结合全局对应矩阵和相对位置信息的古汉语实体关系联合抽取

胡益裕[2]　左家莉[1]　曾雪强[1]　万中英[1]　王明文[1,2]
[1]江西师范大学　计算机信息工程学院　江西　南昌　330022
[2]江西师范大学　数字产业学院　江西　上饶　334000
Email: 329272494@qq.com, {zjl, xqzeng, libby, mwwang}@jxnu.edu.cn

摘要

实体关系抽取是信息抽取领域中一项重要任务，目前实体关系抽取任务主要聚焦于英文和现代汉语领域，关于古汉语领域的数据集构建和方法的研究目前却较少。针对这一问题，本文在研究了开源的《资治通鉴》语料后，人工构建了一个古汉语实体关系数据集，并设计了一种结合全局对应矩阵和相对位置信息的实体关系联合抽取方法。最后通过在本文构建的数据集上进行实验，证明了该方法在古汉语实体关系抽取任务上的有效性。

关键词： 古汉语数据集构建；实体关系联合抽取；全局对应矩阵；相对位置信息

Joint Extraction of Ancient Chinese Entity Relations by Combining Global Correspondence Matrix and Relative Position Information

Yiyu Hu[2]　**Jiali Zuo**[1]　**Xueqiang Zeng**[1]　**Zhongying Wan**[1]　**Mingwen Wang**[1,2]
[1] School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, Jiangxi 330022, China
[2]School of Digital Industry, Jiangxi Normal University, Shangrao, Jiangxi 334000, China
Email: 329272494@qq.com, {zjl, xqzeng,libby, mwwang}@jxnu.edu.cn

Abstract

Entity relation extraction is an important task in the field of information extraction. Currently, entity relation extraction is mainly focused on the fields of English and modern Chinese, but there are few researches on the construction and methods of data sets in the field of ancient Chinese. To solve this problem, this paper constructs an ancient Chinese entity relation dataset by hand after studying the open-source corpus of "Comprehensive Mirror for Aid Government", and designs a joint entity relation extraction method combining global correspondence matrix and relative location information. Finally, experiments are carried out on the dataset constructed in this paper to prove the effectiveness of the proposed method for entity relation extraction in ancient Chinese.

Keywords: Ancient Chinese datasets construct , Joint extraction of entity relationships , global corresponding matrix , relative position information

通讯作者: 左家莉
基金项目: 国家自然科学基金(61866018, 62266023, 62266021); 江西省教育厅科学技术研究项目(GJJ2200330)

1 引言

实体关系抽取(Entity Relation Extraction)任务旨在识别出非结构化文本中的实体和实体与实体之间的语义关系，是信息抽取(Information Extraction, IE)领域中一项重要任务。

实体关系抽取任务的研究工作最早开始于上世纪90年代(Brin, 1998)，随着研究的深入，该任务对于大规模标注数据的需求也在不断上升。为此，近年来，学术界和工业界构建了各种基于英文领域(Riedelet al., 2010; Gardent et al., 2017)和现代汉语领域(Xu et al., 2017)的实体关系数据集，这些工作极大的推动了实体关系抽取任务在上述两个领域中的研究。相较而言，实体关系抽取在古汉语领域中的研究工作则相对较少，其主要原因在于：(1)目前古汉语实体关系标注数据集较少；(2)相较于现代汉语实体关系标注工作而言，古汉语实体关系标注工作在标注原则设定、标注类型选定等方面难度更大，要求标注人员具有扎实的古汉语专业知识。最近，王鑫等人(2021)基于对“二十四史”语料的研究，构建了一份“二十四史”实体关系数据集。然而，通过对该数据集进行分析(表1)，我们发现该数据集具有如下特点：(1)标注规模较小；(2)在数据标注上存在标注稀疏问题，每条标注文本仅标注了少量的实体和关系。上述特点使得实体关系抽取模型较难从“二十四史”数据集中学习到足够的信息，最终导致其在该数据集上的性能较差。

类型	训练集	验证集	测试集
句子数	3113	882	418
每条标注数据标注的关系数	1	1	1
每条标注数据标注的实体数	2	2	2
总实体数	6226	1764	836
总关系三元组数	3113	882	418

Table 1: “二十四史”数据集部分特点描述

而在实体关系抽取方法的研究上，早期的实体关系抽取方法大多是基于特征的方法(Ren et al., 2017; Li and Ji, 2014)。随着深度学习的发展，后来的研究者提出了各种基于深度学习的实体关系抽取方法，这些方法大致可分为基于特殊标签识别的方法(Zheng et al., 2017; Wei et al., 2020; Wang et al., 2020; Zheng et al., 2021; Shang et al., 2022)，以及基于序列生成的方法(Zeng et al., 2018; Sui et al., 2021)。其中，Zheng等人(2021)和Shang等人(2022)采用了一种基于全局对应矩阵(Zheng et al., 2021)的实体关系联合抽取方法，该方法主要通过使用矩阵建模主客体之间(Zheng et al., 2021)、主客体和关系之间(Shang et al., 2022)的关联信息，最终通过该关联信息完成对实体、关系的抽取。目前，该方法在各种英文领域数据集上均取得了SOTA(state-of-the-art)性能。然而表2显示，该方法在目前的古汉语实体关系数据集上容易生成长度较长的异常实体，其主要由于全局对应矩阵是通过实体头尾词来表示完整实体，模型在学习实体信息时主要关注实体的头尾信息，较少关注实体的边界信息，从而使得模型更难精确的识别实体的边界，最终导致模型容易生成表2所示的长度异常实体。

