融合词典信息的古籍命名实体识别研究

康文军　　左家莉*　　揭安全　　罗文兵　　王明文
江西师范大学　计算机信息工程学院　江西　南昌　330022
Email: {kwj, zjl, lwb, mwwang }@jxnu.edu.cn, jieanquan@163.com

摘要

古籍命名实体识别对于古籍实体知识库与语料库的建设具有显著的现实意义。目前古籍命名实体识别的研究较少，主要原因是缺乏足够的训练语料。本文从《资治通鉴》入手，人工构建了一份古籍命名实体识别数据集，以此展开对古籍命名实体识别任务的研究。针对古籍文本多以单字表意且存在大量省略的语言特点，本文采用预训练词向量作为词典信息，充分利用其中蕴涵的词汇信息。实验表明，这种方法可以有效处理古籍文本中人名实体识别的问题。

关键词： 古籍命名实体识别；词典信息；《资治通鉴》实体数据集构建

A Study on the Recognition of Named Entities of Ancient Books Using Lexical Information

Wenjun Kang　Jiali Zuo*　Anquan Jie　Wenbing Luo　Mingwen Wang
School of Computer and Information Engineering, Jiangxi Normal University,
Nanchang, Jiangxi 330022,China
Email: {kwj, zjl, lwb, mwwang }@jxnu.edu.cn, jieanquan@163.com

Abstract

Named entity recognition of ancient texts is of significant practical importance for the construction of a knowledge base and corpus of ancient entities. Currently, there are few studies on named entity recognition of ancient texts, mainly due to the lack of sufficient training corpus. In this paper, a dataset for the recognition of named entity in ancient books is manually constructed, starting from the History as a Mirror, as a way of starting to study the task of named entity recognition in ancient books. The paper uses pre-trained word vectors as lexical information in order to make full use of the lexical information contained in the text, which is characterised by a large number of omissions and single word meanings. Experiments show that this approach can effectively deal with the problem of recognising named entities in ancient texts.

Keywords: named entity recognition of ancient books , lexical information , construction of the History as a Mirror entities dataset

©2023 中国计算语言学大会
根据《Creative Commons Attribution 4.0 International License》许可出版
基金项目： 国家自然科学基金(61866018,62266023); 江西省教育厅研究生创新基金项目(YC2022-s329)
通讯作者： 左家莉

1 引言

1991年，Rau等人 (1991)首次提出了命名实体识别(named entity recognition，NER)任务，随即在1995年11月的第六届MUC会议上，命名实体识别任务正式被确认作为一个明确的概念和重要的研究领域 (刘浏和王东波, 2018)。作为信息抽取的一项子任务，命名实体识别主要是从一些非结构化文本中抽取出具有特定意义的命名实体并且准确归纳到预先定义好的类别，它在关系抽取、机器翻译、知识库构建等自然语言处理下游任务中发挥着重要作用。

先前的NER任务大多集中在英文数据集上，从SIGHAN Bakeoff-2006评测会议开始，中文NER逐渐走进人们的视野。与英文NER相比，由于汉字之间没有明显的分词边界，中文NER需要预先进行中文分词，但是这会造成分词错误传播。因此，之前的中文NER研究大部分基于字级别的方式 (Liu et al., 2010; Gui et al., 2019; Sui et al., 2019)，但是字级别的NER模型无法充分利用词汇以及词汇边界的信息。于是，针对这个问题，早期的一些研究主要通过对LSTM (Hochreiter and Schmidhuber, 1997)、CNN (LeCun and Bengio, 1995)、Transformer (Vaswani et al., 2017)等一些基础神经网络模型进行改进以此来引入外部词典信息，它有效地提升了模型在一些中文基准数据集上的效果 (Zhang and Yang, 2018; Ma et al., 2020; Li et al., 2020; Wu et al., 2022)。自预训练语言模型BERT (Devlin et al., 2019)被提出以来，一些研究开始利用预训练语言模型能够捕捉输入文本间隐式的语义和语法知识的能力，将外部词典信息结合到预训练语言模型中来提升NER模型的性能 (Chang et al., 2021; Nie et al., 2020; Liu et al., 2021)。

近年来，中文NER的研究主要集中在一些有限的领域和实体类型上，但很少对古文等特定领域文本中命名实体进行研究 (陈曙东和欧阳小叶, 2020)。这是由于古籍文本所使用的语言和现代文不同，其文字晦涩难懂且没有一套标准的标注规范，对研究人员的专业能力有一定的要求，而且已经标注并且公布出来的高质量古籍命名实体识别语料很少，因此很多需要靠人工对古籍语料进行预处理和标注。此外，基于深度学习的中文NER模型已经在现代文的数据集上取得了不错的成绩，而专门对古籍命名实体识别模型的研究较少。考虑到引入词典信息可以为模型提供词汇级别的信息，再结合古文多以单字表意且存在大量缩写的语言特殊性，我们认为融合词典信息的模型在古籍NER任务上也是可行的。基于此，我们展开了古籍命名实体识别任务的研究:

（1）本文根据施丁等人 (1994)编纂的《资治通鉴大辞典》，结合开源的《资治通鉴》的现代文翻译版本，研究了古籍命名实体识别的标注原则，并构建了《资治通鉴》古籍命名实体识别数据集，之后还从实体覆盖率和峰度系数两方面对标注数据进行了分析。

