基于岩石文本信息的命名实体识别

杜睿山,陈思路,刘文豪
(东北石油大学 计算机与信息技术学院,黑龙江 大庆 163318)

摘 要:命名实体识别技术是自然语言处理领域的重要任务之一。但岩石文本信息中的命名实体存在边界不清、分词困难、误差传播、计算效率慢等问题。基于岩石文本信息进行知识抽取对油气勘探领域的研究具有重大意义。为此,该文首先构建岩石文本数据集,并提出 Lexicon-BiLSTM-CRF 网络模型应用于非结构化的岩石文本上,该模型首先经过 Lexicon 机制获得每个字符的所有匹配词,从而解决了边界不清、分词困难的问题,在此基础上提升了计算效率。然后通过双向长短期记忆网络(BiLSTM)提取上下文语义特征,将语义向量传入条件随机场(CRF)层并采用维特比算法解码,降低了错误标签的输出概率并预测实体标注标签,最终实现岩石文本的命名实体抽取任务。在构建的岩石文本数据集的基础上进行几组对比实验,验证了该方法在准确率和召回率上具有一定提升。

关键词:命名实体识别;Lexicon;岩石;非结构化文本;条件随机场;知识抽取
中图分类号:TP391.1 **文献标识码**:A **文章编号**:1673-629X(2022)09-0188-05
doi:10.3969/j.issn.1673-629X.2022.09.029

Named Entity Recognition Based on Rock Text Information

DU Rui-shan,CHEN Si-lu,LIU Wen-hao
(School of Computer and Information Technology,Northeast Petroleum University,Daqing 163318,China)

Abstract:Named entity recognition technology is one of the important tasks in the field of natural language processing. However,the named entities in the rock text information have problems such as unclear boundaries,difficult word segmentation,error propagation,and slow calculation efficiency. Knowledge extraction based on rock text information is of great significance to the research in the field of oil and gas exploration. To this end,we first build a rock text data set,and then propose a Lexicon-LSTM-CRF network model to be applied to unstructured rock text. Firstly,the Lexicon mechanism is used to obtain all matching words of each character,so as to solve the problem of unclear boundary and difficult word segmentation,and on this basis,improve the computational efficiency. Then the contextual semantic features are extracted through the bidirectional long-term short-term memory network (BiLSTM),and the semantic vector is passed into the Conditional Random Field (CRF) layer and decoded by the Viterbi algorithm to reduce the output probability of the error label and predict the entity annotation label,and finally realize the rock text Named entity extraction task. Through several comparative experiments on the rock text data set constructed,it is verified that the proposed method has a certain improvement in accuracy and recall.

Key words:named entity recognition;Lexicon;rock;unstructured text;conditional random field;knowledge extraction

0 引 言

自然语言处理属于人工智能与语言学的交叉学科,其中的命名实体识别是实现信息抽取的重要基础任务[1]。为智能化地对岩石薄片信息进行信息抽取以及生成研究对象相关属性等信息,进而揭示具体(比如岩石结构构造、颗粒状态、产状成因等)特征、变化及规律,这在油气勘探、开发、生产的各个阶段发挥着越来越关键的作用[2]。这可以进一步满足研究人员的快速判读需求,实现决策支持的功能。岩石相关信息多以非结构化文本形式存在于书籍和文献中,岩石相关命名实体是从非结构化文本数据中抽取出的一些含有具体特征或描述意义的相关岩石名词。

命名实体识别作为智能问答、知识图谱等自然语言处理下游任务研究的基础工作,一直受到研究者们的关注。命名实体识别的早期方法主要包括基于规则的方法[3]、基于统计的方法以及基于神经网络的方法[4-5]。通过手动创建规则、创建权重或者创建实体和规则之间的一致性可以实现早期的基于规则的方

收稿日期:2021-09-08 **修回日期**:2022-01-12
基金项目:黑龙江省哲学社会科学研究规划项目(19SHE280);东北石油大学引导性创新基金(2020YDL-04)
作者简介:杜睿山(1977-),男,硕士,副教授,CCF 会员(51501M),研究方向为人工智能、机器学习等。

