文章编号:1671-6833(2024)03-0089-07

基于 MacBERT 和 R-Drop 的地质命名实体识别

刘 昕[1], 徐洪珍[1,2], 刘爱华[2], 邓德军[1]

(1.东华理工大学 信息工程学院,江西 南昌 330013;2.东华理工大学 软件学院,江西 南昌 330013)

摘 要:地质命名实体识别中常用的基于 BERT 预训练模型的深度学习方法是基于字的方法,没有利用词信息,且神经网络中的 Dropout 机制会导致训练阶段和推理阶段之间存在不一致性。针对该问题,提出了一种基于 MacBERT 和 R-Drop 的地质命名实体识别模型 MBCR。首先,通过 MacBERT 学习文本特征表示,充分利用字词信息;其次,运用 BiGRU 编码上下文特征,有效提取完整的语义信息;最后,采用 CRF 获取标签间的依赖关系,生成最优标签序列。此外,在训练过程中引入 R-Drop,进一步提升模型的泛化能力。结果表明:与 BiLSTM-CRF、BERT-BiLSTM-CRF 等模型相比,所提 MBCR 模型在 NERdata 数据集上的 $F1$ 值提高了 2.08百分点~4.62百分点,在 Boson 数据集上的 $F1$ 值提高了 1.26百分点~17.54百分点。

关键词:命名实体识别; 地质; MacBERT; BiGRU; R-Drop

中图分类号:TP311 **文献标志码**:A **doi**:10.13705/j.issn.1671-6833.2024.03.002

“深时数字地球”国际大科学计划[1]是由中国科学家发起和主导的国际科学合作计划。该计划以大数据为驱动,建立一个链接地学信息的研究平台,旨在实现地球科学领域重大科学问题的突破。目前,全国地质资料馆已经存储了海量非结构化的地质报告[2],如何高效挖掘其中地质信息是一个重要的研究课题[3]。

地质命名实体识别任务的目标是从给定的地质文本中识别出与地质相关的实体,并将它们归类到预定义的类别[4]。近年来,随着深度学习的发展,预训练模型的研究取得了巨大成功,其中 BERT(bidirectional encoder representations from transformers)模型[5]就是代表之一,在多种自然语言处理任务上均取得了 SOTA(state of the art)的效果。因此,在地质命名实体识别任务上,基于 BERT 模型的深度学习方法是目前的主流方法。尽管这些研究方法已经在地质命名实体识别任务上取得了不错的效果,但是仍然面临两大挑战。首先,这些方法采用的中文 BERT 模型在 MLM(masked language model)任务中使用了字级别的掩码方法,无法捕捉词级别的语义信息。其次,有研究表明神经网络中的 Dropout 机制会导致训练和推理阶段之间存在不一致性,即训练过程中使用的由 Dropout 随机采样的子模型与推理过程中使用的完整模型不一致[6]。

针对上述问题,本文提出了一种基于 MacBERT(masked language model as correction BERT)[7]和 R-Drop(regularized dropout)[8]的地质命名实体识别模型 MBCR(MacBERT-BiGRU-CRF-R-Drop)。首先,通过 MacBERT 在地质文本上学习具有丰富字词特征的向量表示;其次,利用 BiGRU(bidirectional gated recurrent unit)提取向量表示的上文语义信息,并运用 CRF(conditional random field)输出全局最优标签序列;最后,在训练过程中引入 R-Drop,降低由 Dropout 导致的训练和推理阶段之间的不一致性,从而提升模型的泛化能力。

本文的主要创新工作如下:①在地质命名实体识别任务中,首次引入 MacBERT 预训练模型,可以充分利用字词信息;②为地质命名实体识别研究引入 R-Drop 正则化,缓解由 Dropout 造成的训练与推理阶段之间的不一致问题;③在 NERdata 和 Boson 数据集上系统对比了包含 MBCR 在内的多种不同的命名实体识别模型。

收稿日期:2023-09-15;**修订日期**:2023-10-20

基金项目:国家自然科学基金资助项目(62066003);江西省教育厅科技计划项目(GJJ160554);江西省抚州市人才计划项目(2021ED008);江西省网络空间安全智能感知重点实验室室开放项目(JKLCIP202202)

通信作者:徐洪珍(1976—),男,江西抚州人,东华理工大学教授,博士,主要从事机器学习、大数据、云计算研究,E-mail:xuhz@ecut.edu.cn。

引用本文:刘昕,徐洪珍,刘爱华,等. 基于 MacBERT 和 R-Drop 的地质命名实体识别[J]. 郑州大学学报(工学版),2024,45(3):89-95.(LIU X, XU H Z, LIU A H, et al. Geological named entity recognition based on MacBERT and R-Drop[J]. Journal of Zhengzhou University (Engineering Science),2024,45(3):89-95.)