（2）本文使用了融合词典信息的不同命名实体识别模型，来对古籍命名实体识别任务进行研究，并对比了四种不同的预训练语言模型作为特征在各类别实体上的表现，最后用测试集中未登录词(out of vocabulary，OOV)的召回率反映模型在预测古籍文本中未登录词上的性能。

2 相关工作

以往的NER研究大多采用基于规则的方式和基于统计机器学习的方式，但是它们要求研究人员制定合适的特征工程以及具备相关的领域知识 (Zhou and Su, 2002; Takeuchi and Collier, 2002)。随着深度学习的发展，基于深度神经网络的NER模型几乎避免了那些方法的局限性，并在多个基准NER数据集上都取得了显著的性能提升 (Collobert et al., 2011; Lample et al., 2016)。Hammerton (2003)首次尝试使用神经网络模型来研究NER领域的任务。Huang等人 (2015)将双向LSTM-CRF模型用在中文命名实体识别等序列标注任务中。Peng等人 (2016)提出了一种结合中文分词模型的多任务联合训练学习的方式，它帮助模型更好地识别出了在中文社交媒体文本中的实体。尽管基于神经网络模型的方式直接有效地提升了NER模型的性能，但是融入外部词典等更多有用的额外知识仍能够在NER任务上取得较大增益。

本文对古籍NER任务的研究路线，与现有主流的利用外部词典信息来提升字级别NER模型性能的思想一致。Zhang and Yang (2018)首次提出了一种Lattice-LSTM网络结构，它能动态地将句子内所有与词典相匹配的词引入到字级别的中文NER模型中，但是它无法充分利用GPU的并行能力。Ma等人 (2020)通过构建一个SoftLexicon特征，将词典信息和边界信息结合到模型的输入表示层，它有效地避免了设计复杂的结构来融入词典信息，提升了模型的推理速度。Li等人 (2020)提出了一种FLAT-Lattice网络结构，它重新为每个字以及由词典

(c) 2023 中国中文信息学会计算语言学专业委员会

匹配到的词设计两个头尾位置索引，以此将词汇信息引入到Transformer结构，并采用相对位置编码 (Yan et al., 2019)去捕获每个字词之间的距离和方向信息。这种网络结构能够充分利用Transformer模型构建文本间长距离依赖的优势以及具有优越的并行能力。Wu等人 (2022)提出了一个新的InterFormer模块，它可以同时对原始字输入和词汇这两个不同的序列进行建模，从而获得融合了词边界和语义信息的字表征，再利用改进的Transformer模块对获得的字表示进行编码，这种方式有效地降低了FLAT模型的计算消耗，以及可以利用更大的词典进行匹配。

随着大规模预训练语言模型在多项自然语言处理基础任务上刷新了之前最好的结果，NER的研究重心也逐渐转向预训练语言模型。Chang等人 (2021)通过实验表明利用BERT去提取文本的特征向量要优于使用静态词向量的方式。Souza等人 (2019)简单地在BERT顶部添加了一个线性条件随机场层，并经过微调，就在其所用到的NER数据集上获得了出色的性能表现。Beltagy等人 (2019)在科技领域标注数据匮乏的情况下，向大规模预训练语言模型BERT中加入未标注的科技领域数据集进行继续预训练，之后在各种使用科技领域数据集的一些下游任务上取得不错的效果。

一些研究者开始考虑综合上述两种方式的优势，将外部词典信息融入到大规模的预训练语言模型中。Nie等人 (2020)提出了一个语义扩充模块对词典信息进行编码，以及对由词典匹配到的每个词赋予不同的权重，然后利用一个门控模块对BERT作为编码端时输出的原始字信息和扩充语义信息进行控制，它有效地缓解了社交媒体文本中数据稀疏的问题。Diao等人 (2020)通过在预训练语言模型BERT外部额外地添加了一个用来处理N-Gram词典的编码器，以此来显式地融入词语层级的信息。Tian等人 (2020)提出一种键值记忆神经网络的方式，将N-Gram词典的信息以及每个字在匹配到的N-Gram内的位置信息，融入到BERT编码的字信息中，它能够有效地利用特定领域未标注文本的信息来提升模型对未登录词的识别。Liu等人 (2021)将词典信息融入到BERT底层，利用BERT的语言表征能力学习到词典更深层的知识，最后该模型在命名实体识别等序列标注任务中呈现出卓越的性能。

现有古籍NER的研究工作也逐渐转向深度神经网络模型并且结合各种特征进行学习，以此来提升模型在古籍文本实体识别上的效果。徐晨飞等人 (2020)采用BiLSTM-CRF和BERT等四种基础的神经网络模型探究在《方志物产》云南卷语料库上四种实体识别的效果。包振山等人 (2022)提出了一种半监督学习的方法，并结合古籍语言学的特点以及词性等特征，在自建的中医古籍语料上达到了83.28%的效果。张朦等人 (2023)利用双向LSTM模型抽取出《花间集全译》语料的部首、声韵和格律等多个特征并和字向量特征融合，其模型效果达到了85.63%。受Gururangan等人 (2020)提出的基于领域自适应训练思想的启发，王东波等人 (2021)使用精校后的《四库全书》全文作为训练集，在预训练语言模型BERT和RoBERTa框架的基础上使用掩码语言模型任务进行预训练，最终获得了面向古文领域的SikuBERT和SikuRoBERTa预训练语言模型。Wang and Ren (2022)在包含道部、佛部等数十部古籍在内的大规模语料库上基于BERT预训练语言模型进行学习得到了词表更大的古文预训练模型bert-ancient-chinese。