法,但存在着一些缺点,比如可移植性差、维护性差。隐马尔可夫模型[6]、支持向量机[7]、最大熵[8]和条件随机场[9]等对语料库的依赖较大的方法都可作为基于统计的方法实现命名实体识别任务。基于神经网络的方法对特征的依赖更小且更通用,并且广泛用于命名实体识别任务中,如循环神经网络(Recurrent Neural Network,RNN)[10]、长短期记忆网络(Long Short-Term Memory, LSTM)[11]、卷积神经网络(Convolutional Neural Network,CNN)[12]等。近年来,实现命名实体识别采用 BiLSTM 模型结合 CRF 模型实现,可以通过上下文信息的完美结合得到相邻标签之间的依赖关系,并且可以达到良好的效果。Huang 等首次利用此模型进行命名实体识别[13]。丁泽源等在中文医学领域进行命名实体关系抽取[14]。尹学振等针对在互联网公开数据中进行军事领域命名实体识别[15]。

中文命名实体识别是基于深度学习的研究将其转化为序列标注任务,这样的解决方法通常是先进行分词再进行词的分类从而得到命名实体,但这过程中存在着错误传播问题。Zhang 等提出基于 lattice 结构的命名实体识别[16],将单词本身的含义加入基于字向量的模型中从而解决了实体边界不清、误差传播的问题,但由于模型加入词典信息而对不相邻字符增加很多边导致模型过于复杂,所以存在计算效率低的缺点。

该文结合领域专家的意见,基于开放的非结构化文本数据构建了岩石语料集。在此基础上,提出一种岩石相关命名实体识别模型。使得每个字符的所有匹配词合并到字符级别 NER 模型中,进而实现非结构化岩石文本数据的命名实体识别任务。

1　岩石文本的命名实体识别模型

该文构建了基于岩石文本 Lexicon-LSTM-CRF 的 NER 模型,模型自底向上分为以下三个部分:基于 Softlexicon 的字向量表达层和 BiLSTM 层以及 CRF 层。首先,输入序列中的每个字映射为字向量。然后将 Softlexicon 特征组合并加入到字向量的表示中,同时获得每个字符的所有匹配词,这一机制缓解边界不清、分词困难的问题,然后将字向量表示输入到序列编码层,从而提取上下文特征。最终通过 CRF 层,相邻标签之间的依赖关系可以利用特征向量获得,从而降低错误标签的输出概率并输出相应的标签。岩石文本信息的命名实体识别模型如图 1 所示。

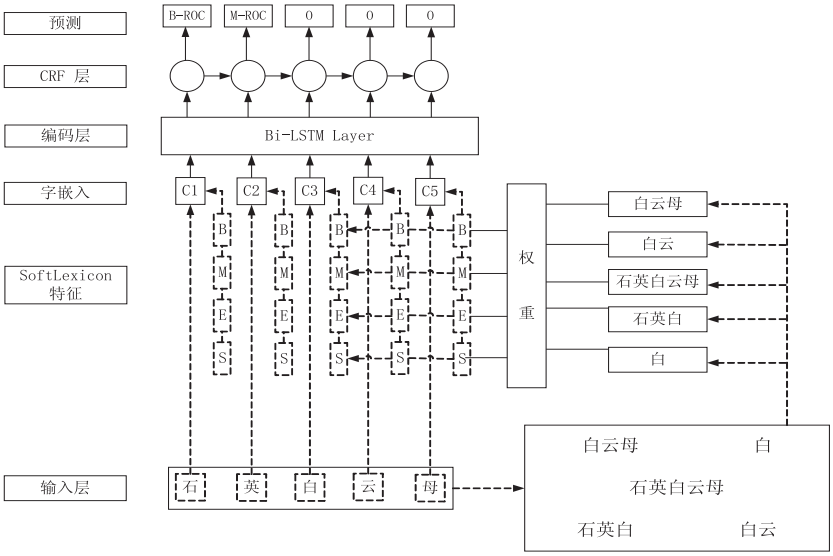

图 1　岩石文本信息的命名实体识别模型

1.1　字向量表示层

简单调节 NER 的字表示层,输入序列中的文字表示为 $\{C_1,C_2,\cdots,C_n\} \in V_c$,$V_c$ 表示字典。每个字 C_i 使用向量表示为:$X_i^c = e^c(C_i)$,e^c 表示字向量查找表。为增强字符表示,引入 bigram,字向量表示为:$X_i^c = [e^c(C_i);e^b(C_i,C_{i+1})]$。接下来把词典信息加到 char embedding 中,输入句子序列 $S = \{C_1,C_2,\cdots,C_n\}$,输入的子串 $W_{i,j} = \{C_i,C_{i+1},\cdots,C_j\}$。