古汉语文本	真实值	预测值
河内公独孤信，南阳公赵贵。	（**赵贵**，任职，南阳公）	（**独孤信，南阳公赵贵**，任职，南阳公）
内牙上都监使章德安数与之争，右都监使李文庆不附于，乙巳，贬德安于处州。	（**章德安**，名，德安） （李文庆，任职，右都监）	（**章德安数与之争，右都监使李文庆**，名，德安）

Table 2: 古汉语数据集上的部分测试结果。其中，“预测值”中粗体字为长度异常实体

鉴于目前的古汉语实体关系数据集存在标注规模较小、标注稀疏等问题，本文研究了开源的《资治通鉴》语料，结合上下文语义和关系触发词，人工重新构建了一份实体和关系更加丰富的古汉语实体关系数据集，并设计了一种基于全局对应矩阵的实体关系联合抽取方法。对于基于矩阵的方法容易生成长度较长的异常实体问题，本文尝试了在全局对应矩阵上引入字与字之间相对位置信息的方法，最终通过大量实验证明了该方法的有效性。

第二十二届中国计算语言学大会论文集，第241页-第252页，哈尔滨，中国，2023年8月3日至5日。

2 相关工作

2.1 实体关系抽取数据集构建

近年来，为了开展实体关系抽取任务而构建的数据集主要有NYT(Riedelet al., 2010)、WebNLG(Gardent et al., 2017)、SemEval[0]等英文数据集，以及Chinese Literature Text(Xu et al., 2017)、DuIE2.0[1](Li et al., 2019)等现代汉语领域数据集。相较而言，目前在古汉语实体关系抽取数据集上的相关研究则相对较少。

最近，王鑫等人(2021)基于对“二十四史”语料的研究，构建了一份“二十四史”实体关系数据集[2]，为古汉语实体关系抽取任务的研究提供了一份数据标注基准。而王一钒等人(2021)则提出了一套由“关系配价标注”、“命名逻辑标注”以及“单一关系存在”原则构成的数据标注原则(Wang et al., 2021)，填补了古汉语实体关系数据集标注工作在标注规范上存在的空白。

2.2 实体关系联合抽取

早期的实体关系联合抽取方法主要是基于特征的方法，如：(Ren et al., 2017; Li and Ji, 2014)，该方法主要是利用设置的特征函数获得数据特征信息，然后通过该特征信息联合识别实体和关系。然而由于该方法在建立特征工程上严重依赖NLP工具和大量人工操作(Zheng et al., 2021)，使得其难以处理数据规模较大的情况。

之后，深度学习技术不断发展，结合深度学习的实体关系联合抽取方法研究受到了广泛关注。其中，Sun等人(2017)选择将实体关系抽取任务建模为序列标注任务，利用一种包含实体、关系信息的标记方案联合建模实体和关系，然而该方案由于只为每个token分配了单个标签，无法处理存在单个token对应多个标签现象的三元组重叠问题。为了解决上述问题，Zeng等人(2018)基于LSTM(Hochreiter and Schmidhuber , 1997)，提出了一种结合实体复制机制的序列到序列模型。该复制机制由于可以对同一token进行多次复制，使得每个token可以参与不同三元组的构建，提升了模型解决三元组重叠问题的能力。

近年来，随着Bert(Devlin et al., 2019)等基于大规模语料的预训练模型的提出，结合预训练模型的实体关系联合抽取方法受到广泛研究。其中，Wei等人(2020)以Bert为编码器，设计了一种“通过主体和关系识别客体”的实体关系联合抽取方法，并取得了新的SOTA性能。然而，由于该模型在实体头尾词匹配时采用“领近匹配”原则(Wei et al., 2020)，导致其无法处理嵌套实体问题。同时还因为该方法采用先识别主体后识别客体的多阶段方式，这使得模型存在错误传播问题。为了解决上述错误传播问题和实体嵌套问题，Wang等人(2020)设计了一种利用大小为n^2(n为输入文本长度)的矩阵(Zheng et al., 2021)提取关系三元组的实体关系联合抽取方法。而Sui等人(2021)则是以Bert为编码器，将实体关系识别任务重新考虑为三元组序列生成任务，通过非自回归的解码方式生成三元组序列，该方式解决了自回归方式(Zeng et al., 2018)生成三元组序列时仍需要按照三元组序列顺序解码的弊端，提升了模型的解码效率。上述两种方法在解决重叠三元组问题上都取得了较好的结果，然而由于两者均是通过不同模块识别实体和关系，不同模块间缺乏信息的交互，从而使得实体和关系之间的相互约束不足，最终导致实体和关系在匹配时出现信息冗余问题(Shang et al., 2022)。