1 相关工作

现有的命名实体识别方法可以分为三类:基于规则的方法、基于统计机器学习的方法和基于深度学习的方法[9]。其中,基于规则的方法需要领域专家对整个文本进行分析来构建规则模板,耗时耗力且难以泛化到其他领域[10]。基于统计机器学习的方法应用相对广泛,如 HMM(hidden Markov model)模型[11]、ME(maximum entropy)模型[12]和 CRF 模型[13]等,但上述方法需要进行复杂的特征工程,对特征选取的质量要求较高。相比于基于统计机器学习的方法,基于深度学习的方法可以自动从文本中提取高阶特征,进一步提高命名实体识别的准确率。在地质领域,目前主要采用的是基于深度学习的方法。张雪英等[14]提出了一种基于深度信念网络的地质命名实体识别模型,并且构建了地质实体信息的标注规范和语料库;Qiu 等[15]针对传统的命名实体识别方法对特征工程依赖严重的问题,提出了一种融入注意力机制的 BiLSTM-CRF 地质命名实体识别方法。

近年来,随着深度学习的蓬勃发展,预训练模型的研究取得了重大进展,其中 Devlin 等[5]提出的 BERT 模型就是代表工作之一。相较于传统的 Word2Vec 模型[16],BERT 能有效处理一词多义的情况,并已被广泛应用于自然语言处理的各种任务。Huang 等[17]提出了一种结合 BERT、BiLSTM(bidirectional long short-term memory)和 CRF 的地质新闻命名实体识别方法,该方法能够从地质新闻数据集中准确地识别出实体的边界和类别;王权于等[18]提出了一种岩土工程命名实体识别模型 BERT-BiGRU-CRF,并且构建了小规模岩土工程命名实体语料库;Yu 等[19]设计了一种基于 BERT 和 CRF 的矿物命名实体识别模型,该模型能够有效地提取出矿物实体信息;Liu 等[20]针对现有的地质命名实体识别方法需要大量标注语料进行训练的问题,提出了一种基于 GeoBERT 的地质命名实体识别模型。

综上所述,在地质命名实体识别领域中,已经有不少研究方法被提出并取得了一定的成果。然而,这些方法都是基于字级别特征表示的方法,未能充分利用词信息,因此难以全面捕捉地质实体的语义信息。此外,基于深度学习的方法所采用的 Dropout 技术会导致训练和推理阶段之间存在不一致性。

2 基于 MBCR 模型的地质命名实体识别

本文提出的地质命名实体识别模型 MBCR 包含 4 个部分:MacBERT 表征学习层、BiGRU 上下文编码层、CRF 标签解码层和 R-Drop 训练算法,其整体结构如图 1 所示。该模型首先通过 MacBERT 表征学习层在地质文本上学习具有字词语义信息的特征向量;其次,使用 BiGRU 上下文编码层对特征向量进行上下文编码;最后,利用 CRF 标签解码层获取概率最大的标签序列。在训练过程中,对每个样本数据采用 R-Drop 训练算法来实现由 Dropout 随机采样的两个子模型间的输出分布趋同,从而减少训练和推理阶段之间的不一致性。

2.1 MacBERT 表征学习层

地质文本中存在大量的多义词,在不同语境中同一个词具有不同的含义。例如,“水平”可以解释为同水平面平行,也可以解释为在某一专业方面所达到的高度;“品位”可以解释为矿石中有用成分的含量,也可以解释为官阶或品质。BERT 模型能够

