CCL23-Eval 任务1系统报告：基于增量预训练与对抗学习的古籍命名实体识别

李剑龙
中国工商银行/ 北京
BISTU-IIIP / 北京
1436631592@qq.com

于右任
BISTU-IIIP / 北京
a154377713@163.com

刘雪阳
BISTU-IIIP / 北京
1239996108@qq.com

朱思文
BISTU-IIIP / 北京
1391911891@qq.com

摘要

古籍命名实体识别是正确分析处理古汉语文本的基础步骤，也是深度挖掘、组织人文知识的重要前提。古汉语信息熵高、艰涩难懂，因此该领域技术研究进展缓慢。针对现有实体识别模型抗干扰能力差、实体边界识别不准确的问题，本文提出使用NEZHA-TCN与全局指针相结合的方式进行古籍命名实体识别。同时构建了一套古文数据集，该数据集包含正史中各种古籍文本，共87M，397,995条文本，用于NEZHA-TCN模型的增量预训练。在模型训练过程中，为了增强模型的抗干扰能力，引入快速梯度法对词嵌入层添加干扰。实验结果表明，本文提出的方法能够有效挖掘潜藏在古籍文本中的实体信息，F1值为95.34%。

关键词： 古籍命名实体识别；增量预训练；快速梯度法

System Report for CCL23-Eval Task 1：GuNER Based on Incremental Pretraining and Adversarial Learning

Jianlong Li
ICBC / Beijing
BISTU-IIIP / Beijing
1436631592@qq.com

Youren Yu
BISTU-IIIP / Beijing
a154377713@163.com

Xueyang Liu
BISTU-IIIP / Beijing
1239996108@qq.com

Siwen Zhu
BISTU-IIIP / Beijing
1391911891@qq.com

Abstract

GuNER is the basic step for analyzing and processing ancient Chinese texts correctly, which is also an important prerequisite for in-depth mining and organizing human knowledge. Due to its high information entropy and difficulty, the technological research progress in ancient Chinese filed is slow. To address the issues of poor anti-interference ability and inaccurate entity boundary recognition in existing entity recognition models, this article proposes a method of combining NEZHA-TCN with global pointer for ancient named entity recognition. At the same time, an ancient text dataset was constructed, which includes various ancient texts from the historical collection, totaling 87M and 397,995 texts, for incremental pretraining of the NEZHA-TCN model. In the process of model training, in order to enhance the anti-interference ability of the model, the fast gradient method is introduced to add interference in the word embedding layer. The experimental results show that the method proposed in this article can effectively mine the entities in the ancient texts, with an F1 value of 95.34%.

Keywords: GuNER , Incremental pretraining , Fast gradient method

第二十二届中国计算语言学大会论文集，第23页-第33页，哈尔滨，中国，2023年8月3日至5日。
(c) 2023 中国中文信息学会计算语言学专业委员会

1 引言

古籍命名实体识别(苏祺 et al., 2023)是当前汉语领域研究的热点问题之一，其旨在通过自然语言处理技术从古汉语文本中抽取出人名、官职名、书籍名等关键信息。然而，古籍命名实体识别领域面临诸多困难。当前古籍文本研究不仅缺乏相应的模型技术支持，而且还面临领域可用训练数据较少的问题，这阻碍了技术的长足发展。

作为汉语理解与分析的关键一环，对古籍文本进行准确高效分析能够为古文分析人员提供技术支持，减轻技术人员的工作量，提升古文在汉语言文学的影响力。但目前，古籍文本分析研究人员少，导致相关工作进展缓慢，算法模型产出滞后。且现有的模型都是沿用在其他领域的模型，导致算法领域特质不鲜明，无法更加高效地对古文文本进行有效分析。针对模型抗干扰能力差，词边界信息难以区分，且开源数据缺乏的问题，本文提出一种基于增量预训练与对抗学习的古籍命名实体识别模型（Ancient Named Entity Recognition Model Based on Incremental Pretraining and Adversarial Learning，ANER-IPAL）用于古籍命名实体识别。

2 相关工作

随着深度学习技术的不断发展，古籍命名实体识别研究主要依托于命名实体识别技术的发展，而命名实体识别研究可以分为三个阶段：传统方法阶段、神经网络阶段和预训练模型阶段。

传统方法阶段主要包括：基于模板的方法和基于统计的方法。基于模板的方法是指利用已建立的规则对句子进行模式匹配，找出句子中对应的实体。这种方法需要语言学家制定相关规则，在数据量较少的情况下可以取得良好的效果。然而，模板中预定义的规则并不适用于领域迁移和未登录词识别场景。因此，基于统计的方法应运而生。基于统计的方法是指利用条件随机场(Conditional Random Field，CRF)、隐马尔可夫模型(Hidden Markov Model，HMM)和最大熵模型(Maximum Entropy Model，MEM)对数据集进行统计和特征建模，并找出文本中的实体。Yang等人(2006)设计了基于HMM的中文命名实体识别算法来识别文本中出现的命名实体，并在当时取得了良好的效果。Duan和Zheng(2011)使用CRF进行中文领域的实体识别模型建模，通过CRF模型获得各标签序列的分数值，并解码得到命名实体识别结果。