最近，在解决重叠三元组和实体嵌套等复杂问题上。Zheng等人(2021)和Shang等人(2022)均采用全局对应矩阵(Zheng et al., 2021)联合建模实体和关系。其中，Zheng等人(2021)选择将实体关系任务划分为实体提取、主客体对齐和关系判断三个子任务，采用先进行潜在关系预测后完成主、客体识别的方式完成实体关系联合抽取。在主、客体对齐任务上，该方法主要利用一个全局对应矩阵学习主体和客体的关联性。虽然该方法提升了模型解决三元组重叠和嵌套实体问题的能力，然而由于该方法是采用先预测潜在关系，后通过潜在关系完成主、客体对齐任务的方式，这使得潜在关系预测阶段出现的错误会传递到主、客体对齐任务上，即模型存在错误传播问题。为了解决上述所说的信息冗余问题和错误传播问题，Shang等人(2022)设计了一种结合特殊标签和全局对应矩阵联合建模实体和关系的方法，该方法解决了上述所说的错误传播问题和信息冗余问题，并在实体关系抽取任务上取得了新的SOTA。

[0]https://github.com/thunlp/OpenNRE/blob/master/benchmark/download_semeval.sh
[1]https://github.com/PaddlePaddle/PaddleNLP/tree/develop/examples/information_extraction/DuIE
[2]https://github.com/jizijing/C-CLUE

然而，本文通过在古汉语数据集上进行了大量实验后发现：全局对应矩阵容易生成长度异常的实体。为了解决该问题，在研究了(Li et al., 2019)引入实体与字符间相对位置信息的方法后，本文设计了一种引入字与字之间相对位置信息的方法，最终通过大量实验证明：引入相对位置信息对于缓解“全局对应矩阵容易生成长度异常的实体”问题的有效性。

3 数据集构建

3.1 数据集的来源

《资治通鉴》是由北宋史学家司马光主编的一部编年体史书，该书记录了从周威烈王二十三年(公元前403年)到五代后周世宗显德六年(公元959年)期间共计1362年的历史。

本文构建的数据集语料来源于古诗文网[3]公开的资治通鉴语料，经过对该语料进行分句、筛选处理后，最终挑选出其中约10000条语句进行标注。目前完成标注的字符总数为76025，句子的平均长度为37.92。

3.2 数据标注准则

本文在实体关系标注类型上，参考了王鑫等人(2021)公布的“二十四史”实体关系数据集上定义的实体关系类型，而在实体和关系的标注原则上，则采用了张欢(2020)针对实体标注提出的简单性原则、易操作性原则、一致性原则。

首先是简单性原则。在实体标注上，本研究将古汉语实体类型简要分为人名(PER)、官职名(JOB)、组织名(ORG)、地名(LOC)四种实体类型，类型数目适中。针对成分复杂的类型，如：人名，本研究并未对其进行细分，对于个别在语义上产生交叉的实体，本文在标注时进行类别统一。在关系标注上，为了降低标注难度，本文选定的均是具有关系触发词或者表义明显的关系类型。所以，本文的实体标注和关系标注符合简单性原则。

其次是易操作原则。本研究为了提升标注工作的易操作性。分别对定义的实体、关系类型和实体、关系标注过程进行了详细的说明（具体见下文），符合易操作性原则。

最后是一致性原则。实体定义类型和关系类型定义是实体关系数据标注的第一步，对于容易混淆的实体类型和关系类型，本研究对其进行合并统一，遵循了一致性原则。

3.3 实体、关系标注说明

本文定义了人名(PER)、官职名(JOB)、组织名(ORG)、地名(LOC)四种古汉语实体类型和“任职”、“隶属于”和“去往”等24种关系类型，具体标注说明如下文所述。

3.3.1 实体标注说明

首先是人名(PER)，在古汉语中，“人名(PER)”成分种类繁多，包括名、字、氏、姓、爵位、排行、谥号、官职等，对于其中以官职和爵位表现的人名(PER)实体，基于对(人名，官职名，任职)这类三元组的考虑，为了维护实体标注的一致性，本研究仍旧将其标注为官职。其次是官职名(JOB)，这一类型的实体表现形式较单一，主要表现为职位名。最后是地名(LOC)和组织名(ORG)，地名(LOC)主要指地理上所定义的“地名”，如：山名、水名和地方名等。组织名(ORG)包括了国家名、氏族名、官署机构名等。

3.3.2 关系标注说明

在古汉语关系标注上，为了降低标注难度，本研究采用了结合关系触发词和上下文语义的标注方法。其中，关系触发词指的是文本中直接表达两个实体间关系的词。例如：“元舆，元褒之兄也。”，通过“兄”这一词可以得出“元舆”是“元褒”的兄长，像“兄”这种可以体现两个实体间关系的词即是关系触发词。下表3为部分关系及其相关触发词介绍。

3.4 数据标注格式

本研究在数据标注格式上，采用了王鑫等人(2021)公布的“二十四史”实体关系数据集上的标注格式，将标注后的数据存储为json文件格式，其中每条标注数据包括了如下内容：古汉语文本(text)、主体(subject)、主体类型(subject_type)、客体(object)、客体类型(object_type)、关系(relation)，具体如图1所示。

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

关系类别	例句
子	守一，仁皎之**子**。
兄	中书令陈淮，徽之**兄**也。
葬于	始安忠武公温峤卒，**葬**于豫章。
弟	略、模，皆越之**弟**也。
升迁	弟晦，亦以皎故累**迁**吏部侍郎。

