理命名实体识别的精度影响最大。因此研究词汇增强模型在地理命名实体识别中是存在重要意义的。对此，本文提出了一种基于先验信息的词汇嵌入方法，将词汇的领域信息加入到字符之中，将其应用于地表水系的知识抽取研究中，提高了命名实体抽取的准确率，具体的模型构建和实验结果将在下文中详细说明。本文提出EXPBERT-BiGRU-CRF模型由扩展嵌入层EXPBERT层、BiGRU层和CRF层组成，具体的模型结构如图1所示。

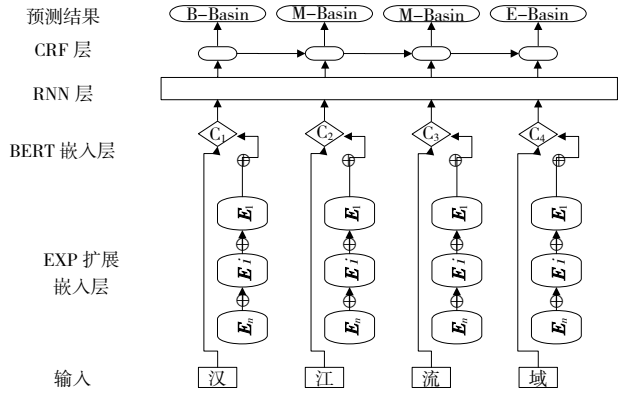


图1 EXPBERT-BiGRU-CRF模型

2 模型构建

2.1 词嵌入层

2.1.1 基于BERT的嵌入层

本文采用基于字嵌入的方法将中文的每个字转换为单独的向量，目的是捕捉更为细粒度的语义信息。BERT是由谷歌提出的基于Transformers的预训练语言模型，其预训练过程通过联合Masked LM（掩码语言模型）和NSP（下句预测）2个任务来学习上下文的文本信息。Masked LM任务的目标是学习字符的上下文语义信息，NSP任务使BERT模型能学习到句子之间的上下文关系，这种关系可以应用到下游的问答或者自然语言推理任务中。BERT结合MLM任务和NSP任务进行联合预训练，为下游的自然语言处理任务打下了良好的基础。

BERT预训练模型的输入层是由三组向量的和组成，分别是采用word piece[14]的方法将中文字符转换为768维度的字符嵌入向量Token Embedding，用于文本分类等任务的分句嵌入向量Segment Embedding，以及区分不同位置中出现同一个字的位置嵌入向量Position Embedding。

BERT模型的输出包括了模型最后隐藏层Last hidden state和池化层Pooler out（表示BERT模型的池化输出），前者包含了文本中每个字符最丰富的语义信息和上下文信息，而后者通常被当作整个文本的向量表示。在命名实体识别任务中，通常采取BERT模型最后隐藏层作为字嵌入的输出向量，如图2所示。

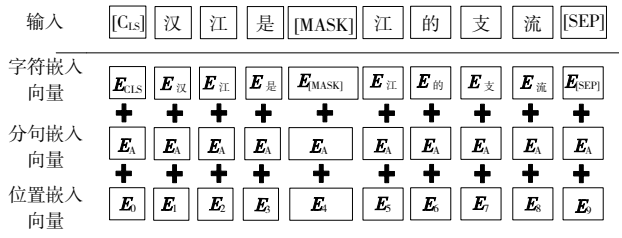


图2　BERT预训练模型输入层

2.1.2　基于先验词典信息的扩展嵌入层

基于字嵌入的命名实体识别方法虽然可以避免中文分词的误差，但是同时也失去了词汇级别的语义信息，而词汇的边界信息通常意味着实体的边界信息。本文在BERT进行字嵌入之后，采用加入先验词汇级特征的方法进行字嵌入的词汇信息增强。基于先验词汇信息的扩展字嵌入的方法，既能保留中文词汇的信息，也能避免中文分词产生的误差。其流程如图3所示。

