引用格式：崔鑫，王琰，侯小刚，周月．基于词汇增强的典型文物命名实体识别算法[J]．中国传媒大学学报（自然科学版），2023，30(02)：51-55.

文章编号：1673-4793(2023)02-0051-05

基于词汇增强的典型文物命名实体识别算法

崔鑫[1]，王琰[2]，侯小刚[2]，周月[3*]

（1. 北京邮电大学计算机学院，北京 100876；
2. 北京邮电大学人工智能学院，北京 100876；
3. 北京邮电大学电子工程学院，北京 100876）

摘要：典型文物的命名实体识别主要从句子中提取出文物名称、朝代、出土地点、馆藏地等类别的实体。典型文物数据具有构词的特殊性，使用现有命名实体识别方法在典型文物数据集上会遇到词边界判断错误等问题。本文提出了一种基于词汇增强的典型文物命名实体识别算法，算法在输入表示层和上下文编码层引入词汇信息，提高了词语领域专业性。算法通过构建文物领域词库，将其作为基于词汇增强的典型文物命名实体识别算法词典，较好地解决了词边界判断错误问题，在典型文物数据集上取得了较好的效果。

关键词：词汇增强；领域词库；命名实体识别

中图分类号：TP391　**文献标识码：**A

A lexicon enhanced named entity recognition algorithm for typical cultural relics

CUI Xin, WANG Yan, HOU Xiaogang, ZHOU Yue*

(Beijing University of Posts and Telecommunications, Beijing 100876)

Abstract: Named entity recognition of typical cultural relics focuses on extracting entities from sentences in categories such as name of cultural relic, dynasty, excavation site, and place of collection. The data of typical cultural relics has the specificity of word construction, and using existing named entity recognition methods on typical cultural relics dataset will encounter problems such as wrong word boundary judgments. The algorithm introduces lexical information in both the input representation layer and the contextual encoding layer to improve the word domain expertise. By constructing a lexicon of heritage domain words, the algorithm is used as a lexicon for the lexically enhanced recognition algorithm of typical heritage named entities, which eventually solves the problem of incorrect word boundary judgement and achieves better results on the typical heritage dataset.

Keywords: lexicon enhanced; domain thesaurus; named entity recognition;

1　引言

文物是中华文化的重要组成部分，对于保护和传承中华文化具有不可替代的作用。本文选取可移动文物中的三类典型文物石刻、陶瓷、青铜器作为研究对象，这些文物是中国文化遗产中较为珍贵且受到广泛关注的部分，对于研究中国古代科技、美学和文化历史等方面具有极为重要的价值。文物数据是指文物各种属性和信息的数字化记录和存储，例如文物的名称、年代、类别、材质、尺寸、形态、寓意、保存状况、

基金项目：国家重点研发计划课题“文化资源大数据服务工程方法与数据加工技术研究”（2021TFF0901701）

作者简介（*为通讯作者）：周月（1986-），男，副教授，博士，主要从事知识图谱、数字化采集等方面的研究。Email：yuezhou@bupt.edu.cn

历史背景等各方面的信息。通过对文物数据的采集、整理和分析,可以更好地了解和挖掘文物的历史文化价值,同时也为文物的保护和传承提供了基础数据支持。通过命名实体识别技术可以从非结构化文本数据中得到实体位置以及实体类型信息,减轻博物馆工作人员人工标注的压力,促进三元组数据的构建。

典型文物数据具有构词的特殊性,比如"四子折桂"表达了石刻的寓意,使用现有的命名实体识别算法很难将"四子折桂"识别为相应的寓意。为了解决该问题,本文提出了一种基于词汇增强的典型文物命名实体识别算法,算法在输入表示层和上下文编码层都引入词汇信息,提高了词语领域专业性。算法通过构建文物领域词库,将其作为基于词汇增强的典型文物命名实体识别算法词典,最终较好地解决了词边界判断错误问题,在典型文物数据集上取得了较好的效果。