Table 3: 部分关系类别及其关系触发词。其中，粗体字为关系触发词

```
{"text": "又诏以太宰颙都督中外诸军事。",
"spo_list": [{"subject": "颙",
              "subject_type": "PER",
              "object": "太宰",
              "object_type": "JOB",
              "relation": "任职"
            }]
}
```

Figure 1: 数据标注格式

4 模型构建

4.1 问题描述

实体关系抽取问题具体描述如下：给定一句输入序列S，S=$(s_1, s_2, ..., s_n)$，其中，s_n表示S中的第n个词，实体关系抽取任务的目标是识别序列S中所有的主体、客体和主、客体的语义关系，并输出(主体，关系，客体)形式的三元组。

为了建模实体关系抽取任务，本文参考了Zheng等人(2021)的多任务学习思路，选择将该任务细分为：主体和客体对齐、实体和关系对齐、实体抽取三个子任务，每个任务的解释如下：(1)主体和客体对齐：目的是得到输入S中所有token间的对应分数，当该对token属于主体和客体的一部分时，其对应分数是最高的；(2)实体和关系对齐：该任务主要是通过矩阵预测输入S中所有token和任务定义的所有关系的对应分数，该对应分数表示了token和所有关系间的关联程度；(3)实体提取：目的是识别出输入S中所有的实体。为了建模三个子任务，本文采用了Zhang等人(2021)提出的全局对应矩阵的方法，为每个任务设计了对应的主体和客体全局对应、实体和关系全局对应、实体头尾全局对应三个模块(图2)，最终通过结合三个模块抽取的信息完成实体关系的联合抽取。

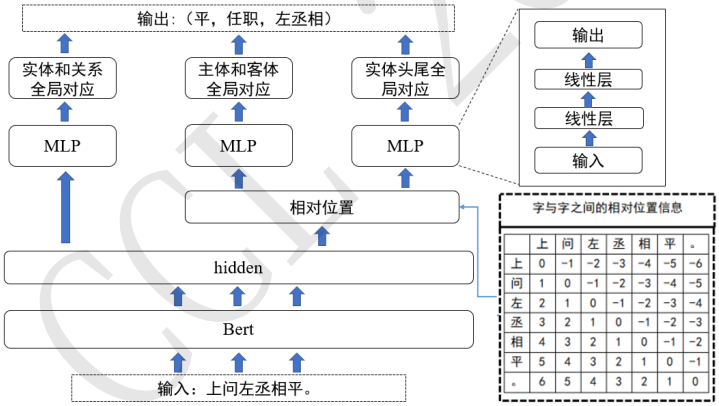

Figure 2: Bert+全局对应矩阵的模型架构

4.2 编码层

给定输入S，S=$(s_1, s_2, ..., s_n)$，s_i表示句子S中第i个词，本部分主要通过一个预训练的Bert作为模型的编码器来获得S对应的向量表示，具体如下公式所示：

$$(h_1, h_2, ..., h_n) = Bert((s_1, s_2, ..., s_n)) \quad (1)$$

其中$(h_1, h_2, ..., h_n)$表示Bert最后一层输出的隐藏层状态，n为输入文本的长度。

4.3 解码层

在本部分，本文将介绍主体和客体全局对应、实体和关系全局对应、实体头尾全局对应三个模块和相对位置信息模块的具体实现细节。

4.3.1 相对位置信息

为了缓解全局对应矩阵的解码方式存在的对实体长度约束不足的问题，本文选择引入字与字之间的相对位置信息。具体做法是：先获得两个输入token间的相对位置，然后通过将其与权重矩阵相乘获得相对位置对应的向量表示$POS_{i,j}$，之后将第i个token对应的hidden与第j个token对应的hidden进行拼接得到结果$Concat(h_i, h_j)$，最后将$POS_{i,j}$与$Concat(h_i, h_j)$相加得到最终输出$h_{i,j}^{pos}$。该过程可以公式化表示为：

$$POS_{i,j} = I_{i,j}W_{pos} + b_{pos} \tag{2}$$

$$h_{i,j}^{pos} = Concat(h_i, h_j) + POS_{i,j} \tag{3}$$

其中，I为图2中的相对位置信息矩阵，$I_{i,j}$表示输入句子中第i个token和第j个token的相对位置，h_i和h_j为输入句子中第i个和第j个词对应的hidden表示，Concat表示拼接。

	上	问	左	丞	相	平	。
上	0	0	0	0	0	0	0
问	0	0	0	0	0	0	0
左	0	0	0	0	0	0	0
丞	0	0	0	0	0	0	0
相	0	0	0	0	0	0	0
平	0	0	1	0	0	0	0
。	0	0	0	0	0	0	0

(a) 主体的开始词(第一列)和客体的开始词(第一行)的全局对应矩阵

	上	问	左	丞	相	平	。
上	0	0	0	0	0	0	0
问	0	0	0	0	0	0	0
左	0	0	0	0	0	0	0
丞	0	0	0	0	0	0	0
相	0	0	0	0	0	0	0
平	0	0	0	0	1	0	0
。	0	0	0	0	0	0	0

