



开放科学  
(资源服务)  
标识码  
(OSID)

# 基于 UIE 系列模型的古籍文本自动标注性能比较研究

王起扬 刘忠宝

北京语言大学信息科学学院 北京 100083

**摘要:** [目的/意义] 古籍标注是古籍信息处理的基础,传统的人工标注方式费时费力,如何高效、准确地对古籍文本进行自动标注成为了一个亟待解决的问题。利用 UIE 系列模型在文本信息抽取方面的优势,结合古籍文本的特点,引入古籍文本的针对性技术,探究该模型在古籍文本自动标注方面的有效性及差异性。[方法/过程] 围绕实体关系抽取与事件论元抽取两类任务进行分析,从准确率、召回率、F1 值等方面对古籍文本的自动标注性能进行比较,以期明晰模型规模、领域适应性以及标注质量之间的关系。[结果/结论] 《二十四史》语料集上的实验结果表明,随着模型规模和训练样本规模的增大,UIE 系列模型的标注性能呈上升趋势,当训练样本规模为 2500 例上下时,实体关系抽取和事件论元抽取的 F1 值均达到最优,分别为 72.08% 和 70.57%。

**关键词:** UIE 系列模型; 古籍文本; 自动标注; 比较研究

**中图分类号:** G256; G35

## Comparative Research on Annotation Performance of Ancient Texts Based on UIE Series Models

WANG Qiyang LIU Zhongbao

School of Information Science, Beijing Language and Culture University, Beijing 100083, China

**Abstract:** [Objective/Significance] The annotation of ancient books is the foundation of ancient texts information processing. Traditional manual annotation methods are time-consuming and laborious. How to efficiently and accurately perform automatic annotation of ancient book texts has become an urgent problem to be solved. This article utilizes the advantages of the Unified Information Extraction models in text information extraction, combined with the characteristics and targeted technology of ancient texts, to explore the effectiveness and differences of this model in automatic annotation of ancient texts. [Methods/Processes] This article analyzes two types of tasks, entity relationship extraction and event argument extraction, and compares the automatic annotation performance of ancient text from the aspects of accuracy, recall, F1 value, in order to clarify the relationship between model size, domain adaptability, and annotation quality. [Results/Conclusions] The experimental results on

**基金项目** 国家社会科学基金重点项目“大数据时代古籍活化赋能文化自信自强的理论、方法与路径研究”(23AZD047)。

**作者简介** 王起扬(1999-), 博士研究生, 主要研究方向为数字人文; 刘忠宝(1981-), 通信作者, 博士, 教授, 主要研究方向为数字人文, E-mail: zbliu@blcu.edu.cn。

**引用格式** 王起扬, 刘忠宝. 基于 UIE 系列模型的古籍文本自动标注性能比较研究[J]. 情报工程, 2025, 11(6): 28-46.

the “Twenty-Four Histories” corpus show that as the model size and training sample size increase, the annotation performance of the UIE series models shows an upward trend. When the training sample size is around 2500, the F1 values for entity relationship extraction and event argument extraction reach their optimal levels, which are 72.08% and 70.57% respectively.

**Keywords:** UIE Series Models; Ancient Texts; Automatic Annotation; Comparative Research

## 引言

随着大数据和人工智能技术的发展，古籍数字化不仅限于文本的扫描和保存，更包括对文本内容的深入挖掘与自动化处理。其中，古籍文本自动标注是古籍信息处理的基础，其效率与准确性直接决定着整个古籍信息处理流程的效能。UIE (Unified Information Extraction)<sup>[1]</sup> 系列模型作为信息抽取领域的重要工具，以其强大的语义理解和信息提取能力，在文本信息抽取方面展现出广阔前景。古籍文本作为文本资源的一个重要组成部分，目前尚未实现自动标注。能否将 UIE 系列模型引入到古籍文本自动标注，是一个值得深入探究的问题。因此，本文首先结合古籍文本的特点，引入古籍文本的针对性技术，随后将《二十四史》作为实验语料，引入 UIE 系列模型，围绕自动标注任务中的实体关系抽取和事件论元抽取两类任务进行实验设计，从准确率、召回率、F1 值等方面对古籍文本的自动标注性能进行比较分析。此外，本文还根据实验结果，探讨了模型规模、领域适应性与标注质量之间的关系。本文研究为古籍文本标注研究提供了有益参考，也为全面提升古籍信息处理性能提供了重要支撑。

## 1 研究进展

传统的古籍信息处理依赖于人工标注，这

种方法不仅效率低下，而且容易受到人为因素的影响。因此，利用计算机技术实现古籍文本的自动标注具有重要的现实意义，自动标注技术是提升古籍信息处理效率的关键。时至今日，尚未出现古籍文本自动标注的工具或模型。然而，近年来出现的 UIE<sup>[1]</sup> 系列模型，以其强大的复杂语义理解与结构化信息抽取能力，为古籍文本自动标注提供了重要借鉴和参考。

目前，UIE 系列模型应用于现代文文本自动标注。皮俊波等<sup>[2]</sup> 将 UIE 框架应用于电网故障处置预案的实体和事件识别中，该研究通过句法分析、ERNIE 3.0 编码及双指针解码等技术手段，有效解决了电网故障处置预案中实体嵌套复杂、区域差异化明显的问题。朱杰等<sup>[3]</sup> 将 UIE 模型引入情感分析领域，提出了一个情感可解释分析方法。该方法不仅判断文本极性，还能给出判断所依据的证据，从而提高了模型的可解释性，这一创新对于理解模型决策过程、增强用户信任具有重要意义。李昌鏢等<sup>[4]</sup> 展示了 UIE 模型在构建特定领域知识图谱方面的能力。该研究通过微调 ERNIE-UIE 模型，实现了对公共卫生事件相关信息的实体关系联合抽取，并成功构建了深圳市公共卫生事件应急管理知识图谱，这一成果不仅验证了 UIE 模型在垂直领域低资源情况下的适用性，还为提升公共卫生事件应急管理提供了有力支持。该文献表明，UIE 模型在构建古籍文本知识图谱方面

也具备潜力，能够帮助研究者快速构建结构化的古籍知识库。

自动标注任务中的实体关系抽取按照文本长度可分为句子级和文档级两种类型，每种类型都有其特定的应用场景和优缺点。任安琪等<sup>[5]</sup>详细阐述了这两种关系抽取方法的现状、所用数据集及其特点，并分析了各自的优势与不足。句子级关系抽取在处理小规模数据集时表现出色，如知识图谱问答系统；而文档级关系抽取则更适用于处理复杂上下文和长距离依赖关系的场景，如新闻事件分析。吴梦成等<sup>[6]</sup>从跨学科视角出发提出了一种创新的方法，该方法通过实体抽取和关系抽取等技术手段，对古农书的非结构化知识进行系统梳理与组织，不仅促进了古农书知识的理解，还为文化遗产与现代应用提供了新的思路。崔斌等<sup>[7]</sup>则聚焦于古代农作物在地理空间上的分布特征，通过文本实体标注等现代信息技术手段，揭示了古代中国典籍中农作物与地点实体间的关联关系，并结合区位商、基尼系数等集聚性测度指标进行了综合分析。