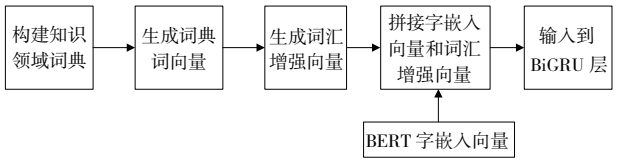


图3　扩展词嵌入层流程

1）构建领域词典。在构建领域词典的过程中，本研究通过专家渠道、百度百科和语料数据库等多种途径收集领域内的词汇信息，并对其进行分类，构建了一个具备先验分类信息的领域词汇词典。

2）构建词典词向量。为了将领域词汇转换为计算机可以处理的形式，本研究采用Word2Vec方法[15]生成词向量。Word2Vec是一种基于神经网络的词向量表示模型，通过预测上下文或目标词来学习词汇的分布式表示，将领域词汇词典转换为高维向量空间中的向量表示，其中每个词在向量空间中的位置反映了其语义信息和词义关系。

表1　词汇增强模型示例

字符	河流	湖泊	流域	水系
汉	{汉江}	{}	{汉江流域}	{}
江	{汉江}	{}	{汉江流域}	{}
流	{}	{}	{汉江流域}	{}
域	{}	{}	{汉江流域}	{}

3）生成词汇增强向量。在SoftLexicon的BMES分类方法中，作者按字符在匹配词汇中的位置，将匹配词分类为“B（匹配词首字符）”、“M（匹配词中间字符）”、“E（匹配词尾字符）”、“S（匹配到单字符词汇）”四类并计算得到四组蕴含词汇边界信息的词向量，并将其融合到字符表示中。在本文研究中则将领域词汇进行预分类，将字符在每一类领域词典中的匹配词汇的加权平均向量作为字符在该分类上的词汇增强向量。具体处理过程如下：对每个字符进行上下文匹配领域字典，生成词汇增强向量。以表1中“汉江流域”的“汉”字为例：在词典的河流类，匹配到“汉江”词汇，在词典的流域类，匹配到了“汉江流域”词汇。“汉”字具有“汉江”和“汉江流域”2个词汇且来自不同分类的先验分类信息。在同一分类中匹配到多个词汇的字符，其对应类的向量为匹配到词汇向量的加权平均值。某个字符 i 的第 t 类的词汇增强向量 $\boldsymbol{E}_t$ 计算公式如式（1）。

$$\boldsymbol{E}_t=\frac{1}{P}\sum_{w\in S_t}p(\boldsymbol{w})e^{w}(\boldsymbol{w}) \tag{1}$$

$$P=\sum_{w\in S_t}p(\boldsymbol{w}) \tag{2}$$

式中，$p(\boldsymbol{w})$ 为字符i匹配词；$\boldsymbol{w}$ 为第 t 类里出现的频次 $e^{w}(\boldsymbol{w})$，表示 $\boldsymbol{w}$ 的词向量；P 为字符i在第 t 类中所有匹配到词 $\boldsymbol{w}$ 的频次之和。

综上，字符i匹配词典所有类的词汇增强向量 $\boldsymbol{E}_{\exp i}$ 计算公式为：

$$\boldsymbol{E}_{\exp i}=\boldsymbol{E}_1\oplus\boldsymbol{E}_2\oplus\ldots\oplus\boldsymbol{E}_n \tag{3}$$

式中，⊕为向量的拼接。

4）拼接BERT字嵌入向量 $\boldsymbol{E}_{\text{berti}}$ 和词汇增强向量 $\boldsymbol{E}_{\text{expi}}$。对于句子T的第 i 个字符，其进行词汇增强后的嵌入向量为：

$$\boldsymbol{E}_i=\boldsymbol{E}_{\text{berti}}\oplus\boldsymbol{E}_{\text{expi}} \tag{4}$$

SoftLexicon方法将字词汇的边界信息融入到字符表示中，并不包含词汇的分类信息，而EXP词汇增强将字符具有的词汇分类信息融入到字符表示中。后者在应用于专业领域的命名实体抽取任务时，能将蕴含词汇先验分类的关键信息融入到字符表示中，在专业