随着深度学习模型的不断发展，命名实体识别的研究方向也从传统方法发展到神经网络方法。神经网络方法不需定义实体抽取规则，它可以自动从文本数据中挖掘潜在特征，并完成命名实体识别任务。由于深度学习方法的高效性和便捷性，近些年，基于此方法的命名实体识别工作如雨后春笋般涌现。Huang等人(2015)提出了一种结合BiLSTM和CRF的中文命名实体识别模型。借助于BiLSTM善于捕捉长距离依赖关系，CRF可以优化序列输出的特点，BiLSTM-CRF模型在命名实体识别任务上取得了良好的效果。Ma和Hovy(2016)提出了BiLSTM-CNN-CRF模型用于实体识别，首先利用CNN获取句子的词级别特征，然后利用BiLSTM获取句子的时序依赖特征，并使用CRF对实体识别结果进行优化。面对模型无法同时关注字词特征的缺陷，Zhang和Yang(2018)提出了Lattice-LSTM模型，该模型通过引入分词结果信息，增加模型输入层的特征信息量；接着利用LSTM提取字词融合信息的隐藏时序特征，进而提升命名实体识别效果。Zhang等人(2019)也提出了一种基于词汇形式的命名实体识别模型，通过结合词级别和字符级别特征，提升了模型在中文命名实体识别任务上的性能。为准确对特定类实体进行准确识别，尼扎木丁等人(2017)使用统计规则对维族人名进行了研究，并获得了优异的命名实体识别结果。马合木提等人(2017)提出了一种基于模糊匹配和语音转换的命名实体识别方法，通过模糊匹配和结合语音模态信息进行实体识别，实验表明，该方法能够有效地识别文本中的命名实体。

由于传统的神经网络模型不能很好地表示句子的语义特征，基于大规模语料库的预训练语言模型应运而生。近年来，随着预训练语言模型的提出，命名实体识别的研究也进入到基于预训练语言模型的时代。廖列法(2023)提出一种基于注意力机制和特征融合的实体识别模型，借助BERT模型的语义表示能力获得了令人满意的实体识别结果。Xu和Li(2021)在生物医学领域的命名实体识别任务中，提出了BERT-BiLSTM-CRF模型，该模型通过关注领域的关键信息，从而获得更好实体识别效果。Li等人(2022)在ALBERT、BiGRU和CRF模型的帮助下，通过预训练语言模型增强句子的信息表达，并使用BiGRU关注文本中的长距离依赖，结合CRF获得更准确的实体序列标签，最终在MARS数据集上取得了良好的性能。郜成胜等人(2020)提出一种基于混合神经网络的命名实体识别方法，通过引入多种深度学习结构来构建命名实体识别模

型，并使用联合任务解码方式，解码得到命名实体识别结果。为了解决当前模型无法从多个角度挖掘更深层次特征的问题，Li和Meng(2021)通过拆解汉字并添加拼写信息来丰富模型的输入，以便模型可以从多个维度提取特征。相关研究人员发现，引入外部知识有助于提升中文命名实体识别的效果。在此基础上，Hu等人(2022)引入了基于BERT模型的知识库实体增强概念，通过结合知识库信息，强化实体边界概念，并在中文命名实体识别任务中取得了良好的性能。Liu等人(2021)提出将外部词典信息添加到BERT模型中，以丰富模型的输入特征，从而获得更好的识别效果。

上述研究工作为古籍实体识别任务带来了新的思路。然而，这些模型往往容易出现实体边界识别不准确和抗干扰能力差，且无法高效关注古籍文本特征的问题。此外，一些模型还依赖于分词结果，这需要额外的分词模型。因此，这些模型很难部署，无法在专业领域广泛使用。为推动古籍命名实体识别研究工作，本文在构建抗干扰能力强和能有效关注关键信息的模型之外，还提出一套能供模型继续预训练的古籍文本数据。

3 基于增量预训练与对抗学习的古籍命名实体识别方法

在古籍文本实体识别方法的构建上，为了更好地关注古籍文本中的关键信息，对其进行更为有效地编码，并准确区分各个实体之间的边界，使用NEZHA-TCN-GP模型进行古籍命名实体识别；为了将模型语义表达能力迁移到古籍文本领域，适应古文文本表达习惯，提出一种基于古籍数据的预训练方法，通过搜集大量古文文本并进行数据处理，实现NEZHA-TCN模型的预训练任务；同时为了增强模型的泛化能力，提出使用对抗学习思路用于NEZHA-TCN-GP模型的训练；并在最后结合规则处理方法，将一些常见的古籍书名和官职名加入规则库，最后得到古籍文本实体识别结果。

