古汉语嵌套命名实体识别数据集的构建和应用研究

谢志强[1], 刘金柱*, 刘根辉[2]
华中科技大学，人文学院，湖北，武汉，430000
xiezhiqiang2020@163.com
1152822887@qq.com
genhuiliu@163.com

摘要

本文聚焦研究较少的古汉语嵌套命名实体识别任务，以《史记》作为原始语料，针对古文意义丰富而导致的实体分类模糊问题，分别构建了基于字词本义和语境义2个标注标准的古汉语嵌套命名实体数据集，探讨了数据集的实体分类原则和标注格式，并用RoBERTa-classical-chinese+GlobalPointer模型进行对比试验，标准一数据集F1值为80.42%，标准二F1值为77.43%，以此确定了数据集的标注标准。之后对比了六种预训练模型配合GlobalPointer在古汉语嵌套命名实体识别任务上的表现。最终试验结果：RoBERTa-classical-chinese模型F1值为84.71%，表现最好。

关键词： 古汉语 ； 嵌套命名实体识别 ； 数据集 ； GlobalPointer

Construction and application of classical Chinese nested named entity recognition data set

Zhiqiang Xie[1], Jinzhu Liu*, Genhui Liu[2]
School of Humanities,Huazhong University of Science and Technology,Hubei,Wuhan,430000
xiezhiqiang2020@163.com
1152822887@qq.com
genhuiliu@163.com

Abstract

This paper focuses on the less studied task of classical Chinese nested entity recognition, and constructs a data set of classical Chinese nested entity with Historical Records as the original corpus.For the fuzzy problem of entity classification caused by the rich meaning of classical Chinese, this paper constructs classical Chinese nested named entity data sets based on two annotation standards: word original meaning and context meaning, and discusses the entity classification principles and annotation formats of the data sets.A comparative experiment with RoBERTa-classical-chinese+globalpointer model are used to determine the annotation standard of the data set. The F1 value of standard one data set is 80.42%, and that of standard two is 77.43%, which determines the annotation standard of the data set. Then, we compares the performance of six pre-training models combined with globalpointer in the task of classical Chinese nested named entity recognition.Finally,the F1 value of RoBERTa-classical-chinese model is 84.71%, which performs best.

Keywords: Classical Chinese , Nested Named Entity Recognition , Data Set , GlobalPointer

基金项目:华中科技大学人文社会科学重大原创性成果培育项目“《册府元龟》引书研究”（21WKFZZX016）
华中科技大学自主创新研究基金专项任务项目“《册府元龟》语料库建设、整理与研究”（2020WKYXZX004）
通讯作者:刘金柱

1 引言

命名实体识别(Named Entity Recognition,NER)作为自然语言处理中的基础任务，是开展信息抽取、构建知识图谱等上游任务的重要一环。关于它的研究最早开始于上世纪60 年代(Grishman and Sundheim, 1995)，具体是指识别出待处理文本中预定义好的命名实体，包括实体边界的识别以及实体类型分类两个任务。之后，随着研究的深入，从实体层级角度又进一步细分为扁平命名实体识别任务和嵌套命名实体识别任务。其中，扁平命名实体识别（Flat NER）任务中，每个实体是最小的单位，不能继续拆分。而嵌套命名实体（Nested NER），又称实体重叠，表示在一个实体的内部还存在着一个或多个其他的实体。

目前，随着基于BERT的自然语言处理预训练模型的演化发展及迁移学习，现代汉语命名实体识别任务得到了相对有效的解决，并在此基础上进一步提升了上游任务的各项指标。但比较而言，古汉语命名实体识别研究则相对较少，仅有少数研究集中在古汉语扁平命名实体识别研究，古汉语嵌套命名实体识别研究则寥寥无几。究其原因，一是相关开源的古汉语高质量语料较少；二是古汉语相较于现代汉语而言，在实体分类原则、标注标准选取等方面处理难度更大，要求研究者除了有工程能力外，还要有扎实的古汉语专业知识做支撑。

针对上述问题，本研究基于人工精校后的《史记》，人工标注构建了古汉语嵌套命名实体识别数据集，并尝试使用GlobalPointer作为解码层配合六种预训练模型，开展面向《史记》的古汉语嵌套命名实体识别研究工作，以期为更细粒度的古汉语实体识别和信息抽取提供经验。