3 古籍命名实体识别数据集

3.1 古籍数据集的选取

《资治通鉴》作为我国第一部编年体的通史，在史学和文学领域有着相当高的研究价值。它由北宋史学家司马光历时19年完成，涵盖了从周朝到后周16朝1362年的历史。随着古籍数字化研究工作日趋成熟，一些平台陆续公布了高质量《资治通鉴》文本，这对后续的古籍研究起到了推动作用。本文选取古诗文网[0]中收录的数字化《资治通鉴》为语料来源，经过分句和人工审校文白对照翻译后，在此语料上开展和完成专有名词的标注工作。

3.2 《资治通鉴》数据集标注

3.2.1 实体标注原则

目前，开源且受到广泛研究的中文命名实体识别基准数据集主要有Ontonotes 4.0(Weischedel et al., 2011)、MSRA(Levow, 2006)、Weibo(Peng and Dredze, 2016)、Resume (Zhang and Yang, 2018)以及Cluener(Xu et al., 2020)等，然而古籍语料从词汇到语法以及实体的构造规律都与现代文语料有所不同，现代文本的标注方法很难直接用于古籍实体标注。于是本文通过参考Ji等人 (2021)构建的“二十四史”命名实体识别数据集以及刘浏 (2018)针对古籍中

[0]https://www.gushiwen.cn/

(c) 2023 中国中文信息学会计算语言学专业委员会

人名、地名、时间三类实体类别和成分划分的探究，并结合《资治通鉴》中待标注的实体为例，制定了一套较为规范的五种类别实体的标注原则。

1、人名(PER)

人名实体是古籍文献中最重要且被研究最多的命名实体之一，它主要指古籍中出现人物的姓名以及一些可以指代人物的词语。与现代人名相比，古代人名种类相对要复杂得多，主要体现在古人除了有姓和名之外，还有字、谥号、爵位、排行、职官、尊称、庙号等，如表1所示。

构成成分	示例
姓+名	**王守澄**恶官者**田全操**
名	独**充国**留屯田
谥号	孝武皇帝、孝惠皇帝
尊称、爵位	沛公、郑伯、恒侯
字	昔鲍叔之于管仲，**子皮**之于**子产**
职官	庶人勇既废，**秦王**已薨

Table 1: 人名主要构成成分和示例

2、地名(LOC)

地名实体是另一种被研究较多的实体，主要分为地名、山川河流名、关隘名等。它多以单字或双字的形式出现，相对人名实体，实体数量较少且不存在大量缩写。此外，它在古籍中出现的位置具有一定的规律，例如一般加在人名前面便于区分标识和避免重名，出现在“于”、“至”、“居”、“迁”、“屯”、“破”和“攻”等一些指示词后面以及伴随在一些“东南西北”方位词附近等，而山水名和关隘名常和“氵”、“阝”等汉字部首相关，如表2所示。

构成成分	示例
人名前面	**新丰**王孝杰从刘审礼击吐蕃
“攻”等指示词	别将陈贞等攻**武陵**
汉字部首	庾亮还**芜湖**、**阳关**三百馀里
方位词	南破**零**、**桂**，东掠**武昌**

Table 2: 地点名主要构成成分和示例

3、官职名(JOB)

官职名实体是指在国家管理和行政工作中承担不同职位、具有不同职权范围和地位等级的一类人群的统称,大体上可以分为中央官职和地方官职两大类，对古籍文本中官职名实体的识别有助于研究当时的官职制度。它同地名实体一样，在古籍中出现的位置具有规律，例如出现在“拜”、“除”、“擢”、“出”、“封”和“迁”等一些表明官职任免升降的词语后面以及出现在人名实体的前面表明其身份等。

4、组织名(ORG)

《资治通鉴》记载了从周威烈王到后周世宗等16朝的历史，因此包含了很多像国家、诸侯国、少数民族部落等政治方面的地点。于是本文对地名实体进行了细分，将它们划分到组织实体。此外，它还包括家族名和官署名等。

5、时间名(TIME)

《资治通鉴》按照时间先后顺序记叙史事，识别出其中的时间实体信息，有助于研究人员梳理历史人物事件发展的脉络。时间实体主要分为月份、季节、年份三种成分，其中季节和月份同现代一样，分为四个季节和十二个月份，年份主要有太岁纪年、干支纪年、采用天子谥号或者尊号的方式纪年以及汉武帝之后的年号纪年。

对于同一个实体名称可以指代不同类别造成的歧义问题，本文结合上下文语境对实体类型进行判断和标注。例如“与刁协帝尽诛王氏”和“以昭仪王氏为德妃”两句中“王氏”在不同语境中分别作为人名和组织名。又如“韩王成又无功”和“臣为韩王送沛公”两句中的“韩王”是一个爵位，都指代的是韩成，但在第一句中“韩王”更倾向于官职名，后一句更倾向于人名。