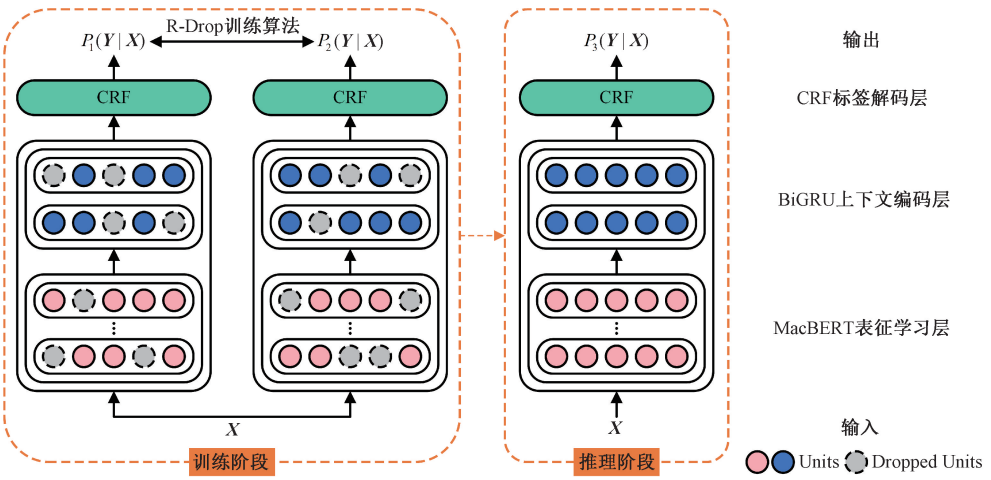

图 1 MBCR 模型结构

Figure 1 MBCR model structure

处理一词多义的情况,但却无法捕捉词级别的语义信息。因此,本文引入 MacBERT 作为模型的表征学习层,学习地质文本特征表示。相比于 BERT,MacBERT 在 MLM 任务中采用了相似词掩码策略。例如:原始文本为“退变为纤闪石,粒度与辉石相近”,BERT 掩码策略为“退[MASK]为纤闪石,粒度与辉石相[MASK]”,MacBERT 掩码策略为“退化为纤闪石,粒度与辉石相似”。该策略减小了预训练和微调任务之间的差距,增强了捕获词级别语义信息的能力。

模型初始输入为地质文本序列 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_k\}$,其中 x_j 为序列 $\boldsymbol{X}$ 的第 j 个字符。序列 $\boldsymbol{X}$ 对应的 MacBERT 输入表示 $\boldsymbol{E}=\{e_1,e_2,\cdots,e_k\}$ 由 3 个嵌入特征相加而成,如图 2 所示。其中,3 个嵌入特征分别为符号嵌入、段嵌入、位置嵌入,符号[CLS]标识序列的开始,符号[SEP]标识序列的结束。向量 $\boldsymbol{E}$ 经过多个双向 Transformer 编码器获得特征向量 $\boldsymbol{A}=\{a_1,a_2,\cdots,a_k\}$。

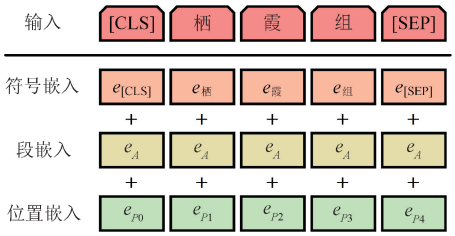

图 2　MacBERT 输入表示

Figure 2　MacBERT input representation

2.2　BiGRU 上下文编码层

地质命名实体识别是一种序列标注任务,通过分析上下文中的语义信息,可以更准确地识别地质实体。前向 GRU 无法获取目标字的下文信息,例如,针对岩石实体“黑云石英片岩”,目标字为“片”,前向 GRU 只能获取“片”的前一个字“英”的特征,无法获取后一个字“岩”的特征,因此,采用 BiGRU 作为模型的上下文编码层。将 MacBERT 层输出的向量输入到 BiGRU 层中,用来提取上下文信息。BiGRU 由前向 GRU 和后向 GRU 构成。相比于 LSTM,GRU 只包含重置门和更新门,结构更精简,在小样本地质数据集上的泛化效果更好。GRU 计算公式[21]为

$$r_t=\sigma(\boldsymbol{W}_{ra}a_t+\boldsymbol{W}_{rh}h_{t-1}+\boldsymbol{b}_r);\tag{1}$$

$$z_t=\sigma(\boldsymbol{W}_{za}a_t+\boldsymbol{W}_{zh}h_{t-1}+\boldsymbol{b}_z);\tag{2}$$

$$\tilde{h}_t=\tanh(\boldsymbol{W}_{\tilde{h}a}a_t+\boldsymbol{W}_{\tilde{h}h}(r_t\odot h_{t-1})+\boldsymbol{b}_{\tilde{h}});\tag{3}$$

