CCL23-Eval 任务1系统报告：基于信息论约束及篇章信息的古籍命名实体识别

张兴华, 刘天昀, 张文源, 柳厅文
中国科学院信息工程研究所
中国科学院大学网络空间安全学院
{zhangxinghua, liutianyun, zhangwenyuan, liutingwen}@iie.ac.cn

摘要

命名实体识别旨在自动识别出文本中具有特定意义的实体（例如，人名、地名），古籍文献中的命名实体识别通过识别人名、书籍、官职等实体，为深度挖掘、组织古汉语人文知识提供重要支撑。现有的中文命名实体识别方法主要聚焦在现代文，但古籍中的实体识别具有更大的挑战，表现在实体的**歧义性**和**边界模糊性**两方面。由于古籍行文简练，单字表达加剧了实体的歧义性问题，句读及分词断句难度的提升使实体边界的识别更具挑战性。为有效处理上述问题，本文提出一种基于信息论及篇章信息的古籍命名实体识别方法。通过检索古籍文本的来源信息融入篇章先验知识，并在同一篇章的古籍文本上采取滑动窗口采样增强，以引入篇章背景信息，有效缓解实体歧义性问题。此外，在信息论视角下，约束实体的上下文信息及实体本身特征的编码，最大程度保留泛化特征，去除冗余信息，缓解实体边界模糊的问题，在词义复杂多样、句读困难的古文典籍中提升命名实体识别性能。最终，在token-wise和span-level感知的命名实体识别基础框架下，本文的方法取得了最优的评测性能。

关键词： 古籍命名实体识别；实体歧义性；实体边界模糊性；信息论；篇章信息

System Report for CCL23-Eval Task 1: Information Theory Constraint and Paragraph based Classical Named Entity Recognition

Xinghua Zhang, Tianyun Liu, Wenyuan Zhang and Tingwen Liu
Institute of Information Engineering, Chinese Academy of Sciences
School of Cyber Security, University of Chinese Academy of Sciences
{zhangxinghua, liutianyun, zhangwenyuan, liutingwen}@iie.ac.cn

Abstract

Named entity recognition (NER) aims to automatically detect specific entity spans with predefined categories (e.g., *person*, *location*), and classical named entity recognition is the important premise to explore and organize classical Chinese humanistic knowledge by recognizing *person*, *book* and *official position* entity. Most existing Chinese named entity recognition methods focus on modern literature, but classical NER possesses significant challenges in entity ambiguity and boundary fuzziness. Due to the concise writing style in classical text, single word expression exacerbates the ambiguity problem, and the increased difficulty of sentences and phrases makes the entity boundaries more challenging. To solve above challenges, this paper proposes the information theory constraint and paragraph based classical NER framework. Specifically, we retrieve

第二十二届中国计算语言学大会论文集，第1页–第13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

{圣宗|PER}/诏/{白|PER}/鞫之，{白|PER}/正/其事。
{虏|PER}/绝/临洮/道，白水军/使/{高柬于|PER}/拒守，{虏|PER}/引去。
{齐王|OFI}/{宪|PER}/白/帝/曰：“{李安|PER}/出自/皂隶，所典/唯/庖厨/而已。
{玄|PER}/又/议/复/肉刑，{琳之|PER}/以为
{觊|PER}/子/{长公|PER}，{澡|PER}/二子/{淹|PER}、{玄|PER}/并在都，驰信/密报
悦性/冲玄，怡神/虚白，餐松/饵术，栖息/烟霞。
{崇成|PER}，本名/{灰|PER}，泰州/{司属司|OFI}/人，{昭祖|PER}/玄孙/也

Table 1: 古典书籍中实体标注样例：PER为人名，OFI表示官职名；单字“白”、“玄”在不同的上下文中具有多重含义，句读采用“/”分隔；句中的字已由繁转简

the source information of classical text to inject the paragraph prior knowledge, and perform the data augmentation via sliding window on text from the same paragraph, introducing background knowledge and alleviating the entity ambiguity issue. In addition, we constrain the feature encoding of entity context and surface name from the perspective of information theory to maximize the general features and reduce redundant information, mitigating the entity boundary fuzziness. Experimental results show that our method based on token-wise and span-level aware NER framework achieves the best performance in classical NER.

Keywords: Classical Named Entity Recognition , Entity Ambiguity , Entity Boundary Fuzziness , Information Theory , Paragraph Information

1 引言

命名实体识别（Name Entity Recognition）任务旨在自动识别出文本中人名、地名、机构名等事件基本构成要素的重要实体。作为一项重要的信息抽取任务，在信息检索 (Fetahu, 2021; Guo, 2009; Mokhtari, 2019; Zhang, 2021; Zhang, 2022)、问答 (Li, 2019; Longpre, 2021)等任务中具有重要意义。古籍文献的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。近年来，学界已有多项研究关注史籍、方志、诗词、中医等类目的古籍命名实体识别，构建了一些针对垂直领域的小型标注数据集，实体标注的体系和规范有所差异，识别范围往往由三种基本实体类别扩充至人文计算研究所需的多种特殊类别，如书名、药物名、疾病名、动植物名等。这些研究所构建针对特殊领域的小型标注数据集，实体类型有差异。总体而言，古籍命名实体识别任务仍旧缺乏可用于模型训练以及评测的公开数据资源，阻碍了技术的长足发展。另一方面，古文字词含义的多样性、行文结构的连续性以及多用繁体字、无句读等特点，也增加了古籍文献命名实体识别任务的复杂和困难程度。