(c) 2023 中国中文信息学会计算语言学专业委员会

3.3 数据集分析

Liang等人 (2021)从实体覆盖率和峰度系数两个角度对中文NER基准数据集进行分析，发现它们中存在着两种可能会影响模型泛化性能的数据偏差，于是本文也将从这两个方面对标注的数据集进行分析。

首先，分别计算验证集和测试集中在训练集出现过的实体比例，比例越高表明会影响模型在预测未见实体上的性能。如表3所示，《资治通鉴》数据集的验证集和测试集中各有43.6%和42.5%的实体在训练集出现过。与Liang等人 (2021)在中文NER基准数据集上计算的实体覆盖率相比，本文划分的数据集实体覆盖率不高。

数据集	验证集实体覆盖率	测试集实体覆盖率
OntoNotes 4.0	50.5%	51.4%
MSRA	55.4%	70.9%
Weibo	49.8%	42.9%
Resume	54.0%	54.4%
Cluener	61.5%	-
资治通鉴	43.6%	42.5%

Table 3: 在验证集和测试集中的实体覆盖率

其次，本文使用峰度系数 (Balanda and MacGillivray, 1998)去度量标注数据集中的fat-head实体，即出现频率较高的实体，高峰度系数意味着它比正态分布具有更多的异常数据，低峰度系数意味着它具有较少的异常值。

$$Kurtosis = \frac{1}{n}\sum_{i=1}^{n}\left[\left(\frac{X_i - \mu}{\sigma}\right)^4\right] \tag{1}$$

其中，X_i是指每个实体类别中各实体出现频次组成的数组，μ是指均值，σ是指标准差。

如表4所示，官职、人名和组织实体间存在着较多的fat-head实体，这是由于官职实体和组织实体的类型相对有限，而人名实体则是由于一些主要人物出现的频次较高。因此，未来我们将尝试增加数据集或使用实体替换算法来缓解这一问题。

实体类别	训练集	验证集	测试集
PER	113.6	21.0	91.8
LOC	58.4	15.2	8.5
JOB	183.9	42.2	59.3
ORG	87.3	12.9	15.0
TIME	14.0	7.6	4.5

Table 4: 不同实体类别的峰度系数

4 模型

本文主要研究古籍中扁平化命名实体 (Yan et al., 2021)的识别，在神经网络模型中，它通常作为一个序列标注任务来处理，也就是对序列中的每一个字分配一个标签。基于深度学习方式的NER模型的一般架构分为三部分：嵌入层、编码层和解码层。首先给定一个输入文本序列$T = (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$，其中$x_i$，i=1，2，...，n对应字级别古籍文本，$y_i$，i=1，2，...，n对应输出序列标签。在嵌入层获得每个原始字的特征向量后，再输入到编码端对上下文信息进行建模。

$$u_1, ..., u_n = Embedding(x_1, ..., x_n) \tag{2}$$

$$h_1, ..., h_n = Encoder(u_1, ..., u_n) \tag{3}$$

(c) 2023 中国中文信息学会计算语言学专业委员会

其中，Embedding()既可以使用静态的词向量，也可以使用基于上下文的词向量，例如BERT等。Encoder()可以是任何一种基础的神经网络模型架构。

在解码端，为了考虑连续标签之间的依赖性，通常采用线性条件随机场(CRF)来对标签序列进行约束和预测。首先将编码层最后的隐藏层状态向量通过一个线性变换层再作为CRF层的输入，然后计算输出标签序列的概率，如式(4)所示。在训练过程中，通过最小化负对数极大似然函数来不断更新模型的参数进行学习，如式(5)所示。在解码阶段，使用维特比算法找到得分最高的标签序列。

$$p(y \mid s) = \frac{\exp\left(\sum_i \left(O_{i,y_i} + T_{y_{i-1},y_i}\right)\right)}{\sum_{\tilde{y}} \exp\left(\sum_i \left(O_{i,\tilde{y}_i} + T_{\tilde{y}_{i-1},\tilde{y}_i}\right)\right)} \tag{4}$$

$$L = -\sum_j \log(p(y \mid s)) \tag{5}$$

其中$O_{i,\tilde{y}_i}$是经过一层线性层后的分数，T是转移分数矩阵，$\tilde{y}$表示所有可能的标签序列。

融合词典信息的NER模型也是基于序列标注的框架，一些研究通过改进不同的神经网络模型架构，将外部词典匹配得到的词汇信息在嵌入层或者编码层和输入文本信息结合，最后在线性CRF层进行解码。如图1所示，LEBERT模型通过词典适配器模块，将词典匹配到的词汇信息整合到预训练语言模型BERT的不同Transformer层中,来充分学习到词汇更深层次的特征。