自动标注任务中的事件论元抽取同样可以按照文本长度或范围进行分类，薛继伟等<sup>[8]</sup>探讨了基于提示学习的篇章级事件论元抽取方法，传统的基于提示学习的方法需要人工手动构建提示模板，这种方法在模板结构固定的情况下容易导致论元抽取错误。针对这一不足，论文提出了一种基于文本触发词自动构建提示模板的方法，并在输入文本中融入事件角色语义信息，以提高模型对文本语义特征的捕获能力。王潞翔等<sup>[9]</sup>提出了结合二维增强融合机制的事件论元抽取模型（W2-ARG），以解决触发词

和论元间缺少交互以及通道内部缺少交互的问题。该方法通过在触发词两边插入标识符引入事件类型信息，并增强触发词和论元的交互。同时，利用膨胀卷积捕获不同距离的单词的语义交互，并通过通道注意力模块增强通道内部的交互。最后，使用拉普拉斯算子突出事件论元在语义空间中的位置特征，显著提升了事件论元抽取的效果。于媛芳等<sup>[10]</sup>针对通用领域事件论元抽取中角色信息利用不足和论元间缺少交互的问题，提出了角色信息引导的事件论元抽取模型。该模型根据角色定义构造角色知识，对角色信息和文本进行独立编码，并采用基于注意力机制的方法获取标签知识增强的文本表示。实验结果表明，角色信息的引导有效地提升了论元抽取的性能，使得该模型的表现优于其他基线模型。刘忠宝等<sup>[11]</sup>以《史记》为例，展示了历史事件抽取与事理图谱构建在历史学研究中的应用价值。该研究在BERT和LSTM-CRF模型的基础上，提出了面向《史记》的历史事件及其组成元素抽取方法，并成功构建了《史记》事理图谱。实验结果表明，该方法在抽取历史事件及其组成元素方面准确率较高，通过事理图谱能够揭示出《史记》中蕴含的深层知识和规律。

随着人工智能技术的不断发展，古籍文本自动标注性能比较研究也将为未来的自然语言处理、机器学习等领域提供新的思路和方法。张琪等<sup>[12]</sup>针对古籍知识库缺乏完善知识溯源机制和部分古汉语文本缺乏触发词的问题，提出了一种将结构化历史知识溯源至史籍原文的方法。该研究通过共指消解、文本蕴涵等技术和方法构建了知识溯源框架，并实验对比了不同

预训练模型和输入策略对知识溯源效果的影响，最终构建的 SHK-Tracer 模型在知识溯源任务中取得了较高的精确率。刘忠宝等<sup>[13]</sup>在梳理古籍信息处理研究进展时指出，该领域历经了从基础数字化、深度数据挖掘到智能系统构建的技术演进，已成为跨学科融合的研究热点，未来应致力于构建开放协同的古籍数字资源共享体系，研发高效适配的古籍信息处理专用模型。古籍文本信息抽取相关研究具有重要作用，是实现古文数字化和信息化的关键一环。基于上述分析，笔者基于 UIE 系列模型，对《二十四史》中实体关系抽取和事件论元抽取问题展开深入研究。

## 2 模型引入

本文首先概述了 UIE 系列模型的基本原理与技术特点，强调了其在深度学习与自然语言处理领域的先进性。随后，设计并实施了一系列实验，采用不同版本的 UIE 模型对《二十四史》中的古籍文本进行自动标注，重点比较了各模型在实体关系抽取与事件论元抽取任务上的性能表现。通过实验评估不同 UIE 系列模型在实体关系抽取和事件论元抽取两类任务的准确率、召回率、F1 值等指标上的表现，并进行比较研究。研究表明，UIE 系列模型在古籍文本的自动标注中展现出了强大的适应性和潜力，但不同版本模型在特定任务上的表现存在差异。此外，本研究还讨论了进一步优化模型性能的可能性，如模型融合、参数调优及针对古籍文本特性的定制化改进等，以期在未来推动古籍自动标注技术的持续发展。

### 2.1 UIE模型简介

UIE<sup>[1]</sup>模型是一种统一的信息抽取框架，它的设计灵感来自近年来自然语言处理(NLP)领域取得的诸多突破，尤其是在实体抽取、关系抽取、事件抽取和情感分析等任务中的成功应用。为了进一步发挥 UIE 的强大功能，PaddleNLP 借鉴该论文<sup>[1]</sup>的方法，基于 ERNIE 3.0 (Enhanced Representation through Knowledge Integration)<sup>[14]</sup>知识增强预训练模型，训练并开源了首个中文通用信息抽取模型 UIE。ERNIE 系列模型自 2019 年起由百度开发，是一种基于深度学习的预训练语言模型。ERNIE 的核心目标是通过引入知识图谱来增强预训练模型的能力，从而满足多种自然语言处理任务需求。

### 2.2 UIE模型的基本结构

图 1 展示了 UIE 模型的基本结构，包括输入编码器、特征提取器和输出解码器三个主要部分。

#### 2.2.1 输入编码器(Encoder)

输入编码器将原始文本转化为向量表示，以捕捉文本中的复杂语义信息。技术细节中设计了结构化抽取语言 (Structured Extraction Language, SEL) 来统一编码异构提取结构，即编码实体、关系、事件统一表示。构建结构化模式提示器 (Structural Schema Instructor, SSI)，采用基于 schema 的 prompt 机制，控制不同的生成需求。编码器的整体框架可以概括为 SSI + Text → SEL，即 SSI 作为输入特定抽取任务的 schema，SEL 将不同任务的抽取结果统一用一种语言表示。