因此，北京大学人工智能研究院和北京大学数字人文研究中心联合组织了古籍命名实体识别评测 (pku, 2023)，基于“二十四史”，设计了涵盖人名、书名、官职名的实体知识体系，构建了覆盖多个朝代的历时、跨领域的数据资源，完善古籍命名实体识别任务的建立。已有命名实体识别的架构大致可以分为基于序列标注 (Lample, 2016; Zhang, 2018; Devlin, 2019)、基于实体span (Tan, 2020; Fu, 2021)和基于文本生成 (Yan, 2021; Zhang, 2022)的框架。基于序列标注的框架将每个序列位置标注为一个标签，比如按照BIO标注，然后采用多层感知机MLP或条件随机场CRF进行解码；基于实体span的方法枚举所有可能的span进行实体分类；而基于文本生成的框架将命名实体识别形式化为文本生成任务，将句子中的目标实体转化为序列进行训练和解码。然而，古籍文本的行文风格相比现代文具有较大差异，古文字词含义的多样性（比如存在大量单字的表达，如表 1所示）以及句读的复杂性给古籍文本中的命名实体识别带来巨大挑战。已有的命名实体识别框架中哪种更适合古籍文本的实体识别任务，以及如何有效处理古籍实体的歧义性及句读困难带来的实体边界模糊问题值得进一步探究。

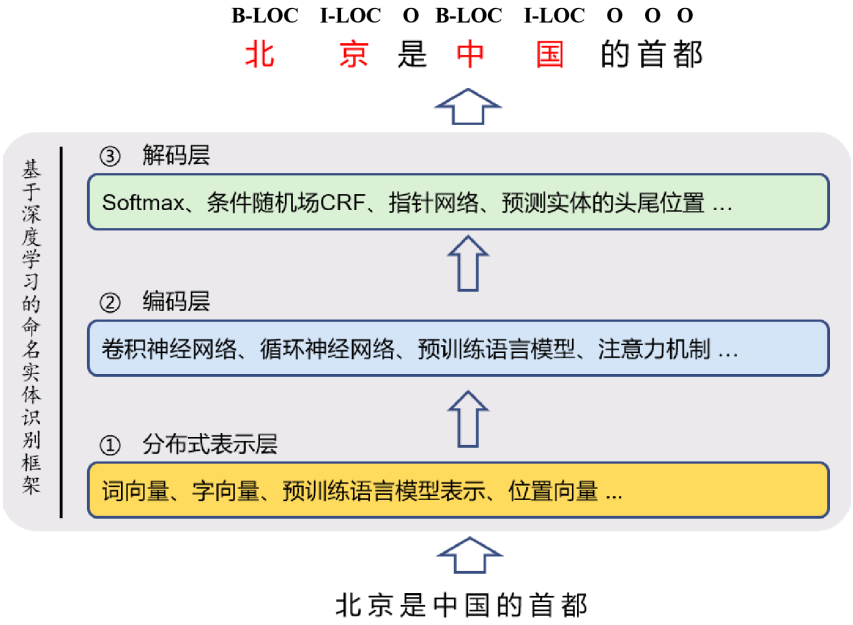

Figure 1: 基于深度学习的命名实体识别框架：分布式表示层、编码层和解码层

为此，本文探究了token-wise感知的序列标注框架和span-level感知的实体识别框架在古籍命名实体识别中的性能表现，并结合古籍的篇章信息，有效处理实体的歧义性问题，因为同一篇古文中的字词含义具有相对一致的表达，如表 1中的“{圣宗|PER}/诏/{白|PER}/鞫之，{白|PER}/正/其事。”，如果在上下文中出现“武白”，同时结合句子中的上下文及“白”的语义信息，更容易识别出上述句子中的人名(PER)实体“白”。与现代文相比，古籍文本的句读或者分词断句更加困难，为实体识别的边界确立带来很大挑战，如何能够动态地综合建模实体的上下文及实体本身的信息，对古籍实体识别具有重要意义。因此，我们从信息论的视角，尽可能保留句子中对识别实体有用的信息，并同时去除冗余信息，以更好地兼顾上下文及实体本身的信息。最终，我们提出的系统框架在古籍命名实体识别任务上取得了最优的性能。

2 方法

为了有效处理古籍文本中实体的歧义性及边界模糊性的问题，本文的研究方法主要包括两种经典的实体识别框架：**Token-wise**感知的序列标注和**Span-level**感知的实体识别框架，并结合古籍的篇章信息有效缓解实体的歧义性问题，在信息论的指导下建立一个可动态考虑实体指称本身及其上下文信息的优化目标，从而更好地建模实体的边界信息，有效应对古籍中分词断句难度大造成的实体边界模糊的问题。

如图1所示，命名实体识别模型整体可以分为3层：分布式表示层、编码层和解码层。分布式表示层用于将文本转化为Embedding向量表示，编码层将Embedding向量经过多层神经网络映射到隐层，得到文本的隐向量表示，用于解码层解码得到每个Token的实体标签。

2.1 分布式表示层&编码层

在预训练语言模型的时代，我们选择BERT系列的语言模型，完成分布式表示层和编码层，可以形式化表示为：

$$\mathbf{H} = \text{BERT}(X) \tag{1}$$

其中，$X = <x_1, x_2, ..., x_n>$表示一个长度为n的句子，$\mathbf{H} = <\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n> \in \mathbb{R}^{n \times d}$为最后一层隐向量表示。

Token-wise感知的序列标注框架（2.2节）和Span-level感知的实体识别框架（2.3节）在分布式表示层和编码层具有相同的结构，而在解码层略有差异，本文将分别在其对应章节进行简要介绍。