2 相关工作

命名实体识别是从句子中提取特定的实体并将其分为对应的类别,比如人名、地名、组织名等,是知识图谱构建的关键步骤,影响之后的关系抽取和知识图谱构建。基于深度学习的命名实体识别方法占据着支配性作用,深度学习采用多层次的处理结构,每一层都会从前一层中抽取部分特征信息,并抽象化表示出更高层次的特征,从而增强数据的表征能力。

基于深度学习的命名实体识别模型主要用到了三类输入表示:单词级别的输入表示、字符级别的输入表示以及混合表示。对于单词级别的输入表示,经过训练,每个单词可以用一个低维度的实值向量表示,Zheng等人[1]采用Word2Vec模型,对于字符级别的输入表示,可以更有效地利用词级别的信息,能够很好的处理词汇溢出(Out-of-vocabulary,OOV)问题,可以对没有见过的单词进行表示,并在语素层面上共享、处理信息。Peters等人提出了ELMo[2]表示,利用深度双向语言模型对大规模语料进行预训练,经原始任务数据集微调,产生适用于命名实体识别等任务的词向量表示。Kuru等人[3]提出了CharNER,将句子视为字符序列,并利用LSTM提取字符级别的表示。除上述两种输入表示,一些研究将附加信息纳入到单词的最终表示中,然后再输入上下文编码层,附加信息包括地名录[4]、词汇相似性[5]、语言依赖性[6]和视觉特征[7]。Devlin等人[8]提出了预训练语言模型BERT,通过无监督的预训练方式学习文本中的双向上下文信息,从而能够更好地理解单词和文本之间的关系。

中文命名实体识别方法通常先使用中文分词工具进行分词,再进行词级别的序列标注,中文分词工具不可避免地会错误地分割句子。一些方法[9,10]使用基于BERT的方法进行命名实体识别,借助预训练语言模型BERT提取通用的包含上下文的文本信息,但是BERT在垂直领域的表现一般,特别是在文物类的文本中表现不佳,BERT提取的信息更加全局,而命名实体识别任务更需要局部信息,因此依然会有词边界判断错误的问题。Zhang和Yang[11]提出了Lattice LSTM,Ma等人[12]提出了SoftLexicon,在基于深度学习的命名实体识别方法的基础上,引入词汇信息,较好地解决了词边界识别错误的问题。SoftLexicon在输入表示层引入词汇信息,Lattice LSTM修改了原有LSTM的结构,在上下文编码层引入了词汇信息。

本文提出了一种结合SoftLexicon与Lattice LSTM的基于词汇增强的典型文物命名实体识别算法,在输入表示层采用SoftLexicon特征进行编码,在上下文编码层采用Lattice LSTM获取上下文语义信息,在输入表示层跟上下文编码层都引入词汇信息,并且构建了文物领域词库,将其作为词典引入基于词汇增强的典型文物命名实体识别算法,较好地解决了词边界判断错误的问题。

3 领域词库构建与典型文物数据集制作

3.1 典型文物数据集制作

典型文物数据集选取了石刻、陶瓷、青铜器三类典型文物,主要数据来源于各地博物馆的官网(比如故宫博物院、山东博物馆),从博物馆官网上爬取到文物的名称、对应图片、对应的文字描述以及来源,具体如表1所示。

在命名实体识别数据集构建中,主要对非结构化的文字描述进行标注。总计收集3128条数据,经过清洗之后的有效数据为3000条,将其划分为训练集2400条、验证集300条以及测试集300条。根据文博专家的指导意见,制定了如表2所示的实体类型。

序列标注的主要方法有BIO、BIOES以及BMES。BMES常用于分词标注,BIO标注缺少显式的单词结尾信息,在Lattice LSTM跟SoftLexicon模型中,需要用到单词结尾的信息,因此在数据集的标注阶段采用BIOES标注法。