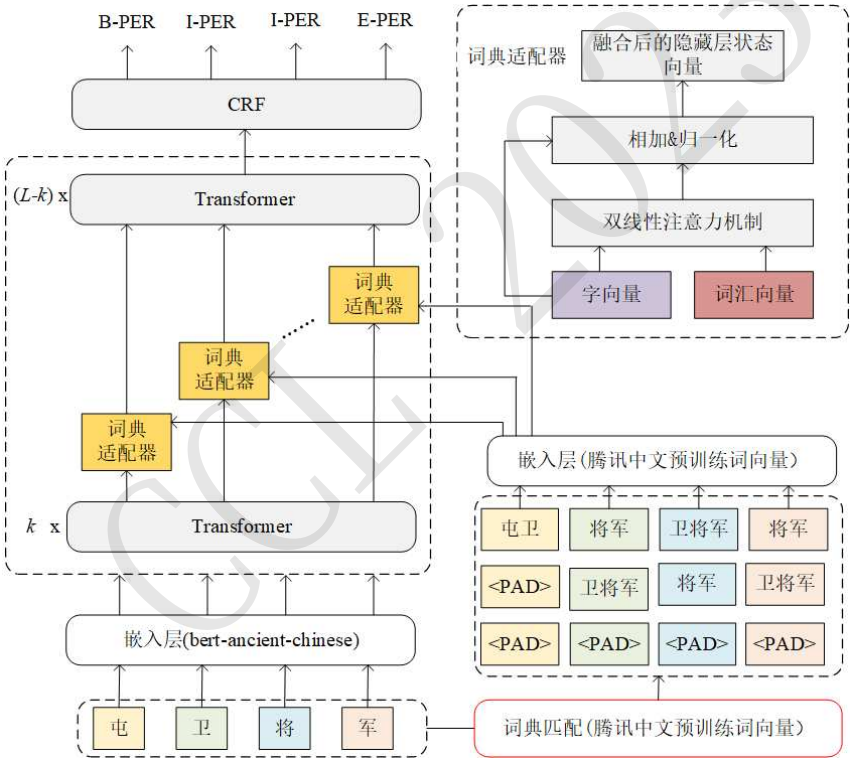

Figure 1: LEBERT(bert-ancient-chinese)模型整体架构

5 实验

5.1 数据集实体统计

本文借助Label Studio[1]进行实体的标注，并参照Resume(Zhang and Yang, 2018)等中文NER基准数据集的格式，将标注好的数据采用BIOES标注模式进行转换。然后，我们将构建的训练语料按照8:1:1的比例划分为训练集、验证集和测试集，再统计划分后实体的数量，

[1]https://labelstud.io/

(c) 2023 中国中文信息学会计算语言学专业委员会

如表5所示。在实验中，我们选取了本文构建的《资治通鉴》NER数据集和Ji等人(2021)公布出来的“二十四史”NER数据集作为研究对象，其中“二十四史”实体数量如表6所示。

实体类别	训练集	验证集	测试集
人名	10915	1271	1377
地名	3250	437	423
官职名	4802	429	403
组织名	1421	296	276
时间名	1743	74	96
总计	22131	2507	2575

Table 5: 《资治通鉴》命名实体数量详情

实体类别	训练集	验证集	测试集
人名	11532	756	859
地名	3625	220	236
官职名	2252	448	349
组织名	2041	4	45
总计	19450	1428	1489

Table 6: “二十四史”实体数量详情

5.2 实验参数设置与评估指标

本文在实验中采用了腾讯中文预训练词向量 (Song et al., 2018)作为外部词典信息，它提供了超过1200万个中文词汇和短语的200维词向量表示，这些词向量表示都是基于大规模语料库进行预先训练得到的。此外，输入文本的最大长度设置为200，batch size大小设置为4，epoch大小设置为50，学习率设置为1×10^{-5}。在评估指标的选择上，本文使用精确率(Precision)、召回率(Recall)以及F1值(F1-Score)来综合考虑这五个实体类别在各个模型上的效果，最后使用未登录实体词的召回率来计算模型在处理未见过实体词上的表现，如式(7)所示。

$$\mathrm{F1-Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \times 100\% \tag{6}$$

$$\mathrm{R}_{OOV} = \frac{\text{模型正确识别的未登录词}}{\text{测试集中未登录词总数}} \times 100\% \tag{7}$$

5.3 实验与分析

5.3.1 在“二十四史”NER数据集上的考察

本文选取了BERT-CRF模型以及融入词典信息的LEBERT模型 (Liu et al., 2021)，来对Ji等人(2021)公开的“二十四史”NER数据集进行考察。此外，我们选取了在中文语料上进行预训练的bert-base-chinese (Devlin et al., 2019)以及在古文语料上进行预训练的sikubert (王东波等人, 2021)分别作为模型的输入特征。

实验结果如表7所示，在使用bert-base-chinese作为输入特征时，LEBERT模型的F1值比BERT-CRF模型提升了1.32%。当换成sikubert预训练语言模型后，两个模型的F1值都进一步提高了，说明这两者的信息有助于对古籍命名实体识别任务的研究。然而，我们通过对“二十四史”NER语料中实体以及文本来源的分析，发现它数据来源于多部不同的史书并且实体标注稀疏，因此我们人工构建了一份《资治通鉴》NER数据集来研究古籍命名实体识别任务。

模型	PER	LOC	JOB	ORG	F1值
BERT-CRF(bert-base-chinese)	67.61%	55.39%	67.44%	48.00%	64.81%
BERT-CRF(sikubert)	67.46%	58.14%	68.22%	50.36%	65.48%
LEBERT(bert-base-chinese)	67.75%	59.65%	69.60%	49.37%	66.13%
LEBERT(sikubert)	70.20%	60.19%	70.26%	57.81%	68.22%