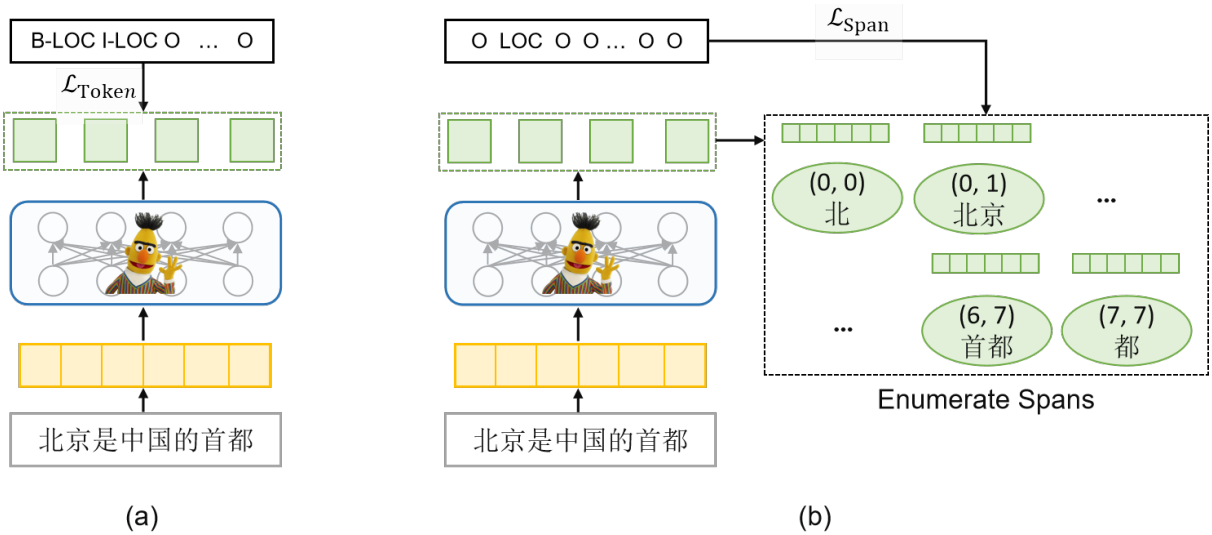

Figure 2: (a) Token-wise感知的序列标注　(b) Span-level感知的实体识别

2.2 Token-wise感知的序列标注框架-解码层

命名实体识别的解码方式有很多种，包括CRF、Softmax、指针网络、头尾实体预测等，在token-wise结构中，我们选择Softmax进行解码。因为BERT系列的预训练语言模型已经学习到足够的句子单元之间的依赖关系，所以BERT+Softmax的精度与BERT+CRF相当，而且推理速度更快。对于指针网络等在BERT上继续引入其它的神经网络结构，从以往的经验来看，调参难度较大，并且收效较小。采用token-wise的序列标注模型结构如图2 (a)所示，针对得到的隐层序列表示$\mathbf{H} = <\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n> \in \mathbb{R}^{n \times d}$，句子中每个字(token) x_i 的组合标签（例如，B-LOC）概率分布为：

$$p(\mathcal{C}^k|x_i) = \frac{\exp\{\mathbf{w}_k^\top \mathbf{h}_i + b_k\}}{\sum_{j=1}^{|\mathcal{C}|} \exp\{\mathbf{w}_j^\top \mathbf{h}_i + b_j\}} \tag{2}$$

其中组合标签指的是实体边界标识集{B, I, O}与实体类型标签集{PER, LOC, ORG}的组合$\mathcal{C}$={B-PER, I-PER, B-LOC, I-LOC, B-ORG, I-ORG, O}。$[\mathbf{w}_k; b_k]$表示第k个组合标签的分类头参数，$p(\mathcal{C}^k|x_i)$表示字(token) x_i 属于第k个类别的概率。其最终的优化目标基于交叉熵损失进行计算：

$$\mathcal{L}_{\text{Token}} = -\frac{1}{|\mathcal{D}|} \sum_{X_i \in \mathcal{D}} \sum_{x_j \in X_i} \sum_{k=1}^{|\mathcal{C}|} y_{j,k} \log(p(\mathcal{C}^k|x_j)) \tag{3}$$

其中$y_{j,k}$为y_j中的第j个元素，y_j为token x_j的one-hot标签，$\mathcal{D}$为训练语料。

2.3 Span-level感知的实体识别框架-解码层

在编码得到句子中每个token的隐层序列表示$\mathbf{H} = <\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n> \in \mathbb{R}^{n \times d}$之后，本文穷举句子中的所有满足最大实体长度限制的实体span $S = s_1, s_2, ..., s_m$，然后为每个span赋予实体的类别标签$t \in \mathcal{T}$ ={PER, LOC, ORG, O}。以“北京是中国的首都”为例，最大实体长度限制为2，那么穷举的实体span集合S ={北, 北京, 京, 京是, ..., 首都, 都}，以及对应的标签Y ={O, LOC, O, O, ..., O, O}。

每个实体span的表示由两部分组成：**边界表示**和**span长度表示**。第i个实体span的边界表示由实体的开始token表示$\mathbf{h}_i^s$和结束token表示$\mathbf{h}_i^t$拼接得到：$\mathbf{h}_i^{bd} = [\mathbf{h}_i^s; \mathbf{h}_i^t]$；span长度表示为了编码实体的长度信息，每个实体长度len_i均对应一个与$\mathbf{h}_j$维度相同的可优化的向量表示$\mathbf{l}_i^{len}$。最终每个实体span的表示通过拼接两部分的表示得到：$\mathbf{h}_i^{span} = [\mathbf{h}_i^{bd}; \mathbf{l}_i^{len}]$。那么句子中每个span的实

Figure 3: 检索融入古籍文本的来源信息

体类型标签（例如，LOC）概率分布为：

$$p(\mathcal{T}^k|s_i) = \frac{\exp\{\mathbf{w}_k^\top \mathbf{h}_i^{span} + b_k\}}{\sum_{j=1}^{|\mathcal{T}|} \exp\{\mathbf{w}_j^\top \mathbf{h}_i^{span} + b_j\}} \tag{4}$$

其中$[\mathbf{w}_k; b_k]$表示第k个类别标签的分类头参数，$p(\mathcal{T}^k|s_i)$表示实体span s_i 属于第k个类别的概率。其最终的优化目标基于交叉熵损失进行计算：

$$\mathcal{L}_{\text{Span}} = -\frac{1}{|\mathcal{D}|} \sum_{X_i \in \mathcal{D}} \sum_{s_j \in S_i} \sum_{k=1}^{|\mathcal{T}|} y_{j,k} \log(p(\mathcal{T}^k|s_j)) \tag{5}$$

其中$y_{j,k}$为y_j中的第j个元素，y_j为span s_j的one-hot标签，$\mathcal{D}$为训练语料，S_i通过对句子X_i按照最大长度限制穷举span得到。