表1 典型文物数据集示例

图像	文物名称	文字描述	来源
	郭季妃门扉石刻	郭季妃门扉石刻，东汉，高124厘米，宽97厘米。此石扉为1920年左右出土于山西离石，经北京大学收藏，20世纪50年代拨交故宫博物院。	故宫博物院
	磁山文化红陶盂	红陶盂，新石器时代磁山文化，高15.3厘米，口径15.3厘米，底径11.3厘米。盂是一种盛汤浆或饭食的器皿，红陶夹砂能够提高陶器的耐热程度，因此这件红陶盂可能是一件炊器。	故宫博物院
	祖辛方鼎	祖辛方鼎，商代，高23、腹纵15.6、横13.9、足高8.4厘米，山东长清小屯遗址出土。方鼎四隅饰有扉棱，增加了整器的灵动之气，腹部饰兽面纹，以云雷纹为地，足饰阴线蝉纹，铸工精美。	山东博物馆

表2 典型文物数据集的8种实体类型

实体类型	实体描述	示例
文物名称	文本中出现的文物名称	郭季妃门扉石刻
朝代	文物所属的朝代	南宋、清朝
出土地点	文物出土地点	石桥镇新屋嘴村
博物馆	文物所在的博物馆	故宫博物馆、四川泸县宋代石刻博物馆
颜色	文物的颜色	海棠红、胭脂红
纹样	文物包含的纹样	龙纹、云纹
寓意	文物的文化内涵	安全保障、东方之神
其他	以上7类实体中不包含的实体	《汉书·游侠传》、琵琶

3.2 领域词库构建

典型文物数据集中的文本有很多文物领域的专有名词和领域词汇，比如："四子折桂"、"北方七宿"、"磁山文化"等。使用常见的中文分词工具对文物语料进行分词，往往无法进行准确地切分，影响语义信息的提取。文物领域词库的丰富性和准确性影响着命名实体识别以及之后的关系抽取，因此，非常有必要制作文物领域的领域词库。

本文主要研究的是文物领域的知识图谱构建，因此主要关注与文物名称、朝代、出土地点、博物馆、纹样、寓意等有关的细分领域词库。通过收集输入法词库、百科类词库、以及一些细分领域的词库，再加入人工筛选，以及领域专家提供部分种子词语，得到种子领域词库。在构建种子领域词库的过程中，主要参考了THUOCL词库、搜狗输入法词库以及DomainWords-Dict词库中一些细分领域词库，具体如表3所示。

表3 构建种子词库所需的领域词库

领域词库名称	细分领域
THUOCL词库	成语、地名
搜狗输入法词库	城市信息、考古
DomainWordsDict词库	地点名称、考古挖掘

本文利用词向量技术扩充领域词库，采用腾讯AI Lab提供的包含800万词汇的中文词向量，对种子领域词库中的纹样、朝代、寓意等词语，计算语义相似的前10个词，具体示例如表4所示。以"龙纹"为例，可以通过词向量技术获得相似词"风纹"、"云纹"以及

“龙风纹”,但是也会出现一些噪声词,比如“纹饰”、“夔龙”,所以还需要进行人工筛选。

表4 词向量相似词扩展示例

原词	相似词
龙纹	风纹、云纹、龙凤纹、纹饰、夔龙纹、云龙纹、螭龙、螭龙纹、螭纹、夔龙
清朝	在清朝、满清、清王朝、清朝后期、明朝、清朝时期、大清王朝、清朝末期、大清朝、满清王朝
富贵平安	富贵吉祥、平安富贵、吉祥如意、吉祥富贵、富贵长寿、富贵满堂、吉祥平安、平安如意 五福临门、福禄

借助已有领域词库构建种子词库以及通过词向量技术对种子词库进行扩充,最终得到15000个文物领域的词语,部分例子如表5所示。

表5 文物领域词库示例

类型	领域词汇
文物名称	妇好鸮尊、云纹铜禁、四羊方尊
朝代	清朝、明朝、南宋
出土地点	河南省安阳市、四川省泸州市
博物馆	中国国家博物馆、故宫博物院
纹样	龙纹、风纹、卷草纹
寓意	多子多福、吉祥如意

4 算法框架

如图1所示,基于词汇增强的典型文物命名实体识别算法可以分为输入表示层、上下文编码层以及标签解码层。输入表示层采用SoftLexicon,上下文编码层采用Lattice LSTM,标签解码层采用CRF,输入表