Table 7: 在“二十四史”NER数据集上的实验结果

5.3.2 在《资治通鉴》NER数据集上的研究

本文做了三组实验来对比目前加入词典信息的不同模型在古籍命名实体识别任务上的性能，第一组选取了不加入任何外部信息的基准模型，BiLSTM-CRF、BERT-

CRF和TENER (Yan et al., 2019)；第二组选取了传统的融入词典信息的模型，Lattice-LSTM (Zhang and Yang, 2018)、FLAT (Li et al., 2020)、NFLAT (Wu et al., 2022)和Lexicon-AugmentedNER (Ma et al., 2020)；第三组选取了将词典信息融入预训练语言模型中的模型，WMSEG (Tian et al., 2020)和LEBERT (Liu et al., 2021)。此外，本文使用谷歌发布的中文预训练语言模型bert-base-chinese (Devlin et al., 2019)抽取输入文本的特征。

实验结果如表8所示，相比于基准模型，加入词典信息有助于提升基础神经网络模型在古籍实体识别上的效果。在不使用预训练语言模型的情况下，LexiconAugmentedNER模型效果最优，F1值为73.21%。在使用预训练语言模型后，古籍实体识别的效果达到了77.26%，由于预训练模型BERT在处理中文文本时通常聚焦于字级别的输入，当加入词级别的信息后，模型在古籍文本实体识别上的效果进一步提升了，F1值高达81.24%。

模型	精确率	召回率	F1值
BiLSTM-CRF	68.82%	62.20%	65.34%
TENER	71.22%	65.36%	68.17%
BERT-CRF(bert-base-chinese)	77.59%	76.93%	77.26%
Lattice-LSTM	72.25%	65.51%	68.72%
FLAT	75.62%	68.65%	71.96%
NFLAT	75.24%	69.75%	72.39%
LexiconAugmentedNER	77.86%	69.09%	73.21%
WMSEG(bert-base-chinese)	79.97%	80.16%	80.06%
LEBERT(bert-base-chinese)	81.24%	81.24%	81.24%

Table 8: 在标注的《资治通鉴》NER数据集的实验结果

5.3.3 在不同预训练语言模型上古籍各类别实体识别的效果

为了研究使用不同的预训练语言模型能否进一步提升模型在识别古籍文本实体上的效果，我们选取了在《资治通鉴》NER数据集上F1值最高的LBERT模型作为基准模型，然后分别加入chinese-bert-wwm (Cui et al., 2021)、sikubert (王东波等人, 2021)和bert-ancient-chinese (Wang and Ren, 2022)这三个不同的预训练语言模型作为特征进行对比实验。

实验结果如表9所示，从模型的F1值来看，换成bert-ancient-chinese预训练语言模型后模型的性能最优，F1值上升了2.79%，其次是sikubert预训练语言模型F1值上升了1.02%，chinese-bert-wwm预训练语言模型F1值上升了0.46%，这表明特定领域上的预训练语言模型能够进一步帮助模型提高对古籍实体的识别，尤其是利用了更大规模的古籍语料文本进行预训练的语言模型。从各类别实体识别的效果来看，模型对人名实体和地点实体识别的F1值提升最多，分别上升了3.63%和4.48%，这说明相对于其它实体类型而言，它们由于实体构成多样且词汇用法较现代多有不同，需要来自更多特定领域的信息来帮助模型准确地识别出来。

模型	PER	LOC	JOB	ORG	TIME	F1值
LEBERT(bert-base-chinese)	85.18%	80.77%	79.72%	67.51%	74.85%	81.24%
LEBERT(chinese-bert-wwm)	86.22%	81.26%	77.73%	67.60%	74.44%	81.70%
LEBERT(sikubert)	86.57%	83.30%	79.06%	66.06%	77.27%	82.26%
LEBERT(bert-ancient-chinese)	88.81%	85.25%	79.64%	68.20%	77.53%	84.03%

Table 9: LEBERT模型在不同预训练语言模型上的实验结果

5.3.4 在预测未见过实体词上的效果

未登录词的识别是命名实体识别模型中的一个挑战，因此本文通过计算测试集中未登录词的召回率，来研究融入词典信息的NER模型在缓解古籍文本未登录词上的效果。从表10可以看出，加入词典信息和古文领域预训练语言模型都能够提升模型在识别OOV词上的性能。

(c) 2023 中国中文信息学会计算语言学专业委员会

模型	R_{OOV}
测试集未登录实体数量	1480
BERT-CRF(bert-base-chinese)	78.04%
WMSEG(bert-base-chinese)	82.84%
LEBERT(bert-base-chinese)	82.77%
LEBERT(sikubert)	84.86%
LEBERT(bert-ancient-chinese)	86.96%

Table 10: 在测试集未见过实体词上的召回率

6 案例分析

相对其它实体类型，人名实体构成灵活且常常被省略姓氏，因此模型在识别某些人名实体上难度较高，为此我们挑选出模型对人名实体识别的样例进行分析。如表11示例一所示，在某些情况下，LEBERT模型在识别一些被省略姓氏的人名实体上表现较好。