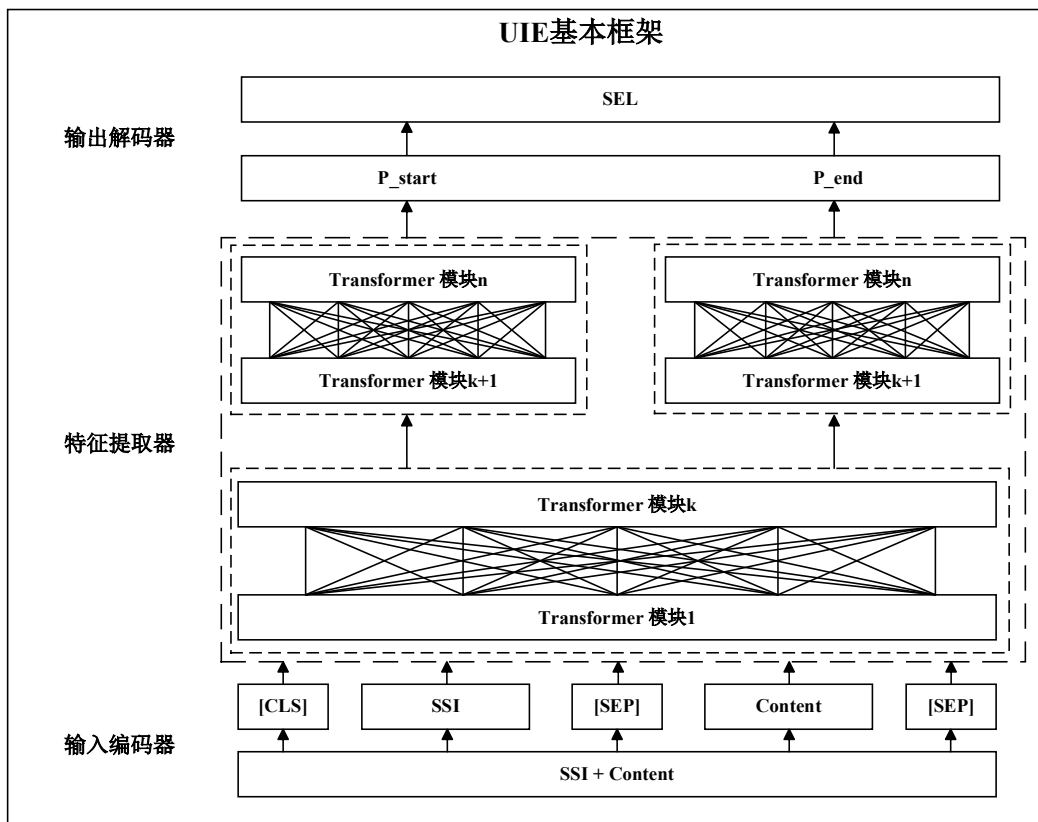


图1 UIE 基本框架

### 2.2.2 特征提取器

UIE 模型通常采用基于 Transformer 架构的预训练模型，如 BERT、ERNIE 等。在特征提取器内部，多个 Transformer 层堆叠起来，逐层提取输入数据的深层次特征。每一层都会根据前一层的输出和当前层的参数进行计算，最终生成包含丰富语义信息的特征表示。特征提取器的设计直接影响模型的性能和鲁棒性。

### 2.2.3 输出解码器 (Decoder)

输出解码器根据结构化抽取语言的规范，将输出的特征表示转换为具体的结构化输出，如实体、关系、事件等。在解码过程中，UIE 模型通常采用自回归生成方式，根据当前已生成的部分和编码器输出的特征表示，预测并生

成下一个部分，直到生成完整的目标结构。

## 2.3 实体关系抽取和事件论元抽取

本文的核心任务是利用 UIE 系列模型强大的预训练知识迁移能力和结构化生成能力进行古籍的实体关系抽取和事件论元抽取。然而，古籍文本（如文言文、特定历史时期的文献）具有显著区别于现代文本的特性，如用词生僻、通假字异体字繁多、句法结构简练特殊、缺乏现代标点、存在大量历史专名与特定表达范式等。这些特性对通用信息抽取模型构成严峻挑战，可能导致实体识别模糊、关系模式误判、事件理解偏差等问题。为有效应对古籍文本的特殊性，提升模型在古籍场景下的鲁棒性和准

确性，在 UIE 基础流程之上，引入并实施以下关键针对性技术与策略。

### （1）古籍文本预处理与增强。

①异体字/通假字规范化：构建针对目标古籍语料（如《二十四史》特定篇章）的异体字、通假字对照映射表。在输入模型前，将文本中的异体字、通假字系统性地替换为标准正体字（例如，将“邊”替换为“边”），显著降低模型理解难度，提高实体识别的统一性。

②句读补充：针对原始古籍文本常无标点或仅有简单句读的情况，集成基于规则和统计模型的古籍专用断句方式，在输入前添加符合文言文语感的现代标点符号（特别是逗号、句号），为模型提供更清晰的语法边界信息，极大促进了长句、复杂句的理解与实体/事件边界判定。

③领域词典融入：集合大规模古籍实体/事件词典（如历史人物、地名、官职名、典籍名等）作为外部知识源。在输入处理阶段，利用词典对文本进行实体候选词标注、词向量增强，引导模型更准确地识别古籍特有的命名实体/事件。

（2）Prompt 数据集工程优化与古籍知识注入。

①古籍语境化 Prompt 数据集设计：摒弃通用 Prompt 数据集模板，针对古籍中的特定关系（如“封爵”“谥号”“师承”“任职”）和事件类型（如“册封”“征伐”“灾异”），设计包含古籍语境线索和关系/事件关键词的 Prompt。例如，关系抽取 Prompt 不是“人物—人物关系”，而是具体化为“历史人物间的官职任免关系”或“家族谱系中的父子兄

弟关系”，并提供典型关系词的示例（如“拜为”“迁”“父”“兄”）。事件抽取 Prompt 则明确事件类型（如“科举事件”）并提示关键论元角色（如“科年”“登第者”“及第等级”）。

②实体类型与关系模式适配：根据古籍内容特点，扩展和定制实体类型体系（如增加“年号”“谥号”“爵位”“典籍”等类型）和关系/事件模式（如“撰著”“赐谥”“避讳”等）。这些定制化后的数据集模式通过 Prompt 清晰地传递给模型，指导其学习古籍特有的信息结构。

### （3）模型微调与领域自适应。

①古籍语料预训练/微调：在通用预训练 UIE 系列模型基础上，使用特定目标领域的古籍语料进行持续预训练和领域自适应微调，使模型能够深度吸收古籍词汇、句法和语义知识，提升对古汉语表达习惯的建模能力。

②困难样本挖掘与增强：针对古籍中频发的实体歧义（同名异人、同人异名）、关系稀疏、事件表述隐晦等问题，在训练过程中侧重挖掘和人工校验困难样本，并利用基于古籍规则的数据增强（如替换历史同义词、模拟句法变体）策略，扩充模型对古籍复杂场景的覆盖能力。

结合上述针对性技术，使用 UIE 系列模型处理古籍信息抽取任务的流程描述如图 2 所示。

实体关系抽取首先由 UIE 模型识别文本中的命名实体（如人名、地名、机构名等），随后进一步分析这些实体之间的语义联系，从而抽取它们之间的特定关系（如出生地、任职机构、亲属关系等）。这一过程依赖于模型对文本上下文的理解能力，以及对大量预训练数据中实体关系模式的学习。通过构建实体与关系之间的映射关系，UIE 模型能够生成结构化

的关系三元组。

事件论元抽取关注于文本中描述的事件及其参与者（即论元）的识别与提取。UIE 模型首先识别出文本中的事件触发词，这些触发词通常指示了某个事件的发生。随后，模型根据事件类型与上下文信息，抽取出与该事件相关

的论元（如事件的主体、客体、时间、地点等）。这一过程要求模型具备对事件结构的深入理解，以及对不同事件类型下论元角色的准确判断。本文专注于事件论元抽取效能，不直接对比触发词的识别效果，更多地关注于论元本身的识别、分类和关系理解等方面。