2 相关工作

2.1 嵌套实体数据集构建的相关工作

近年来，为了开展嵌套命名实体识别任务而构建的数据集主要有ACE语料(Mitchell et al., 2005)、GENIA语料(Zhou et al., 2004)、SciERC语料库(Ringland et al., 2019)、KBP2015语料库、NNE数据集(Ringland et al., 2019)、人民日报语料库、CADEC(Karimi et al., 2015)等。

2.2 古汉语命名实体识别相关工作

古汉语命名实体识别研究按照其历史发展进程主要分为基于规则和词典匹配的命名实体识别、基于统计机器学习的命名实体识别、基于深度学习的命名实体识别。

2.2.1 基于规则和词典匹配

基于规则和词典匹配，其思想在于观察特定领域文本中的语法规则，从而归纳并设计出特定实体的提取规则以完成提取。曾艳和侯汉清(2008)基于N元语法提出了一种古文自动抽词方法，使用N-gram 对古文语料自动分词，利用抽词词典和停用词词典匹配人名、地名、书名、官职名等词汇，最终对N 元组进行过滤并人工判别选词；朱锁玲等人(2011)通过统计《古今地名对照表》等资料中的古代地名，构建了地名词典，并从《大埔县志》等地方志中抽取地名的上下文特征构造地名识别规则库，以《方志物产》作为语料，对物产地名进行实体识别。皇甫晶等人(2013)通过手动构建规则，对纪传体古汉语文献进行姓名的实体识别。

2.2.2 基于统计机器学习

基于统计机器学习的方法，是将实体识别任务细化成一个多分类问题或者序列标注问题，即将该任务转化成为一个基于字符的分类问题，通过已标注的数据训练模型，将不同字符映射成为不同标签的过程。黄水清等人(2015)在基于先秦古汉语语料库基础上，使用条件随机场模型构建特征，对地名实体进行识别。同样基于条件随机场模型进行古汉语命名实体识别的有李娜等(2018)、王东波等(2018)、袁悦等(2019)。

2.2.3 基于深度学习

深度学习方法也是基于多分类问题和序列标注问题两个任务，但是通过完成一个比统计机器学习更加复杂的建模过程，达到一个更好的任务效果。一个标准的NER深度学习模型一般由输入层、编码层和解码层三层结构建模成。其中，新增的复杂编码层需要解决特征抽取的问题，以捕获实体上下文的特征表示。而经典的特征器包括卷积神经网络（CNN）、循环神经网络（RNN）、递归神经网络、Transformer神经网络(Lample et al., 2016)以及语言模型网络等。崔竟烽等人(2020)通过人工标注数据集，使用深度学习的BERT模型对菊花古典诗词进命名实体识别。刘忠宝等人(2020)使用BERT+Bi-LSTM+CRF对《史记》中的实体进行了识别。

3 古汉语嵌套命名实体识别任务研究

3.1 任务定义

嵌套命名实体识别问题可以形式化表示为：给定一个序列$X = \{x_1, x_2, ..., x_n\}$，其中$X_n$表示序列的第n个词条，预测该序列的标签$Y = \{y_1, y_2, ..., y_n\}$。与非嵌套命名实体识别不同的是，嵌套命名实体识别的词条标签Y_n表示多标签而不是单标签，$Y_n = \{y_n^1, y_n^2, ..., y_n^m\}$，其中m为嵌套层数。

3.2 数据集来源

《史记》，二十四史之一，最初称为《太史公书》、《太史公记》、《太史记》，是西汉史学家司马迁撰写的纪传体史书，是中国历史上第一部纪传体通史，作品中撰写了上至上古传说中的黄帝时代，下至汉武帝太初四年间共3000多年的历史。

本文工作所采用的语料来源为2014年中华书局点校本。精校的《史记》版本已经去除了文本中的特殊字符，并在此基础上进行了分行，目前标注完成的数据集字符数为57015，平均句子长度19.97。

3.3 数据集标注

3.3.1 标注原则的确定

数据集标注遵循张欢(2020)提出的简单性原则、易操作性原则、一致性原则。

其一，简单性原则。本研究将古汉语实体划分为五大类，包括“人（PER）”、“地点（LOC）”、“官职（JOB）”、“书（BOOK）”和“时间（TIME）”。“人（PER）”这类实体没有根据其构造成分细化。剔除了“组织（ORG）”，将其合并到“地点（LOC）”，合并“朝代（DYN）”和“年号（REI）”为“时间（TIME）”，类别数量适中，遵循了简单性原则。