领域中相较于SoftLexicon方法拥有更好的精度，两者具体的精度对比详见3.5节。

2.2 BiGRU层

引入词典信息的字嵌入层接下来将输入到RNN层，本文采用的门控循环单元（GRU）网络是RNN网络的一种经典变体，一个GRU单元由更新门 z_t 和重置门 r_t 组成，GRU单元的概念图如图4所示。

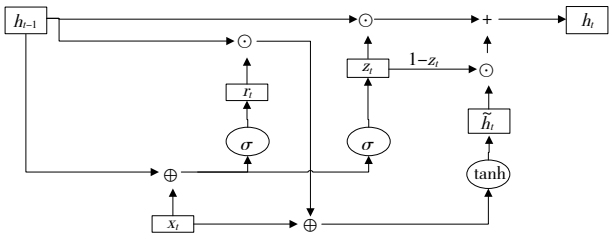


图4 GRU单元概念图

图4中⊙为向量的数乘，⊕为向量拼接，+为向量和。

式（5）中，重置门 r_t 将上一时刻隐藏状态 h_{t-1} 与当前时刻 x_t 拼接后输入Sigmoid函数，其中 W_r 表示重置门权重，由于Sigmoid的存在，其输出门控信号值为0~1之间，用于控制前一时刻的隐藏层信息有多少被保留。

$$r_t=\sigma\left(W_r\times\left[h_{t-1},x_t\right]\right) \tag{5}$$

式（6）中，更新门 z_t 由上一时刻隐藏状态 h_{t-1} 与当前时刻 x_t 拼接后输入Sigmoid函数，其中 W_z 表示更新门权重，输出门控阈值用于选择有多少信息被传递到当前单元的隐藏层中。

$$z_t=\sigma\left(W_z\times\left[h_{t-1},x_t\right]\right) \tag{6}$$

式（7）中，当前候选记忆 $\tilde{h}_t$ 由两部分组成，一部分是通过重置门信号决定的上一时刻状态信息 h_{t-1}，另一部分是当前输入信息 x_t，重置门输出信号为0时，表示完全遗弃来自过去状态的信息，其中 $W_{\tilde{h}}$ 表示权重信息。

$$\tilde{h}_t=\tanh\left(W_{\tilde{h}}\times\left[r_t\times h_{t-1},x_t\right]\right) \tag{7}$$

最终输出由更新门控制，更新门的门控阈值参数决定上一时刻的隐藏状态 h_{t-1} 遗忘程度，另一部分是当前候选记忆加入到输出的程度，两者分别由 $1-z_t$ 和 z_t 控制

$$h_t=\left(1-z_t\right)\times h_{t-1}+z_t\times\tilde{h}_t \tag{8}$$

GRU单元拥有更少的门控数量和更少的参数，能有效加快模型的训练速度和减少模型的大小。本文采用双向的GRU网络来学习上下文的语义信息，BiGRU网络的结构如图5所示，嵌入层的字符嵌入向量输入到双层的GRU网络，分别学习字符的上下文信息，由输出层输出预测标签的概率。

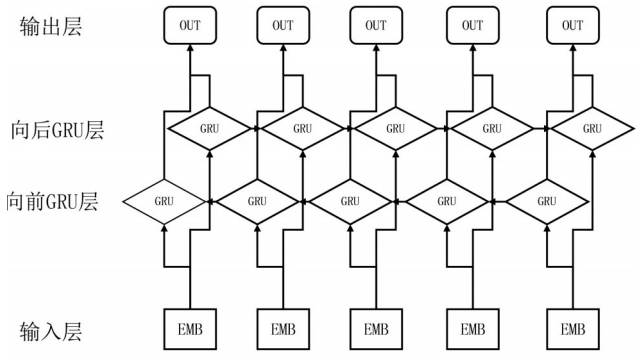


图5 BiGRU网络层示意图

2.3 CRF层

条件随机场（CRF）是根据给定的一组输入序列条件 X 下另一组输出标签 Y 的条件概率分布模型 $P(Y|X)$。在命名实体标注的任务中，CRF可以单独作为模型解决命名实体识别任务，也能作为深度学习模型中的标签解码层，用于捕捉标签间常具有的强依赖性，增加输出相邻标签的约束，有效减少错误判别的几率。BiGRU层输出的向量经CRF层解码得到最终预测标签。具体解码过程如下：