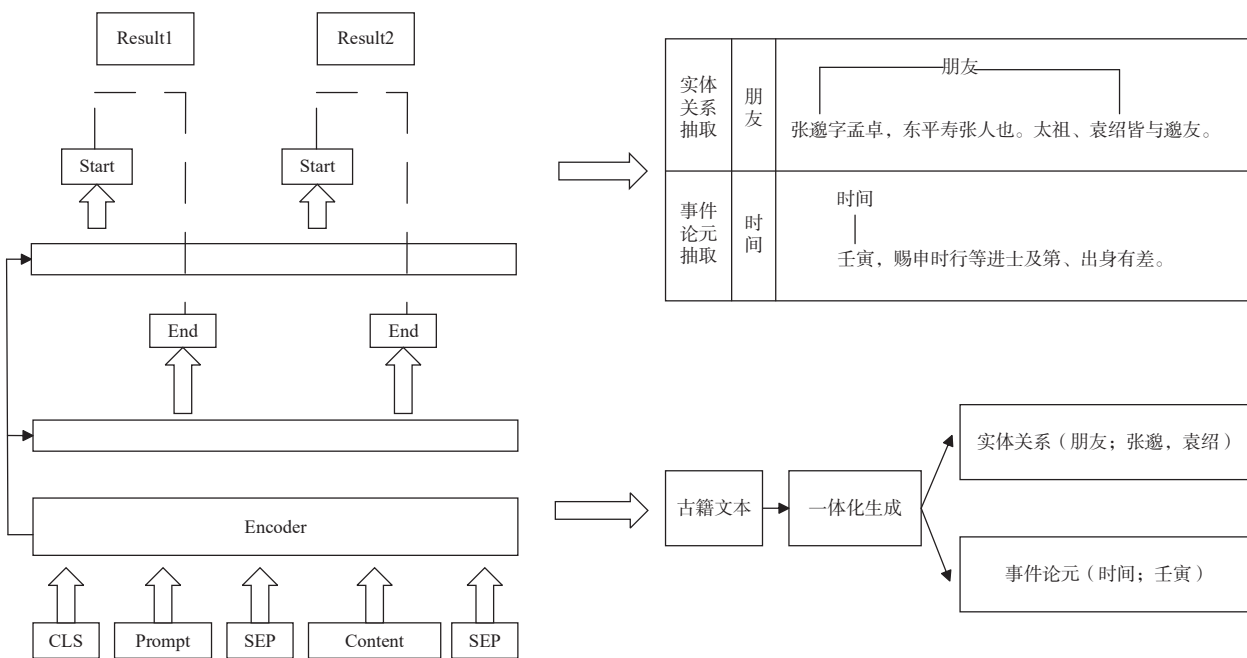


图 2 实体关系抽取和事件论元抽取流程

下面举例说明 UIE 模型进行实体关系抽取和事件抽取的过程。

(1) 实体关系抽取

以“张邈字孟卓，东平寿张人也。太祖、袁绍皆与邈友。”为例，UIE 模型进行实体关系抽取的过程如下。

①输入处理：

**Content:** 输入文本为“张邈字孟卓，东平寿张人也。太祖、袁绍皆与邈友。”。

**Prompt:** 在输入文本前，根据任务需求拼接上特定的提示语（Prompt），以指示模型进

行关系抽取。此例中，Prompt 包含“历史人物关系”等指示信息。

CLS（分类标记）与 SEP（分隔符）：虽然 UIE 模型主要基于 Seq2Seq 架构，不直接涉及 CLS 和 SEP 在 BERT 等模型中的典型用法，但在模型内部处理时，可视为隐式地进行了类似的功能划分，以区分不同部分的输入。

② Encoder 阶段：输入文本（包括 Prompt 和 Content）被送入 Encoder 部分，Encoder 通常采用 Transformer 结构，负责捕捉文本中的上下文信息和语义特征。Encoder 输出每个 token

的隐层表示，这些表示蕴含了丰富的语义和上下文信息。

③ Decoder 阶段：Decoder 利用 Encoder 的输出，结合 SEL 语法（如“Spotting”和“Associating”操作），生成目标结构化信息。在本例中，Decoder 会识别出“张邈”和“袁绍”为实体，并将它们之间的关系标记为“朋友”。

④结果输出：输出结构化的关系信息，如“张邈—朋友—袁绍”。

#### （2）事件论元抽取

以“壬寅，赐申时行等进士及第、出身有差。”为例，UIE 模型进行事件论元抽取的过程如下。

①输入处理与 Encoder 阶段同实体关系抽取。

② Decoder 阶段（针对事件论元）：Decoder 根据 SEL 语法，定位到事件的时间论元“壬寅”以及其他事件论元。

③结果输出：输出结构化的事件论元信息，如“时间：壬寅；受封者：申时行等；受封内容：进士及第”。

UIE 模型通过统一的 Text-to-Structure 生成架构和 SEL 语法，实现了对多种信息抽取任务的统一建模和高效处理。在实体关系抽取和事件论元抽取任务中，UIE 模型能够准确识别并结构化输出关键信息，极大地提高了信息抽取的效率和准确性。同时，通过引入 Prompt 机制，模型能够灵活地适应不同的抽取需求，展现出强大的泛化能力。

通过系统性地引入古籍文本预处理与增强、Prompt 数据集工程优化与古籍知识注入、模型领域自适应微调等针对性技术，本文显著提升了 UIE 模型在复杂古籍文本上的信息抽取性能。

这些技术有效克服了古籍语言特性带来的主要障碍，是本工作相对于简单应用通用 UIE 模型的核心创新点。统一的结构化生成架构（SEL）结合这些古籍适配技术，为古籍深度信息自动化处理提供了高效且鲁棒的解决方案。

## 3 实验设计和实验结果分析

### 3.1 数据集来源与预处理

本文选取了多个代表性古籍文本数据集作为实验对象，覆盖不同历史时期与内容领域。实验的古籍文本自动标注包括两类任务：一类是实体关系抽取，另一类是事件论元抽取。数据来源于 GitHub 上的“<https://github.com/jizijing/C-CLUE>”与“<https://github.com/Lyn4ever29/GuwenEE>”，分别针对实体关系抽取与事件论元抽取任务。这些语料库基于《二十四史》构建，通过大规模语言模型与人工标注相结合的方式形成标注语料。

对所有数据集进行统一预处理，包括文本清洗、去除停用词等步骤。例如，图 3 中的原始数据去除“object\_type”“subject\_type”后得到图 4 中的数据示例；图 5 中的原始数据“壬寅，赐申时行等进士及第、出身有差。”处理后得到 3 个事件论元“时间”“受封者”“受封内容”，结果如图 6 所示。以上内容经过整理和统一格式后构成数据子集，其中获得实体关系数据共 4405 条，事件论元数据共 4654 条。之后将数据集划分，以 500 条为单位，按照约 7:2:1 的比例划分训练集、验证集和测试集，不同大小的训练集、验证集和测试集的数量如表 1 所示。