由于使用 ExSoftword 方法将会导致无法加载预训练模型,且会缺失匹配信息,则该文使用 Softlexicon 方法避免以上缺点。对于一个输入句子 S 的一个字符,它的所有匹配词分为 BMES 四个类,得到 4 个词集合,具体见式(1)~式(4):

$$B(c_i) = \{w_j, \forall W_{j,k} \in L, i < k \leqslant n\} \tag{1}$$

$$M(c_i) = \{w_j, \forall W_{j,k} \in L, 1 \leqslant j < i < k \leqslant n\} \tag{2}$$

$$E(c_i) = \{w_{j,i}, \forall W_{j,i} \in L, 1 \leqslant j < i\} \tag{3}$$

$$S(c_i) = \{c_i, \exists c_i \in L\} \tag{4}$$

其中,L 表示词典,S 表示词集合。接下来对词集合做压缩,将每个类别的 word embedding 压缩为一个

embedding,词加权解决了词平均性能一般的缺点。训练数据和验证数据构成静态数据集,使用每个词在一个静态数据集上出现的频率作为静态权重,以替代 attention 动态权重从而加快训练速率。词平均表示为 $V^S(S) = |S|^{-1}\sum_{w\in s} e^w(w)$,加权表示为 $V^S(S) = (4\sum_{w\in s} z(w)e^w(w))/Z$,其中 $Z = \sum_{w\in B\cup M\cup E\cup S} z(w)$。对于一个词集合的权重归一化,会考虑所有词集合,如果包含子串 w 的子串 a 被匹配,w 的频率不会增加。最后,将词典信息合并到字符表示上作为该层的输出,具体公式见式(5)、式(6)。

$$e^S(B,M,E,S) = [V^S(B),V^S(M),V^S(E),V^S(S)] \tag{5}$$

$$X^C = [X^C;e^S(B,M,E,S)] \tag{6}$$

1.2 BiLSTM 层

1997 年,Hochreiter 等提出一种特定形式的循环神经网络——长短时记忆网络 LSTM[17]。该模型的输入层为输入 X_t,隐藏层输出为 h_t,输入门 i_t、输出门 o_t、遗忘门 f_t 以及记忆控制器 C_t 等四部分组成每个 LSTM 记忆单元。LSTM 记忆单元如图 2 所示。

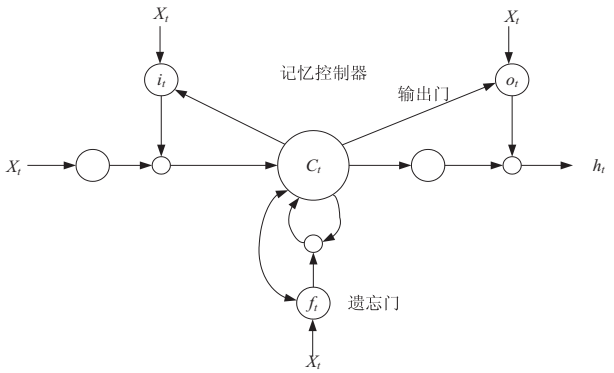

图 2 LSTM 记忆单元结构

由于 LSTM 局限于只能计算过去的上下文信息,未来的上下文信息对岩石文本信息的实体特征提取同样重要,故可以采用 BiLSTM 神经网络模型[18]。BiLSTM 模型通过顺序和逆序对输入的序列进行计算并输出两个隐藏层的向量并拼接得到最终的输出向量。该文结合词典信息,对字符之间的依赖关系进行建模引入序列建模层。这一层的通用架构包括双向长短期记忆网络(BiLSTM)、卷积神经网络(CNN)和变换器(Vaswani et al.,2017)。在这项工作中,用一个单层的 BiLSTM 实现了这个层。这里,精确地展示了正向 LSTM 的定义,具体见公式(7)~公式(9):