2.4 篇章信息利用

考虑到古籍中同一篇文章具有相对一致的表达，比如同一篇中相对集中地用单字人物缩写，并且上下文中会出现人物的全称，如表1中的人名"白"在上下文中会多次提到该人名，且会提及其全称"武白"，如果能够考虑篇章信息，对解决实体歧义性问题将有很大增益。因此，本文设计了两种数据增强的策略：融入篇章先验信息和篇章内滑动窗口采样。

融入篇章先验信息：为了将同一篇章内的文本进行关联，本文在中华经典古籍库[0]中去查找每句话的来源，如图3所示。通过将检索到的来源"新唐书>卷九十六>列传第二十一>杜如晦"拼接在原句子"帝曰："玄龄、如晦不以旧进，特其才可与治天下者，师合欲以此离间吾君臣邪？"斥岭表。"后面，引入篇章先验信息，隐式地增强篇章内的关联。增强后的文本采用token-wise感知的序列标注框架进行编码-解码，值得注意的是，拼接的来源对应的token不参与序列标注模型的训练和推理。

篇章内滑动窗口采样：通过前面步骤，可以将训练语料中的句子按照来源出处聚合成一个篇章，如图4所示，即来自同一来源的句子拼接合并在一起。通过设置最大的窗口长度以及移动的步长，在聚合的篇章中不断滑动窗口进行数据增强。通过在同一篇章文本上进行滑动增强，可以显式增加篇章内的关联，获得更加丰富的语义信息。增强后的文本同样采用token-wise感知的序列标注框架进行训练。

[0]https://publish.ancientbooks.cn/docShuju/platform.jspx

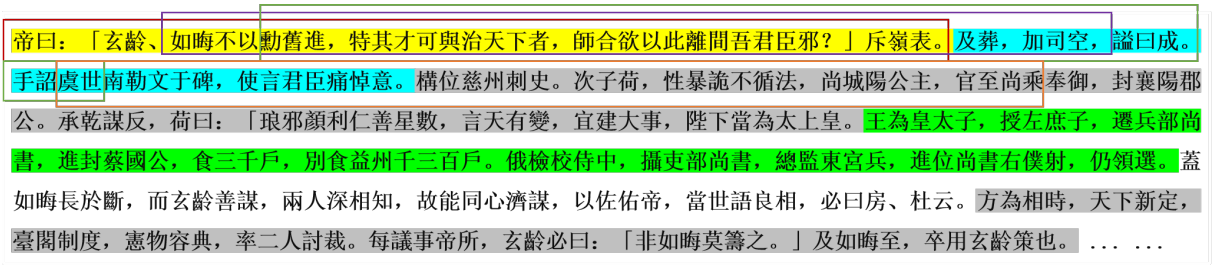

Figure 4: 篇章内跨句滑动窗口采样（部分示例），不同的句子用不同颜色标识

2.5 信息论视角下的实体识别

与现代文的行文风格相比，古籍文本的句读或者分词断句的难度更大，这为实体识别中实体边界的检测带来巨大挑战。因此，本工作采用基于span的框架，来建模实体span-level的信息，同时从信息论的视角显式地约束实体特征的表达：最大化实体上下文特征与实体surface name特征之间的互信息，增强泛化特征的编码；最小化冗余信息，防止模型过分记忆实体的surface name或句子中的某些偏置。

为此，考虑两个具有相同类型的实体s_1和s_2，同时具有相同的上下文，但其实体指称不同。那么s_1与其隐层状态表示$\mathbf{h}_1^{span}$的互信息$I(s_1;\mathbf{h}_1^{span})$可以通过链式法则分解得到：

$$I(s_1;\mathbf{h}_1^{span}) = \underbrace{I(\mathbf{h}_1^{span};s_2)}_{\text{泛化特征}} + \underbrace{I(s1;\mathbf{h}_1^{span}|s_2)}_{\text{冗余信息}} \tag{6}$$

其中$\mathbf{h}_1^{span}$和$\mathbf{h}_2^{span}$为实体span s_1和s_2的表示，$I(\mathbf{h}_1^{span};s_2)$表示非特定实体的信息，$I(s1;\mathbf{h}_1^{span}|s_2)$表示$s_1$中特有的实体信息，这部分信息并不能从$s_2$中得到。

因此，针对两个正样本样例来说，任何包含两个句子中所有共享信息的特征$\mathbf{h}^{span}$也一定会包含特定的标签信息，所以句子中的特定信息是冗余的，因此我们的优化目标应该是最大化共享信息$I(\mathbf{h}_1^{span};s_2)$，同时最小化句子特定信息$I(s1;\mathbf{h}_1^{span}|s_2)$。

如Wang (2022)证明，$I(\mathbf{h}_1^{span};s_2)$的下界是$I(\mathbf{h}_1^{span};\mathbf{h}_2^{span})$，因此最大化$I(\mathbf{h}_1^{span};\mathbf{h}_2^{span})$近似最大化$I(\mathbf{h}_1^{span};s_2)$，基于InfoNCE可以实现上述目标：

$$\mathcal{L}_{general} = -\frac{1}{|\mathbf{D}|}\sum_{j=1}^{m}\log\frac{\exp(sim(\mathbf{h}_1^{span},\mathbf{h}_2^{span})/\tau)}{\sum_{k=1}^{m}\exp(sim(\mathbf{h}_1^{span},\mathbf{h}_k^{span})/\tau)} \tag{7}$$

其中$\mathbf{D}$为整个构造的正负样本对语料（正样本通过实体指称的替换得到），m为每条样本的正负样本数量，τ为温度系数。同样地，$I(s1;\mathbf{h}_1^{span}|s_2)$的上界为：

$$\mathcal{L}_{specific} = \mathbb{E}_{s_1,s_2}\mathbb{E}_{\mathbf{h}_1^{span},\mathbf{h}_2^{span}}[D_{JS}[p(\mathbf{h}_1^{span}|s_1)||p(\mathbf{h}_2^{span}|s_2)]] \tag{8}$$