人名实体存在一定的歧义性，如表11示例二所示，“定兴”在不同的语境下可以是人名实体，也可以是地名实体。尽管我们结合了上下文语境对实体进行标注，但融入外部词典的模型仍无法正确识别出某些实体在当下语境的实体类型，这是由于“定兴”在词典中同时是人名或地名，它无法提供给模型对应的上下文信息。当把bert-ancient-chinese古文预训练语言模型作为LEBERT模型的输入特征后，它能有效地缓解这种歧义问题，这可能是它在大规模古文领域数据集上进行预训练时，学习到了这种上下文的信息。

示例一： 引入词典信息的NER模型实体抽取结果	
古籍文本片段	上欲遣淮南太守**戴僧静**将兵讨子响，**僧静**面启曰...
实际标注标签	B-PER I-PER E-PER B-PER E-PER
Lattice-LSTM	B-PER I-PER E-PER O S-PER
FLAT	B-PER I-PER E-PER O O
LEBERT(bert-base-chinese)	B-PER I-PER E-PER B-PER E-PER
古籍文本片段	**子如**曰：“**消难**亦通**子如**妾，此事正可掩覆。...”
实际标注标签	B-PER E-PER B-PER E-PER B-PER E-PER
Lattice-LSTM	O S-PER O O O O
FLAT	O S-PER O O O O
LEBERT(bert-base-chinese)	B-PER E-PER B-PER E-PER B-PER E-PER
示例二： 加入古文领域预训练语言模型实体抽取结果	
古籍文本片段	... 应募隶屯卫将军**云定兴**，说**定兴**多赍旗鼓为疑兵，...
实际标注标签	B-PER I-PER E-PER B-PER E-PER
LEBERT(bert-base-chinese)	O B-LOC E-LOC B-LOC E-LOC
LEBERT(sikubert)	O B-LOC E-LOC B-LOC E-LOC
LEBERT(bert-ancient-chinese)	B-PER I-PER E-PER B-PER E-PER
古籍文本片段	...今若杀**山阳**，与雍州举事，...则霸业成矣！**山阳**持疑不进...
实际标注标签	B-PER E-PER B-PER E-PER
LEBERT(bert-base-chinese)	B-LOC E-LOC B-LOC E-LOC
LEBERT(sikubert)	B-LOC E-LOC B-LOC E-LOC
LEBERT(bert-ancient-chinese)	B-PER E-PER B-PER E-PER

Table 11: 测试集中各模型预测示例

7 总结与未来工作

本文首先考察了“二十四史”NER数据集，发现它语句来源分散且标注稀疏，因此我们选取《资治通鉴》作为研究语料，并对数据集中五个类别的实体进行标注，人工构建了一份《资治

(c) 2023 中国中文信息学会计算语言学专业委员会

通鉴》命名实体识别数据集。此外，本文采用不同方式融入词典信息的模型，以及加入特定领域的预训练语言模型作为特征，来研究古籍命名实体识别任务。实验表明，加入词典信息能够帮助模型识别出被省略姓氏的人名实体，并提升模型在预测未见过实体上的性能。然而，融合词典信息的模型不具备消歧能力，无法解决一词多义的现象。

我们通过对标注数据集的分析，发现某些实体类别存在一些fat-head实体，会对模型的泛化能力造成影响，因此我们将增加语料的规模和语料的来源，并标注更多的实体类型。在古籍标注的过程中，我们观察到现代文翻译很好地对古文进行了补充，因此这些翻译中也包含了许多的信息。在未来，我们将尝试利用这些翻译的信息，来对古籍命名实体识别任务进行研究。

参考文献

K. Balanda and H. MacGillivray. 1988. Kurtosis: a critical review. *The American Statistician*, 42:111–119.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: a pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3615–3620.

Ronan Collobert, Jason Weston, Leon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research 12(Aug)*:2493–2537.

Yuan Chang, Lei Kong, Kejia Jia, Qinglei Meng. 2021. Chinese named entity recognition method based on bert. *2021 IEEE International Conference on Data Science and Computer Application*, pages 294-299.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. In *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 3504-3514.

Shizhe Diao, Jiaxin Bai, Yan Song, Tong Zhang, and Yonggang Wang. 2020. ZEN: pre-training chinese text encoder enhanced by n-gram representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4729–4740.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Tao Gui, Ruotian Ma, Qi Zhang, Lujun Zhao, Yu-Gang Jiang, and Xuanjing Huang. 2019. Cnn-based chinese ner with lexicon rethinking. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 4982–4988.

Zhiheng Huang, Wei Xu, and Kai Yu 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

James Hammerton. 2003. Named entity recognition with long short-term memory. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 172–175.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Zijing Ji, Yuxin Shen, Yining Sun, Tian Yu, and Xin Wang. 2021. C-CLUE: a benchmark of classical chinese based on a crowdsourcing system for knowledge graph construction. *Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction*.

Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. FLAT: chinese ner using flat-lattice transformer. In *Proceedings of the 58th Annual Meeting of the Associa-tion for Computational Linguistics*, pages 6836–6842.