(b) 主体的结尾词(第一列)和客体的结尾词(第一行)的全局对应矩阵

Figure 3: 主体和客体全局对应

4.3.2 主体和客体全局对应

在获得了相对位置信息模块的输出后，为了建模主体和客体的全局对应任务，本文选择将主体和客体的全局对应任务细分为主、客体开始词对应和主、客体结尾词对应两个子任务，并为每个子任务设计了对应的主、客体开始词全局对应矩阵(图3(a))和主、客体结尾词全局对应矩阵(图3(b))。该模块的具体实现如下公式所示：

$$h_{i,j}^{so} = MLP(h_{i,j}^{pos}) \tag{4}$$

$$P_{i,j}^{soh} = sigmoid(h_{i,j}^{so}W_{soh} + b_{soh}) \tag{5}$$

$$P_{i,j}^{sot} = sigmoid(h_{i,j}^{so}W_{sot} + b_{sot}) \tag{6}$$

W_{soh}、W_{sot}、b_{soh} 、b_{sot} 为可训练的参数，MLP为多层感知机，$h_{i,j}^{pos}$为相对位置模块的输出。

4.3.3 实体和关系全局对应

为了建模实体和关系的全局对应任务，本文选择将其细化为主体开始词和关系的全局对应以及客体开始词和关系的全局对应两个子任务，并设置了主体开始词和关系的全局对应矩阵(图4(a))和客体开始词和关系的全局对应矩阵(图4(b))。

该模块的步骤可以形式化表示为如下公式：

$$h_i^{er} = MLP(h_i) \tag{7}$$

$$P_i^{sr} = sigmoid(h_i^{er}W_{sr} + b_{sr}) \tag{8}$$

$$P_i^{or} = sigmoid(h_i^{er}W_{or} + b_{or}) \tag{9}$$

其中，h_i为输入句子中第i个词对应的隐藏层状态表示，W_{sr}、W_{or}、b_{sr}、b_{or}为可训练的参数，MLP为多层感知机。

	...	...	任职	...	...	...	...
上	0	0	0	0	0	0	0
问	0	0	0	0	0	0	0
左	0	0	0	0	0	0	0
丞	0	0	0	0	0	0	0
相	0	0	0	0	0	0	0
平	0	0	1	0	0	0	0
。	0	0	0	0	0	0	0

(a) 主体的开始词(第一列)和关系(第一行)的全局对应矩阵

	...	...	任职	...	...	...	...
上	0	0	0	0	0	0	0
问	0	0	0	0	0	0	0
左	0	0	1	0	0	0	0
丞	0	0	0	0	0	0	0
相	0	0	0	0	0	0	0
平	0	0	0	0	0	0	0
。	0	0	0	0	0	0	0

(b) 客体的开始词(第一列)和关系(第一行)的全局对应矩阵

Figure 4: 实体和关系全局对应

4.3.4 实体头尾全局对应

本部分的目的是识别出输入文本中的所有实体。为了建模实体信息，本部分设计了实体头尾全局对应矩阵(图5)，通过该矩阵联合建模实体头尾信息。本模块的工作过程可以表示为如下：

$$h_{i,j}^{ht} = MLP(h_{i,j}^{pos}) \tag{10}$$

$$P_{i,j}^{ht} = sigmoid(h_{i,j}^{ht} W_{ht} + b_{ht}) \tag{11}$$

其中，W_{ht}、b_{ht}为可训练的参数，MLP为多层感知机，$h_{i,j}^{pos}$为相对位置模块的输出。

	上	问	左	丞	相	平	。
上	1	0	0	0	0	0	0
问	0	0	0	0	0	0	0
左	0	0	0	0	1	0	0
丞	0	0	0	0	0	0	0
相	0	0	0	0	0	0	0
平	0	0	0	0	0	0	0
。	0	0	0	0	0	0	0

Figure 5: 实体头尾全局对应。其中，实体开始词(第一列)和结尾词(第一行)的全局对应矩阵

4.4 损失函数

本研究采用交叉熵损失函数作为模型的损失函数，模型最终的损失函数主要由L^{so}、L^{er}和L^{ht}三部分构成，其具体表示如下：$L = L^{so} + L^{er} + L^{ht}$，假设输入是由n个token组成的序列，上述公式中每部分的解释如下所示。

首先是L^{so}，其包括了L^{soh}、L^{sot}，两者具体计算过程如下公式所示：

$$L^{so} = 0.5 \times (L^{soh} + L^{sot}) \tag{12}$$

$$L^{soh} = \frac{-1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_{i,j} log(P_{i,j}^{soh}) + (1 - y_{i,j}) log(1 - P_{i,j}^{soh}) \tag{13}$$

$$L^{sot} = \frac{-1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_{i,j} log(P_{i,j}^{sot}) + (1 - y_{i,j}) log(1 - P_{i,j}^{sot}) \tag{14}$$

上述公式中，$P_{i,j}^{soh}$表示第i个token为主体开始词和第j个token为客体开始词的条件概率；$P_{i,j}^{sot}$表示第i个token为主体结尾词和第j个token为客体结尾词的条件概率。

