研究论文

基于知识蒸馏的半监督古籍实体抽取*

唐　朝[1,2]　陈　波[3]　谭泽霖[4,5]　赵小兵[3,5]

[1]（中央民族大学哲学与宗教学学院 北京 100081）

[2]（中央民族大学国家安全研究院 北京 100081）

[3]（中央民族大学信息工程学院 北京 100081）

[4]（中央民族大学中国少数民族语言文学学院 北京 100081）

[5]（国家语言资源监测与研究少数民族语言中心 北京 100081）

摘要：**【目的】**通过知识蒸馏将来源于无监督数据的额外知识以训练数据的形式注入学生实体抽取模型，缓解古籍实体抽取任务有监督数据稀缺的问题。**【方法】**使用大语言模型作为生成式知识教师模型，在无监督语料上进行知识蒸馏；基于《左传》和GuNer的有监督数据构造词典知识教师模型蒸馏词典知识，共同构建半监督古籍实体抽取数据集，将古籍实体抽取任务转换为序列到序列任务，再微调mT5、UIE等预训练模型。**【结果】**在《左传》和GuNer数据集上抽取4类实体的F1值分别达到89.15%和95.47%，与使用古籍语料增量微调的基线模型SikuBERT和SikuRoBERTa相比，分别提升8.15和9.27个百分点。**【局限】**未加入实体额外信息，受限于大模型生成的数据质量。**【结论】**本文方法在低资源情境下，利用预训练大语言模型和词典资源的知识优势，将知识有效蒸馏到学生实体抽取模型，能显著提升古籍实体抽取的效果。

关键词：命名实体识别　半监督学习　大语言模型　知识蒸馏

分类号：TP393

DOI：10.11925/infotech.2096-3467.2024.0722

引用本文：唐朝，陈波，谭泽霖等. 基于知识蒸馏的半监督古籍实体抽取[J]. 数据分析与知识发现，2025，9（7）：118-129.(Tang Chao, Chen Bo, Tan Zelin,et al. Semi-Supervised Ancient Classic Entity Extraction Based on Knowledge Distillation[J]. Data Analysis and Knowledge Discovery, 2025, 9(7): 118-129.)

1　引　言

中国古代典籍承载了古代社会政治、经济、文化等多方面的重要信息，对于理解传统文化、思想、道德观念等具有不可替代的价值[1]。随着数字人文研究逐渐兴起，国内相关领域的学者在古代文献的信息抽取研究方面逐渐活跃起来。古籍数量庞大且类型丰富，人工标注不仅成本高昂，且耗费大量时间。这一现实加剧了古籍实体抽取任务所面临的困难，使得有监督数据的稀缺问题更为突出[2]。

古籍实体抽取和通用实体抽取发展路线基本相同，前期工作围绕传统方法展开，使用机器学习算法如条件随机场（Conditional Random Field，CRF）模型[3]等进行自动分词、词性标注、实体标注等任务。随着深度学习在计算机视觉领域的成功应用，结合长短期记忆网络（Long Short-Term Memory，LSTM）[4]等方法也逐渐向古籍实体抽取领域迁移。英语和现代汉语预训练语言模型的巨大成功，推动了数字人文领域的古籍语言模型发展，例如SikuBERT和SikuRoBERTa等模型[5]。语言模型的

通讯作者（Corresponding author）：赵小兵（Zhao Xiaobing），ORCID：0000-0003-1217-8650，E-mail：nmzxb_cn@163.com。

*本文系国家社会科学基金项目（项目编号：22&ZD035）的研究成果之一。

The work is supported by the National Social Science Fund of China (Grant No. 22&ZD035).

实体嵌入可以用来表示实体的语义信息和上下文关系[6]，为实体抽取任务提供更丰富的特征信息。但是，这些方案仍然无法摆脱需要大量人工标记样本的困境，标注实体类型多样化的缺失进一步限制了传统监督学习方法在古籍实体抽取任务中的应用。

本文提出一种基于教师模型蒸馏的半监督古籍实体抽取方法，利用知识蒸馏，将源自无监督数据的外部知识转化为训练数据，注入学生实体抽取模型，缓解古籍实体抽取任务中有监督数据稀缺带来的问题。具体而言，外部知识可归纳为两类：一是词典知识，利用预定义的古籍词典，将其中的实体信息作为额外补充；二是生成式知识，通过大语言模型生成新的古籍实体样本扩充训练数据集，为学生实体抽取模型提供样本多样性和上下文信息。这两种知识源通过协同作用，提高其抽取性能。本文工作归结如下：

（1）提出一种知识蒸馏的半监督古籍实体抽取方法，利用知识蒸馏将教师模型中的知识以训练数据的形式注入学生实体抽取模型。

（2）构建词典知识教师模型和生成式知识教师模型，分别将词典中的实体先验知识和大语言模型中的实体知识蒸馏出来。训练基于解码器架构的和基于编码器-解码器架构的两类学生模型。

（3）在《左传》和GuNER两个公开数据集上进行实证，本文提出的半监督方法的性能超过所有基线模型。

2　相关工作

2.1　数字人文领域的传统实体抽取方法

随着预训练语言模型的发展，相关研究以简繁体中文BERT作为编码层结合下游任务针对不同领域的语料进行持续微调，获得了良好的性能。

第一种方法是将实体抽取建模为序列标注问题，对句子中经过向量化表达的字词单元进行分类，得到每个字符串的实体标注，这类工作主要通过提升网络模型的表达能力获得更好的实验表现。肖瑞等[7]使用BiLSTM+CRF模型在医案古籍上进行中医药名抽取，能够有效处理古籍序列数据中的长期依赖关系，并考虑到标签之间的依赖关系，从而提高标注的准确性。谢靖等[8]使用Flat-Lattice增强的SikuBERT预训练模型对中医文献进行深度加工和知识标注，识别的语料为《黄帝内经·素问》。在整个数据集上依次尝试SikuBERT、SikuRoBERTa等语言模型，结合Flat-Lattice最终F1值达到89%。谢志强等[9]针对古汉语嵌套命名实体抽取任务，采用简繁体转换的殆知阁古文文献语料训练得到的GuwenBERT和RoBERTa-Classical-Chinese作为基线模型，采用全局指针将实体的开始位置和结束位置视为整体进行判别，结果证明全局指针对嵌套实体有较好的抽取性能。

第二种方法是针对标注数据的稀缺进行数据增强。刘鑫[10]针对弱监督标注数据中存在错误标注、影响以包（Package）为级别的关系抽取的问题，提出基于双注意力机制的弱监督深度学习模型，在人工标注的测试集上，联合训练后F1值达到92%。王士权等[11]在中国计算语言学大会古籍命名实体识别评测任务中利用古籍相关的领域数据和任务数据对BERT-wwm-ext进行持续预训练，再基于Pair-Wise投票的置信实体筛选算法得到候选实体，对候选实体利用上下文增强策略进行实体识别修正，获得95.87%的F1抽取表现。