定义发射分数矩阵 $\boldsymbol{E}$ 和转移分数矩阵 $\boldsymbol{T}$，其中发射分数矩阵 $\boldsymbol{E}$（状态矩阵）来自于BiGRU层输出，转移分数矩阵 $\boldsymbol{T}$ 表示标签之间的转移概率，在随机初始化后随着模型训练而自动更新。一段标签序列的得分由发射得分和转移得分组成。以长度为4的预测标签序列{START，B-River，E-River，O，O，END}为例，该标签序列的发射得分如式（9）。

$$\begin{aligned}\text{Emissionscore}=&\boldsymbol{E}_{0,\text{START}}+\boldsymbol{E}_{1,\text{B-River}}+\boldsymbol{E}_{2,\text{B-River}}+\\&\boldsymbol{E}_{3,\text{B-River}}+\boldsymbol{E}_{4,\text{B-River}}+\boldsymbol{E}_{5,\text{END}}\end{aligned} \tag{9}$$

式中，$\boldsymbol{E}_{\text{i, type}}$ 为发射矩阵 $\boldsymbol{E}$ 中字符i的type类标签得分。

该标签序列的转移得分如式（10）。

$$\begin{aligned}\text{Trasitionscore}=&\boldsymbol{T}_{\text{START}->B-\text{River}}+\boldsymbol{T}_{B-\text{River}->E-\text{River}}+\\&\boldsymbol{T}_{E-\text{River}->O}+\boldsymbol{T}_{O->O}+\boldsymbol{T}_{O->\text{END}}\end{aligned} \tag{10}$$

式中，$\boldsymbol{T}_{\text{typea}->\text{typeb}}$ 为转移矩阵中从a类标签转移至b类标签的得分。

该标签序列的最终得分为

$$S=\text{Emissionscore}+\text{Transitionscore} \tag{11}$$

CRF的目标即是找到得分最高的预测标签序列，通过维特比算法可以得到分数最高的输出，即最终的预测标签，根据预测标签即可得到地表水系命名实体的抽取成果。

3　实验与分析

3.1　数据来源

本文采用的数据集来源自维基百科，包含中国河流列表和中国湖泊列表下的所有地表水系词条，有效数据共46万字符，约9 648条文本。实体的标签及含义示例如表2所示。

表2　实体标签说明及示例

实体标签	标签类型及含义	示例
RIVER	河流，包括自然河流和人工沟渠	长江
LAKE	湖泊，包括天然湖泊和人工水库	东湖
BASIN	流域，即分水线所包围的河流集水区	汉江流域
RIVERSYS	水系，即流域内所有河流、湖泊等各种水体组成的水网系统	清江水系

3.2　数据标注与分析

数据标注采用BMES的标注格式，人工标注平台采用开源Doccano[22]平台，标注界面如图6所示。标注完成数据中有效标签为32 780个，包含实体10 510个，其中河流实体7 494个，湖泊实体2 280个，水系实体220个，流域实体516个。