3.1 NEZHA-TCN-GP模型

为了更好地对文本进行编码，在选取基线模型时，使用NEZHA-Chinese-Base模型（后面称为NEZHA模型）对古籍文本进行词向量的获取，同时在模型最后一层加上两层的时序卷积神经网络，用于挖掘潜藏在古籍文本中的局部时序关联语义信息，提升模型对句子特征的编码能力。为了提升模型的抗干扰能力，使用FGM在词嵌入层添加干扰信息。同时在解码层上使用全局指针网络对文本中的实体进行位置解码，最后解码得到实体信息。相关模型的架构如图1所示:

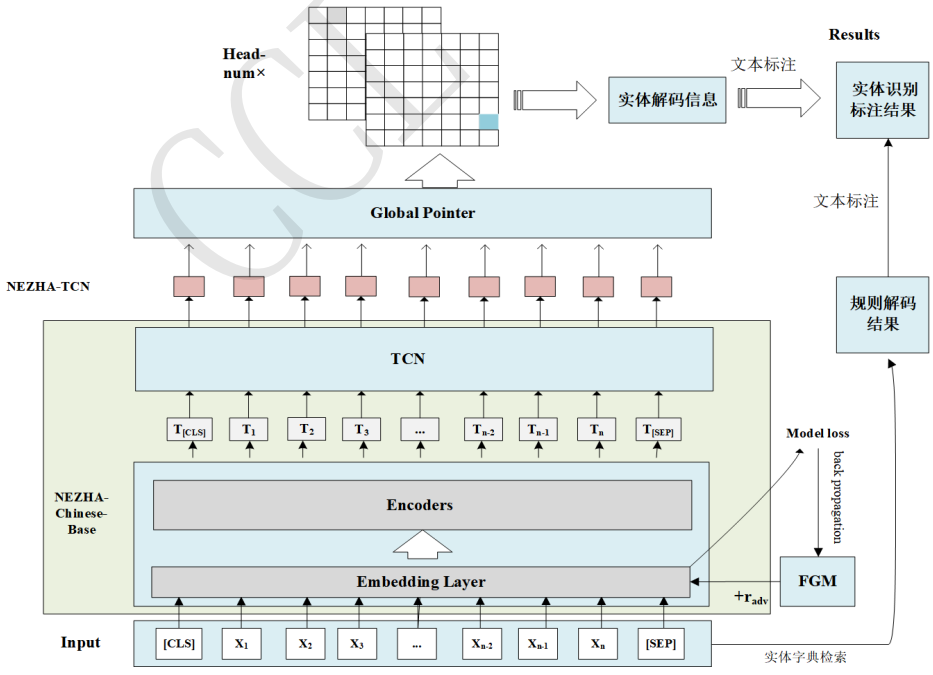

Figure 1: 模型整体架构图

在模型的编码层使用NEZHA模型对相应的古籍文本进行编码处理。与BERT模型不

同，NEZHA模型使用相对位置编码词向量，该方式能够让模型更好地挖掘文本中的字符关联信息。通过使用相对位置的正弦函数计算输出和attention的得分。该想法源于Transformer中使用的函数式绝对位置编码。在训练时通过引入混合精度训练方式进行模型的训练，完成预训练过程的加速。

值得注意的是，为了更好地关注古籍文本中的潜藏的局部特征和时序关系，我们在NEZHA模型的后面加上了两层时序卷积网络，该网络通过膨胀式卷积神经网络对文本中的特征信息进行增强关注，TCN 模型以CNN 模型为基础，并有如下两个特点：

- 序列建模: 传统的卷积神经网络并不能关注潜藏在文本中的时序信息，导致模型对文本时序关系建模能力差。而TCN模型通过设计时序卷积模块，关注文本中的时序信息，增强网络的时序信息建模，能够深层次地挖掘文本中的关联信息。
- 历史记忆: 时序卷积神经网络通过使用空洞卷积和残差模块完成网络的建模，让模型能够提升长时序文本建模能力，关注时序跨度大的关联关系信息，从而提升模型的性能。

同时时序卷积神经网络支持并行计算。与在RNN 中对后续时间步的预测必须等待其前任完成的情况不同，卷积可以并行完成，因为每一层都使用相同的滤波器。因此，在训练和评估中，长输入序列可以在TCN 中作为一个整体进行处理，而不是像在RNN 中那样按顺序处理。TCN还具有更大的局部感受视野，TCN 可以通过多种方式改变其感受野大小。TCN模型的构造如图2所示。