第三种方法是将实体抽取转换为片段分类的方法，该类方法基于跨度（Span），其抽取策略是先枚举文本中所有可能存在的连续文本片段，然后对这些片段预测类别，最终得到实体标签。这类方法的预测结果是某个实体的边界以及它的实体类别。其代表工作为Zhu等[12]提出的实体边界平滑方法和Wang等[13]提出的构建标签空间对实体进行联合抽取方法。

除了采用额外的网络结构提高性能，本文还利用有限的标注数据，结合教师模型生成的伪标签对学生模型进行微调，缓解上述方法在特定实体抽取场景的限制。学生模型为编码-解码结构，与只使用Transformer编码器的序列标注方法相比，引入解码器能够有效地利用来自编码器的上下文古籍信息，并通过自身的注意力机制关注输入序列的不同部分，更好地生成目标实体序列。

2.2　基于生成式语言模型的实体抽取方法

大型生成式语言模型的优异性能不仅在机器翻译、问答系统等传统自然语言处理任务中得以体现，并且业界已使用大语言模型（Large Language

Model,LLM)进行实体抽取任务。这类工作的思想是将实体抽取任务转换为多轮对话,通过对大模型的问答获取句子中的实体信息。Wei等[14]将零样本信息抽取任务转变为一个两阶段框架的多轮对话,第一步用于识别类型,第二步用于识别指定类型的值。通过提示模板填充任务定义,然后调用ChatGPT接口,取得结果后进行规则解析,结构化相应答案,最终在NYT11-HRL数据集上取得了超越完全监督的实验表现(F1达到51.3%)。领域内也已推出基于古籍语料的大模型,如张君冬等[15]构建了面向中医古籍的Huang-Di大模型,在Ziya-LLaMA-13B-V1开源模型基础上,通过继续预训练、有监督微调、直接偏好优化的全流程训练的步骤构建中医古籍生成式对话大语言模型,该模型具备中医知识理解力以及中医古籍对话能力。

综上,学术界在通用领域的命名实体抽取任务中已经取得了显著成果,借助GPT4、讯飞星火等大模型的卓越推理和计算能力,在部分数据集上的表现甚至超越了传统的有监督算法[16]。然而,在古籍信息抽取细分领域,大型模型在实体抽取任务上未能取得完全成功,通用场景和特定场景的实体抽取仍然存在不小的性能差距。由于语言结构的特殊性,传统方法依然占据主导地位。本文提出的数据蒸馏方法借助大模型的能力提高学生模型[17]的泛化能力。

3 基于知识蒸馏的半监督古籍实体抽取

如图1所示,本文的知识蒸馏框架主要通过引入词典知识和大模型对无标记古籍文本进行软标签标注,分别构建词典知识教师模型和生成式知识教师模型。框架分为三部分:首先,词典知识模型通过引入额外的实体词典来扩充有监督数据集,将实体词典内蕴含的知识以数据驱动的方式注入下游模型;然后,将生成式语言模型作为教师模型进行数据蒸馏,教师模型对无标记文本进行预测,得到句子中实体的软标签,将语言模型中获取的知识信息以半监督数据的方式传递;最后,在半监督数据集上微调两种不同结构的学生模型,逐步提升学生模型的性能,实现下游任务的识别。

与传统的知识蒸馏过程不同,本文虽涉及采用教师模型和学生模型在数据集上进行训练与微调以实现知识传递[18],但本研究利用的大语言模型(LLM)已在大规模语料上完成预训练,具备丰富先验知识。相较之下,直接采用该LLM预检索召回的实体作为知识载体,可以在大幅简化传递流程的同时提升效率。在此机制中,教师模型的核心职能是提供知识指导,而非通过常规训练监督学生模型的优化过程。

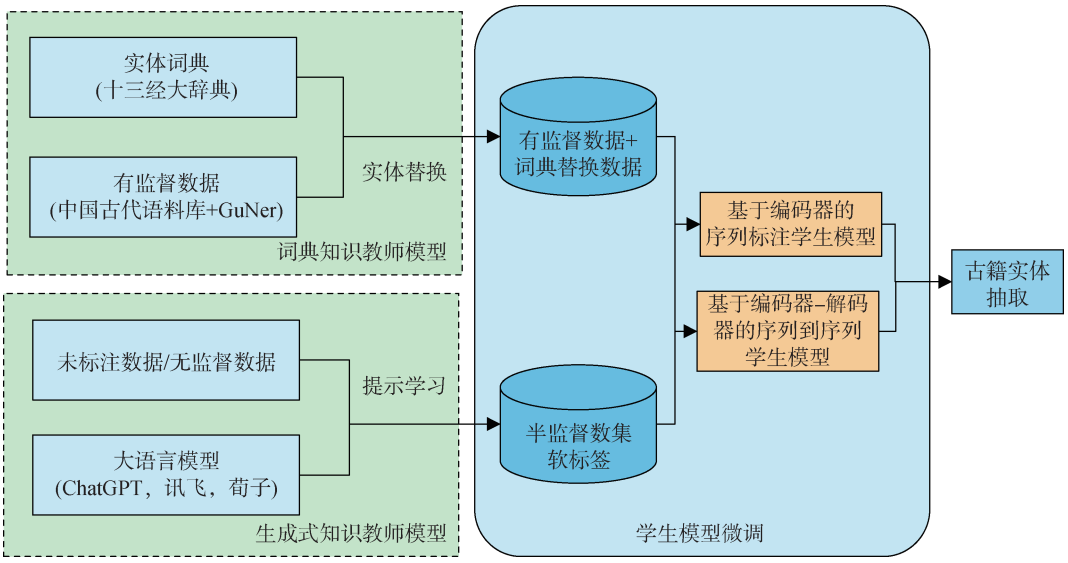


图1 知识蒸馏框架

Fig.1 Framework of Knowledge Distillation

3.1　教师模型

利用词典知识教师模型和生成式知识教师模型将有利于古籍实体抽取的知识注入自标注数据中。

(1)词典知识教师模型

使用词典作为一种知识源,将词典中的实体作为额外的标签进行监督训练[19]。从实体词典中获取抽取类型的实体,采用实体替换引入更多的变化和多样性,利用实体替换生成的数据训练学生模型,使学生模型能够更好地适应不同的实体上下文,从而提高泛化能力、对噪声和错误的抵抗能力,并降低对具体实体的过度依赖。