第二十二届中国计算语言学大会论文集，第241页-第252页，哈尔滨，中国，2023年8月3日至5日。

其次L^{er}包括了L^{sr}、L^{or}，两者具体计算过程如下公式所示:

$$L^{er} = 0.5 \times (L^{sr} + L^{or}) \tag{15}$$

$$L^{sr} = \frac{-1}{n \times n^r} \sum_{i=1}^{n} \sum_{j=1}^{n^r} y_{i,j} log(P_{i,j}^{sr}) + (1 - y_{i,j}) log(1 - P_{i,j}^{sr}) \tag{16}$$

$$L^{or} = \frac{-1}{n \times n^r} \sum_{i=1}^{n} \sum_{j=1}^{n^r} y_{i,j} log(P_{i,j}^{or}) + (1 - y_{i,j}) log(1 - P_{i,j}^{or}) \tag{17}$$

上述步骤中，$P_{i,j}^{sr}$表示第i个token为主体开始词，并且与第j个关系存在关联的条件概率；$P_{i,j}^{or}$表示第i个token为客体开始词，并且与第j个关系存在关联的条件概率。

最后是L^{ht}，其为4.3.4中实体头尾全局对应矩阵对应的损失，具体的计算过程如下所示:

$$L^{ht} = \frac{-1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_{i,j} log(P_{i,j}^{ht}) + (1 - y_{i,j}) log(1 - P_{i,j}^{ht}) \tag{18}$$

其中，$P_{i,j}^{ht}$表示第i个token为实体开始词以及第j个token为实体结尾词的条件概率。上述公式中n为输入序列的长度，n^r为定义的关系数。

5 实验

5.1 实验的评估指标和参数设置

在实验中，本文选择Adam(Kingma and Ba, 2014)作为模型优化器，将学习率设置为5×10^{-5}，输入序列的最大长度为100，dropout设置为0.1，batch_size设置为8，epoch设置为100，主体和客体全局对应、实体和关系全局对应、实体头尾全局对应三个模块的阈值均设置为0.5，MLP中使用的激活函数是ReLU(Glorot et al., 2011)。此外，本研究以精确率(Precision)、召回率(Recall)、F1值(F1-Score)为实验的评估指标。

5.2 古汉语数据集描述

在数据集划分上，本文选择以7：2：1的比例，将古汉语数据集划分为训练集、验证集和测试集。下表为每个数据集的实体类型分布情况(表4)和三元组的分布情况(表5)描述。

实体类型	训练集	验证集	测试集
人名(PER)	8446	3167	1634
官职(JOB)	3828	1535	740
地名(LOC)	1752	664	346
组织(ORG)	474	142	104
总计	14500	5408	2824

Table 4: 各个实体类型在训练集、验证集和测试集上的样本分布情况

数据集类型	三元组数
训练集	7242
验证集	2754
测试集	1412
总计	11408

Table 5: 三元组分布情况

5.3 实验过程与结果分析

5.3.1 预训练模型选择

为了选择一个合适的预训练模型参与本文的实验，本文挑选了Guwen-Bert[4]、RoBERTa-classical-chinese(Koichi Yasuoka, 2022)、SiKuBERT、SiKuRoBERTa(Wang et al., 2022)和bert-base-chinese-ner[5]五个预训练模型，然后让它们分别与OurModel模型进行搭配组合，最后将组合后的模型分别在古汉语数据集上进行实验。其中，Guwen-Bert、RoBERTa-classical-chinese、SiKuBERT和SiKuRoBERTa是在大量古汉语语料上训练的预训练模型，而bert-base-chinese-ner则是基于命名实体识别任务训练的预训练模型，本次实验均只使用了它们最后一

[4]https://github.com/ethan-yt/guwenbert
[5]https://github.com/ckiplab/ckip-transformers

层输出的隐藏层表示。实验的最后结果如表6所示。从最后的F1值上看，bert-base-chinese-ner+OurModel的组合在古汉语数据集上的表现要优于其它组合，OurModel为本文构建的模型。

预训练模型	测试结果
Guwen-bert	65.0
RoBERTa-classical-chinese	65.5
SiKuBERT	64.9
SiKuRoBERTa	65.7
bert-base-chinese-ner	**67.0**

Table 6: 各类预训练模型在OurModel上的F1(%)值对比

5.3.2 不同方法在古汉语实体关系数据集上的性能对比

为了探究本文提出的方法和基线方法在古汉语实体关系数据集上的性能表现，本文选取了基于其它解码方式的CasRel (Wei et al., 2020)和SPN4RE(Sui et al., 2021)，以及采用矩阵解码方式的OneRel(Shang et al., 2022)、PRGC(Zheng et al., 2021)和TPLinker(Wang et al., 2020)五种基线模型，与OurModel(本文的方法)进行对比实验。为了便于对比，各个方法使用的预训练模型均为5.3.1中F1值最好的bert-base-chinese-ner，最终各个方法在古汉语数据集上的结果如表7所示。

模型名称	精确率	召回率	F1值
CasRel	48.0	33.0	39.1
SPN4RE	51.1	40.6	45.2
OneRel	62.4	47.3	53.8
PRGC	69.0	52.1	59.4
TPLinker	74.8	**58.4**	65.6
OurModel	**81.6**	56.8	**67.0**