其二，易操作性原则。实体的标注包括实体边界和实体类型两个环节。本研究标注实体时，以实体开头为起，至实体结尾为止，进行全选，之后选择实体类型，即完成标注。此外，标注实体均是拆分至最小实体单位为止。标注方式遵循了易操作性原则。

其三，一致性原则。实体类型定义是进行实体识别任务的第一个步骤，通常情况下，研究者都会根据研究的具体需求和侧重来确定需要识别的实体类型。因此，在实体标注时，对于实体类型的确定存在一定的主观性。本研究标注实体时遵循一致性的原则，针对容易混淆的实体标签，选择对这些标签进行合并处理，具体后述。

3.3.2 古汉语实体体系的构建

本研究将古汉语实体划分为五大类，包括“人（PER）”、“地点（LOC）”、“官职（JOB）”、“书（BOOK）”和“时间（TIME）”。

第一类实体是“人（PER）”：一般的，在命名实体识别任务中，“人（PER）”这类实体主要指语料中的人名。但是，古汉语“人（PER）”这类实体的构造成分较之现代汉语要更为复杂，“人（PER）”成分种类繁多，包括名、字、氏、姓、爵位、排行、谥号、官职等。如表1举例所示:

“人（PER）”实体构造成分	举例
名	鄭伯克**段**于鄢
氏+姓	晉獻公欲以**驪姬**爲夫人
官职+名+尊称	**司徒皇父**帅师御之
谥号+排行	**襄仲**欲立之
官职	日夜望**將軍**至，豈敢反乎

Table 1: “人（PER）”实体构造成分及举例

本研究虽是古汉语嵌套命名实体识别，但对于“人（PER）”这类实体的构造成分并不细分。原因有二：一是标注数据集有限，过于复杂的实体划分会导致数据稀疏问题，从而影响最终模型的效果。二是“人（PER）”构造成分的划分存在争议，且本研究的目的也不是人名考证。因此，“人（PER）”这类实体应当遵循简单性原则，除“官职（JOB）”这类区分度高的

实体外，剩余皆统一合并到“人（PER）”中。表1中的“司徒皇父”可以通过嵌套标注的方式清晰的划分出实体结构，即“人[官职+名]”，但还出现很多单个表“官职（JOB）”的词汇代指“人（PER）”的情况，如“ 日夜望將軍至，豈敢反乎 ”。又如“诸侯”一词，用“官职（JOB）”代指“人（PER）”或“地点（LOC）”。

第二类实体是“地点（LOC）”：本研究中“地点（LOC）”除了地理意义上的“地名”，还包含“山名”、“水名”等。除此之外，由于《史记》是记载了上至上古传说中的黄帝时代，下至汉武帝太初四年间共3000多年的历史，所以包含很多诸如未统一的“诸侯国”、“部落”、“族群”、“某一氏族”等具有政治意义的地点，这些词细分应为“组织（ORG）”，但遵循简单性原则，同时为了保证样本分布的均衡，本研究将这些词也归类到“地点（LOC）”。

第三类实体是“官职（JOB）”：指在国家机构中担任一定职务的官吏，上至中央大员，下至地方小吏。但正如前文“人（PER）”实体界定所述，有很多表官职的词语实际语境中用来代指“人（PER）”，因此在标注时结合上下文进行了判别，尽可能保证实体归类的准确性。同时，也设计了根据字词本义的标注方式，用于对比补充实验。

第四类实体是“书（BOOK）”：包含书籍、诗歌、文章等。整体样本量较少，但界限分明。

第五类实体是“时间（TIME）”：只包括“朝代（DYN）”和“年号（REI）”，并不包含“季节”、“月份”等实体，是本研究人为规定的狭义“时间（TIME）”。“朝代（DYN）”是界定某一个统一政权时期的名词，但由于《史记》的历史跨度有限，导致“朝代（DYN）”这类实体数量极小，并不适合单独成为一个实体类别。而用“年号（REI）”纪年，是从汉武帝开始的，因此“年号（REI）”数量更是稀少，于是合并两类实体为“时间（TIME）”。