$$h_t=(1-z_t)\odot h_{t-1}+z_t\odot\tilde{h}_t。\tag{4}$$

式中:a_t 表示 t 时刻的输入;r_t 和 z_t 分别表示 t 时刻重置门和更新门的输出;$\tilde{h}_t$ 和 h_t 分别表示 t 时刻的候选隐藏状态和隐藏状态;$\boldsymbol{W}$ 和 $\boldsymbol{b}$ 分别表示权重矩阵和偏置向量;σ 表示 Sigmoid 函数;$\odot$ 表示 Hadamard 乘积。

2.3　CRF 标签解码层

BiGRU 层能编码上下文语义信息,但忽略了地质实体标签间的依赖关系。因此,在模型的顶部使用 CRF 层,学习地质实体相邻标签间的约束关系,保证预测标签序列的合理性。

对于地质文本序列 $\boldsymbol{X}=\{x_1,x_2,\cdots,x_k\}$,其标签序列 $\boldsymbol{Y}=\{y_1,y_2,\cdots,y_k\}$ 的得分为

$$S(\boldsymbol{X},\boldsymbol{Y})=\sum_{j=0}^{k}T_{y_j,y_{j+1}}+\sum_{j=1}^{k}O_{j,y_j}。\tag{5}$$

式中:$T_{y_j,y_{j+1}}$ 表示由标签 y_j 转移到标签 y_{j+1} 的概率;O_{j,y_j} 表示第 j 个字符预测为标签 y_j 的分数。

通过 Softmax 函数对标签序列得分 $S(\boldsymbol{X},\boldsymbol{Y})$ 进行归一化,得到标签序列 $\boldsymbol{Y}$ 的概率分布为

$$P(\boldsymbol{Y}\mid\boldsymbol{X})=\frac{\mathrm{e}^{S(\boldsymbol{X},\boldsymbol{Y})}}{\sum_{\tilde{\boldsymbol{Y}}}\mathrm{e}^{S(\boldsymbol{X},\tilde{\boldsymbol{Y}})}}。\tag{6}$$

式中:$\tilde{\boldsymbol{Y}}$ 表示所有可能的标签序列。

最后,通过 Viterbi 算法[22]获得概率最大的标签序列:

$$\boldsymbol{Y}^*=\operatorname{argmax}P(\boldsymbol{Y}\mid\boldsymbol{X})。\tag{7}$$

在模型训练过程中,CRF 损失为

$$L_{\mathrm{CRF}}=-\log P(\boldsymbol{Y}\mid\boldsymbol{X})。\tag{8}$$

2.4　R-Drop 训练算法

考虑到训练阶段使用的由 Dropout 生成的子模型与推理阶段使用的完整模型之间存在不一致性,本文引入 R-Drop 训练算法。该算法通过最小化双向 KL 散度来正则化子模型间的输出分布,从而缓解了训练和推理阶段之间的不一致问题。

将地质文本序列 $\boldsymbol{X}$ 分别输入到由 Dropout 采样的 2 个子模型中,获得其标签序列 $\boldsymbol{Y}$ 的 2 个概率分布 $P_1(\boldsymbol{Y}\mid\boldsymbol{X})$ 和 $P_2(\boldsymbol{Y}\mid\boldsymbol{X})$,其双向 KL 散度损失为

$$L_{\mathrm{KL}}=\frac{1}{2}(D_{\mathrm{KL}}(P_1(\boldsymbol{Y}\mid\boldsymbol{X})\parallel P_2(\boldsymbol{Y}\mid\boldsymbol{X}))+(D_{\mathrm{KL}}(P_2(\boldsymbol{Y}\mid\boldsymbol{X})\parallel P_1(\boldsymbol{Y}\mid\boldsymbol{X}))。\tag{9}$$

因此,模型整体的损失为

$$L=L_{\mathrm{CRF}}^1+L_{\mathrm{CRF}}^2+\alpha L_{\mathrm{KL}}。\tag{10}$$

式中:α 为 L_{KL} 的权重系数。

算法 1　R-Drop 训练算法。

输入:训练数据集 $\boldsymbol{D}=\{(\boldsymbol{X}_i,\boldsymbol{Y}_i)\}_{i=1}^{n}$;

输出:模型参数 θ。

① 初始化模型参数 θ;