给定有监督数据集D_y,y_i是对应的实体。给定词典$H=\{(e_{1t1},e_{2t1},e_{kt1}),(e_{1t2},e_{2t2},e_{kt2}),\cdots,e_{kt}\}$,$e_{kt}$为实体类型$t$下的第$k$个实体。针对原数据集的实体$y_i$随机挑选字典中长度一致的$e_{kt}$进行替换,将选择的实体$e_{kt}$替换到原始文本序列$S_i$的对应位置上,得到替换后的文本序列$s_{i'}=(s_{1'},s_{2'},\cdots,s_{i'})$,新的自标注数据集为$D'_y=\{(s_{i'},y_{i'})|i=1,2,\cdots,n\}$。以人物实体为例,对于原数据中长度为2的实体“隱公”,从词典中随机抽取长度匹配的人物实体进行替换,例如“文侯”、“斗丹”。词典H生成的示例为:

原句:癸未/时间,葬 宋穆公/人物。

生成样本1:戊午/时间,葬 郑子华/人物。

生成样本2:闰月/时间,葬 赵景子/人物。

(2)生成式知识教师模型

利用预训练模型对大量未标注古籍文本进行实体自标注,将预训练模型学习到的实体抽取知识注入自标注数据。具体地,利用通用大模型,将古籍实体自标注建模为两阶段对话生成(如图2左侧所示):第一阶段对话,在提示中提供对实体抽取的任务提示[20],“你现在是一名精通古籍的专家,并且十分了解句子中的实体”指定模型的角色,给定古籍平行翻译输入及可能的实体类型,并以历史对话的方式引入少样本学习[21],加入任务描述要求模型生成句子中存在的实体类型;第二阶段对话,根据第一阶段对话得到的实体类型,进一步提问,要求模型生成对应类型的实体。

古籍大模型采用指示学习方法(Instruct Learning)进行构建。只需添加任务描述(如图2右侧所示),提供更明确的指令“请提取以下文本中的人物名/地理名/时间名”,结合古籍原文即可直接提取相应类型的实体。通过上述方法生成融入知识的古籍实体抽取自标注数据。

以“初,郑武公娶于申,曰武姜,生庄公及共叔段。”为例,其翻译为“初,郑武公在申地娶了一妻子,叫武姜,她生下庄公和共叔段。”,针对通用大模型提供少样本“宋辟公薨,子剔成立”,存在的实体类型是[人物,地理,时间]。第二阶段对话得到的结果为人物实体(郑武公,武姜,庄公,共叔段)、地理实体(申)以及时间实体为(初)。对于古籍大模型,单轮对话直接通过指令输出实体。本文重新设计针对古籍的实体问答链,具体步骤如图2所示。

本文使用的生成式语言模型分别是荀子古籍大模型①、讯飞星火大模型②和ChatGPT③,其中通用大模型的句子输入为白话文平行翻译,古籍大模型的句子输入为古籍原文。为减少教师模型的“噪声”,采用基于投票规则的集成策略[22],出现在句子中的去重实体经过处理后至少被两个模型标记投票才会保留到学生模型的微调数据集中。

以《元史》列传·卷七十四为例,原文为“伯琦仪观温雅,粹然如玉,虽遭时多艰,而善于自保”,对应的平行翻译为“伯琦态度温雅,肤白如玉,即使一生遭到很多磨难,但善于自己保重”。

荀子大模型输出实体为{人物:伯琦},ChatGPT输出实体为{人物:(周伯琦,自己)},星火大模型输出实体为{人物:伯琦}。

在所有的模型输出中,实体“伯琦”被保留,而“周”虽出现在ChatGPT的实体中,但原句中并无此独立成分,且“自己”未被其他两个模型(荀子和星火)识别输出,因此二者均未被保留。

3.2　学生模型

使用教师模型构建的自标注数据微调学生古籍

①https://github.com/Xunzi-LLM-of-Chinese-classics/XunziALLM.

②https://www.xfyun.cn/solutions/xinghuoAPI.

③https://openai.com/chatgpt/.

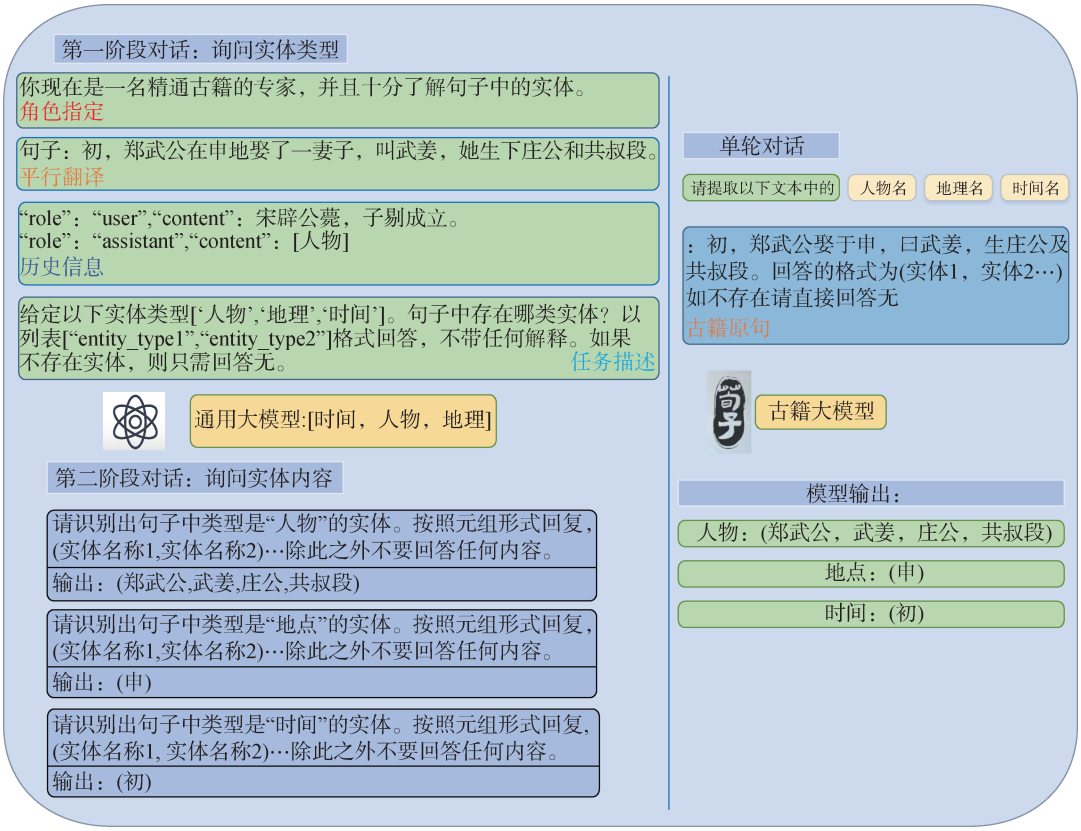


图 2　生成式知识教师模型

Fig.2　Generative Teacher Model with Knowledge

命名实体抽取模型，实现知识蒸馏。学生模型采用两种模型架构对古籍中的实体进行抽取，基于编码器序列标注模型[23]和基于编码器-解码器的序列到序列模型[24]。两种模型的输入都是古籍句子，基于编码器的序列标注模型输出为BIO标记，基于编码器-解码器序列到序列模型直接输出抽取的实体序列，两种模型的输入输出及结构如图3所示。