以上是对本研究中古汉语各类实体的界定范围和界定原因的说明。对于实体分类界限模糊的问题，本研究设计了两个标注标准：一方面主要是根据字词本义进行标注，同时也尝试了结合上下文具体语境对字词进行标注，构建了2个数据集，开展对比补充实验，根据最终的F1值来考察2种标注方式的优劣。

3.3.3 数据集标注格式

数据集标注的格式参考2021年阿里天池发布的中文医疗信息处理挑战榜CBLUE(Chinese Biomedical Language Understanding Evaluation)包含的中文医学命名实体识别任务的数据集(Hongying et al., 2020)。

具体的格式为：整个数据集为一整个json，里面每一条数据为一个json，内部是句子、实体的起始位置、实体类别和实体。举例如下：


```
{
  "text": "而蚩尤最爲暴，莫能伐。",
  "entities": [
    {
      "start_idx": 1,
      "end_idx": 2,
      "type": "PER",
      "entity": "蚩尤"
    }
  ]
},
```


Figure 1: 扁平命名实体的标注格式

```
{
  "text": "懿王之時，王室遂衰，詩人作刺。",
  "entities": [
    {
      "start_idx": 1,
      "end_idx": 1,
      "type": "JOB",
      "entity": "王"
    },
    {
      "start_idx": 0,
      "end_idx": 1,
      "type": "PER",
      "entity": "懿王"
    }
  ]
},
```


Figure 2: 嵌套命名实体的标注格式

4 模型选择

本研究基于目前主流的方式进行命名实体识别模型的训练，即采用基于深度学习方式分别搭建六个预训练模型，并尝试将GlobalPointer作为解码层，以优化模型性能。

4.1 古汉语预训练模型

在命名实体识别研究中，基于深度学习的方式逐渐取代基于统计机器学习的方法，主要是由于深度学习模型通过构建更为复杂的编码层，表现出更为显著的特征抽取性能。尤其是在BERT出现之后，基于深度学习的方式更是成为了新的基准方式，并出现一系列基于BERT的改进变体模型，RoBERTa就是其中的典型代表。RoBERTa模型不仅继承了BERT模型的优势，而且用更大的单次训练样本数和更多的数据训练模型，移除了NSP(Next Sentence Prediction)任务，同时采用动态编码，极大提升了BERT的模型性能。

但上述这些工作主要集中于面向英语和现代汉语的模型训练，阎覃(2020)开发的GuwenBERT模型首次将此项工作迁移到古汉语领域，GuwenBERT模型是基于殆知阁古文文献语料训练，其中包含15694本古文书籍，字符数1.7B。所有繁体字均经过简体转换处理，结合现代汉语RoBERTa权重和大量古文语料，将现代汉语的部分语言特征向古代汉语迁移以提升表现。

本次的任务是古汉语的嵌套命名实体识别，如果使用GuwenBERT模型就需要将古汉语的文献转换成现代汉语的文献，但繁简转换会出现很多问题，如一简对多繁、引发歧义等。王东波等(2021)以校验后的高质量《四库全书》全文语料作为训练集，基于BERT深度语言模型框架，构建了面向古文智能处理任务的SikuBERT和SikuRoBERTa预训练语言模型。第二年，Koichi Yasuoka等(2022)完成了RoBERTa-classical-chinese模型，并且扩展了繁体词表。SikuRoBERTa和RoBERTa-classical-chinese两个预训练语言模型的开发解决了繁简转换这一问题。因此，本研究主要侧重选用SikuRoBERTa和RoBERTa-classical-chinese两个模型进行训练。

4.2 GlobalPointer

在命名实体识别研究中，通过编码层对实体进行抽象语义表示，生成相应的标签序列，而解码层则用于预测实体的边界以及实体的类型，是整个实体识别模型的最后阶段。常用的标签解码器包括MLP+softmax层、条件随机场(CRF)、循环神经网络解码器、指针网络（Pointer Network）几种类型。

其中，MLP+softmax层、条件随机场(CRF)、循环神经网络解码器主要用于扁平命名实体识别，指针网络（Pointer Network）除了用于扁平命名实体识别外，也适用于嵌套命名实体识别。指针网络解码器采用循环解码的结果，它将序列标注问题转化为先分界再分类两个子任务，其中先分界指找出实体的开始位置和结束位置，然后对识别出的实体块进行类型识别，之后再继续找下一个块的结束位置，再将这个块进行分类，一直循环直到序列结束。但指针网络