② 随机抽样数据对 $(\boldsymbol{X}_i,\boldsymbol{Y}_i)\sim\boldsymbol{D}$;

③ 将数据 $\boldsymbol{X}_i$ 重复输入模型两次,获得输出分布 $P_1(\boldsymbol{Y}_i \mid \boldsymbol{X}_i)$ 和 $P_2(\boldsymbol{Y}_i \mid \boldsymbol{X}_i)$;

④ 使用式(8)计算得到 CRF 损失 L_{CRF}^1 和 L_{CRF}^2;

⑤ 使用式(9)计算得到双向 KL 散度损失 L_{KL};

⑥ 通过最小化式(10)中的损失 L 来更新模型参数 θ;

⑦ 重复②~⑥,直到模型收敛。

3 实验

3.1 数据集

为了验证模型的有效性和泛化能力,本文在地质领域命名实体识别数据集 NERdata[2] 和通用领域命名实体识别数据集 Boson 上进行了实验。

NERdata 是由马凯等[2]基于 4 份区域地质调查报告构建的数据集,包含 6 类实体:岩石(Rock)、地质构造(Geological structure)、矿物(Mineral)、地点(Location)、地质时间(Geological time)、地层(Stratum)。

Boson 是由 Boson 中文语义开放平台发布的数据集,包含 6 类实体:人名(Person)、产品(Product)、时间(Time)、机构(Organization)、地名(Location)、公司(Company)。

将 NERdata 和 Boson 数据集中的文本以句号进行切分,分别得到了 10 796 和 9 875 条语句。按照 6:2:2的比例将各个数据集划分为训练集、验证集、测试集。NERdata 数据集实体统计如表 1 所示,Boson 数据集实体统计如表 2 所示。

表 1 NERdata 数据集实体统计

Table 1 Entity statistics of NERdata dataset

实体类别	实体数量		
	训练集	验证集	测试集
岩石	6 003	1 985	1 860
地质构造	841	273	246
矿物	2 925	1 004	1 059
地点	1 458	486	412
地质时间	1 148	388	334
地层	1 861	573	585

表 2 Boson 数据集实体统计

Table 2 Entity statistics of Boson dataset

实体类别	实体数量		
	训练集	验证集	测试集
人名	2 838	1 000	942
产品	2 216	803	760
时间	2 500	833	801
机构	1 581	538	532
地名	2 754	926	801
公司	1 412	450	451

实验采用命名实体识别任务中常用的 BIO 标注规范,如表 3 所示。

表 3 BIO 标注示例

Table 3 BIO annotation examples

文本序列	标注结果
具复理石韵律特征	具/O 复/B-Mineral 理/I-Mineral 石/I-Mineral 韵/O 律/O 特/O 征/O
碎屑岩,碳酸盐岩	碎/B-Rock 屑/I-Rock 岩/I-Rock,/O 碳/B-Rock 酸/I-Rock 盐/I-Rock 岩/I-Rock

3.2 评价指标

实验采用精确率 P、召回率 R 和 $F1$ 值作为模型性能的评价指标,具体计算公式为

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

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

$$F1 = \frac{2 \times P \times R}{P + R}。 \tag{13}$$

式中:TP 表示真正例的数量;FP 表示假正例的数量;FN 表示假负例的数量。当且仅当实体的边界和类型都被正确识别,才认为该实体被正确识别。

3.3 实验环境及设置

本文采用 Python 3.8.0 和 PyTorch 1.12.0 作为实验环境,使用 Chinese-MacBERT-base 版本的 MacBERT,其包含 12 个双向 Transformer 编码器。模型的超参数:MacBERT 隐藏层维度为 768;BiGRU 隐藏层维度为 256;学习率为 10^{-5};*Batch size* 为 32;Dropout 丢弃神经元的比例 D_{p} 为 30%。

3.4 不同模型性能对比

为了验证 MBCR 模型的有效性,本文在 NERdata 和 Boson 数据集上系统对比了 BiLSTM-CRF[23]、BERT-BiLSTM-CRF[17]、BERT-BiGRU-CRF[18]、MacBERT-BiLSTM-CRF[24]和 MBCR 这 5 种不同的命名实体识别模型。不同模型在 NERdata 和 Boson 数据集上的对比结果如表 4 所示。