表 1 数据集比例划分

训练集	验证集	测试集
500	150	100
1000	300	200
1500	450	300
2000	600	400
2500	750	500
3000	900	600
3500	1150	700

```
{ "text": "十七年而崩。三年丧毕，禹亦乃让舜子，如舜让尧子。诸侯归之，然后禹践天子位。尧子丹朱，舜子商均，皆有疆土，以奉先祀", "spo_list": { "predicate": "子", "object_type": "PER", "subject_type": "PER", "object": "商", "subject": "舜" }
```

图 3 实体关系原始数据示例

```
{ "content": "十七年而崩。三年丧毕，禹亦乃让舜子，如舜让尧子。诸侯归之，然后禹践天子位。尧子丹朱，舜子商均，皆有疆土，以奉先祀", "result_list": { "text": "商", "start": 44, "end": 44, { "text": "舜", "start": 15, "end": 15 }, "relation": "子" }
```

图 4 实体关系转换后的数据示例

```
{ "text": "壬寅，赐申时行等进士及第、出身有差。", "id": "3c19138cde7d4ac19d70aac88b2acfaa", "event_list": [ { "event_type": "政治/分封事件", "trigger": "赐", "trigger_start_index": 3, "arguments": [ { "argument_start_index": 0, "role": "时间", "argument": "壬寅", "alias": [] }, { "argument_start_index": 4, "role": "受封者", "argument": "申时行等", "alias": [] }, { "argument_start_index": 8, "role": "受封内容", "argument": "进士及第", "alias": [] }, "class": "政治" ] }
```

图 5 事件论元原始数据示例

```
{ "content": "壬寅，赐申时行等进士及第、出身有差。", "result_list": [ { "text": "壬寅", "start": 0, "end": 1 }, "arguments": "时间" }
{ "content": "壬寅，赐申时行等进士及第、出身有差。", "result_list": [ { "text": "申时行等", "start": 4, "end": 7 }, "arguments": "受封者" }
{ "content": "壬寅，赐申时行等进士及第、出身有差。", "result_list": [ { "text": "进士及第", "start": 8, "end": 11 }, "arguments": "受封内容" }
```

图 6 事件论元转换后的数据示例

### 3.2 实验设置

基于 PyTorch 深度学习框架进行多模型选择，满足精度、速度要求，表 2 列举了 UIE 的全系列模型参数及其关键属性，包括结构、语言和预训练 / 微调的主要领域。考虑到模型

大小、计算资源限制以及领域相关性，从表 2 中选择了 uie-nano、uie-micro、uie-mini、uie-medium 这四个通用中文领域模型分别训练微调，进行古籍文本自动标注性能比较实验，并在相应的测试集上进行性能评估。它

们分别代表了 UIE 系列模型中不同性能和资源占用的选项，在保持较高性能的同时，逐渐改变模型体积和资源消耗，从而满足了不同场景下的需求。具体来说，uie-nano 是 UIE 系列中体积最小、资源占用最低的模型，它适用于对资源要求极高或需要在受限环境下运行的场景。uie-micro 和 uie-mini 在体积和资源占用上略大于 uie-nano，但性能有所提升，适用于对性能有一定要求，同时保持较低资源占用的场景。uie-medium 则适用于常

规场景，在保证性能的同时，降低资源消耗。通过比较这四个通用领域模型，可以清晰地看到 UIE 系列模型在性能和资源占用上的多样化选择，有助于更全面地了解 UIE 系列模型的特点和适用场景。在进行实验时保持实验的一致性和可比性，包括使用相同的模型、相同的超参数设置、相同的评估指标等。另外，对实验结果进行分析和解释时，考虑到训练集大小对模型性能的影响，并注意可能存在的其他因素。实验参数设置如表 3 所示。

表 2 UIE 系列模型参数及领域说明

模型	结构及领域说明	语言
uie-base	12-layers, 768-hidden, 12-heads 通用领域（在大规模通用中文文本上预训练）	中文
uie-base-en	12-layers, 768-hidden, 12-heads 通用领域（在大规模通用英文文本上预训练）	英文
uie-medical-base	12-layers, 768-hidden, 12-heads 生物医学领域（在生物医学中文文本上微调）	中文
uie-medium	6-layers, 768-hidden, 12-heads 通用领域（较小规模通用中文模型）	中文
uie-mini	6-layers, 384-hidden, 12-heads 通用领域（轻量级通用中文模型）	中文
uie-micro	4-layers, 384-hidden, 12-heads 通用领域（轻量级通用中文模型）	中文
uie-nano	4-layers, 312-hidden, 12-heads 通用领域（超轻量级通用中文模型）	中文
uie-m-large	24-layers, 1024-hidden, 16-heads 多语言通用领域（大规模多语言通用模型）	中、英文
uie-m-base	12-layers, 768-hidden, 12-heads 多语言通用领域（多语言通用模型）	中、英文

表 3 实验参数设置

learning_rate	batch_size	max_seq_len	num_epochs	logging_steps	valid_steps	device
1e-5	16	512	100	10	100	gpu

### 3.3 评估指标

在模型评估部分采用精确率 (Precision)、召回率 (Recall) 以及 F1 分数 (F1 Score) 作为核心评价指标, 以全面客观地衡量各个 UIE 系列模型的古籍文本自动标注性能。精确率反映了模型预测为正类的样本中, 真正为正类的比例, 是衡量模型预测准确性的重要指标。具体而言, 高精确率表明模型在识别正样本时犯错较少, 即预测为正类的结果中大部分是真正的正类。召回率则关注于模型能够成功找出所有正类样本的能力, 即实际为正类的样本中被模型正确预测为正类的比例。高召回率意味着模型能够更全面地捕捉到所有相关的正类样本, 减少漏检。F1 分数是精确率与召回率的调和平均数, 为两者提供了一个平衡的视角。F1 分数越高, 说明模型在精确率和召回率之间取得了更好的平衡, 整体性能更为优异。因此, 通过计算并比较不同模型在精确率、召回率及 F1 分数上的表现, 能够系统地评估各模型的优劣, 为后续的模型优化与选择提供有力依据。具体如式 (1) 至式 (3) 所示, 其中 TP 表示模型预测正确的正例, FP 表示模型误报的正例, FN 表示模型漏报的正例。

$$P = \frac{TP}{TP + FP} \times 100\% \quad (1)$$

$$R = \frac{TP}{TP + FN} \times 100\% \quad (2)$$

$$F1 = \frac{2 \times P \times R}{P + R} \times 100\% \quad (3)$$

### 3.4 实验结果及分析

#### 3.4.1 基线模型

在相同语料库上选择 Guwenbert-base、

roberta-classical-chinese-base、SikuBERT、SikuRoberta 四个预训练模型作为基线进行性能对比:

(1) GuwenBERT-base<sup>[15]</sup>: 由学界发布的首批古汉语 BERT 模型之一, 在《四库全书》等通用古籍语料上训练, 专注于文言文词汇和句法表征, 作为通用古籍基线覆盖广泛历史时期, 但未针对特定典籍优化。

(2) roberta-classical-chinese-base<sup>[16]</sup>: 基于 RoBERTa 架构优化的古汉语模型, 采用动态掩码和更大批次训练, 覆盖先秦至明清经典文献, 强调经典文献适应性, 适合高规范性文本 (如经史子集)。

(3) SikuBERT<sup>[17]</sup>: 专为《四库全书》设计的领域模型, 融合繁体字、异体字特征, 增强古籍实体识别能力, 作为领域专用模型解决《四库全书》中繁体/异体字和古术语的嵌入表示问题。

(4) SikuRoberta<sup>[17]</sup>: SikuBERT 的 RoBERTa 升级版, 通过全文掩码策略提升上下文建模能力, 适用于长篇章古籍分析, 改进了长距离依赖建模, 对篇章级古籍任务 (如事件链抽取) 有显著优势。

基线模型实验结果如表 4 所示。

为明确 UIE 在古籍标注任务中的定位, 将其与当前主流大模型进行特性对比, 如表 5 所示。

①在任务适配性方面, UIE 作为信息抽取专用框架, 其编码模板直接对齐实体/关系抽取目标, 因此在结构化标注任务上显著优于通用生成式模型 (如 ChatGPT、Claude) 和纯理解模型 (如 ERNIE 3.0、LLaMA)。②在资源

表 4 基线模型实验结果

		Guwen-BERT	roberta-clas-sical-chinese-base	Siku-BERT	Siku-RoBERTa
事件关系抽取	Precision	27.10	58.72	63.09	59.30
	Recall	27.10	28.34	48.30	45.17
	F1	38.77	38.23	54.71	51.28
事件论元抽取	Precision	21.42	20.05	45.56	40.40
	Recall	11.96	8.15	43.26	39.61
	F1	14.50	11.57	44.38	40.00

表 5 UIE 与主流大模型在古籍标注任务中的特性对比

模型类型	代表模型	古籍任务适配性及领域迁移
UIE 系列	uie-micro/uie-medium	支持结构化抽取，微调即可适配
古籍专用基线	SikuRoberta/GuwenBERT	针对古籍进行预训练
中文通用大模型	ERNIE 3.0、ChatGLM3	Prompt 工程适配，句法表征不足
生成式大模型	GPT-3.5、Claude、XuanYuan	依赖指令，需大量古籍微调数据

效率方面，相比百亿参数级大模型（如 GPT-3.5、PanGu- $\Sigma$ ），UIE（如 uie-nano/micro）为轻量级模型，在古籍场景可实现倍速推理，满足低资源部署需求。③在领域迁移方面，尽管大模型（如 ChatGLM3、XuanYuan）在泛中文任务表现优异，但其在古籍术语、通假字、特殊句法上的表征能力弱于古籍专用模型（如 Siku-BERT）和轻量微调后的 UIE，突显出领域适配的必要性。

综上所述，本文未直接采用大模型方法进行古籍文本的实体关系抽取与事件论元抽取，主要基于以下三个方面：①任务结构性错配：古籍实体关系抽取与事件论元抽取需严格遵循预定义 Schema，而生成式大模型的自由文本输出格式存在不可控、标准化难的问题。UIE 的 Prompt 模板机制天然适配结构化抽取需求。②领域知识缺陷：主流大模型在文言虚词（如“之乎者也”语义功能）、通假字（如“蚤→早”）、典故等古籍特性上表征薄弱，微调所需的高质量对齐数据（古籍原文—标注）稀缺，而 UIE

通过小样本微调即可激活模型古籍语义理解能力（本研究实验显示 uie-nano 模型增加 500 条样本可使 F1 提升 20.9%）。③落地可行性瓶颈：百亿级大模型推理需  $\geq 80\text{GB}$  显存，而古籍研究机构普遍缺乏高性能计算资源。UIE-micro 仅需 4GB 显存，在嵌入式设备（如 Jetson Nano）上仍可实现在实时抽取，符合古籍数字化轻量化部署的刚需。

### 3.4.2 实体关系抽取

本文设计了样本大小递增策略，以 500 条数据为一个增量单位，从 500 条样本逐步扩展至 3000 条样本，以全面评估不同规模数据集对实验结果的影响。针对实体关系抽取任务，选取四个 UIE 系列模型，即 uie-nano、uie-micro、uie-mini 及 uie-medium，这些模型在复杂度与性能之间存在差异。实验基于预先处理完毕的数据集进行，旨在探究各模型在不同样本量条件下的实体关系抽取能力。表 6 详细记录了这一系列实验的结果，为深入理解模型性能随样本规模变化的趋势提供了数据支持。

表 6 实体关系抽取实验结果

Uie系列模型 样本大小	Uie-nano	Uie-micro	Uie-mini	Uie-medium
500	Precision=0.40426	Precision=0.46721	Precision=0.48062	Precision=0.60000
	Recall=0.36076	Recall=0.36076	Recall=0.39241	Recall=0.47468
	F1=0.38127	F1=0.40714	F1=0.43206	F1=0.53004
1000	Precision=0.59868	Precision=0.62712	Precision=0.63245	Precision=0.65552
	Recall=0.58333	Recall=0.59295	Recall=0.61218	Recall=0.62821
	F1=0.59091	F1=0.60956	F1=0.62215	F1=0.64157
1500	Precision=0.63474	Precision=0.65899	Precision=0.64798	Precision=0.69676
	Recall=0.59623	Recall=0.59833	Recall=0.60460	Recall=0.62971
	F1=0.61489	F1=0.62719	F1=0.62554	F1=0.66154
2000	Precision=0.72143	Precision=0.72469	Precision=0.72193	Precision=0.75132
	Recall=0.64434	Recall=0.65072	Recall=0.66667	Recall=0.67943
	F1=0.68071	F1=0.68571	F1=0.69320	F1=0.71357
2500	Precision=0.72843	Precision=0.73907	Precision=0.75231	Precision=0.74966
	Recall=0.64536	Recall=0.65664	Recall=0.68728	Recall=0.69424
	F1=0.68439	F1=0.69542	F1=0.71833	F1=0.72088
3000	Precision=0.76310	Precision=0.78335	Precision=0.79412	Precision=0.79625
	Recall=0.62867	Recall=0.61490	Recall=0.62614	Recall=0.65636
	F1=0.68939	F1=0.68898	F1=0.70020	F1=0.71957

由图 7、图 8、图 9 直观展示实验结果。

由图 7 所示的实验结果可以看出，针对最优表现的 uie-medium 模型进行深入分析，当训练集样本量从 500 条扩充至 3000 条时，Precision 值显著提升，由初始的 60.01% 跃升至

79.63%，增幅高达 19.62%，明确揭示了训练集样本数量的增加与 Precision 值提升之间的显著正相关关系。该发现强调了在实体关系抽取任务中，增加训练数据对于提升模型预测准确性的重要性。实验中所使用的四个 uie 系列模型