的设计在做实体识别时，会出现训练和预测不一致的问题，即训练的时候开始位置和结束位置是孤立的，预测的时候开始位置和结束位置是有联系的，所以开始位置和结束位置都预测正确实体才算预测正确，这就导致了不一致。

针对上述不一致的问题，苏剑林(2022)利用全局归一化的思路来进行命名实体识别，设计了可以无差别地识别嵌套实体和非嵌套实体的GlobalPointer。它的主要思想体现在它将开始位置和结束位置视为一个整体去进行判别。具体的，设长度为n的序列输入t经过编码后得到向量序列$\{h_1, h_2, ..., h_n\}$,通过变换$q_{i,\alpha} = W_{q,\alpha}h_i + b_{q,\alpha}$和$k_{i,\alpha} = W_{k,\alpha}h_i + b_{k,\alpha}$，我们可以得到序列向量序列$\{q_{1,\alpha}, q_{2,\alpha}, ..., q_{n,\alpha}\}$和$\{k_{1,\alpha}, k_{2,\alpha}, ..., k_{n,\alpha}\}$，它们是识别第$\alpha$种类型实体所用的向量序列。此时可以定义

$$s_\alpha(i,j) = q_{i,\alpha}^T k_{j,\alpha} \tag{1}$$

作为从i到j的连续片段是一个类型为α的实体的打分。也就是说，用$q_{i,\alpha}$与$k_{j,\alpha}$的内积，作为片段$t_{|i:j|}$是类型为α的实体的打分。

接下来的关键是损失函数的设计。苏剑林(2022)提出一个用于多标签分类的损失函数，特别适合总类别数很大但目标类别数较小的多标签分类问题。该函数为

$$log(1 + \sum_{(i,j)\in P_\alpha} e^{-s_\alpha(i,j)}) + log(1 + \sum_{(i,j)\in Q_\alpha} e^{s_\alpha(i,j)}) \tag{2}$$

其中P_α是该样本的所有类型为α的实体的开始和结束位置集合，Q_α是该样本的所有非实体或者类型非α的实体的开始和结束位置集合，只需要考虑i⩽j的组合，即

$$\begin{aligned} \Omega &= \{(i,j) \mid 1 \le i \le j \le n\} \\ P_\alpha &= \{(i,j) \mid t_{|i:j|}\text{是类型为}\alpha\text{的实体}\} \\ Q_\alpha &= \Omega - P_\alpha \end{aligned} \tag{3}$$

5 实验

5.1 模型参数设置及模型构建

本次实验采用RoBERTa-classical-chinese、SikuRoBERTa、SikuBERT、RoBERTa-wwm-ext、BERT-wwm-ext和GuwenBERT六个预训练语言模型，并在训练过程中进行微调，获取字嵌入向量，并完成特征抽取，以捕获实体上下文的特征表示，然后输入GlobalPointer进行解码，获得预测标签序列。下图3展示了以RoBERTa-classical-chinese+GlobalPointer搭建的模型框架，其他模型与之类似。