$$\begin{bmatrix} i_t \\ f_t \\ o_t \\ \xi_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} \left(W \begin{bmatrix} x_t^c \\ h_{t-1} \end{bmatrix} + b \right) \tag{7}$$

$$C_t = C_t \odot i_t + C_{t-1} \odot f_t \tag{8}$$

$$h_t = o_t \odot \tanh(C_t) \tag{9}$$

其中,σ 是 element-wise sigmoid 函数,$\odot$ 表示元素 element-wise product。W 和 b 是训练参数。在前向和后向 LSTM 的第 i 步串联隐藏状态 $h_i = [\overrightarrow{h_i};\overleftarrow{h_j}]$ 组成 C_i 的上下文相关表示。这一层的输出融合了语料的语义特征,各类实体标签最高者为预测结果。将结果作为 CRF 层的输入。

1.3 CRF 层

在序列建模层的顶部,通常应用序列条件随机场,它是一种用来标记和切分序列化数据的统计模型[19]。即在给定观测序列下,计算输出标记序列的条件概率分布,见公式(10)。

$$p(y \mid s;\theta) = \frac{\prod_{t=1}^{n} \varnothing_t(y_{t-1},y_t \mid s)}{\sum_{y'\in y_s}\prod_{t=1}^{n} \Phi_t(y'_{t-1},y'_{t-1} \mid s)} \tag{10}$$

句子 s 的所有可能标签序列用 y_s 表示,其中 $\varnothing_t(y_{t-1},y_t \mid s) = \exp(W_{y',y}^{\mathrm{T}}h_t + b_{y',y})$;$W_{y'}^{\mathrm{T}}$,$y$ 和 $b_{y',y}$ 是对应标签对 (y',y) 的训练参数;θ 代表模型参数。标签预测是在给定输入序列 S 的条件概率最大的情况下进行的,使用 Veterbi 算法对得到的条件概率进行最大可能序列路径求解。

CRF 层有效地考虑了上下文依赖,在 BiLSTM 层之后增加了 CRF 层,因此实体识别模型利用上下文信息的组合来有效地考虑标签依赖。

有效降低了错误标签的输出概率并实现预测实体标注标签。利用训练好的模型,对语料进行实体标注,在 CRF 层,转移矩阵作为参数,更新 BiLSTM 中的参数与 CRF 中转移概率矩阵时使用最大似然估计作为真实标记序列的概率从而标注实体类型,最终输出标注结果。从而完成岩石文本非结构化数据信息的命名实体识别任务。

2 模型实验及分析

通用领域的命名实体识别具有稳定的类别、规范的结构、统一的命名规则,主要包括人名、地名、组织名称等实体。相比而言,岩石相关文本信息的命名实体分类更为复杂,使用相关教材和文献制作数据集是非常好的数据来源。其中《矿物岩石学》、《简明岩石学》、《矿物学》等教材包含着有价值的岩石相关的实体信息。例如矿物实体、岩石实体、各类属性实体。为了弥补岩石相关文本的开放命名实体识别语料库的不足,该文基于教科书的非结构化数据构建了一个语料库,为基于开放数据研究岩石信息的命名实体识别奠定基础。结合领域专家的专业知识和已有的文献资料

确定岩石相关文本信息的命名实体划分类别,并将实体的模糊边界与实体的简化表示相结合,将原始未标注语料的语料按照字级别进行标注。最终构建包含 26 784 个句,17 个类别的语料集。

2.1　实体标注与分类

语料库是以原始语料文本为原材料,通过标注任务导向的操作方法从而形成带有语言学信息标注的语料文本。该文结合岩石文本语料本身特点,采用 BMOES 标注与自定义标注标签相结合的方式进行标注。针对其专业术语多、歧义少的特点,采用简洁、高效的 BMOES 标注机制,领域专家参与共同标注。BMOES 标注是针对数据集中的每个实体进行字级别的位置标注,命名实体的开始用 B 表示,命名实体的内部用 M 表示,命名实体的尾部用 E 表示,单个命名实体用 S 表示,不属于命名实体中的字用 O 表示。将岩石信息分为 17 大类,标注形式具体如表 1 所示。