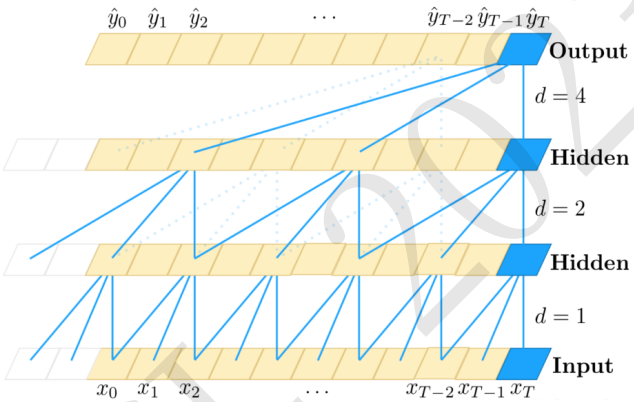

Figure 2: TCN模型结构图

为了更好地识别实体的边界，本模型使用全局指针识别句中实体。与使用CRF作为解码层的模型不同，全局指针将实体识别任务建模为子串提取任务，这种建模方式可以更准确地识别实体信息。对于长度为n的句子，句子中连续片段的最大数量为n(n+1)/2。然后，模型需要从这些片段中选择实体。假设句子中实体总数为k，实体类别数为m。全局指针可以将任务建模为在句子中选择k个实体并对每个实体进行m分类的任务。因此，对于句子：“迈尔万出生于中国。”，可以维护一个维度数为[Num-head, L, L]的矩阵，其中Num-head表示实体类别总数，L为句子的长度。全局指针旨在从上述句子中提取“迈尔万”与“中国”，并将其识别为名称与位置实体。对于上述句子，其包含名称实体“迈尔万”和位置实体“中国”。对于CRF方法，句子的标签解码过程可使用图3表示；对于全局指针模型，句子中的实体信息可使用二维矩阵进行表示，如图4所示。

在图3中，命名实体识别任务被建模为一个标签序列预测任务，并利用CRF获得概率最大的预测标签序列。在图3中，深黄色和深蓝色的部分表示句中实体。句子包含两个实体类别，其中Num-head值为2，每个头代表一种实体类别。因此，对于句子中的一个实体，当起始位置为i，结束位置为j时。坐标(i, j)位置用“1”标记，其他位置使用“0”标记。

与CRF相比，全局指针可以规避字符级别的标签错误。此外，全局指针可以更准确地识别实体的边界。在全局指针层前，通过模型编码层和对抗学习层已经得到了经过扰动的句子编码信息，可表示为$H = [h_1, h_2, ..., h_n]$。对于每一个词向量，其经过全连接层，可得到$q_{i,c}$和$k_{j,c}$，其计算方式可以使用公式(1)和(2)表示。

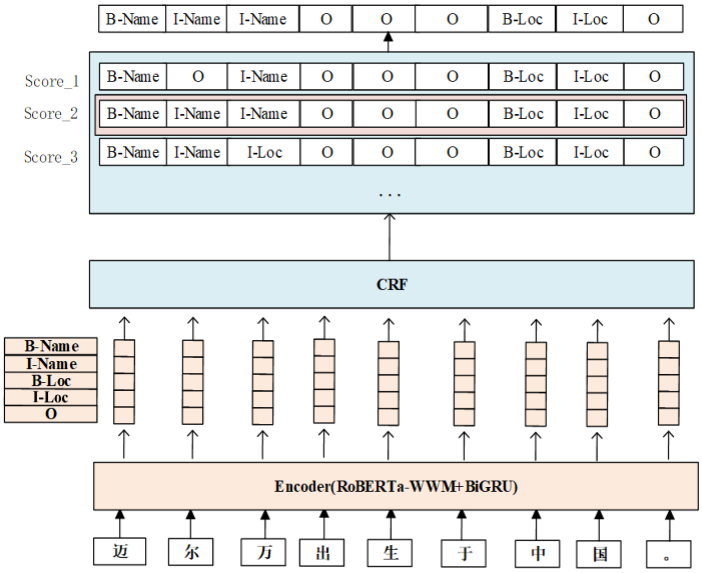

Figure 3: 基于CRF的实体识别解码结构

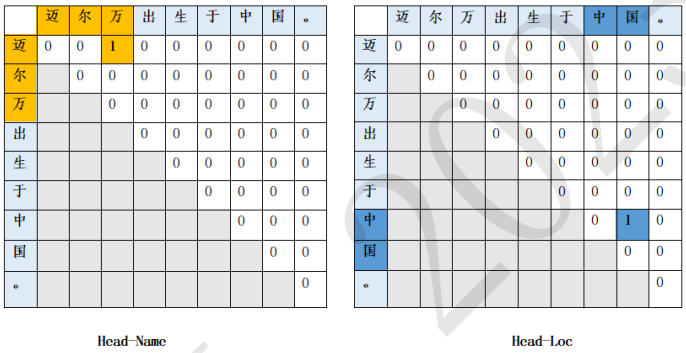

Figure 4: 基于全局指针的实体识别解码结构

$$q_{i,c} = w_{q,c}h_i + b_{q,c} \tag{1}$$

$$k_{j,c} = w_{k,c}h_j + b_{k,c} \tag{2}$$

上述式子中，$q_{i,c}$和$k_{j,c}$用于全局指针评分函数的构造，c表示某一实体类别，$w_{q,c}$和$w_{k,c}$表示权重参数，$b_{q,c}$和$b_{k,c}$表示偏置参数。根据公式(1)和(2)进行评分函数的构造，如公式(3)所示。