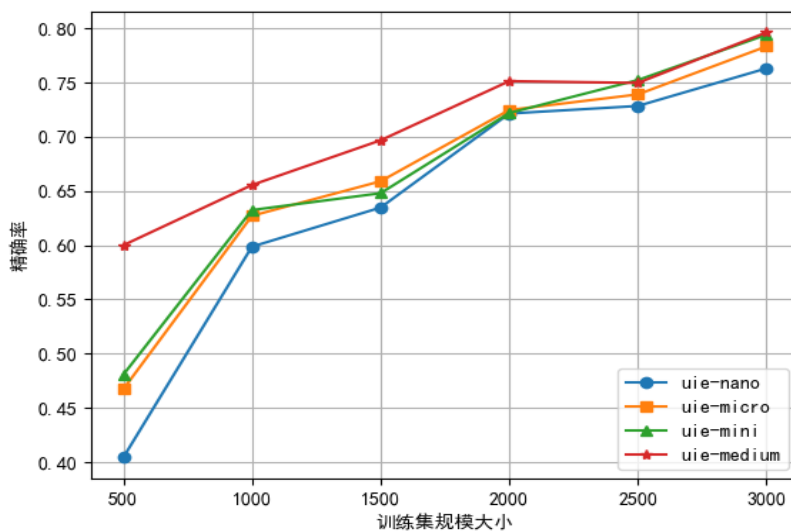


图 7 实体关系抽取 Precision 对比

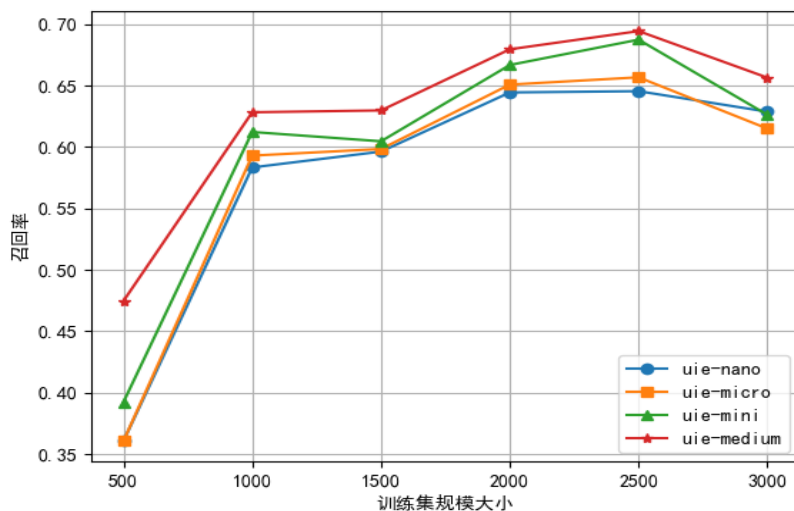


图 8 实体关系抽取 Recall 对比

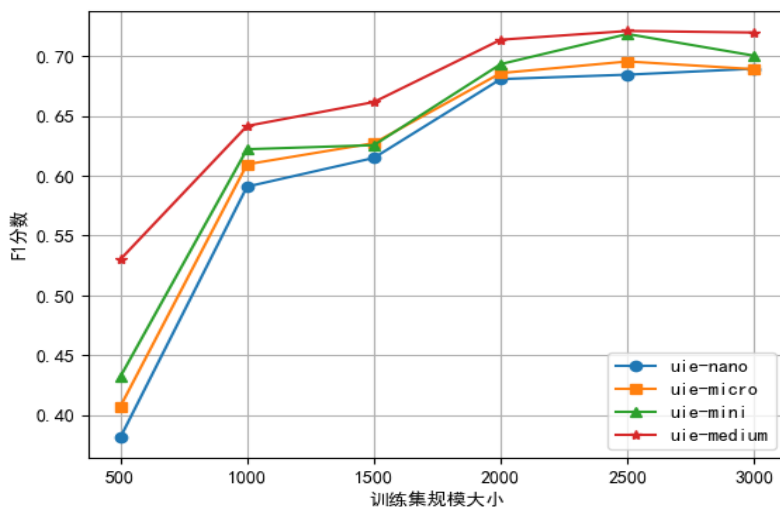


图 9 实体关系抽取 F1 值对比

均展现出随着训练集样本量增长而不断提升的精确率，这有力验证了这些模型在更大规模数据集上提升预测准确性的能力。因此可以推断，通过扩大训练集规模，可以有效提升模型对正类样本识别的精度，进而增强整体实体关系抽取的性能表现。

根据图 8 所呈现的实验结果，观察到在训练集规模从初始增加至 2500 条样本的过程中，四个 uie 系列模型的 Recall 值均展现出随训练集样本量增长而上升的趋势。然而，当训练集

样本量进一步扩展至 3000 时，这四个模型的 Recall 值却出现了不同程度的下降，此现象亦间接导致了 F1 分数的降低。上述现象表明，在训练集样本量增加至某一临界点（实验中为 2500 条样本）后，Recall 值达到其峰值，其中 uie-medium 模型在训练集样本量为 2500 条时取得了最高的 Recall 值，即 69.42%。这一发现揭示了 Recall 值并非单调递增至训练集规模，而是存在一个最优区间，超出此区间后，模型全面捕捉正类样本的能力反而有所下降。因此，

可以推断 Recall 值的大小并非完全依赖于训练集规模的无限扩大,而是受到模型学习能力、数据分布特性及模型泛化能力等多重因素的共同影响。

根据图 9 所展示的实验数据分析,随着训练集样本数量的递增,四种模型在实体关系抽取任务中的 F1 值均呈现出上升趋势。进一步细致考察图表趋势,发现当训练集样本量分别为 2000 条、2500 条及 3000 条时,各模型的 F1 值虽有所波动但总体保持相近水平。在样本量为 2500 条时, uie-medium 模型取得了本次实验中的最高 F1 值,达到了 72.08%,标志着其在此条件下的综合性能最优。值得注意的是,在样本量从 2000 条增加至 2500 条的过程中, uie-mini 模型的 F1 值相较于其他模型有显著的提升,增幅达到 2.51%,表明该模型在此范围内

对训练数据量的增加更为敏感。然而,不论训练集样本量如何变化, uie-medium 模型始终保持着高于其他模型的 F1 值,这充分验证了其在实体关系抽取任务中的稳定性和优越性。

综上所述,实验结果不仅揭示了训练集样本量对模型性能的正向影响,还强调了不同模型在不同样本规模下的表现差异及最优选择,为实际应用中的模型选择与训练策略提供了重要参考。

### 3.4.3 事件论元抽取

在事件论元抽取任务中,样本大小递增策略和 UIE 系列模型选取保持不变,样本最大值为 3500。本实验依托于预处理后的数据集进行,旨在探究各 UIE 系列模型在不同样本量条件下的事件论元抽取能力,实验结果如表 7 所示。