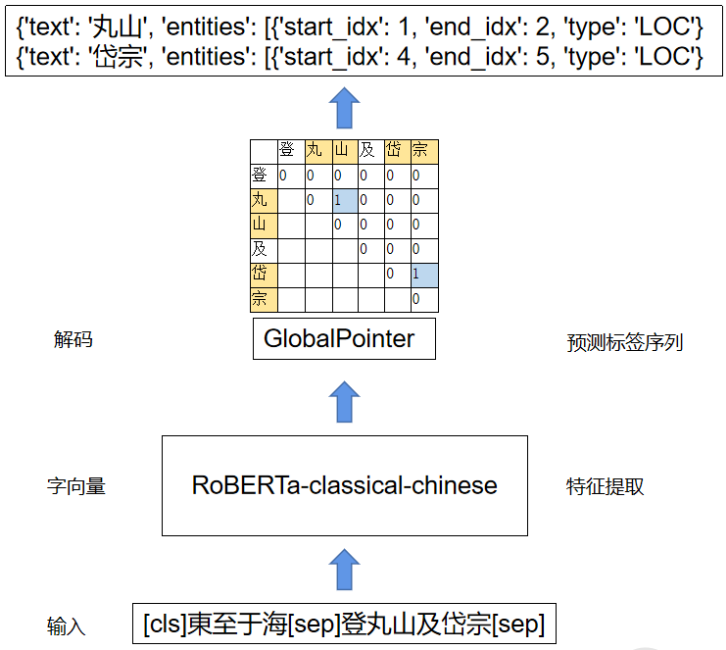

Figure 3: RoBERTa-classical-chinese+GlobalPointer模型的整体结构

同时，使用Adam优化器(Da, 2014)，所有模型的学习率（Learning Rate）均设置为$2e^{-5}$，输入序列最大长度（Maxlen）为128，每批训练大小（Batch Size）为16，迭代次数（Epochs）为50，训练时保存效果最优的模型。

5.2 测评指标及评价方法

为了评估以上六个预训练语言模型对古汉语实体识别的效果，本次实验采用精确率(precision)、召回率(recall)和F1值(F1 score)作为评估标准。

$$\text{精确率} = \frac{\text{正确预测为正的数量}}{\text{所有预测为正的数量}} \tag{4}$$

$$\text{召回率} = \frac{\text{正确预测为正的数量}}{\text{实际为正的数量}} \tag{5}$$

$$F1\text{值} = \frac{2 \times \text{精确率} \times \text{召回率}}{\text{精确率} + \text{召回率}} \tag{6}$$

精确率越高，代表模型对负样本的区分程度越高；召回率越高，代表模型对正样本的识别程度越好；F1值越高，代表模型越稳定。

5.3 实验结果及分析

5.3.1 实验一：不同数据规模和比例下的实验结果

为了确定实验中训练集、验证集和测试集的占比，同时考察不同数据集规模对实验结果的影响，将训练集、验证集和测试集以25%的比列增量多轮试验。实验中，训练集、验证集和测试集之间的比例暂定为7:1.5:1.5。由于只是考察数据集规模对实验的影响，且之后实验的预训练模型主要以RoBERTa类型为主，所以在实验一中，主要选取了RoBERTa-classical-chinese+GlobalPointer模型。以下表2展示的是数据集规模对古汉语嵌套实体识别的实验结果:

语料比列	P	R	F1
25%	68.92%	67.65%	**68.05%**
50%	79.60%	63.98%	**70.42%**
75%	72.98%	73.29%	**72.61%**
100%	85.35%	84.44%	**84.71%**

Table 2: 不同数据集规模的实验结果

之后为了进一步确定和考察不同比例对于古汉语嵌套实体识别的影响，将全量数据按照8:1:1划分后，进行了补充实验，以下表3-5展示了不同比例划分下的实体数量分布情况和不同比例下的实验结果:

实体种类	训练集	验证集	测试集
人（PER）	3686	280	621
书（BOOK）	83	2	0
地点（LOC）	2424	214	450
官职（JOB）	1833	175	465
时间（TIME）	349	48	65
实体总数	8375	719	1601

Table 3: 8:1:1数据集实体数量分布

实体种类	训练集	验证集	测试集
人（PER）	3390	446	751
书（BOOK）	77	6	2
地点（LOC）	2208	363	517
官职（JOB）	1651	264	558
时间（TIME）	330	57	75
实体总数	7656	1136	1903

Table 4: 7:1.5:1.5数据集实体数量分布

模型	各数据集占比	P	R	F1
RoBERTa-classical-chinese	8:1:1	**85.35%**	**84.44%**	**84.71%**
RoBERTa-classical-chinese	7:1.5:1.5	82.28%	79.47%	80.42%

Table 5: 不同数据集比列的实验结果

根据表2得知，随着数据集规模的扩大，F1值也随之增长。25%到50%，50%到75%，F1值呈平缓稳定的增长，由75%到100%后，F1值显著增长，可见扩大数据集，模型的学习效果会更高，当数据集达到一个阈值，会迎来效果的质变。所以，在合理的范围内，应尽可能的扩充数据集。根据这一结论，训练集、验证集和测试集之间的比例也由7:1.5:1.5调整到了8:1:1，效果如表5所示，F1值有所提高。