$$s_c(i,j) = q_{i,c}^T k_{j,c} \tag{3}$$

式(3)中，S_c表示句子中位置i到位置j字符为实体类型c的分数。为了关注句子中各词位置信息，使用RoPE(Shaw et al., 2018)显式地添加位置信息。RoPE是一个变换矩阵，其计算方式满足方程: ，因此可以在评分函数中显式地添加位置信息得到式(4)。

$$s_c(i,j)' = (R_i q_{i,c})^T (R_j k_{j,c}) = q_{i,c}^T R_i^T R_j k_{j,c} = q_{i,c}^T R_{j-i} k_{j,c} \tag{4}$$

公式(4)中，表示添加位置信息后的评分函数，表示位置编码矩阵，和表示词向量经线性变换后可用于评分的输出。

同时，添加位置信息的评分函数，可用于评估句子中对应位置实体属于c类型的分数。因此，全局指针模型的损失函数可以使用评分函数进行构造，计算方式如公式(5)。

$$loss = log(1 + \sum_{(i,j)\in P_c} e^{-s_c(i,j)}) + log(1 + \sum_{(i,j)\in Q_c} e^{s_c(i,j)}) \tag{5}$$

在式(5)中，只需要考虑i_j=j的情况，且公式满足条件(6)和(7)。

$$\Omega = (i,j)|1 \le i \le j \le n \tag{6}$$

$$Q_c = \Omega - P_c \tag{7}$$

在公式(6)和(7)中，i和j代表实体在句子中的起始和结束位置，n代表句子的长度，P_c代表实体集合，Q_c表示实体类型不是c的实体集合。

3.2 基于古籍数据的预训练

为了让模型更好地对古籍文本进行语义表达，本文构建了一套古籍文本，为领域内数据扩充提供支持；同时为了强化模型在古籍文本上的字符级映射能力，本文提出一种结合MLM预训练方式的NEZHA-TCN模型，该模型使用NEZHA-Chinese-Base作为基础模型，并使用时序卷积神经网络充分挖掘潜藏在古籍文本中的字符级别关联关系。

3.2.1 古籍数据的获取与处理

为了更好地将模型参数微调至古籍文本领域，本文搜集了大量的古籍文本，用于NEZHA-TCN模型的预训练。文本数据包含24史中的所有文本信息。由于古籍命名实体识别任务的文本长度基本都是在100左右，且最大长度不超过128。因此将相关的文本信息进行处理，按照逗号、句号等信息将相关的文本切分为长度大于20小于128的长度，这样能使模型更好的学习古籍文本间的关联信息。相关数据处理流程如图5所示。

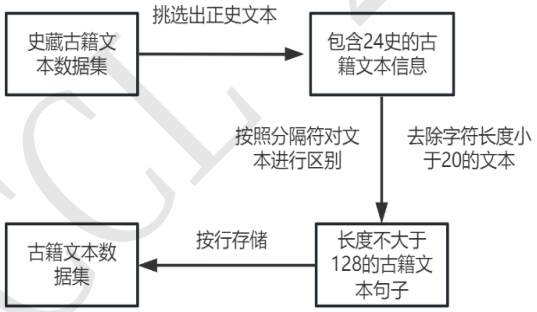

Figure 5: 古籍文本预训练数据处理流程

从图5中可以看到，本文搜集的古籍文本数据需要经过切分拼接处理，并将相关的数据处理成与训练数据类似的格式，即保持字符的繁体形式和长度特征，古籍数据集相关信息如下表所示。

从表中可以看出，本文提出的古籍数据集，一共包含正史文本中的24部书籍，同时文本被处理成长度接近于100个字符的繁体中文文本。在处理过程中，我们还舍弃了长度值小于20个字符的文本，防止文本过短带来模型性能的影响。数据集一共包含近40万条文本，各条文本按书籍内出现顺序排序。

在数据处理过程中，由于提供的训练数据为繁体字，在实验过程中，我们发现直接使用繁体字进行模型的训练和预测较简体字效果要好，因此在模型预训练过程中，我们将相关的简体字转化为繁体字进行NEZHA-TCN模型的预训练。在预训练过程中，使用正常单字掩码方式进行模型的预训练，并将掩码概率设置为15%，使得模型能够更好地关注文本间的信息。