（1）基于编码器的序列标注模型

以“己酉，大赦”为例，模型首先通过分词器将文本转换为字编码，进而转换为输入的向量。经过向量化表达后得到词嵌入，再经过编码器进行特征提取，进而获得每个位置的隐层表示。模型输出标注序列，其中每个位置对应一个隐层向量。将每个位置的隐层向量通过一个全连接层进行线性映射，转换为一个维度与实体类别数量相同的向量。最终输入Softmax函数转换为概率分布，归一化得到每个BIO标记类别的概率值，模型输出标签：[B-time，I-time，O，O，O]。

给定半监督古籍数据集 $D=\{D_y,D_n\}$，D_n 为教师模型获得的自标注数据集。对应的标注序列为 $B=(b_1,b_2,\cdots,b_i)$。s_i 为数据集 D 第 i 条数据，其实体标签为 b_i。

给定输入的句子 s_i，经过词向量表达，再通过编码器（Encoder）得到输出，如公式（1）所示。

$$E_t = Encoder(Embedding(s_i)) \tag{1}$$

使用一层分类器层作为模型的结束，最终的输出如公式（2）所示。

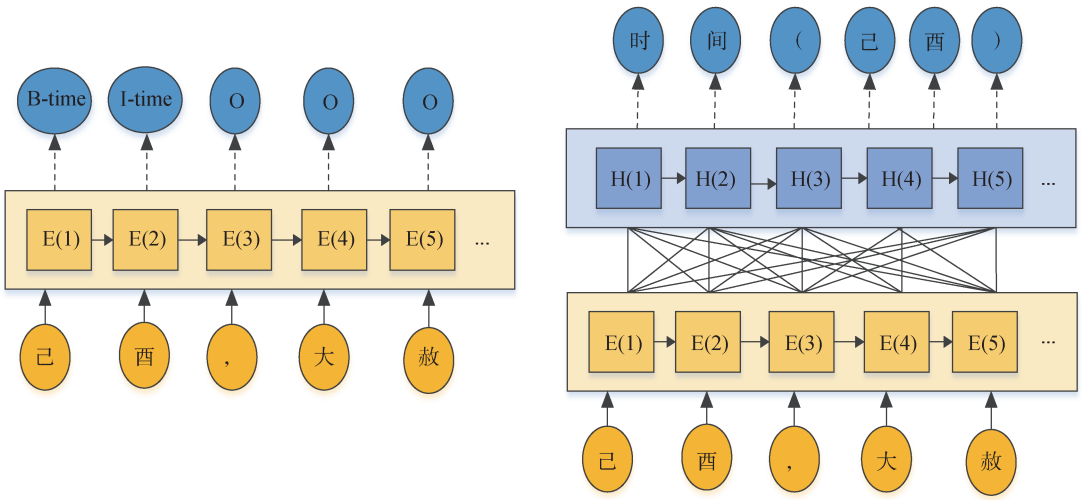
$$\hat{y} = \mathrm{softmax}(w^{\mathrm{T}}Et + b) \tag{2}$$

其中，$\hat{y}$ 为模型预测的BIO标签，w 为全连接层的权重，b 为偏置。损失函数如公式（3）所示。在训练过程中利用梯度下降最小化该损失函数。

$$Loss = -\sum_{i}^{B} b_i \log(p(x_i = \hat{y})) \tag{3}$$

（2）基于编码器-解码器的序列到序列模型

该模型的输入与基于编码器的序列标注模型相同，仍然为古籍的原句“己酉，大赦”。其结构本质是



(a) Transformer-Encoder only　　(b) Transformer-Encoder&Decoder

图 3　学生模型

Fig.3　Students Model

在 Transformer 的编码器基础上加入解码器，通过注意力机制计算解码器当前位置的字与编码器输出之间的注意力分数，聚焦于古籍句子中与当前位置相关的信息。在得到编码器的所有中间层表示 E_t 后，经过解码器（Decoder）进行解码。依然使用全连接层进行线性映射。不同于基于编码器的序列标注模型，未归一化分数不再是 BIO 标签的输出概率，而是词表维度的输出，即每个词汇的概率。其中每一维代表一个单词的得分，分类器层将得分归一化并选取最高概率的字符，得到的解码序列即实体，模型直接输出：时间（己酉）。

4　古籍实体抽取实验

4.1　数据集

本文在有监督数据集的基础上构建了两个自标注数据集。通过不同组合形式的训练数据微调学生模型，并在两个测试数据集上进行评估。词典知识教师模型和生成式知识教师模型获得的数据仅被用于在训练阶段微调学生模型，通过从有监督数据中随机抽取样本构建验证集和测试集。为确保数据一致性，进行以下预处理：①统一文本表示：使用 OpenCC 工具将古籍中的繁体字转换为简体字；②去除生僻字干扰：通过正则表达式过滤生僻字，在数据蒸馏阶段去除，防止对下游模型产生干扰；③实体标记统一：采用 PER（人物）、TIME（时间）、OFI（官职）、LOC（地理）标签表示实体类型。

（1）有监督数据集

古籍实体抽取的有监督数据集来自《左传》和《二十四史》，共包含 4 类实体类型，其中人物、地理、时间实体来自石民等[25]构建的中国古代语料库①，官职实体来自“古籍命名实体识别 2023”（GuNer 2023）数据集②。针对《左传》数据集，将词性标注数据进行转换，得到人物（对应词性 nt）、时间（对应词性 nr）、地理（对应词性 ns）三类实体，获得 8 699 条数据，其中 1 999 条作为测试数据集；针对 GuNer 数据集（受限于词典，书籍实体未纳入），直接将该数据集中的人物实体和官职实体抽取出来，获得 2 346 条数据，其中 200 条作为测试数据集。数据集的实体统计和示例如表 1 所示。

（2）基于生成式知识模型构造的数据集

用于构建自标注数据的原始语料为文言文-现代文平行语料③，共计 972 467 句。使用的荀子古籍

①中国古代语料库：https://catalog.ldc.upenn.edu/LDC2017T14.

②GuNer2023：https://guner2023.pkudh.org/.

③https://github.com/NiuTrans/Classical-Modern.git.