5.3.2 实验二：不同标注标准下的实验结果

前文提到实体分类界限模糊的问题，基于此，本次实验设计了两个标注标准：一是根据字词本身含义进行标注。二是结合具体语境含义对字词进行标注。在全量数据下，使用RoBERTa-classical-chinese+GlobalPointer模型，按照7:1.5:1.5的比例对两个数据集进行了对比实验，根据最终的F1值来敲定最终标注的标准。

第二十一届中国计算语言学大会论文集，第406页-第416页，南昌，中国，2022年10月14日至16日。

标注标准	P	R	F1
标准一	**82.28%**	**79.47%**	**80.42%**
标准二	77.17%	78.17%	77.43%

Table 6: 不同标注标准的实验结果

根据表6得知，标准一在精确率、召回率和F1值上都要优于标准二。标准一的标注原则是根据字词本身含义进行标注，如“官职”在语境中代指“人”，依旧标注为“人（PER）”。通过本次对比实验也可以看出，就本研究使用的语料和条件设置下，对于字词的引申含义、具体语境含义等这些并非字词本身义项的情况出现时，依旧要按照字词原本含义进行标注，这一点是不同于某一字词的不同义项应用在不同场景这种情况的，所以并不能按照同一标准对待。

5.3.3 实验三：不同模型下的实验结果

在确定数据集标注采用的标准后，我们又增加了SikuRoBERTa、SikuBERT、RoBERTa-wwm-ext、BERT-wwm-ext、GuwenBERT5个预训练语言模型，与GlobalPointer搭配，展开不同模型维度的对比试验。其中，Siku系列、RoBERTa-classical-chinese和GuwenBERT都是面向古文开发的的预训练语言模型，剩余是面向现代汉语开发的预训练语言模型。具体实验结果如下表7：

模型	P	R	F1
RoBERTa-classical-chinese	**85.35%**	**84.44%**	**84.71%**
SikuRoBERTa	85.02%	83.73%	84.18%
SikuBERT	85.43%	82.12%	83.66%
RoBERTa-wwm-ext	81.38%	81.65%	81.29%
BERT-wwm-ext	79.66%	83.72%	81.13%
GuwenBERT	78.98%	77.20%	77.84%

Table 7: 模型结果

表7比较了RoBERTa-classical-chinese、Siku系列模型、RoBERTa-wwm-ext、BERT-wwm-ext和GuwenBERT六个模型在同一数据集下的效果。

从表7的结果来看，SikuRoBERTa较之SikuBERT，RoBERTa-wwm-ext较之BERT-wwm-ext，F1值分别提升0.52%，0.16%。说明RoBERTa预训练模型性能要优于BERT预训练模型，RoBERTa模型不仅继承了BERT模型的优势，而且用更大的单次训练样本数和更多的数据训练模型，移除了NSP任务，采用动态编码。这一系列的改进措施促使了RoBERTa的效果优于BERT。

RoBERTa-classical-chinese和Siku系列模型在精确率、召回率、F1值三个指标上都要高于RoBERTa-wwm-ext和BERT-wwm-ext两个模型。这是由于，RoBERTa-classical-chinese基于殆知阁古文文献语料训练的语言模型，Siku系列模型是基于校验后的高质量《四库全书》全文语料训练的，均是针对古汉语专门训练的预训练语言模型，而RoBERTa-wwm-ext和BERT-wwm-ext是面向现代汉语训练的预训练语言模型。所以RoBERTa-classical-chinese和Siku系列模型更加适合本次的古汉语实体识别任务。

观察对比发现，GuwenBERT预训练语言模型的F1值在本次实验较之其他模型低。虽然GuwenBERT预训练语言模型采用古汉语语料进行训练，但是所有语料均被转换为简体，繁体转换为简体后，会出现一简对多繁的问题，从而引发歧义，而且部分语料未经分句、没有标点符号，则进一步放大了这一问题，最终导致模型性能降低。

对于古汉语预训练语言模型的对比。平均F1值RoBERTa-classical-chinese和SikuRoBERTa相差无几，RoBERTa-classical-chinese高出SikuRoBERTa 0.53%。两类模型同属于RoBERTa，结构相似，出现这个差距主要在于RoBERTa-classical-chinese训练所采用的语料数量要大于Siku系列模型，但由于语料并非精加工，甚至存在未分句，未标点的语料，使得最终F1值只是略高于Siku系列模型。