表 4 不同模型性能对比

Table 4 Performance comparison of different models %

模型	NERdata			Boson		
	P	R	$F1$	P	R	$F1$
BiLSTM-CRF[23]	66.04	71.31	68.57	67.08	63.63	65.31
BERT-BiLSTM-CRF[17]	68.29	72.95	70.55	80.01	80.38	80.20
BERT-BiGRU-CRF[18]	68.66	73.04	70.78	80.28	81.18	80.72
MacBERT-BiLSTM-CRF[24]	68.97	73.40	71.11	80.82	82.37	81.59
MBCR	71.31	75.18	73.19	82.49	83.21	82.85

实验结果表明,MBCR 模型性能优于其他对比模型。MBCR 在 NERdata 数据集上的精确率、召回率和 $F1$ 值分别为 71.31%、75.18% 和 73.19%,在 Boson 数据集上的精确率、召回率和 $F1$ 值分别为 82.49%、83.21% 和 82.85%。虽然 BERT、MacBERT 预训练模型参数量大,导致模型整体复杂度高,但由于其可以生成考虑语境的嵌入向量,因此能够显著提高模型的表现。相比于 BERT,MacBERT 减少了预训练和微调任务之间的差距,增强了捕获词级别语义信息的能力,从而使模型能够更准确地识别出实体的类型和边界。由于 BiGRU 的结构更精简,参数更少,因此在小样本数据集上的泛化效果更好。R-Drop 仅增加了双向 KL 散度损失,没有进行结构修改。与 MacBERT-BiLSTM-CRF 相比,MBCR 在 NERdata 和 Boson 数据集上的 $F1$ 值分别提升了 2.08 百分点和 1.26 百分点,具有显著的优势。

3.5　消融实验

在 BiGRU 的基础上进行实验,验证了 MBCR 中各个组成部分对模型性能的影响,在 NERdata 和 Boson 数据集上的消融实验结果如表 5 所示。引入 CRF 来获取标签间的依赖关系,有助于提升模型性能。采用 MacBERT 可以生成高质量动态嵌入向量,从而使模型能够有效识别出实体信息。加入 R-Drop 后,模型的精确率、召回率和 $F1$ 值均有明显提升,验证了 R-Drop 的有效性。

表 5　消融实验结果

Table 5　Ablation experiment results　%

模型	NERdata			Boson		
	P	R	$F1$	P	R	$F1$
BiGRU	67.58	67.55	67.56	66.65	62.70	64.62
BiGRU+CRF	67.61	72.02	69.75	68.29	64.59	66.39
BiGRU+CRF+MacBERT	69.10	73.80	71.37	81.33	82.60	81.96
BiGRU+CRF+MacBERT+R-Drop	71.31	75.18	73.19	82.49	83.21	82.85

3.6　Dropout 的影响

实验探索了 Dropout 对 MBCR 模型性能的影响,在 NERdata 和 Boson 数据集上的实验结果如表 6 所示。Dropout 在训练过程中随机丢弃一定比例的神经元来执行隐式集成,有助于提升 MBCR 的泛化能力。但是,MBCR 的 $F1$ 值不会随着 Dropout 丢弃神经元的比例 D_p 的增加而完全提高。当 D_p 为 50%时,MBCR 在 NERdata 和 Boson 数据集上的 $F1$ 值均低于 D_p 为 30%时的 $F1$ 值。

表 6　Dropout 对模型性能的影响

Table 6　Effect of Dropout on model performance　%

D_p	NERdata			Boson		
	P	R	$F1$	P	R	$F1$
0	67.97	71.04	69.47	79.05	79.94	79.49
10	70.40	74.76	72.51	82.39	83.16	82.77
30	71.31	75.18	73.19	82.49	83.21	82.85
50	71.01	72.53	71.76	77.71	75.37	76.52

3.7　α 的影响

本节实验探索了 R-Drop 的权重 α 对 MBCR 模型性能的影响,在 NERdata 和 Boson 数据集上的实验结果如表 7 所示。α 太小会导致模型过于复杂,容易发生过拟合。相反,α 太大会使得模型过于简单,存在欠拟合的风险。当 α 为 5 时,MBCR 在 NERdata 和 Boson 数据集上的性能最出色。