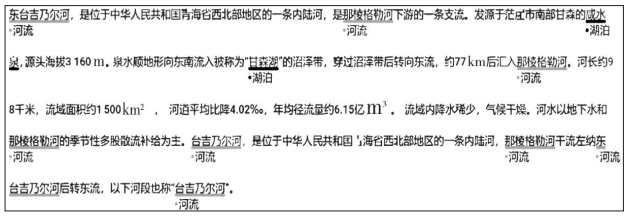


图6　Doccano运行界面图

BMES的标注方法如表3所示，B表示实体的开头字符，M表示实体的中间字符，E表示实体的结尾字符，S表示单个实体字符，O表示不带标签的字符。

表3　BMES标注方法示例

汉	B-RIVER
江	E-RIVER
位	O
于	O
长	B-BASIN
江	M-BASIN
流	M-BASIN
域	E-BASIN

本实验使用数据集按照80%：10%：10%的比例将数据集划分为训练集、验证集和测试集。领域命名实体的数据集中每一行由两列组成，数据的两列之间采用制表符隔开，其中第一列为文本内容，第二列是BMES格式的标签序列，采用空行来分隔句子。

3.3　实验环境和超参数配置

实验的计算机采用配置为Windows11操作系统，CPU为i7-10700，内存32 GB，显卡采用RTX2080，显存8 GB，Python-3.8，torch-1.81，CUDA-11.1。BERT预训练模型采用Google的BERT-BASE-Chinese模型。训练参数中优化器采用AdamW优化器，学习率为5e-5，训练Epoch次数为15，RNN隐藏层的单元数量为128，在RNN网络层之后加入Dropout为0.5的层来避免过拟合。

3.4　评价标准

本实验的评价标准（表4）包括准确率 P，召回率 R 和 F_1 值，具体计算公式如式（12）、（13）、（14）。

表4　评价标准

	实际为真	实际为假
预测为真	*TP*	*FP*
预测为假	*FN*	*TN*

$$P=\frac{TP}{TP+FP}\times 100\% \tag{12}$$

$$R=\frac{TP}{TP+FN}\times 100\% \tag{13}$$

$$F_1=\frac{P\times R\times 2}{P+R}\times 100\% \tag{14}$$

3.5　实验结果与分析

为了验证本文提出的EXPBERT-BiGRU-CRF模型的有效性，本实验采用BERT-CRF模型、BiGRU-CRF模型、BiLSTM-CRF模型和SoftLexicon-BERT-BiGRU-CRF基线模型加以比较，上述模型在自建地表水系数据集上运行，其准确率 P，召回率 R 和 F_1 值如下表5所示。

表5　模型实验结果

模型名称	P	R	F_1
BiGRU-CRF	86.21	48.24	60.78
BiLSTM-CRF	86.10	43.33	55.98
BERT-CRF	96.52	84.18	89.72
BERT-BiGRU-CRF	95.28	86.70	90.67
SoftLexicon-BERT-BiGRU-CRF	95.88	86.87	91.00
EXPBERT-BiGRU-CRF	97.52	94.58	95.94

由表5可知，BiGRU相较于BiLSTM的召回率更高，F_1 值也更高，由于GRU的单元相较于LSTM更简单，因此GRU的训练速度更快且参数更少，对于较短的文本和小样本的数据上性能更为优秀。BERT预训练语言模型对于命名实体识别的任务性能提升巨大，在准确度和召回率上都有大幅度提升，达到了89.72%的 F_1 值，具有实用价值。Bert+BiGRU的组合牺牲了一些准确度，提升了召回率，F_1 值也有小幅度提升，对于单纯BERT线性输出的模型性能有些许提

升，达到了90.67%的 F_1 值。SoftLexicon的扩展特征全方位提升了模型的能力，综合 F_1 值达到了91%。EXP扩展特征对模型的准确度和召回率都有显著的提升，尤其是在召回率上提升明显，在自建的地表水系数据集上的 F_1 值达到了95.94%。

4 结 语

本文根据维基百科的水系数据构建了地表水系的百科数据集和领域词典，针对地理命名实体识别任务提出了基于先验词典信息的词汇增强模型，并采用与BERT字嵌入结合的改进方法，在字嵌入过程中，将词汇的先验分类信息作为增强特征信息加入到嵌入向量中，并在后续网络层采用了BiGRU和CRF的网络结构来构建模型，在构建数据集上取得了良好效果。与已有的模型相比较，其准确率 P ，召回率 R 和 F_1 值都有所提高。但是在一些方面上还存在需改进之处：首先是由于人工成本较高，构建数据集数量不够大，水系和流域的标签数量较少；其次是在流域和水系实体中，存在典型的嵌套实体，但本实验没有考虑嵌套实体的识别，在以后的研究中，还可以在预训练模型的选择和模型性能优化上开展工作。