(c) 2023 中国中文信息学会计算语言学专业委员会

Wei Liu, Xiyan Fu, Yue Zhang, and Wenming Xiao. 2021. Lexicon enhanced chinese sequence labeling using bert adapter. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 5847–5858.

Zhangxun Liu, Conghui Zhu, and Tiejun Zhao. 2010. Chinese named entity recognition with a sequence labeling approach: based on characters, or based on words? In *Advanced intelligent computing theories and applications. With aspects of artificial intelligence, Springer*, pages 634–640.

Guanqing Liang and Cane Wing-Ki Leung. 2021. Improving model generalization: a chinese named entity recognition case study. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 992–997.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*,pages 260–270.

Yann LeCun and Yoshua Bengio. 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*,3361(10):1995.

Gina-Anne Levow. 2006. The third international Chinese language processing bakeoff: ord segmentation and named entity recognition. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 108–117.

Ruotian Ma, Minlong Peng, Qi Zhang, Zhongyu Wei, and Xuanjing Huang. 2020. Simplify the usage of lexicon in chinese ner. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5951–5960.

Yuyang Nie, Yuanhe Tian, Xiang Wan, Yan Song, and Bo Dai. 2020. Named entity recognition for social media texts with semantic augmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 1383–1391.

Nanyun Peng and Mark Dredze. 2016. Improving named entity recognition for chinese social media with word segmentation representation learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 149–155.

L. F. Rau. 1991. Extracting company names from text. *The Seventh IEEE Conference on Artificial Intelligence Application*, pages 29-32.

Yan Song, Tong Zhang, Yonggang Wang and Kai-Fu Lee. 2021. ZEN 2.0: continue training and adaption for n-gram enhanced text encoders. *arXiv preprint arXiv:2105.01279*.

Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. Directional skip-gram: explicitly distinguishing left and right context for word embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–180.

Dianbo Sui, Yubo Chen, Kang Liu, Jun Zhao, and Shengping Liu. 2019. Leverage lexical knowledge for chinese named entity recognition via collaborative graph network. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3830–3840.

Fábio Souza, Rodrigo Nogueira, Roberto Lotufo. 2019. Portuguese named entity recognition using BERT-CRF. *arXiv preprint arXiv:1909.10649*.

Koichi Takeuchi and Nigel Collier. 2002. Use of support vector machines in extended named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002*.

Yuanhe Tian, Yan Song, Fei Xia, Tong Zhang, and Yonggang Wang. 2020. Improving chinese word segmentation with wordhood memory networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8274–8285.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

(c) 2023 中国中文信息学会计算语言学专业委员会

Shuang Wu, Xiaoning Song, Zhenhua Feng, and Xiaojun Wu. 2022. NFLAT: non-flat-lattice transformer for chinese named entity recognition. *arXiv preprint arXiv:2205.05832*.

Pengyu Wang and Zhichen Ren. 2022. The uncertainty-based retrieval framework for ancient chinese cws and pos. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 164–168.

Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, et al. 2011. OntoNotes release 4.0. *Web Download. Philadelphia: Linguistic Data Consortium*.

Liang Xu, Qianqian Dong, Cong Yu, Yin Tian, Weitang Liu, Lu Li, and Xuanwei Zhang. 2020. CLUENER2020: Fine-grained name entity recognition for chinese. *arXiv preprint arXiv:2001.04351*.

Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021. A unified generative framework for various ner subtasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 5808–5822.

Hang Yan and Bocao Deng and Xiaonan Li and Xipeng Qiu. 2019. TENER: adapting transformer encoder for named entity recognition. *arXiv preprint arXiv:1911.04474*.

Yue Zhang and Jie Yang. 2018. Chinese ner using lattice lstm. In *Proceedings of the 56th Annual Meeting of the Associa-tion for Computational Linguistics*, pages 1554–1564.

GuoDong Zhou and Jian Su. 2002. Named entity recognition using an hmm-based chunk tagger. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 473–480.

包振山,宋秉彦,张文博,孙超. 2022. 基于半监督学习和规则相结合的中医古籍命名实体识别研究. 中文信息学报, 36(6): 90-100.

陈曙东,欧阳小叶. 2020. 命名实体识别技术综述. 无线电通信技术,46(03):251-260.

刘浏. 2018. 古汉语典籍中的实体知识挖掘研究. 南京大学,DOI:10.27235/d.cnki.gnjiu.2018.001041.

刘浏,王东波. 2018. 命名实体识别研究综述. 情报学报,37(03):329-340.

施丁, 沈志华, 陈东林, 和龚. 1994. 资治通鉴大辞典. 吉林人民出版社.

苏棋, 胡韧奋, 诸雨辰, 严承希, 王军. 2021. 古籍数字化关键技术评述. 数字人文研究1(3): 83-88.

王东波，刘畅，朱子赫，刘江峰，胡昊天，沈思，李斌. 2021. SikuBERT与SikuRoBERTa：面向数字人文的《四库全书》预训练模型构建及应用研究.

徐晨飞, 叶海影, 包平. 2018. 基于深度学习的方志物产资料实体自动识别模型构建研究. 数据分析与知识发现, 4(8): 86-97.

张朦, 刘忠宝. 2023. 数字人文环境下融入多特征的词命名实体识别. 计算机系统应用,32(3):300-308.

CCL 2023