其中D_{JS}表示JS散度(Jensen-Shannon divergence)，$p(\mathbf{h}_1^{span}|s_1)$和$p(\mathbf{h}_2^{span}|s_2)$均通过均值和方差进行刻画（类似变分自编码器），$\mathcal{L}_{specific}$的目标是确保$\mathbf{h}^{span}$的不变性。

2.6 训练及推理

训练优化：本文提出的系统中，包含token-wise的序列标注模型$\mathcal{M}_{token}$及其融入篇章先验信息的模型$\mathcal{M}_{token_prior}$、滑动窗口采样模型$\mathcal{M}_{token_slide}$，信息论视角下的Span-level的模型$\mathcal{M}_{span_info}$。其中token-wise系列的模型$\mathcal{M}_{token}$、$\mathcal{M}_{token_prior}$和$\mathcal{M}_{token_slide}$的优化损失为公式3$\mathcal{L}_{\text{Token}}$，Span-level感知的模型$\mathcal{M}_{span_info}$优化损失为$\mathcal{L}_{\text{Span}} + \mathcal{L}_{\text{general}} + \mathcal{L}_{\text{specific}}$。

融合推理：考虑到两种模型架构不同的优势，本文保留$\mathcal{M}_{span_info}$（最大span长度设为10）预测的长度大于等于4的实体，而$\mathcal{M}_{token}$、$\mathcal{M}_{token_prior}$、$\mathcal{M}_{token_slide}$和$\mathcal{M}_{span_info}$（最大span长度设为4）集成后保留长度小于4的实体，最后将结果集成得到每条句子中最终的预测答案。

3 实验

3.1 数据集

训练集以“二十四史”为基础语料，包含13部书中的22卷语料，随机截断为长度约100字的片段，标注了人名（PER）、书名（BOOK）、官职名（OFI）三种实体，总计15.4万字（计标点）。训练集有2347条数据，测试集224条。在线下模型训练中，随机选取20%的验证集用于超参数的调优。

3.2 实现细节

本文中的模型架构为RoBERTa(Liu, 2019)，预训练参数来自于Hugging Face中在古文数据上预训练的roberta-classical-chinese-large-char[1]。对于模型$\mathcal{M}_{token}$、$\mathcal{M}_{token_prior}$和$\mathcal{M}_{token_slide}$，学习率为2e-5，batch大小设置为16；$\mathcal{M}_{token_slide}$的滑动窗口大小为100，滑动步长为4。$\mathcal{M}_{span_info}$模型的学习率设置为1e-5，batch大小为32，最大实体长度分别设置为4和10进行实验。

3.3 实验结果

队伍	F1值
KDSec_IIE	96.15
翼智团	95.82
北京信息科技大学智能信息处理研究所	95.34
小新	95.08
wzjj98	94.34

Table 2: 线上评测性能（F1值）

线上结果 如表2所示，最终我们提出的方法在线上评测中取得了96.15%的F1值，一方面我们有效缓解了古籍中实体的歧义性和边界模糊性问题，另一方面，充分发挥了不同实体识别框架的优势，比如token-wise感知的序列标注框架受限于长实体的识别，而span-level的框架弥补了该缺陷。

架构	模型	P	R	F1
Token-wise	$\mathcal{M}_{token}$	93.01	92.88	92.94
	$\mathcal{M}_{token\_prior}$	92.59	94.74	93.65
	$\mathcal{M}_{token\_slide}$	92.87	93.88	93.37
Span-level	$\mathcal{M}_{span\_info}$(最大实体长度4)	92.29	88.53	90.37
	$\mathcal{M}_{span\_info}$(最大实体长度10)	93.80	94.74	94.26

Table 3: 变体模型线下实体识别性能

各种变体模型的性能 表3给出了我们构建的两大类五种模型，在线下构造的验证集中实体识别的性能（精准度P，召回率R及F1值）。可以看出，信息论视角下Span-level的实体识别框架性能最优，而当最大长度设置为4时，其性能出现了明显下降，因为验证集中存在相当数量长度大于4的实体。相比基准的token-wise感知的序列标注模型$\mathcal{M}_{token}$，引入篇章先验信息及篇章内滑动窗口数据采样的方法$\mathcal{M}_{token_prior}$和$\mathcal{M}_{token_slide}$均取得了显著提升，证明了古籍命名实体识别任务中，篇章信息的重要性。

不同实体长度下的性能 图5展示了token-wise和span-level两类框架在不同长度实体上的F1值，其中token-wise的模型是$\mathcal{M}_{token}$，Span-level的模型是$\mathcal{M}_{span_info}$(最大实体长度10)。从图中可以看出，当实体长度较小时，两种框架的性能相当，而Span-level的框架在长实体上展现出了更

[1]https://huggingface.co/KoichiYasuoka/roberta-classical-chinese-large-char

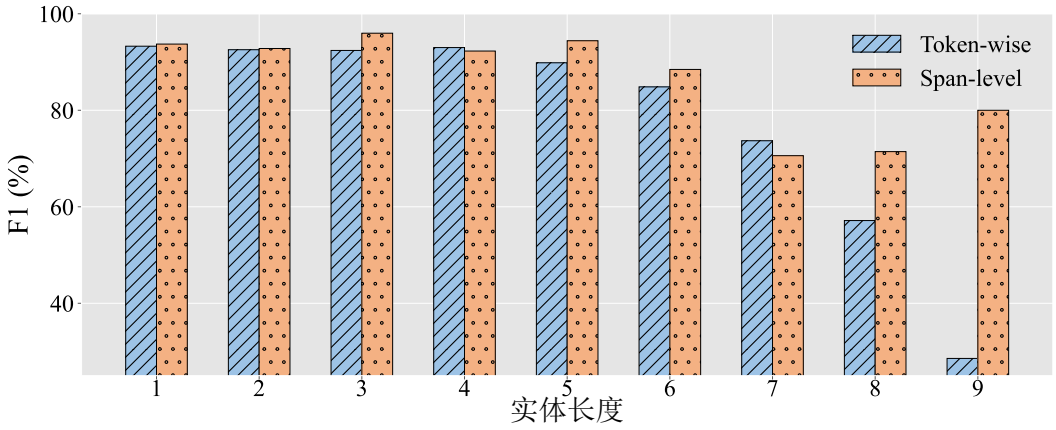

Figure 5: 两类框架在不同长度实体（K=1,2,3, ..., 9）上的性能

大的优势，由于其可以获取更长距离的依赖关系，因此选择这两类框架进行融合是可以互补取得更有效性能的。

架构	模型	P	R	F1
Token-wise	$\mathcal{M}_{token}$	92.25	92.11	92.18
	$\mathcal{M}_{token\_prior}$	91.57	93.69	92.62
	$\mathcal{M}_{token\_slide}$	91.65	92.65	92.15
Span-level	$\mathcal{M}_{span\_info}$(最大实体长度4)	91.40	87.68	89.50
	$\mathcal{M}_{span\_info}$(最大实体长度10)	92.90	93.83	93.37