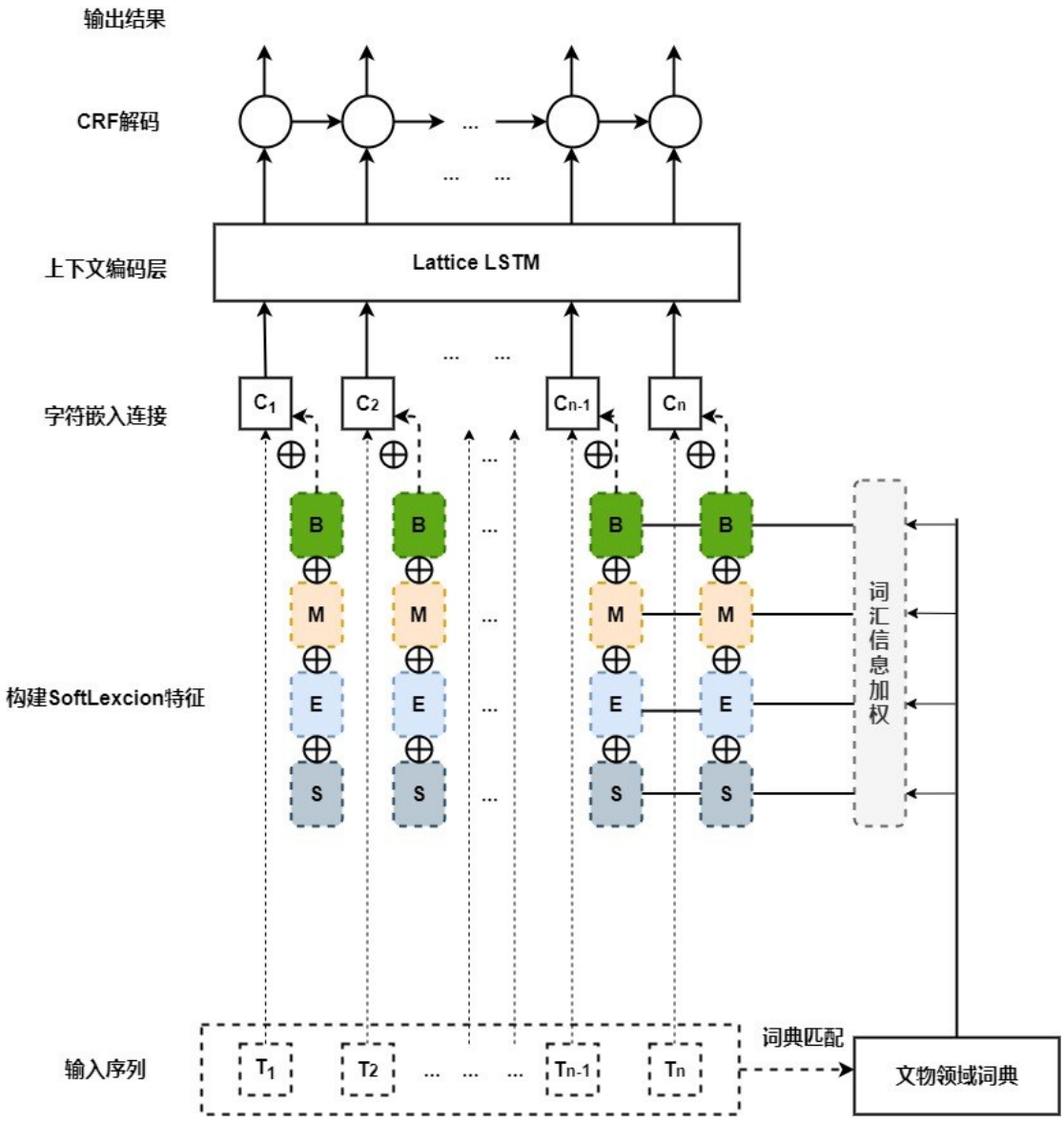

图1 基于词汇增强的典型文物命名实体识别算法框架

示层跟上下文编码层都引入了词汇信息，以增强命名实体识别模型鉴别词边界的能力。

5 实验结果

本文分别对比了BERT+BiLSTM+CRF模型、BERT+CRF模型、Lattice LSTM模型、SoftLexicon模型以及SoftLexicon+Lattice LSTM+ CRF模型（本文方法）。实验结果如表6所示。