表1　有监督数据集

Table 1　Supervised Dataset

数据集	语料来源	实体	实体数(含重复)
中国古代语料库	《左传》	人物	10 662
		地理	5 199
		时间	1 417
GuNer	《二十四史》	人物	6 670
		官职	3 363

大模型为Xunzi-Qwen-7B-CHAT,星火大模型为V3版本,ChatGPT为GPT-3.5 Turbo-1106。使用通用大模型提问时采用无标注的现代文翻译,再对原句子进行回标。例句:‘乃留建业。’,‘于是孙登留在建业。’,{‘人物’:[‘孙登’],‘地理’:[‘建业’],‘时间’:[]}。该句抽取的人物为孙登,地理为建业,回标时就会省略掉“孙登”。基于两类方法共构建43万条有实体数据,根据实体数量和实体类型对数据集进一步筛选,最终得到3万条数据用于训练。无监督数据集的统计如表2所示。

表2　无监督数据集

Table 2　Unsupervised Dataset

数据	语料来源	实体	实体数(含重复)
文言文现代平行语料	《史记》《资治通鉴》《汉书》等	人物	492 474
		地理	173 151
		时间	173 151
		官职	61 095

(3)基于词典知识模型构造的数据集

该部分数据由词典知识生成,使用的词典为《十三经大辞典》,共14 695个实体,其中人物、时间、地理和官职实体分别为4 081、440、1 444和919个。原始文本包含中国古代语料和GuNer语料共11 045条,其中2 191条未包含任何实体。考虑到单字实体容易造成歧义,如地理中的“叶”“冯”“汉”,官职中的“子”“公”“士”“王”等,仅对原语料中长度大于等于2的实体进行同类替换。除去无实体的句子和所有实体全部为单字的句子,中国古代语料库和GuNer可供使用的样本分别为6 451条和1 566条,对每个句子随机选取实体进行替换,每条原始数据生成两条样本,分别得到12 902条和3 132条样本。

①https://github.com/universal-ie/UIE.git.

4.2　实验环境和实验参数

采用Ubuntu操作系统,深度学习框架为TensorFlow2.5。配置CPU为e5-2680 v4,在内存32GB以及显存32GB的Tesla V100上进行训练。训练参数如表3所示。

表3　训练参数

Table 3　Configuration of Training Parameters

参数	取值
Epoch	15/20
Batch_Size	32
Learning_Rate	1e-5/2e-5
Weight_Decay	0.01
Warmup_Steps	100
Maxlen	512

基于编码器的序列标注学生模型采用古文处理预训练模型SikuBERT和SikuRoBERTa。两者均在BERT-Base[26]和RoBERTa[27]基础上使用《四库全书》语料增量预训练得到。

基于编码器-解码器的序列预测学生模型为Mengzi(中文预训练模型T5)[28]、mT5[29](多语言版本T5)和UIE[30](信息抽取预训练模型T5)。T5模型将自然语言处理(Natural Language Processing,NLP)的多种任务转化为“文本到文本”的问题,使单一模型能够处理多种类型的NLP任务。mT5在T5的基础上,增加了对中文的支持。Mengzi重新训练了分词器,使用更大规模的中文语料进行微调以适应中文下游任务。UIE①是基于T5的通用信息抽取预训练模型,将不同的信息抽取任务统一建模为“文本到结构”范式。与序列标注模型不同,Mengzi、mT5和UIE都没有在繁体古籍语料上进行增量微调,整体参数如表4所示。

4.3　评价标准

实体抽取任务的评价指标采用精确率(Precision)、召回率(Recall)和F1值,其中F1值为Micro F1-Score。

4.4　《左传》和GuNer抽取结果

本文在《左传》数据集上进行时间、地理、人物实体的抽取评测,在GuNer数据集上进行人物、官职实

表4　语言模型训练参数

Table 4　Training Configuration for Language Models

模型名称	SikuBERT	SikuRoBERTa	Mengzi-T5-Base	mT5-Base	UIE-Char-Small
词表大小	29 791	29 791	32 128	250 112	18 118
隐层数量	12	12	12	12	8
注意力头	12	12	12	6	6
注意力随机 Dropout	0.1	0.1	0.1	0.1	0.1
序列最大长度	512	512	512	512	128
训练语料	《四库全书》	《四库全书》	300GB 互联网语料	多语言数据集 C4	英文维基百科中的所有纯文本，Wikidata 收集的结构化数据
训练方式	增量训练	增量训练	重新训练	重新训练	增量训练
分词器	字粒度	字粒度	字粒度	字节对编码	字粒度
结构	Encoder-Only	Encoder-Only	Encoder & Decoder	Encoder & Decoder	Encoder & Decoder

体的评测。基线模型为 SikuBERT 和 SikuRoBERTa，整体抽取结果如表 5 和表 6 所示。

表5　《左传》抽取实验结果

Table 5　Experimental Result of ZuoZhuan Dataset

模型	精确率(%)	召回率(%)	F1 值(%)
SikuBERT	\	\	81.59
SikuRoBERTa	\	\	81.33
SikuBERT+词典知识	81.73	88.95	85.18
SikuBERT+词典知识+大模型知识	85.15	86.43	85.79
SikuRoBERTa+词典知识	84.18	86.21	85.18
SikuRoBERTa+词典知识+大模型知识	84.74	88.65	86.20
Mengzi	79.65	78.56	79.10
mT5	81.34	83.18	82.24
mT5+词典知识	84.94	88.42	86.65
mT5+词典知识+大模型知识	85.20	86.80	85.99
UIE	89.43	88.37	88.90
UIE+词典知识	88.11	88.09	88.11
UIE+词典知识+大模型知识	88.98	89.32	**89.15**

表6　GuNer抽取实验结果

Table 6　Experimental Result of GuNer Dataset

模型	精确率(%)	召回率(%)	F1 值(%)
SikuBERT	82.97	89.71	86.20
SikuRoBERTa	86.25	89.06	87.63
UIE	89.86	87.24	88.53
UIE+词典知识	85.49	86.73	86.10
UIE+词典知识+大模型知识	96.25	94.71	**95.47**

实验结果表明，采用知识蒸馏能够显著提升学生模型在《左传》数据集的实体抽取性能。具体而言，经过微调后的编码器学生模型 SikuBERT 和 SikuRoBERTa 的初始 F1 值分别为 81.59% 和 81.33%，引入词典知识后可以达到 85.18%，结合生成式语言知识后分别提升至 85.79% 和 86.20%，知识蒸馏使序列标注学生模型提升约 4 个百分点。对于编码器-解码器学生模型，mT5 和 UIE 初始 F1 值为 82.24% 和 88.90%，采用知识蒸馏后为 85.99% 和 89.15%，分别提升 3.75 和 0.25 个百分点。

知识蒸馏可以显著提升稀疏类型的实体抽取性能，相比人物、地理、时间标注数据较为丰富的《左传》数据集，在 GuNer 数据集上，UIE 的 F1 值从 88.53% 提升到 95.47%，提升了 6.94 个百分点，相比基线模型的提升约 9 个百分点，这表明教师模型的知识融入对于稀疏类型的实体抽取任务具有显著的促进作用。

此外，采用编码器-解码器的学生模型相比只使用编码器的学生模型表现更优，在 T5 类模型中只有 Mengzi 初始 F1 值低于 80%，而 UIE 模型在不采用任何知识蒸馏的情况下在《左传》和 GuNer 上的表现就可以达到 88.90% 和 88.53%，相较于基线模型存在 7.31 和 2.33 个百分点的性能优势。UIE 在大规模的数据源中学习到了更通用的信息抽取能力，向低资源任务迁移时仍然保持了其有效性。