Table 4: 变体模型线下实体边界检测性能

实体边界检测的性能 表4描述了线下验证集上各变体模型边界检测的性能，实体边界检测指的是实体左右边界的识别，不考虑实体类型，也就是说当实体的左右边界完全正确时，我们认为整个实体预测正确。可以看出基于信息论显式建模泛化特征、消除冗余特征（$\mathcal{M}_{span\_info}$，最大实体长度设置为10），可以在一定程度上缓解古籍文本中实体边界模糊的问题。

3.4 案例研究

如表5所示，相比基础的序列标注模型$\mathcal{M}_{token}$，本文采用的其它变体$\mathcal{M}_{token_prior}$，$\mathcal{M}_{token_slide}$以及$\mathcal{M}_{span_info}$在缓解实体歧义性和边界模糊性挑战上均有自己的优势。如第1个例子中的“白”，由于$\mathcal{M}_{token}$不能关联到训练语料中同一篇章内有关“武白”的句子，因此其不能将句子中所有的“白”识别出来，而$\mathcal{M}_{token_prior}$拼接篇章先验信息后，便可以将所有人名“白”识别出来。对于第2个例子，由于训练语料中存在大量“曰”字后面接人名的样本，因此$\mathcal{M}_{token}$倾向于将“悉编掣逋”等官职名(OFI)错误预测为人名(PER)，而训练语料中与之同一篇章内的句子存在“曰”字后面接官职名的样本，通过篇章内的滑动窗口采样，强化了上述模式，$\mathcal{M}_{token_slide}$从而缓解了实体上下文模式的歧义性。从最后两个例子可以看出，相比$\mathcal{M}_{token}$，信息论视角下的Span-level的模型$\mathcal{M}_{span_info}$可以缓解实体边界的模糊性问题，例如“唐柳芳”本身是指唐代的柳芳，而$\mathcal{M}_{token}$不能清晰地识别实体的左边界，也许仅学习到“{PER}有言”这样的模式，造成结果错误。

3.5 LLM在古籍命名实体识别上的表现

通过实验发现，LLM在该任务上可以表现出不错的性能，但仍然与专有小模型存在差距。利用原始训练数据以LoRA (Hu, 2022)方式微调ChatGLM-6B 20个Epoch，在测试集中可以达到83%左右的性能，但需要大量的后处理进行格式对齐。此外，本文进行了ChatGPT (gpt3.5-turbo-0315)在零样本、1个同源样本和5个同源样本下的Few-shot性能测试（同源样本指prompt中的例子选自与验证集相同的文章），如表6所示。

挑战	模型	推理结果
实体歧义性	$\mathcal{M}_{token}$	{圣宗|PER}诏{白|PER}鞫之，白正其事。使高丽还，权中京{留守|OFI}。时{慎行|PER}诸子皆处权要，以白断百姓分籍事不直，坐左迁。×
	$\mathcal{M}_{token\_prior}$	{圣宗|PER}诏{白|PER}鞫之，{白|PER}正其事。使高丽还，权中京{留守|OFI}。时{慎行|PER}诸子皆处权要，以{白|PER}断百姓分籍事不直，坐左迁。✓
	$\mathcal{M}_{token}$	{都护|OFI}一人，曰{悉编掣逋|PER}；又有{内大相|OFI}曰{曩论掣逋|PER}，亦曰{论莽热|PER}，{副相|OFI}曰{曩论觅零逋|PER}，{小相|OFI}曰{曩论充|PER}，各一人；×
	$\mathcal{M}_{token\_slide}$	{都护|OFI}一人，曰{悉编掣逋|OFI}；又有{内大相|OFI}曰{曩论掣逋|OFI}，亦曰{论莽热|OFI}，{副相|OFI}曰{曩论觅零逋|OFI}，{小相|OFI}曰{曩论充|OFI}，各一人；✓
实体边界模糊性	$\mathcal{M}_{token}$	别封{真定县男|OFI}，行并州{刺史|OFI}。{显祖|PER}受禅，别封{朝陵县|OFI}，又封{霸城县|OFI}，加位特进。×
	$\mathcal{M}_{span\_info}$	别封{真定县男|OFI}，行并州{刺史|OFI}。{显祖|PER}受禅，别封朝陵县，又封霸城县，加位特进。✓
	$\mathcal{M}_{token}$	{唐柳芳|PER}有言：帝定祸乱，而{房|PER}、{杜|PER}不言功；×
	$\mathcal{M}_{span\_info}$	唐{柳芳|PER}有言：帝定祸乱，而{房|PER}、{杜|PER}不言功；✓

Table 5: 案例分析

方案	ROUGE-1	ROUGE-2	ROUGE-L	BLEU
零样本	96.36	81.10	84.82	69.11
1个同源样本示例	94.34	81.99	85.34	72.11
5个同源样本示例	94.72	84.78	87.39	75.67

Table 6: ChatGPT在不同示例样本数量下的性能表现

由于LLM生成的结果存在一些未见字符或者格式不一致的问题，因此我们采用生成式指标进行评估。随着同源示例样本增加时，ROUGE-2、ROUGE-L、BLEU指标都随之增加，这也验证了我们之前的猜想；但ROUGE-1却出现了降低和波动，分析实验结果发现，样例增多后会导致模型生成与示例相关的内容，造成性能指标的波动和降低。[2]