古籍文本参数信息	数据描述
数据集大小	87M
文本最小长度	20字符
数据总量	397995条
文本来源	24史文本
最大长度	128字符
词典字符数量	21128

Table 1: 古籍预训练数据文本信息

3.2.2 古籍数据的预训练

在对NEZHA-TCN模型进行预训练时，使用单字掩码的方式进行字符的掩码操作，在古文句子中随机选定15%的字符，在选定的字符中，80%的字符被替换为“[MASK]”，10%的字符被随机替换为其他单词，其余10%保持不变。实验结果表明，沿用BERT预训练方法中的掩码机制可以提高模型的泛化能力和句子语义建模能力。

在预训练过程中，古籍文本总数将近40万条，batch-size设置为32，文本最大长度为128，并保存模型训练到第500,000轮次时使用模型进行下游任务的精调处理。更加具体的预训练参数如下表所示。

预训练参数	参数值
掩码概率	15%
掩码方式	单字掩码
随机数种子	42
批尺寸大小	32
学习率	5e-5
最大长度	128
训练步数	500,000

Table 2: 实验环境配置

3.3 对抗学习

为了提高模型的抗干扰能力，使用快速梯度法(Fast Gradient Method，FGM)(Miyato et al., 2016)在模型训练过程中添加干扰。FGM可以获得更好的对抗样本，提高模型的性能。使用FGM进行模型训练，其过程包括两个步骤:

- 最大化内部损失函数值: 为了在模型训练过程中往损失值增加的方向引入扰动，并在优化空间中找到最大的影响函数，内部损失函数值应最大化。
- 最小化任务判别损失函数值：在对模型添加干扰后，模型的输出分布也能与原始分布保持一致。因此，在上述最大化内部损失函数值的情况下，该模型在外部需要找到最优的模型参数，任务判别损失函数应最小化。

对抗训练过程中的最小-最大公式描述如公式(8)所示。

$$\min_{\theta} E_{(x,y)\sim D}[\max_{r_{adv}\in S} L(\theta, x + r_{adv}, y)] \tag{8}$$

在式子(8)中，表示模型的参数。E表示在对抗性学习过程中的期望值。D表示数据集信息。L表示被扰动的神经网络的损失函数。x和y分别表示输入和输出。r_{adv}表示对模型添加的扰动，S表示扰动空间。对于FGM算法，输入数据的梯度可表示为式(9)。在对抗学习过程中，扰动值的计算可用公式(10)表示。

$$g = \nabla_x L(\alpha, x, y) \tag{9}$$

$$r_{adv0} = \alpha sgn(g) \tag{10}$$

式(10)中，表示添加扰动的概率，sgn为阶跃函数。与快速梯度下降法(Fast Gradient Sign Method，FGSM)不同，FGM模型添加的扰动信息使用梯度二范数进行计算，这可以让模型得到更好的泛化能力。FGM的扰动值计算方式如式(11)所示。

$$r_{adv} = \alpha g / \|g\|_2 \tag{11}$$

式(11)中，α表示添加扰动的概率，r_{adv}表示所添加的扰动量。为了更好地解释模型中FGM的计算流程，可使用图6对模型训练过程如何添加相应的扰动进行描述。

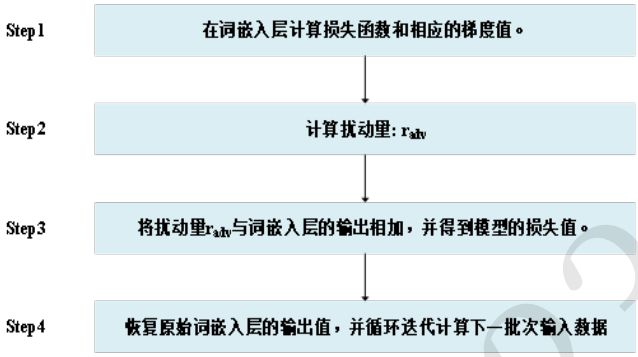

Figure 6: FGM对抗训练流程图

在对模型编码和相应的扰动量的使用下，模型在词嵌入层得到了经过添加扰动量的词向量信息。在模型编码层中经过embedding层后的词向量信息可以使用公式12表示。

$$h_{ea} = h_e + h_{adv} \tag{12}$$

在式(12)中，h_{ea}表示添加扰动后的词向量表示，h_e表示经过词嵌入层得到的词向量输出，r_{adv}表示添加的扰动量。接着将得到的词向量h_{ea}送入到模型编码层中，得到经过关键信息增强和语义信息优化的字符编码向量$H = [h_1, h_2, ..., h_n]$。