6 总结与展望

本文通过使用六种预训练语言模型实现了古汉语嵌套命名实体识别的初探，检验了古汉语预训练模型RoBERTa-classical-chinese和SikuRoBERTa配合GlobalPointer模型在古汉语嵌套命名实体识别任务上的良好性能。同时探讨了古汉语嵌套实体数据集的实体分类原则、数据集的格式和确定数据集的标注标准。

由于原始语料本身的特点，使得本次实验所用的数据集存在规模较小、各实体类别样本分布不均匀的不足，对于模型的最终性能造成了一定的影响。同时，由于鲜有开源的古汉语预训练模型，与现代汉语模型对比效果有失公平，所以无法进行更多模型的对比。因此，下一步工作将围绕扩充不同类别古文语料，建立更大规模、分布更均匀的高质量标注数据集，同时尝试并优化算法，以期进一步提升古汉语嵌套识别模型的性能，为古汉语信息抽取和知识图谱的建立提供帮助。

参考文献

Kingma Da. 2014. A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ralph Grishman and Beth Sundheim. 1995. Appendix c: Named entity task definition (v2. 1). In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pages 317–332.

Zan Hongying, Li Wenxin, Zhang Kunli, Ye Yajuan, Chang Baobao, and Sui Zhifang. 2020. Building a pediatric medical corpus: Word segmentation and named entity annotation. In *Workshop on Chinese Lexical Semantics*, pages 652–664.

Sarvnaz Karimi, Alejandro Metke-Jimenez, Madonna Kemp, and Chen Wang. 2015. Cadec: A corpus of adverse drug event annotations. *Journal of biomedical informatics*, 55:73–81.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Alexis Mitchell, Stephanie Strassel, Shudong Huang, and Ramez Zakhary. 2005. Ace 2004 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 1:1–1.

Nicky Ringland, Xiang Dai, Ben Hachey, Sarvnaz Karimi, Cecile Paris, and James R Curran. 2019. Nne: A dataset for nested named entity recognition in english newswire. *arXiv preprint arXiv:1906.01359*.

Huan Zhang, Yuan Zong, Baobao Chang, Zhifang Sui, Hongying Zan, and Kunli Zhang. 2020. 面向医学文本处理的医学实体标注规范(medical entity annotation standard for medical text processing). In *Proceedings of the 19th Chinese national conference on computational linguistics*, pages 561–571.

Guodong Zhou, Jie Zhang, Jian Su, Dan Shen, and ChewLim Tan. 2004. Recognizing names in biomedical texts: a machine learning approach. *Bioinformatics*, 20(7):1178–1190.

刘忠宝，党建飞，张志剑. 2020. 《史记》历史事件自动抽取与事理图谱构建研究. 图书情报工作, 64(11):116.

安孝一. 2022. Transformers のbert は共通テスト『国』をり受け解析するをるか. 洋学へのコンピュタ利用第*33* 回研究セミナ, 33:3–34.

崔竞烽，郑德俊，王东波，李婷婷. 2020. 基于深度学习模型的菊花古典诗词命名实体识别. 情报理论与实践, 43(11):150.

曹艳，侯汉清. 2008. 古籍文本抽词研究. 图书情报工作, 52(01):132.

朱锁玲，包平. 2011. 方志类古籍地名识别及系统构建. 中国图书馆学报, 37(3):118–124.

李娜，包平. 2018. 面向数字人文的馆藏方志古籍地名自动识别模型构建. 图书馆, (5):67–73.

袁悦，王东波，黄水清 ，李斌. 2019. 不同词性标记集在典籍实体抽取上的差异性探究. 现代图书情报技术, 003(003):57–65.

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.

王东波，高瑞卿，沈思，李斌. 2018. 面向先秦典籍的历史事件基本实体构件自动识别研究. 国家图书馆学刊, 27(1):65–77.

皇甫晶，王凌云. 2013. 基于规则的纪传体古代汉语文献姓名识别. 图书情报工作, 57(03):120.

苏剑林. 2022. Efficient globalpointer：少点参数，多点效果, Jan. https://spaces.ac.cn/archives.

阎覃. 2020. Guwenbert: 古文预训练语言模型(古文bert). 2020-11-22]. https://github.com/Ethan-yt/guwenbert.

黄水清，王东波，何琳. 2015. 基于先秦语料库的古汉语地名自动识别模型构建研究. 图书情报工作, 59(12):135.