4.5　不同类型实体性能评估

为进一步对比分析模型的实体抽取性能，选择知识蒸馏后的 mT5、UIE 和基线模型进行横向比较，《左传》数据集的每类实体抽取结果如图 4 所示，基线模型三类实体抽取性能如表 7 所示。

表7　基线模型在《左传》数据集上的F1值

Table7　F1-Score of the Baseline Models in ZuoZhuan

模型	实体	Epoch=3	Epoch=10
SikuBERT	地理	82.04	81.82
	人物	78.48	79.38
	时间	91.77	92.35
SikuRoBERTa	地理	81.58	80.29
	人物	78.65	79.80
	时间	89.88	91.67

在《左传》的实验结果表明，知识蒸馏后的UIE模型在不同类型实体的抽取任务中表现最为均衡且较优，其精确率、召回率和F1值均保持在85%以上。其中识别性能最好的实体仍然是人物实体，精确率和F1值可以达到90%。而时间实体的F1值（87.00%）虽略低于其他两类实体，但较基线模型（F1值为89.88%）仅低约2.88个百分点，差距相对较小。在编码器-解码器结构的学生模型中，UIE表现优于mT5。

在GuNer数据集上也进行了性能评估，选取知识蒸馏后的UIE和基线模型进行对比，官职实体和人物实体抽取结果如图5所示。

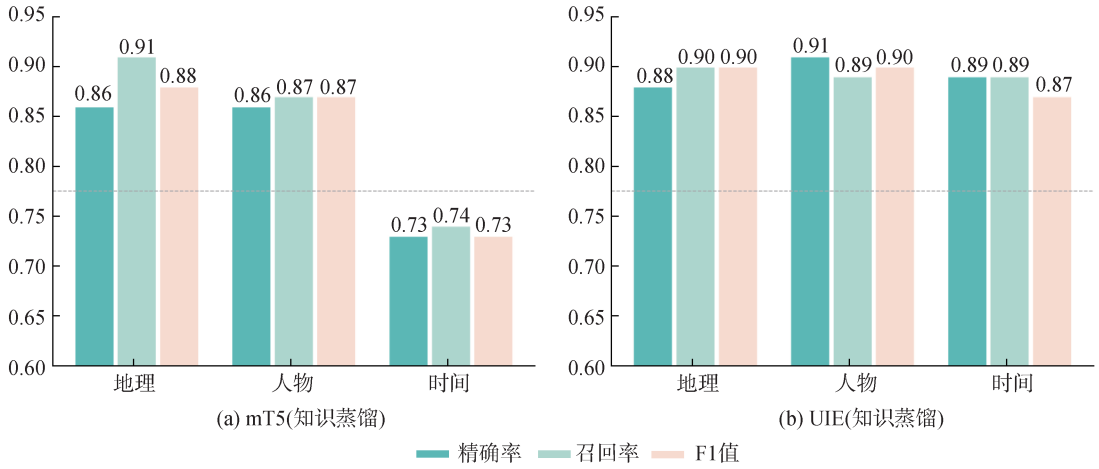


图4　《左传》数据集的每类实体抽取表现

Fig.4　Experimental Performance of Various Entities in ZuoZhuan

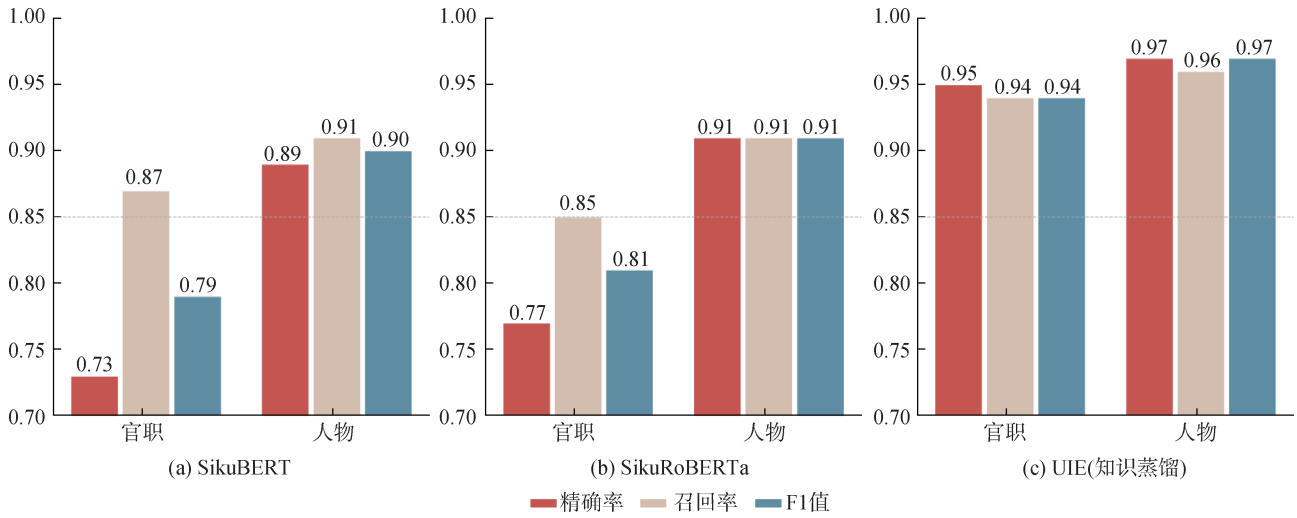


图5　GuNer每类实体抽取表现

Fig.5　Experimental Performance of Various Entities in GuNer

使用知识蒸馏对官职类实体抽取提升效果最明显。基线模型在 GuNer 上的实体抽取表现明显波动，SikuRoBERTa 模型进行人物实体抽取的 F1 值超过 91%，而对官职实体仅在 81% 左右。由于人物实体的数量约为官职实体的两倍，模型对人物实体较为敏感。这从侧面验证了 UIE 模型的鲁棒性，采用知识蒸馏后不仅提高了实体抽取的性能，还弥补了实体间的性能差距，官职与人物实体的 F1 值仅相差 3 个百分点。UIE 相比基线模型性能更优，所有指标均在 94% 以上，提升较为明显。

4.6　不同数量的生成式知识蒸馏数据的影响

为评估不同数量的生成式知识教师模型蒸馏数据的影响，在融入词典知识的基础上，调整生成式教师模型的数据量。随机抽取 5 000、10 000 和 15 000 条数据进行对比实验，其中抽取性能最好的 UIE 模型在全量数据集和 GuNer 数据集上的抽取结果分别如图 6 和图 7 所示。