3.4 结合规则

在研究中，我们发现，模型中存在一些常见的预测错误，比如漏标，错标的情况出现，为了更好地整理预测结果，我们结合相应的规则进行结果的矫正输出。由于测试集数据量少，此种方法会有一定的效果。

同时在模型的预测中，我们还注意到本文提出的模型会无差别地识别嵌套实体和非嵌套实体，而在真实数据中，没有嵌套实体地出现，因此在数据输出处理时，我们使用如下规则选择嵌套实体中的某一个实体，保证模型能够以最大概率输出，得到最好的结果。嵌套实体的留取规则为：将所有实体按照实体初始位置进行升序排序，按照实体结束位置进行降序排序，去除后续嵌套的实体。

4 实验

4.1 实验数据集

本文主要使用古籍命名实体识别数据集(苏祺 et al., 2023)进行模型的微调，在去除提供训练数据集中不存在实体的句子后，数据集一共包含2137文本，3种类型实体，分别为：BOOK（书籍名）、PER（人名）和OFI（官职名称）。三类实体分布不均衡，其中BOOK类型实体最少。在训练数据处理过程中，使用BIO方式对数据进行标注，其中“B”表示实体开头字符，“I”表示实体非开头字符，“O”表示非实体元素。为了更好地验证本文提出模

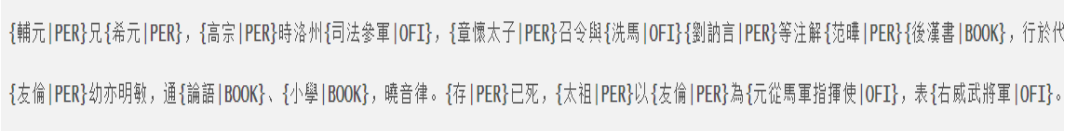

Figure 7: 数据标注示例图

型的有效性，在实验中，随机选取前2000条数据样本进行训练，后137条数据作为验证集，测试集数据224条。该数据集的原始标注情况如图7所示。

从图7中可以看到，在原始标注信息中使用大括号将相应的实体进行标注，并使用相关分隔符标识实体的类型。同时在训练数据中，所有出现的实体没有嵌套情况出现，因此在后续实体解码过程中，可以直接将嵌套实体进行规则化处理。

4.2 评价指标

在实验中，选择Micro-F1作为评价模型性能的主要指标，并在文中将其记为F1值。同时使用Recall和Precision作为辅助指标查看模型的效果，相关指标的计算方式如式(13)、(14)和(15)所示。

$$Recall = \frac{|S \cap G|}{|G|} * 100\% \tag{13}$$

$$Precision = \frac{|S \cap G|}{|S|} * 100\% \tag{14}$$

$$F_1 = \frac{2 * Recall * Precision}{Recall + Precision} * 100\% \tag{15}$$

在上述公式中，G 表示数据集中所有实体集合，可以表示为$G = \{g_1, g_2, g_3, ..., g_n\}$. S 表示模型预测实体的结果集合，可以表示为$S = \{s_1, s_2, s_3, ..., s_n\}$. 任意一个元素$G$ and S 包含实体和相应的实体类型。

4.3 实验环境与参数

古籍命名实体识别研究使用Linux系统进行实验，同时使用Python编程语言进行模型代码编写，计算资源为GPU，显存大小为16g，更加具体的实验环境如下表所示。

环境名称	参数值
操作系统	Linux
编程语言	Python3.7
CPU	i5-9300h
内存大小	16g
GPU型号	GeForce RTX 2080 Ti
Pytorch	1.10.0
Transformers	4.9.2

Table 3: 实验环境配置

从上表中可以看出，本实验使用PyTorch深度学习框架进行模型的训练与测试，同时结合第三方资源库Transformers进行预训练语言模型框架代码的开发。在本节中，各数据集上的模型参数配置如下表所示。

从表中可以看出，Batch-size设置为48，Max-len为128，Num-head表示各个数据集实体类别总数。模型使用分段学习率进行参数调整，预训练模型部分使用微调策略进行训练，学习率为5e-5，全局指针层使用更大的学习率进行参数调整，学习率为1e-3。在模型训练中，对抗学习的扰动概率为0.25。

参数名称	参数值
Max-len	128
Batch-size	48
Type-num	3
Learning rate 1	5e-5
Learning rate 2	1e-3
α	0.25
随机数种子	42

Table 4: 模型参数

4.4 实验结果分析

本节主要对本文提出的方法进行实验验证。在实验中使用相关数据集进行消融实验，以验证本文提出的模型相较于其他模型的优势。消融实验的具体参数有：LSTM、TCN、预训练、对抗学习、结合规则。因此在基线模型的选取上，将NEZHA-Chinese-Base+全局指针模型作为基线模型，后续将其称为NEZHA。各种模型在测试集上的F1值如下表所示。