4 相关工作

已经有多种研究范式和模型结构用于命名实体识别任务。早期的工作基于序列标注的范式，将句子序列中每个token标注为一个组合标签，从而派生出不同的标注规范，比如BIO或BIOES，结合实体类型标签，形成形如B-PER、I-PER的实体标签。之后将命名实体识别任务视为token的分类任务，衍生出有监督机器学习的序列标注模型。隐马尔可夫模型HMM (Bikel, 1999; Zhou, 2002) 能够捕捉现象的局部性，但难以捕捉序列远距离信息。最大熵模型MEM (Borthwick, 1998; Tsai, 2004) 通过已有先验知识的前提下选择熵最大的概率分布，确定某实体类型。条件随机场CRF (McCallum, 2003; Krishnan, 2006) 能够捕获数据的全局分布，改善长距离依赖的问题。在深度学习方法中，BiLSTM+CRF (Huang, 2015; Lample,

[2]输入ChatGPT的Instruction为“我希望你充当古籍命名实体识别专家，将以下输入文本中的所有的人名（PER），书籍名（BOOK），官职名（OFI）进行提取，返回结果将识别的实体替换实体—类别的格式，其中类别使用PER、BOOK、OFI标识。不要输出解释，更不要更改我的原始文本，只输出最终结果即可，你明白了吗？”。同源样本示例以“如*的返回结果为*”的方式添加到“其中类别使用PER、BOOK、OFI标识之后”

2016)或BERT+CRF是一种经典的命名实体识别架构，但由于BERT (Devlin, 2019)强大的上下文编码能力，CRF带来的性能增益降低，但推理速度变慢，因此BERT+Softmax的模型结构凭借其可靠的性能和效率，逐步取代了BERT+CRF。

与序列标注不同的是，基于实体span的方法将命名实体识别视为文本跨度分类问题，已经成为主流方法之一。基于预训练语言模型，相关研究工作 (Sohrab, 2018; Luan, 2019; Wadden, 2019) 通过连接span或聚合单词的表示，将他们接入线性分类器进行实体类型预测。 (Yu, 2020) 采用双仿射分类器融合实体首尾边界的表示进行span分类。同时，一些研究方法通过增加边界监督信息改善基于span的框架。(Zheng, 2019; Tan, 2020) 通过多任务学习检测实体边界，(Shen, 2021) 在span的预测后进行边界回归，而 (Li, 2022) 设计了两种词对关系辅助span分类。

近年来，端到端的生成式抽取框架直接将命名实体识别任务转换为文本生成任务，将句子中的目标实体转为序列进行训练和解码。(Dan, 2016)首先应用生成式框架在该任务中，输入句子并输出实体起始位置、长度和类型。(Straková, 2019)通过将嵌套实体标签建模为多标签解决嵌套命名实体识别的问题。(Hang, 2021)基于BART (Lewis, 2020)及指针网络的思想实现实体跨度及类型序列的生成，而(Lewis, 2022) 则基于T5 (Colin, 2020)生成实体序列，通过循环一致性训练的方法提升了无监督命名实体识别的能力。

现有的研究方法大多集中在英文及中文现代文的命名实体识别，缺少对古文的相关研究。而古文相比现代文具有更大的实体歧义性和边界模糊性，因此本文通过引入古文篇章知识及信息论，更好地缓解实体指称本身的歧义性，权衡实体指称本身及其上下文特征的编码。

5 总结

本文介绍了我们解决古籍命名实体识别任务的框架，从缓解古籍实体的歧义性和边界模糊性入手，在token-wise和span-level感知的两种框架基础上，融入古籍的篇章信息，并从信息论的视角编码泛化特征，去除冗余信息，提升古籍文本中实体识别的性能。在相关古籍文本上进行评测时，我们发现有几个未来亟待探究或解决的问题:

- 实体边界检测和类型预测的难度差异：我们将实体识别任务分解为两个子任务：实体的边界检测和类型预测分别进行训练，通过评估发现，边界检测的性能要远低于实体类型的预测，这也证明了古籍高难度的句读或分词断句，为实体的边界检测带来巨大挑战
- 同一句中具有较强的模式一致性：如下面这句话:

  “{萧惟信|PER}，字{耶宁|PER}，楉特部人。五世祖{霞赖|PER}，{南府宰相|OFI}。曾祖{乌古|PER}，{中书令|OFI}。祖{阿古只|PER}，{知平州|OFI}。”

  其具有很强的一致性的同位模式，通过单次解码很大概率会漏掉上述模式中的某些实体，因此强化句子中该同位模式的关联是一个值得探究的问题
- 大模型在实体识别中的表现有待提升：正如我们在前面评估的那样，大模型在实体识别任务上的性能有较大的提升空间，由于其生成的不确定性以及幻想等问题，造成非原句实体出现在生成结果中的现象；同时经过评测发现，大模型对实体的边界识别不够精准，如何构建指令或设计策略，使大模型更高效地适配到实体识别任务上值得探究

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)*.

Xiao Wang and Shihan Dou and Limao Xiong and Yicheng Zou and Qi Zhang and Tao Gui and Liang Qiao and Zhanzhan Cheng and Xuanjing Huang. 2022. *MINER: Improving Out-of-Vocabulary Named Entity Recognition from an Information Theoretic Perspective*. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

Xinghua Zhang and Bowen Yu and Yubin Wang and Tingwen Liu and Taoyu Su and Hongbo Xu. 2022. *Exploring Modular Task Decomposition in Cross-Domain Named Entity Recognition*. Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval.

Besnik Fetahu, Shervin Malmasi, Anjie Fang, and Oleg Rokhlenko. 2021. *Gazetteer Enhanced Named Entity Recognition for Code-Mixed Web Queries.* Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.

Jiafeng Guo, Gu Xu, Xueqi Cheng, and Hang Li. 2009. *Named entity recognition in query.* Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval.