表 1　命名实体标注类别

实体类别	实体开始	实体内部	实体结尾	单个实体
组成矿物(ROC)	B-ROC	M-ROC	E-ROC	S-ROC
岩石(CRAG)	B-CRAG	M-CRAG	E-CRAG	S-CRAG
粒状物(GRA)	B-GRA	M-GRA	E-GRA	S-GRA
晶体(JIN)	B-JIN	M-JIN	E-JIN	S-JIN
晶系(JX)	B-JX	M-JX	E-JX	S-JX
颜色(COL)	B-COL	M-COL	E-COL	S-COL
光泽(SHINE)	B-SHINE	M-SHINE	E-SHINE	S-SHINE
造岩矿物(KW)	B-KW	M-KW	E-KW	S-KW
体(TI)	B-TI	M-TI	E-TI	S-TI
形(SHA)	B-SHA	M-SHA	E-SHA	S-SHA
状(SHAPE)	B-SHAPE	M-SHAPE	E-SHAPE	S-SHAPE
角(ANG)	B-ANG	M-ANG	E-ANG	S-ANG
面(FAC)	B-FAC	M-FAC	E-FAC	S-FAC
线(LIN)	B-LIN	M-LIN	E-LIN	S-LIN
光(SHN)	B-SHN	M-SHN	E-SHN	S-SHN
作用(REF)	B-REF	M-REF	E-REF	S-REF
性质(PRO)	B-PRO	M-PRO	E-PRO	S-PRO

2.2　语料集统计

针对已获取的原始语料数据,应用命名实体标注、分类机制,实施对原始语料的实体标注,最终形成了岩石文本语料集,具体各类实体数目如表 2 所示。

表 2　岩石文本语料集实体统计

实体类别	实体数目/个	实体类别	实体数目/个
组成矿物	509	形	49
岩石	162	状	258
粒状物	32	角	36

续表 2

实体类别	实体数目/个	实体类别	实体数目/个
晶体	136	面	153
晶系	26	线	30
颜色	307	光	80
光泽	36	作用	45
造岩矿物	87	性质	192
体	381		

2.3　实验及结果分析

综上所述,由于目前没有开放的岩石语料库,该文手动构建了中文岩石文本实体语料库用于研究。实验中随机划分 70% 的语料作为训练集,20% 的语料作为验证集,10% 的语料作为测试集。命名实体识别模型的超参数设置如表 3 所示。

表 3　实验参数设置

参数名称	参数值
字向量维度	100
Dropout	0.5
迭代次数	20
学习率	0.001 5
隐藏单元数量	300

设置了以下 5 组实验,应用准确率 P、召回率 R 与 F 值进行模型评估。在构建好的岩石文本数据集上,比较了上述三种命名实体识别的有效性。实验结果如表 4 所示。

表 4　实体识别模型效果对比

模型	准确率/%	召回率/%	F1 值/%
BiLSTM	87.67	89.20	88.43
BiLSTM-CRF	90.91	92.76	91.83
Lattice-LSTM	92.76	90.03	91.37
Lattice-LSTM-CRF	94.09	95.19	94.64
Lexicon-BiLSTM-CRF	95.52	96.73	96.10

从实验结果可以看出,BiLSTM-CRF 相比于不引入 CRF 层的 BiLSTM 模型准确率提高 1.24%,F1 值提高 1.4%;Lattice-LSTM-CRF 相比于不引入 CRF 层的 Lattice-LSTM 模型准确率提高 1.33%,F1 值提高 3.27%;加入 CRF 层可以充分考虑实体的逻辑性和顺序性,从而提升了准确率与 F1 值,证实了此机制可以降低错误标签的输出概率,有助于标签的预测。此外,提出的 Lexicon-BiLSTM-CRF 的方法 F1 值可以达到 96.10%。在中文的数据集上,该方法比 Lattice-LSTM 中文实体识别模型的效果还要好,主要原因在于 lexicon 解决了词典无需重复多次调用的缺点,性能得以提升。