表 7　α 对模型性能的影响

Table 7　Effect of α on model performance

α	NERdata			Boson		
	P/%	R/%	$F1$/%	P/%	R/%	$F1$/%
1	70.61	73.60	72.08	81.68	82.81	82.24
3	70.71	74.04	72.34	82.17	83.11	82.64
5	71.31	75.18	73.19	82.49	83.21	82.85
7	70.84	74.62	72.68	81.96	82.88	82.42
9	70.79	74.29	72.50	81.58	82.65	82.11

4　结论

针对地质命名实体识别问题,本文提出了一种基于 MacBERT 和 R-Drop 的地质命名实体识别模型 MBCR。在 NERdata 数据集和 Boson 数据集上的实验结果表明,MBCR 模型性能明显优于其他模型。

由于目前公开的带标注的地质命名实体识别数据集规模较小,并且存在标注不统一和实体位置偏差的问题,这导致了所有实验模型的评价指标整体偏低。因此,未来的研究工作将围绕两个方面开展:一方面,优化模型结构,进一步提升模型的泛化能力;另一方面,完善地质命名实体识别语料库,规范地质实体标注。

参考文献:

[1]　WANG C S, HAZEN R M, CHENG Q M, et al. The

deep-time digital earth program: data-driven discovery in geosciences[J]. National Science Review, 2021, 8(9): nwab027.

[2] 马凯, 田苗, 谭永健, 等. 基于四份区域地质调查报告构建的命名实体识别试验数据集研发[J]. 全球变化数据学报(中英文), 2022, 6(1): 78-84, 237-243.
MA K, TIAN M, TAN Y J, et al. Development of a named entity recognition dataset based on four regional geological survey reports[J]. Journal of Global Change Data & Discovery, 2022, 6(1): 78-84, 237-243.

[3] QIU Q J, XIE Z, WU L, et al. Automatic spatiotemporal and semantic information extraction from unstructured geoscience reports using text mining techniques[J]. Earth Science Informatics, 2020, 13(4): 1393-1410.

[4] 储德平, 万波, 李红, 等. 基于ELMO-CNN-BiLSTM-CRF模型的地质实体识别[J]. 地球科学, 2021, 46(8): 3039-3048.
CHU D P, WAN B, LI H, et al. Geological entity recognition based on ELMO-CNN-BiLSTM-CRF model[J]. Earth Science, 2021, 46(8): 3039-3048.

[5] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2018-10-11)[2023-03-15]. https://arxiv.org/abs/1810.04805.

[6] ZOLNA K, ARPIT D, SUHUBDY D, et al. Fraternal dropout[EB/OL]. (2017-10-31)[2023-03-15]. https://arxiv.org/abs/1711.00066.

[7] CUI Y M, CHE W X, LIU T, et al. Pre-training with whole word masking for Chinese BERT[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504-3514.

[8] LIANG X B, WU L J, LI J T, et al. R-drop: regularized dropout for neural networks[J/OL]. (2021-10-29)[2023-03-15]. https://arxiv.org/abs/2106.14448.

[9] LIU P, GUO Y M, WANG F L, et al. Chinese named entity recognition: the state of the art[J]. Neurocomputing, 2022, 473: 37-53.

[10] 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.

[11] ZHANG J, SHEN D, ZHOU G D, et al. Enhancing HMM-based biomedical named entity recognition by studying special phenomena[J]. Journal of Biomedical Informatics, 2004, 37(6): 411-422.

[12] SAHA S K, SARKAR S, MITRA P. Feature selection techniques for maximum entropy based biomedical named entity recognition[J]. Journal of Biomedical Informatics, 2009, 42(5): 905-911.

[13] SUN C J, GUAN Y, WANG X L, et al. Rich features based conditional random fields for biological named entities recognition[J]. Computers in Biology and Medicine, 2007, 37(9): 1327-1333.

[14] 张雪英, 叶鹏, 王曙, 等. 基于深度信念网络的地质实体识别方法[J]. 岩石学报, 2018, 34(2): 343-351.
ZHANG X Y, YE P, WANG S, et al. Geological entity recognition method based on deep belief networks[J]. Acta Petrologica Sinica, 2018, 34(2): 343-351.

[15] QIU Q J, XIE Z, WU L, et al. BiLSTM-CRF for geological named entity recognition from the geoscience literature[J]. Earth Science Informatics, 2019, 12(4): 565-579.

[16] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of the 26th International Conference on Neural Information Processing Systems: Volume 2. New York: ACM, 2013: 3111-3119.