模型名称	F1（%）
NEZHA	91.15
NEZHA-LSTM	91.90
NEZHA-TCN	92.32
Nezha-TCN+预训练	93.35
Nezha-TCN+预训练+对抗学习	94.22
ANER-IPAL	95.34

Table 5: 模型消融实验结果

从上表中可以看到，基线模型使用NEZHA-Chinese-Base模型作为词向量编码器，并结合使用全局指针对文本中的实体进行预测，其F1值也达到了90%以上，这表明:NEZHA-Chinese-Base在本任务上有着较好的语义表征能力，全局指针解码器也能很好的解决古籍命名实体识别任务。在模型加上TCN模块后，模型能够有效挖掘文本中的语义关联信息，较基线模型在F1值上提升了1.17%。同时在本文提出的古籍数据上进行继续预训练能够有效提升模型对古文文本的实体识别效果，F1值较无增量预训练的方法也有所提升。在训练过程加上对抗学习能够有效提升模型的泛化性能，在训练数据较少的情况下提升较为明显。消融实验结果表明，使用TCN、预训练、对抗学习和结合规则这几种方法都能够有效提升古籍命名实体识别的预测效果，且最后的F1值为95.34%。

5 总结

本文从算法构建和数据出发，不仅为古籍文本领域构建了一套可用于古籍文本预训练的数据，还构建了一整套用于古籍命名实体识别研究的算法。实验结果表明，本文提出的方法能够有效地将预训练语言模型的能力进行场景迁移，同时还能够有效且稳定地关注古籍文本中的关键特征信息，对提升古籍文本实体识别准确率有较好的效果。

致谢

感谢北京信息科技大学智能信息处理研究所对本工作的支持。

参考文献

苏祺,王莹莹,邓泽琨,杨浩,王军. 2023. CCL23-Eval 任务1总结报告: 古籍命名实体识别(GuNER2023).

Hongkui Y, Huaping Z, Qun L. 2006. Chinese Named Entity Recognition Based on Cascading Hidden Markov Model. *Journal of communication*, 27(2):87–94.

Duan H, Zheng Y. 2011. A Study on Features of The Crfs-Based Chinese Named Entity Recognition. *International Journal of Advanced Intelligence*, 3(2):287–294.

Huang Z, Xu W, Yu K. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv preprint*, arXiv:1508.01991.

Ma X, Hovy E. 2016. End-To-End Sequence Labeling via Bi-Directional LSTM-CNNs-CRF. *arXiv preprint*, arXiv:1603.01354.

Y Zhang, J Yang. 2018. Chinese NER Using Lattice LSTM. *arXiv preprint*, arXiv:1603.01354.

Y Zhang, J Yang. 2019. Chinese Named Entity Recognition Augmented with Lexicon Memory. *arXiv preprint*, arXiv:1912.08282.

塔什甫拉提·尼扎木丁,汪昆,艾斯卡尔·艾木都拉. 2017. 统计与规则相结合的维吾尔语人名识别方法. 自动化学报, 43(04):653–664.

热合木·马合木提,于斯音·于苏普,张家俊. 2017. 基于模糊匹配与音字转换的维吾尔语人名识别. 清华大学学报*(*自然科学版*)*, 57(02):188–196.

廖列法,谢树松. 2023. 基于注意力机制特征融合的中文命名实体识别. 计算机工程, 1-10[2023-03-11]. DOI:10.19678/j.issn.1000-3428.0064432.

Xu L, Li J. 2021. Biomedical Named Entity Recognition Based on BERT and BiLSTM-CRF. *Computer Engineering and Science*, 43(10):1873–1879.

Junhuai L, Miaomiao C, Huaijun W, et al. 2022. Chinese Named Entity Recognition Method Based on ALBERT-BGRU-CRF. *Computer Engineering*, 48(06):89–94.

郜成胜,张君福,李伟平. 2020. 一种基于混合神经网络的命名实体识别与共指消解联合模型. 电子学报, 48(03):442-448.

Li J, Meng K. 2021. MFE-NER: Multi-Feature Fusion Embedding for Chinese Named Entity Recognition. *arXiv preprint*, arXiv:2109.07877.

Hu J, Hu Y, Liu M, et al. 2021. Chinese Named Entity Recognition Based on Knowledge Base Entity Enhanced BERT Model. *Journal of Computer Applications*, 42(9):2680–2685.

Liu W, Fu X, Zhang Y, et al. 2021. Lexicon Enhanced Chinese Sequence Labeling Using BERT Adapter. *arXiv preprint*, arXiv:2105.07148.

Shaw P, Uszkoreit J, Vaswani A. 2018. Self-Attention with Relative Position Representations. *arXiv preprint*, arXiv:1803.02155.

Miyato T, Dai A M, Goodfellow I. 2016. Adversarial Training Methods for Semi-Supervised Text Classification. *arXiv preprint*, arXiv:1605.07725.