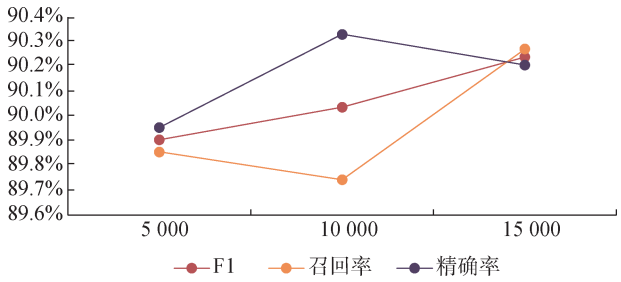


图 6　不同蒸馏数据数量下 UIE 在全量数据集的表现

Fig.6　Experimental Performance of UIE Under Different Numbers of Distillation Data in All Dataset

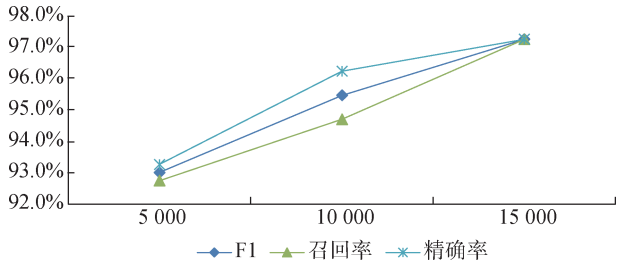


图 7　不同蒸馏数据数量下 UIE 在 GuNer 数据集上的表现

Fig.7　Experimental Performance of UIE Under Different Numbers of Distillation Data in GuNer

从抽取结果来看，随着大模型蒸馏的数据增多，UIE 模型的抽取性能也有所提高。在不同训练数据量下，模型的 F1 值分别为：5 000 条时 89.90%，10 000 条时 90.03%，15 000 条时 90.23%，数据量增加至 15 000 条时相较 5 000 条提升约 0.3 个百分点。此外，在 GuNer 数据集上，模型性能由 93.01% 显著提升至 97.26%，提升幅度达 4.25 个百分点。受限于随机采样数据的质量，在全量数据上模型的精确率和召回率并不总是线性提升，召回率先降后升，精确率先升后降。这说明模型从原数据集学习到的知识，会被蒸馏数据中产生的“噪音”所影响。随着数据量的增加，干扰会得到修正。而 GuNer 数据集中的官职实体往往具备更高的一致性，模型获得的知识置信度更高，其性能呈现出持续、稳定的提升趋势。

5　结　语

针对有监督数据样本稀缺问题，本文使用知识蒸馏技术进行缓解，通过词典进行实体替换，将模型训练数据泛化为更一般化的形式，将词典中的知识融入模型。再通过大模型进行实体扩充，丰富实体的多样化。这一策略既实现了知识的有效传递与融合，也提高了学生模型的泛化能力。两个数据集上的实验结果表明，上述策略是有效的。但是本文未引入更多实体信息用于提升识别能力，未来将进一步加入章节名称、上下文等额外信息开展研究。

参考文献：

[1]　王永友，骆丹．习近平关于历史文化遗产保护利用的重要论述研究[J]．文化软实力，2023，8(2)：14-22.(Wang Yongyou, Luo Dan. On Xi Jinping's Important Expositions on the Protection and Utilization of Historical and Cultural Heritage[J]. Cultural Soft Power, 2023, 8(2): 14-22.)

[2]　刘耀，李冠霖，李浣青．面向中医古籍的单篇文本知识标引与结构解析技术[J]．图书情报工作，2022，66(24)：118-127.(Liu Yao, Li Guanlin, Li Huanqing. Knowledge Indexing and Structural Analysis Techniques for Single Text of Ancient Chinese Medical Books[J]. Library and Information Service, 2022, 66(24): 118-127.)

[3]　王铮．基于CRF的古籍地名自动识别研究——以《三国演义》为例[D]．南宁：广西民族大学，2008.(Wang Zheng. Conditional Random Fields Based Location Name Recognition in Ancient Chinese——Take the “Romance of the Three Kingdoms” as an Example[D]. Nanning: Guangxi Minzu University, 2008.)

[4]　苏祺，胡韧奋，诸雨辰，等．古籍数字化关键技术评述[J]．数字人文研究，2021，1(3)：83-88.(Su Qi, Hu Renfen, Zhu Yuchen,

et al. Key Technologies for Digitization of Ancient Chinese Books [J]. Digital Humanities Research, 2021, 1(3): 83-88.)

[5] 王东波, 刘畅, 朱子赫, 等. 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[J]. 图书馆论坛, 2022, 42(6): 31-43. (Wang Dongbo, Liu Chang, Zhu Zihe, et al. Construction and Application of Pre-Trained Models of Siku Quanshu in Orientation to Digital Humanities[J]. Library Tribune, 2022, 42(6): 31-43.)

[6] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[OL]. arXiv Preprint, arXiv: 1301.3781.

[7] 肖瑞, 胡冯菊, 裴卫. 基于BiLSTM-CRF的中医文本命名实体识别[J]. 世界科学技术-中医药现代化, 2020, 22(7): 2504-2510. (Xiao Rui, Hu Fengju, Pei Wei. Chinese Medicine Text Named Entity Recognition Based on BiLSTM-CRF[J]. Modernization of Traditional Chinese Medicine and Materia Medica-World Science and Technology, 2020, 22(7): 2504-2510.)

[8] 谢靖, 刘江峰, 王东波. 古代中国医学文献的命名实体识别研究——以Flat-Lattice增强的SikuBERT预训练模型为例[J]. 图书馆论坛, 2022, 42(10): 51-60. (Xie Jing, Liu Jiangfeng, Wang Dongbo. Study on Named Entity Recognition of Traditional Chinese Medicine Classics: Taking SikuBERT Pre-Training Model Enhanced by the Flat-Lattice Transformer for Example[J]. Library Tribune, 2022, 42(10): 51-60.)

[9] 谢志强, 刘金柱, 刘根辉. 古汉语嵌套命名实体识别数据集的构建和应用研究[C]//第21届中国计算机语言学大会论文集. 2022: 406-416. (Xie Zhiqiang, Liu Jinzhu, Liu Genhui. Construction and Application of Classical Chinese Nested Named Entity Recognition Data Set[C]//Proceedings of the 21st Chinese National Conference on Computational Linguistics. 2022: 406-416.)

[10] 刘鑫. 基于弱监督深度学习的中医文本关系抽取研究[D]. 唐山: 华北理工大学, 2020. (Liu Xin. Research on Traditional Chinese Medicine Texts Relation Extraction Based on Weakly Supervised Deep Learning[D]. Tangshan: North China University of Science and Technology, 2020.)