参考文献

[1] 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

[2] 张雪英,张春菊,吴明光,等. 顾及时空特征的地理知识图谱构建方法[J]. 中国科学(信息科学学),2020, 50(7):1 019-1 032

[3] 蒋秉川,万刚,许剑,等. 多源异构数据的大规模地理知识图谱构建[J]. 测绘学报,2018,47(8):1 051-1 061

[4] 李璇,吴雷,姜梦,等. 基于轨迹大数据的司机出行知识图谱构建与实现[J]. 地理空间信息,2023,21(1):32-35

[5] 陆锋,余丽,仇培元. 论地理知识图谱[J]. 地球信息科学学报,2017,19(6):723-734

[6] Liu X, Zhang S, Wei F, et al. Recognizing Named Entities in Tweets[C]//The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, Portland, Oregon, USA, 2011

[7] 双哲,孙蕾. 基于改进的隐马尔可夫模型在网页信息抽取中的研究与应用[J]. 计算机应用与软件,2017,34(2):42-47

[8] 俞鸿魁,张华平,刘群,等. 基于层叠隐马尔可夫模型的中文命名实体识别[J]. 通信学报,2006(2):87-94

[9] 何炎祥,罗楚威,胡彬尧. 基于CRF和规则相结合的地理命名实体识别方法[J]. 计算机应用与软件, 2015, 32(1): 179-185

[10] 翟社平,段宏宇,李兆兆. 基于BILSTM_CRF的知识图谱实体抽取方法[J]. 计算机应用与软件, 2019, 36(5):269-274

[11] 曾青霞,熊旺平,杜建强,等. 结合自注意力的BiLSTM-CRF的电子病历命名实体识别[J]. 计算机应用与软件,2021,38(3): 159-162

[12] Jin Y, Xie J, Guo W, et al. LSTM-CRF Neural Network with Gated Self Attention for Chinese NER[J]. IEEE Access, 2019, 7: 136 694-136 703

[13] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of Deep Bidirectional Transformers for language Understanding[C]//Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Techno- logies (NAACL-HLT), 2019: 4 171-4 186

[14] 屈丹丹,杨涛,朱垚,等. 基于字向量的BiGRU-CRF肺癌医案四诊信息实体抽取研究[J]. 世界科学技术-中医药现代化, 2021, 23(9):3 118-3 125

[15] 杨飘,董文永. 基于BERT嵌入的中文命名实体识别方法[J]. 计算机工程,2020,46(4):40-44

[16] 王天一,孟小亮,张华. 一种面向生态环境领域的知识图谱构建方法[J]. 地理空间信息,2023,21(1):14-19

[17] Li X, Yan H, Qiu X, et al. FLAT: Chinese NER Using Flat-lattice Transformer[J]. 2020. DOI:10.18653/v1/2020.acl-main,611

[18] Zhang Y Yang J. Chinese NER Using Lattice LSTM[C]// The 56th Annual Meeting of the Association for Computational Linguistics (ACL),2018.DOI:10.18653/v1/P18

[19] Peng M, Ma R, Zhang Q,et al. Simplify the Usage of Lexicon in Chinese NER[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA:Association for Computational Linguistics,2020:5 951-5 960

[20] Chen A, Yin C. CRW-NER: Exploiting Multiple Embeddings for Chinese Named Entity Recognition[C]// 4th IEEE International Conference on Artificial Intelligence and Big Data (ICAIBD),2021: 520-524

[21] 李发东,王海起,孔浩然,等. 联合词汇增强的中文细粒度地理命名实体识别模型研究[J]. 地球信息科学学报,2023, 25(6):1 106-1 120

[22] Mikolov T, Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and Their Compositionality[J]. Advances in Neural Information Processing Systems, 2013,26-30