Table 7: 五种基线模型和OurModel在古汉语数据集上的精确率(%)、召回率(%)和F1值(%)。其中，OurModel为本文构建的模型

任务名称	模型类别	精确率	召回率	F1值
r	TPLinker	88.1	**63.5**	**73.8**
	OurModel	**91.9**	59.9	72.5
(s,o)	TPLinker	77.8	**60.6**	68.1
	OurModel	**85.1**	59.0	**69.7**

Table 8: OurModel和TPLinker在不同任务上的精确率(%)、召回率(%)和F1值(%)对比。其中r表示关系，s表示主体，o表示客体

从表8的结果可知，首先，相较于基于其它解码方式的方法来说，基于矩阵解码方式的方法在古汉语数据集上均获得了更好的结果。本文认为这是由于全局对应矩阵本质上是让模型学习实体与实体或者实体与关系的内在关联，相较于其它方式来说，更能缓解实体与实体、实体与关系匹配出现的冗余问题。

其次，OurModel在古汉语数据集上的F1值比最好的基线模型TPLinker要高出1.4%。为了进一步探究OurModel在古汉语数据集中表现优于基线模型的原因，在主客体对识别和关系抽取两个任务上，本文对比了基线模型中F1值最好的TPLinker和本文提出的模型的性能，结果如表9所示。

从表8中我们还可以得知，本文的方法虽然在关系抽取任务上比TPLinker方法在F1值上低1.3%，但是在主客体对识别任务上则高出1.6%，且在主客体对识别任务上本文的方法在精确率上比TPLinker高出7.3%，于是本文推测OurModel优于其它方法的原因可能是：相对位置信息可以提升实体识别的精度。

5.3.3 消融研究

为了检验相对位置信息对OurModel最终性能的影响和探究引入相对位置信息方法的泛化能力，本文分别在两个古汉语数据集上对相对位置信息模块进行消融实验。其实验结果如表9和表10所示。

从表9和表10我们可以看出，当引入相对位置信息时，在《资治通鉴》数据集上，模型的精确率上升了9.7%，召回率上升1.8%，F1值上升了4.7%。而在“二十四史”数据

	精确率	召回率	F1值
不加相对位置信息	71.9	55.0	62.3
加相对位置信息	**81.6**	**56.8**	**67.0**

Table 9: 在《资治通鉴》数据集上，OurModel在两种情况下的精确率(%)、召回率(%)和F1值(%)对比

	精确率	召回率	F1值
不加相对位置信息	9.2	**37.6**	14.9
加相对位置信息	**12.0**	36.8	**18.1**

Table 10: 在“二十四史”数据集上，OurModel在两种情况下的精确率(%)、召回率(%)和F1值(%)对比

集(Wang., 2021)上，模型的精确率上升了2.8%，F1值上升了3.2%。从表11中可以看出，当不加入相对位置信息时，模型容易生成长度较长的错误实体。综合上述结果，本文得出：当模型缺乏相对位置信息时，会极大削弱模型对实体长度的约束，导致模型预测出长度过长的错误实体，从而使得模型的精确率和召回率下降，这也进一步验证了相对位置信息能提升实体识别的精度的假设。同时，结合表9和表10的结果，我们可知：加入相对位置信息的方法在两个古汉语数据集上均使得模型最终的性能有所提升，这证明了引入相对位置信息的方法在古汉语实体关系抽取任务上的泛化性。

输入文本	实际标注的三元组	加入相对位置信息	不加入相对位置信息
魏相州刺史中山文庄王熙，英之子也，与弟给事黄门侍郎略、司徒祭酒纂。	(纂,任职,司徒祭酒) (熙,任职, 文庄王)	(纂,任职,**司徒祭酒**) (熙,任职,**文庄王**)	(纂,任职,**司徒祭酒纂，皆为清河王**) (熙,任职,**中山文庄王**)
祜官至尚书左仆射，爵新平王。	(祜,任职,尚书左仆射) (祜,任职, 新平王)	(祜,任职,**尚书左仆射，爵新平王**)	(祜,任职,**尚书左仆射**) (祜,任职, **新平王**)

Table 11: 两种不同情况下OurModel的结果对比

6 总结与展望

本文针对目前古汉语实体关系数据集存在的标注稀疏问题，基于对《资治通鉴》语料和目前在古汉语数据标注的研究，人工标注了一份实体关系数更加丰富的古汉语实体关系数据集，并构建了结合全局对应矩阵和相对位置信息的古汉语实体关系联合抽取模型。针对全局对应矩阵的方法容易生成较长实体的问题，本文尝试引入了字与字之间的相对位置信息的方法，最后通过在古汉语数据集上的实验证明了该方法的有效性。

然而本研究目前构建的《资治通鉴》具有标注规模较小、个别关系样本数较少等特点，这对模型最终的性能产生了一定程度的影响。此外，本研究所使用的引入相对位置信息的方法目前仅在两个古汉语数据集上进行了验证，这使得本文在对该方法的泛化能力的研究上存在不足。因此，下一步，本研究将继续研究不同类别的古汉语语料，以期建立更大规模的古汉语实体关系抽取标注数据集，进一步提升古汉语实体关系抽取任务的性能。

参考文献

Brin S. 1998. Extracting Patterns and Relations from the World Wide Web. In *International workshop on the world wide web and databases.*(pp. 172-183). Berlin, Heidelberg: Springer Berlin Heidelberg.