Shekoofeh Mokhtari, Ahmad Mahmoody, Dragomir Yankov, and Ning Xie. 2019. *Tagging Address Queries in Maps Search.* Proceedings of the AAAI Conference on Artificial Intelligence.

Ningyu Zhang, Qianghuai Jia, Shumin Deng, Xiang Chen, Hongbin Ye, Hui Chen, Huaixiao Tou, Gang Huang, Zhao Wang, Nengwei Hua, and Huajun Chen. 2021. *AliCG: Fine-grained and Evolvable Conceptual Graph Construction for Semantic Search at Alibaba.* Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21).

Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019. *Entity-Relation Extraction as Multi-turn Question Answering.* Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. *Entity-Based Knowledge Conflicts in Question Answering.* Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.

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

Yue Zhang and Jie Yang. 2018. *Chinese NER using lattice LSTM.* Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *BERT: Pre-training of deep bidirectional transformers for language understanding.* Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).

Chuanqi Tan, Wei Qiu, Mosha Chen, Rui Wang, and Fei Huang. 2020. *Boundary enhanced neural span classification for nested named entity recognition.* Proceedings of Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020).

Jinlan Fu and Xuanjing Huang and Pengfei Liu. 2021. *SpanNer: Named Entity Re-/Recognition as Span Prediction.* Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).

Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang and Xipeng Qiu. 2021. *A Unified Generative Framework for Various NER Subtasks.* Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers).

Shuai Zhang, Yongliang Shen, Zeqi Tan, Yiquan Wu and Weiming Lu. 2022. *De-Bias for Generative Extraction in Unified NER Task.* Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

Edward J Hu and yelong shen and Phillip Wallis and Zeyuan Allen-Zhu and Yuanzhi Li and Shean Wang and Lu Wang and Weizhu Chen. 2022. *LoRA: Low-Rank Adaptation of Large Language Models.* International Conference on Learning Representations.

Yinhan Liu and Myle Ott and Naman Goyal and Jingfei Du and Mandar Joshi and Danqi Chen and Omer Levy and Mike Lewis and Luke Zettlemoyer and Veselin Stoyanov. 2019. *Roberta: A robustly optimized bert pretraining approach.* arXiv preprint arXiv:1907.11692.

Daniel M Bikel and Richard Schwartz and Ralph M Weischedel. 1999. *An algorithm that learns what's in a name.* Machine learning.

GuoDong Zhou and Jian Su. 2002. *Named entity recognition using an HMM-based chunk tagger.* Proceedings of the 40th annual meeting of the association for computational linguistics.

Andrew Borthwick and John Sterling and Eugene Agichtein and Ralph Grishman. 1998. *Description of the MENE Named Entity System as used in MUC-7*. Proceedings of the Seventh Message Understanding Conference (MUC-7), Fairfax, Virginia, April 29-May 1, 1998.

Richard Tzong-Han Tsai and Shih-Hung Wu and Cheng-Wei Lee and Cheng-Wei Shih and Wen-Lian Hsu. 2004. *Mencius: A Chinese named entity recognizer using the maximum entropy-based hybrid model*. International Journal of Computational Linguistics & Chinese Language Processing, Volume 9, Number 1, February 2004: Special Issue on Selected Papers from ROCLING XV.

Andrew McCallum and Wei Li. 2003. *Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons*. Proceedings of the 7th Conference on Natural Language Learning, Edmonton, May 31-Jun 1, 2003.

Vijay Krishnan and Christopher D Manning. 2006. *An effective two-stage model for exploiting non-local dependencies in named entity recognition*. Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the association for computational linguistics.

Mohammad Golam Sohrab and Makoto Miwa. 2018. *Deep Exhaustive Model for Nested Named Entity Recognition*. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.

Yi Luan and Dave Wadden and Luheng He and Amy Shah and Mari Ostendorf and Hannaneh Hajishirzi. 2019. *A general framework for information extraction using dynamic span graphs*. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).

David Wadden and Ulme Wennberg and Yi Luan and Hannaneh Hajishirzi. 2019. *Entity, Relation, and Event Extraction with Contextualized Span Representations*. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).

Juntao Yu and Bernd Bohnet and Massimo Poesio. 2020. *Named Entity Recognition as Dependency Parsing*. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.

Changmeng Zheng and Yi Cai and Jingyun Xu and Ho-fung Leung and Guandong Xu. 2019. *A Boundary-aware Neural Model for Nested Named Entity Recognition*. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).

Chuanqi Tan, Wei Qiu, Mosha Chen, Rui Wang, and Fei Huang. 2020. *Boundary enhanced neural span classification for nested named entity recognition*. In Proceedings of the AAAI conference on artificial intelligence, volume 34, pages 9016–9023.

Yongliang Shen, Xinyin Ma, Zeqi Tan, Shuai Zhang, Wen Wang, and Weiming Lu. 2021. *Locate and label: A two-stage identifier for nested named entity recognition*. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 2782–2794.

Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022. *Unified named entity recognition as word-word relation classification*. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 10965–10973.

Dan Gillick, Cliff Brunk, Oriol Vinyals, and Amarnag Subramanya. 2016. *Multilingual language processing from bytes*. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1296–1306.

Jana Straková and Milan Straka and Jan Hajic. 2019. *Neural Architectures for Nested NER through Linearization*. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.

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 (Volume 1: Long Papers), pages 5808–5822.

Mike Lewis and Yinhan Liu and Naman Goyal and Marjan Ghazvininejad and Abdelrahman Mohamed and Omer Levy and Veselin Stoyanov and Luke Zettlemoyer. 2020. *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension.* Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.

Andrea Iovine and Anjie Fang and Besnik Fetahu and Oleg Rokhlenko and Shervin Malmasi. 2022. *CycleNER: an unsupervised training approach for named entity recognition.* Proceedings of the ACM Web Conference 2022.

Colin Raffel and Noam Shazeer and Adam Roberts and Katherine Lee and Sharan Narang and Michael Matena and Yanqi Zhou and Wei Li and Peter J. Liu. 2020. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.* Journal of Machine Learning Research.

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

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* Proceedings of NAACL-HLT.