[17] HUANG C, WANG Y Z, YU Y Q, et al. Chinese named entity recognition of geological news based on BERT model[J]. Applied Sciences, 2022, 12(15): 7708.

[18] 王权于, 李振华, 涂志鹏, 等. 基于BERT-BiGRU-CRF模型的岩土工程实体识别[J]. 地球科学, 2023, 48(8): 3137-3150.
WANG Q Y, LI Z H, TU Z P, et al. Geotechnical named entity recognition based on BERT-BiGRU-CRF model[J]. Earth Science, 2023, 48(8): 3137-3150.

[19] YU Y Q, WANG Y Z, MU J Q, et al. Chinese mineral named entity recognition based on BERT model[J]. Expert Systems with Applications, 2022, 206: 117727.

[20] LIU H, QIU Q J, WU L, et al. Few-shot learning for name entity recognition in geological text based on GeoBERT[J]. Earth Science Informatics, 2022, 15(2): 979-991.

[21] CHO K, VAN M B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[EB/OL]. (2014-06-03)[2023-03-15]. https://arxiv.org/abs/1406.1078.

[22] STRUBELL E, VERGA P, BELANGER D, et al. Fast and accurate entity recognition with iterated dilated convolutions[EB/OL]. (2017-02-07)[2023-03-15]. https://arxiv.org/abs/1702.02098.

[23] HUANG Z H, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[EB/OL]. (2015-08-09)[2023-03-15]. https://arxiv.org/abs/1508.01991.

[24] CUI Z J, YUAN Z M, WU Y F, et al. Intelligent recommendation for departments based on medical knowledge graph[J]. IEEE Access, 2023, 11: 25372-25385.

Geological Named Entity Recognition Based on MacBERT and R-Drop

LIU Xin[1], XU Hongzhen[1,2], LIU Aihua[2], DENG Dejun[1]

(1. School of Information Engineering, East China University of Technology, Nanchang 330013, China; 2. School of Software, East China University of Technology, Nanchang 330013, China)

Abstract: The commonly used deep learning methods based on BERT pre-trained model in geological named entity recognition were character-based approaches, and could not utilize word-level information. Additionally, the dropout mechanism in neural networks might cause inconsistency between the training and inference stage. To address this issue, a geological named entity recognition model MBCR based on MacBERT and R-Drop was proposed. Firstly, MacBERT was used to learn text feature representations, which could fully utilize character and word information. Then, BiGRU was employed to encode context features, effectively extracting complete semantic information. Subsequently, CRF was adopted to capture dependencies between labels and generate the optimal label sequence. Moreover, R-Drop was introduced during the training process to further enhance the model's generalization capabilities. Compared with BiLSTM-CRF, BERT-BiLSTM-CRF, and other models, the proposed MBCR model improved the *F*1-score on the NERdata dataset by 2.08-4.62 percentage points and on the Boson dataset by 1.26-17.54 percentage points.

Keywords: named entity recognition; geology; MacBERT; BiGRU; R-Drop

---

(上接第 45 页)

Object Detection and Recognition Algorithm Based on YOLOv5 and the Fusion of Attention and Multistage Features

WANG Yu, BI Yu, SHI Jiantong, XIAO Hongbing, SUN Mei

(School of Computer and Artificial Intelligence, Beijing Technology and Business University, Beijing 100048, China)

Abstract: To tackle the problem of low accuracy of detection and recognition for object in complex scenes, YOLOv5 object detection and recognition algorithm based on attention and multistage feature fusion(AMFF) was proposed in this study. The main ideas included adding the proposed dual space directions pyramid split attention (DSD-PSA) mechanism to the backbone network of the traditional YOLOv5s model to enhance the learning of the feature map space and channel information, adopting multistage feature fusion(MFF) structure in the bottleneck network to fuse the features of different branches, increasing richness of the feature and improving the ability to cope with complex scenes. In addition, C3Ghost module and depthwise separable convolution were used to replace C3 module and common convolution to reduce the number of parameters and the complexity of network. Compared with the traditional YOLOv5s algorithm, the mean average accuracy of the proposed algorithm in the VOC2007+2012 data set reached 85%, and the mean average accuracy of the smart retail cabinet commodity identification data set reached 97.2%, which verified the effectiveness and feasibility of the proposed algorithm.

Keywords: deep learning; YOLOv5s; object detection; multistage feature fusion; attention mechanism