3 结束语

岩石实体识别中存在实体界限不清晰、实体种类丰富、数量大等问题,该文面向教材文献等非结构化数据进行岩石命名实体识别。结合领域专家的专业知识,建立了岩石相关命名实体分类规则,并构建了基于教材文献等非结构化文本数据的语料集。以岩石相关非结构化信息抽取为对象,提出了针对岩石文本信息的 Lexicon-BiLSTM-CRF 模型抽取方法。该模型利用字向量的优势,在基于字向量的模型中加入单词本身的含义。除此之外,该模型在训练过程中保存了所有可能匹配单词的同时利用 attention 机制自动给单词赋权重,进而提高了运行效率和准确率。在岩石文本数据集上通过实验对比,分析并验证了基于 BiLSTM-CRF、Lattice- BiLSTM-CRF、Lexicon-BiLSTM-CRF 的实体识别模型的有效性。下一步将用该方法在油气勘探、开发、生产领域的其他类型语料上进行广泛的训练和测试,提高模型的泛化能力。

参考文献:

[1] SANG E F,DE MEULDER F. Introduction to the CoNLL-2003 shared task:language-independent named entity recognition[J]. arXiv:cs/0306050,2003.

[2] 杨玉茹,程合生,潘卫红,等. 薄片鉴定在油气勘探开发中快速评价案例分析[J]. 地质科技情报,2015,34(2):171-173.

[3] 刘 浏,王东波. 命名实体识别研究综述[J]. 情报学报,2018,37(3):329-340.

[4] 邱 莎,阿 圆,王付艳,等. 基于统计的中文地名自动识别研究[J]. 计算机技术与发展,2011,21(11):35-38.

[5] 王 栋,李业刚,张 晓,等. 基于准循环神经网络的中文命名实体识别[J]. 计算机工程与设计,2020,41(7):2038-2043.

[6] BENGIO Y,SCHWENK H,SENÉCAL J,et al. Neural probabilistic language models[J]. The Journal of Machine Learning Research,2003,3(6):1137-1155.

[7] 陈 霄,刘 慧,陈玉泉. 基于支持向量机方法的中文组织机构名的识别[J]. 计算机应用研究,2008,25(2):362-364.

[8] VAN DEN BOSCH A. Using induced rules as complex features in memory-based language learning[C]//Proceeding of the 4th computational natural language learning and the second learning language in logic workshop (CoNLL2000 and LLL2000). Lisbon:ACM,2000:73-78.

[9] 孙 晓,孙重远,任福继. 基于深层条件随机场的生物医学命名实体识别[J]. 模式识别与人工智能,2016,29(11):997-1008.

[10] COLLOBERT R,WESTON J,BUTTOU L,et al. Natural language processing from scratch[J]. The Journal of Machine Learning Research,2011,12(1):2493-2537.

[11] KHALIFA M,SHAALAN K. Character convolutions for arabic named entity recognition with long short-term memory networks[J]. Computer Speech & Language,2019,58(11):335-346.

[12] YU Sheng,CHENG Yun,XIE Li,et al. A novel recurrent hybrid network for feature fusion in action recognition[J]. Journal of Visual Communication and Image Representation,2017,49(11):192-203.

[13] HUANG Z,XU W,YU K. Bidirectional LSTM - CRF models for sequence tagging[J]. Computer Science,2015,4(1):1-10.

[14] 丁泽源,杨志豪,罗 凌,等. 基于深度学习的中文生物医学实体关系抽取系统[J]. 中文信息学报,2021,35(5):70-76.

[15] 尹学振,赵 慧,赵俊保,等. 多神经网络协作的军事领域命名实体识别[J]. 清华大学学报:自然科学版,2020,60(8):648-655.

[16] ZHANG Y,YANG J. Chinese NER using lattice LSTM[J]. arXiv:1805.02023,2018.

[17] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation,1997,9(8):1735-1780.

[18] GRAVES A,SCHMIDHUBER J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures[J]. Neural Networks,2005,18(5-6):602-610.

[19] LAFFERTY J, MCCALLUM A, PEREORA F. Conditional random fields: probabilistic modes for segmenting and labeling sequence data[C]//Proceedings of the 18th international conference on machine learning (ICML2001). Williamstown:ICML,2001:282-289.