表6 实验结果

算法	精确率	召回率	F1值
BERT+BiLSTM+CRF	0.824	0.791	0.807
BERT+CRF	0.822	0.789	0.805
Lattice LSTM	0.835	0.811	0.823
Lattice LSTM*	0.857	0.836	0.846
SoftLexicon	0.844	0.822	0.833
SoftLexicon*	0.859	0.843	0.851
本文方法	0.861	0.832	0.846
本文方法*	0.878	0.852	0.865

对于Lattice LSTM方法、SoftLexicon方法以及本文算法，本文使用两种词典分别进行实验，无‘*’标记符表示使用Lattice LSTM提出的词库，‘*’标记符表示使用本文制作的文物领域词库。实验结果表明，BERT+CRF模型与BERT+BiLSTM+CRF的效果差别不大，这是由于BERT强大的上下文编码能力可以提取出需要的信息，BiLSTM只是在BERT的基础上选择有效的信息进行处理。引入词汇信息的方法有明显的提升。输入表示层SoftLexicon和上下文编码层Lattice LSTM都引入词汇信息优于分别在两层单独引入词汇信息的效果。

6 结论

为了解决文物领域数据构词特殊性导致实体边界识别错误的问题，本文构建了文物领域词库，并提出了一种基于词汇增强的典型文物命名实体识别算法。首先，在输入表示层采用SoftLexicon，引入词汇信息；其次，在上下文编码层采用Lattice LSTM，在输入表示层的基础上再次引入词汇信息；最后，在标签解码层采用CRF解码，获取最终的标签。实验结果表明，使用本文构建的文物领域词库，基于词汇增强的命名实体识别方法在典型文物数据集上有较好的表现。

本文在构建典型文物数据集时，主要数据来源于相关博物馆官网的图文对数据，只对文本数据进行了算法处理。目前并没有高精度的文物多模态命名实体识别数据集，未来会考虑在领域专家的指导下对图文数据进行多模态标注，融合图像文本信息，进行命名实体识别。

参考文献（References）：

[1] Zheng S，Wang F，Bao H，et al. Joint extraction of entities and relations based on a novel tagging scheme[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics（Volume 1：Long Papers）. 2017：1227-1236.

[2] Peters M E，Neumann M，Iyyer M，et al. Deep contextualized word representations[C]//Proceedings of NAACL-HLT. 2018：2227-2237.

[3] Kuru O，Can O A，Yuret D. Charner：Character-level named entity recognition[C]//Proceedings of COLING 2016，the 26th International Conference on Computational Linguistics：Technical Papers. 2016：911-921.

[4] Liu T，Yao J G，Lin C Y. Towards improving neural named entity recognition with gazetteers[C]//Proceedings of the 57th annual meeting of the association for computational linguistics. 2019：5301-5307.

[5] Ghaddar A，Langlais P. Robust lexical features for improved neural network named-entity recognition[J]. arXiv preprint arXiv：1806.03489，2018.

[6] Jie Z，Lu W. Dependency-guided LSTM-CRF for named entity recognition[J]. arXiv preprint arXiv：1909.10148，2019.

[7] Lu D，Neves L，Carvalho V，et al. Visual attention model for name tagging in multimodal social media[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics（Volume 1：Long Papers）. 2018：1990-1999.

[8] Devlin J，Chang M W，Lee K，et al. BERT：Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//Proceedings of NAACL-HLT. 2019：4171-4186.

[9] 谢腾，杨俊安，刘辉. 基于BERT-BiLSTM-CRF模型的中文实体识别[J]. 计算机系统应用，2020，29(07)：48-55. Y

[10] Hu S，Zhang H，Hu X，et al. Chinese named entity recognition based on BERT-CRF model[C]//2022 IEEE/ACIS 22nd International Conference on Computer and Information Science（ICIS）. IEEE，2022：105-108.

[11] 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）. 2018：1554-1564.

[12] Ma R，Peng M，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. 2020：5951-5960.

编辑：龙学锋