Yee Seng Chan and Dan Roth. 2011. Exploiting syntactico-semantic structures for relation extraction. In *Proceedings of ACL.*

Diederik P. Kingma and Jimmy Lei Ba. 2014. A method for stochastic optimization. *arXiv preprint arXiv:1412.6980.*

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., Volume 1 (Long and Short Papers), pages 4171–4186.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini 2017. Creating training corpora for nlg micro-planners. In *Proceedings of ACL.*

Xavier Glorot,Antoine Bordes, and Yoshua Bengio 2011. Deep Sparse Rectifier Neural Networks. *Journal of Machine Learning Research*, pages 315-323.

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

Ziran Li, Ning Ding, Zhiyuan Liu, Hai-Tao Zheng, and Ying Shen 2019. Chinese Relation Extraction with Multi-Grained Information and External Linguistic Knowledge. In *Proceedings of ACL.*

Qi Li and Heng Ji. 2014. Incremental joint extraction of entity mentions and relations. In *Proceedings of ACL.*

Shuangjie Li, Wei He, Yabing Shi, Wenbin Jiang, Haijin Liang, Ye jiang, Yang Zhang, Yajuan Lyu, and Yong Zhu 2019. DuIE: A Large-scale Chinese Dataset for Information Extraction. *Natural Language Processing and Chinese Computing: 8th CCF International Conference*, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8. Springer International Publishing, pages 791-800.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes* 30(1):3–26.

Xiang Ren, Zeqiu Wu, Wenqi He, Meng Qu, Clare R Voss, Heng Ji, Tarek F Abdelzaher, and Jiawei Han. 2017. Cotype: Joint extraction of typed entities and relations with knowledge bases. In *Proceedings of the 26th International Conference on World Wide Web*. pages 1015-1024.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *ECML-PKDD.*

Rink, Bryan, and Sanda Harabagiu. 2010. Utd: Classifying semantic relations by combining lexical and semantic resources. In *Proceedings of the 5th International Workshop on Semantic Evaluation.*pages 256–259.

Dianbo Sui, Yubo Chen, Kang Liu, Jun Zhao, Xiangrong Zeng, and Shengping Liu. 2021. Joint Entity and Relation Extraction with Set Prediction Networks. *arXiv preprint arXiv:2011.01675.*

Xin Wang, Zijing Ji, Yuxin Shen, Qingyan Guo, Yang Sun, Guanzhong Liu, Zijun Wang, Yining Sun, and Tian Yu. 2021. C-CLUE: A Benchmark of Classical Chinese Based on a Crowdsourcing System for Knowledge Graph Construction. In*Proceedings of CCKS.*

Yu-Ming Shang, Heyan Huang, and Xian-Ling Mao. 2022. OneRel: Joint Entity and Relation Extraction with One Module in One Step. In *Proceedings of AAAI.*

Koichi Yasuoka, Christian Wittern, Tomohiko Morioka, Takumi Ikeda, Naoki Yamazaki, Yoshihiro Nikaido, Shingo Suzuki, Shigeki Moro, Kazunori Fujita. 2022. Designing Universal Dependencies for Classical Chinese and Its Application. *Journal of Information Processing Society of Japan*, 63(2): 355-363.

Zhepei Wei, Jianlin Su,Yue Wang,Yuan Tian, and Yi Chang. 2020. A Novel Cascade Binary Tagging Framework for Relational Triple Extraction. In *Proceedings of ACL.*

Yucheng Wang, Bowen Yu, Yueyang Zhang, Tingwen Liu, Hongsong Zhu, and Limin Sun. 2020. TPLinker: Single-stage Joint Extraction of Entities and Relations Through Token Pair Linking. In *Proceedings of COLING.*

Jingjing Xu, Ji Wen, Xu Sun, Qi Su. 2017. A Discourse-Level Named Entity Recognition and Relation Extraction Dataset for Chinese Literature Text. *arXiv:1711.07010.*

Xiangrong Zeng, Daojian Zeng, Shizhu He, Kang Liu, and Jun Zhao. 2018. Extracting relational facts by an end-to-end neural model with copy mechanism. In *Proceedings of ACL.*

Suncong Zheng, Feng Wang, Hong yunBao, Yuexing Hao, Peng Zhou, and Bo Xu. 2017. Joint extraction of entities and relations based on a novel tagging scheme. In *Proceedings of ACL.*

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 CCL.*

Hengyi Zheng, Rui Wen, Xi Chen, Yifan Yang, Yunyan Zhang, Ziheng Zhang, Ningyu Zhang, Bin Qin, Ming Xu, and Yefeng Zheng. 2021. PRGC: Potential Relation and Global Correspondence Based Joint Relational Triple Extraction. In *Proceedings of ACL.*

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. 2003. Kernel methods for relation extraction. *Journal of machine learning research*, 3(Feb):1083–1106.

王东波，刘畅，朱子赫，刘江峰，胡昊天，沈思，李斌. 2022. 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. 图书馆论坛,42(06):31-43.

王一钒，李博，史话，苗威，姜斌. 2021. 古汉语实体关系联合抽取的标注方法. 数据分析与知识发现,5(9):63-74.