[11] 王士权, 石玲玲, 蒲璐汶, 等. CCL23-Eval任务1系统报告:基于持续预训练方法与上下文增强策略的古籍命名实体识别[C]//第22届中国计算机语言学大会论文集. 2023: 14-22. (Wang Shiquan, Shi Lingling, Pu Luwen, et al. System Report for CCL23-Eval Task 1: Named Entity Recognition for Ancient Books Based on Continual Pre-Training Method and Context Augmentation Strategy[C]//Proceedings of the 22nd Chinese National Conference on Computational Linguistics. 2023: 14-22.)

[12] Zhu E W, Li J P. Boundary Smoothing for Named Entity Recognition[OL]. arXiv Preprint, arXiv: 2204.12031.

[13] Wang Y J, Sun C Z, Wu Y B, et al. UniRE: A Unified Label Space for Entity Relation Extraction[OL]. arXiv Preprint, arXiv: 2107.04292.

[14] Wei X, Cui X Y, Cheng N, et al. ChatIE: Zero-Shot Information Extraction via Chatting with ChatGPT[OL]. arXiv Preprint, arXiv: 2302.10205.

[15] 张君冬, 杨松桦, 刘江峰, 等. AIGC赋能中医古籍活化: Huang-Di大模型的构建[J]. 图书馆论坛, 2024, 44(10): 103-112. (Zhang Jundong, Yang Songhua, Liu Jiangfeng, et al. AIGC Empowering the Revitalization of Ancient Books on Traditional Chinese Medicine: Building the Huang-Di Large Language Model[J]. Library Tribune, 2024, 44(10): 103-112.)

[16] Han R, Peng T, Yang C, et al. Is Information Extraction Solved by ChatGPT? An Analysis of Performance, Evaluation Criteria, Robustness and Errors[OL]. arXiv Preprint, arXiv: 2305.14450.

[17] Ho N, Schmid L, Yun S Y. Large Language Models are Reasoning Teachers[OL]. arXiv Preprint, arXiv: 2212.10071.

[18] Fukuda T, Suzuki M, Kurata G, et al. Efficient Knowledge Distillation from an Ensemble of Teachers[C]//Proceedings of the 18th Annual Conference of the International Speech Communication Association. ISCA, 2017: 3697-3701.

[19] 徐秋荣. 基于特征融合的中文命名实体识别[D]. 上海: 华东师范大学, 2022. (Xu Qiurong. Chinese Named Entity Recognition Based on Feature Fusion[D]. Shanghai: East China Normal University, 2022.)

[20] Schick T, Schütze H. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference[C]// Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2021: 255-269.

[21] Wang Y Q, Yao Q M, Kwok J T, et al. Generalizing from a Few Examples: A Survey on Few-Shot Learning[J]. ACM Computing Surveys, 2020, 53(3): Article No.63.

[22] You Z, Feng S L, Su D, et al. SpeechMoE2: Mixture-of-Experts Model with Improved Routing[OL]. arXiv Preprint, arXiv: 2111.11831.

[23] Liu W, Fu X Y, Zhang Y, et al. Lexicon Enhanced Chinese Sequence Labeling Using BERT Adapter[OL]. arXiv Preprint, arXiv: 2105.07148.

[24] Yan H, Gui T, Dai J, et al. A Unified Generative Framework for Various NER Subtasks[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021: 5808-5822.

[25] 石民, 李斌, 陈小荷. 基于CRF的先秦汉语分词标注一体化研究[J]. 中文信息学报, 2010, 24(2): 39-45. (Shi Min, Li Bin, Chen Xiaohe. CRF Based Research on a Unified Approach to Word Segmentation and POS Tagging for Pre-Qin Chinese[J]. Journal of Chinese Information Processing, 2010, 24(2): 39-45.)

[26] Vaswani A, Shazeer N, Parmar N, et al. Attention is All You Need [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 6000-6010.

[27] Liu Y, Ott M, Goyal N, et al. RoBERTa: A Robustly Optimized BERT

Pretraining Approach[OL]. arXiv Preprint, arXiv: 1907.11692.
[28] Zhang Z S, Zhang H Q, Chen K M, et al. Mengzi: Towards Lightweight yet Ingenious Pre-Trained Models for Chinese[OL]. arXiv Preprint, arXiv: 2110.06696.
[29] Raffel C, Shazeer N, Roberts A, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer[J]. The Journal of Machine Learning Research, 2020, 21(1): 5485-5551.
[30] Lu Y, Liu Q, Dai D, et al. Unified Structure Generation for Universal Information Extraction[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. 2022: 5755-5772.

作者贡献声明：

赵小兵：提出研究思路，设计研究方案；
谭泽霖：采集、清洗和分析数据；
唐朝：进行实验，撰写论文；
陈波：论文最终修订。

利益冲突声明：

所有作者声明不存在利益冲突关系。

支撑数据：

[1] 唐朝 . 半监督古籍实体蒸馏数据集及代码 .https://github.com/Greenhorntc/Ancient_Classic_Ner.

收稿日期：2024-07-22
收修改稿日期：2024-10-13

Semi-Supervised Ancient Classic Entity Extraction Based on Knowledge Distillation

Tang Chao[1, 2]　Chen Bo[3]　Tan Zelin[4, 5]　Zhao Xiaobing[3, 5]

[1](School of Philosophy and Religious Studies, Minzu University of China, Beijing 100081, China)
[2](Institute of National Security, Minzu University of China, Beijing 100081, China)
[3](School of Information Engineering, Minzu University of China, Beijing 100081, China)
[4](School of Chinese Ethnic Minority Languages and Literatures, Minzu University of China, Beijing 100081, China)
[5](National Language Resource Monitoring and Research Center of Minority Languages, Beijing 100081, China)

Abstract: [**Objective**] This work aims to address the challenge of scarce supervised data in classical Chinese entity extraction by leveraging knowledge distillation techniques to inject knowledge from unsupervised external sources into a student model. [**Methods**] A large language model is utilized as a generative knowledge teacher model to perform knowledge distillation on unsupervised corpora. Additionally, a dictionary knowledge teacher model is built using supervised data from the ZuoZhuan and GuNer datasets. The knowledge distilled from both teachers is integrated to compile a semi-supervised dataset for classical Chinese entity extraction. The task is then reformulated as a sequence-to-sequence problem, and pre-trained models such as mT5 and UIE are fine-tuned on this dataset. [**Results**] On the ZuoZhuan and GuNer datasets, the proposed method achieves F1-Score of 89.15% and 95.47%, respectively, outperforming the baseline models SikuBERT and SikuRoBERTa, which were incrementally fine-tuned on classical Chinese corpora, by 8.15% and 9.27% in F1-Score. [**Limitations**] The method does not incorporate additional entity type information, and the quality of data pre-retrieved by the LLMs may affectt extraction results. [**Conclusions**] In low-resource settings, the proposed approach effectively distills the knowledge advantages of pre-trained large language models and dictionary resources into the student entity extraction model, significantly improving the performance on classical Chinese entity extraction tasks.

Keywords: Named Entity Recognition　Semi-supervised Learning　LLMs　Knowledge Distillation