表 7 事件论元抽取实验结果

Uie 系列模型 样本大小	Uie-nano	Uie-micro	Uie-mini	Uie-medium
500	Precision=0.47368	Precision=0.50000	Precision=0.52381	Precision=0.54545
	Recall=0.29670	Recall=0.36145	Recall=0.36264	Recall=0.39560
	F1=0.36486	F1=0.41958	F1=0.42857	F1=0.45860
1000	Precision=0.52109	Precision=0.54255	Precision=0.56338	Precision=0.59211
	Recall=0.37963	Recall=0.37500	Recall=0.43478	Recall=0.48913
	F1=0.43824	F1=0.44348	F1=0.49080	F1=0.53571
1500	Precision=0.58824	Precision=0.60684	Precision=0.61304	Precision=0.63900
	Recall=0.51852	Recall=0.52593	Recall=0.52222	Recall=0.57037
	F1=0.55118	F1=0.56349	F1=0.56400	F1=0.60274
2000	Precision=0.61656	Precision=0.63608	Precision=0.65421	Precision=0.69309
	Recall=0.55833	Recall=0.56303	Recall=0.58824	Recall=0.60088
	F1=0.58601	F1=0.59733	F1=0.61947	F1=0.64370
2500	Precision=0.71012	Precision=0.71574	Precision=0.72886	Precision=0.75448
	Recall=0.60526	Recall=0.63371	Recall=0.65843	Recall=0.66292
	F1=0.65351	F1=0.67223	F1=0.69185	F1=0.70574
3000	Precision=0.68272	Precision=0.68944	Precision=0.72473	Precision=0.74675
	Recall=0.60088	Recall=0.61439	Recall=0.62177	Recall=0.63653
	F1=0.63919	F1=0.64976	F1=0.66931	F1=0.68725
3500	Precision=0.73039	Precision=0.73031	Precision=0.74554	Precision=0.76722
	Recall=0.58203	Recall=0.59766	Recall=0.61230	Recall=0.63086
	F1=0.64783	F1=0.65736	F1=0.67239	F1=0.69239

从图 10 的实验结果可以看出,在事件论元抽取任务中, Precision 值总体上随着训练集样本数量的增加而逐步提升。然而,当训练集样本量超过 2500 条后,四个 uie 系列模型的 Precision 值均展现出小幅度的波动现象。这一趋势与实体关系抽取任务中的 Precision 值变化趋势存在差异,其背后可能的原因在于事件论元类别的显著增加以及类别间复杂性的加剧导致 Precision 值在达到一定训练规模后出现不稳定波动。尽管如此,从整体趋势来看,事件论元抽取与实体关系抽取在 Precision 值的变化上仍保持着一定的相似性,即随着训练集样本量的不断扩充, Precision 值均呈现出持续上升的趋势,这进一步验证了训练数据规模对于提升模型预测精确度的积极作用。

基于图 11 所展示的实验结果,观察到在事件论元抽取任务中,当训练集样本量达到 2500 条时, Recall 指标呈现出峰值状态, uie-medium 模型在此条件下实现了 66.29% 的最大 Recall 值。训练集样本量超过 2500 条后,所有 uie 系列模型的 Recall 值均呈现出下降趋势,这一现象与先前在实体关系抽取任务中观察到的 Recall 值变化趋势一致。因此,可以合理推断在训练集样本量约为 2500 条时,实体关系抽取任务和事件论元抽取任务的模型均能在识别所有正例方面达到最佳性能。然而,一旦训练集样本量进一步增加, Recall 值却出现逆转,表明模型在识别正例的全面性上有所减弱,这暗示了模型在超过这一特定样本量后,其泛化能力或是对新样本的捕捉能力可能受到了限制。这一发现为优化模型训练策略及样本选择提供了重要的依据。

根据图 12 所展示的实验数据分析,发现在事件论元抽取任务中,当训练集样本量达到 2500 条时, F1 分数达到其最优值,具体表现为 uie-medium 模型取得了 70.57% 的 F1 分数。训练集样本量超出 2500 条后, F1 分数并未显著下降,而是趋于一个相对稳定的区间,对于四个 uie 系列模型而言,大致维持在 65%~70% 之间。这一稳定区间与在实体关系抽取实验中观察到的 F1 分数稳定区间一致,表明在训练集样本量超过一定阈值后,模型在综合评估精度与召回率方面的性能表现趋于平稳。此发现不仅加深了对模型性能随训练集规模变化规律的理解,也为后续模型调优及训练数据规划提供了参考。

#### 3.4.4 实验小结

从图 9 和图 12 的实验结果中可以看出,在实体关系抽取和事件论元抽取中,样本数量对 F1 性能的影响是显著的。与基线模型的横向对比结果表明, UIE 系列模型具有明显的优越性。UIE 系列模型之间的纵向对比结果表明,无论是通过精确度 (P 值)、召回率 (R 值) 还是 F1 分数来衡量, uie-medium 模型相较于其他三个 uie 系列模型均展现出了显著的优势。这一优越性能的核心原因可归因于 uie-medium 模型采用了更深层次的隐藏层 (hidden layers) 结构。具体而言,更多的隐藏层数量不仅增强了模型对输入数据的非线性变换能力,还促进了复杂特征的有效提取与组合,从而提升了模型在处理特定任务时的泛化能力和准确性。与当前主要大语言模型相比, UIE 系列模型在低资源场景下表现突出,仅需少量标注数据即可达到较高 F1 值。以下基于整体性能对比、精确率与召

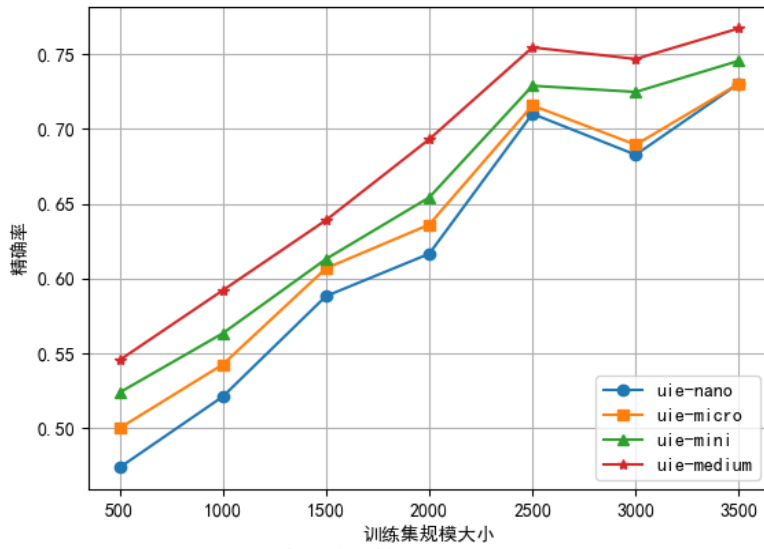


图 10 事件论元抽取 Precision 对比

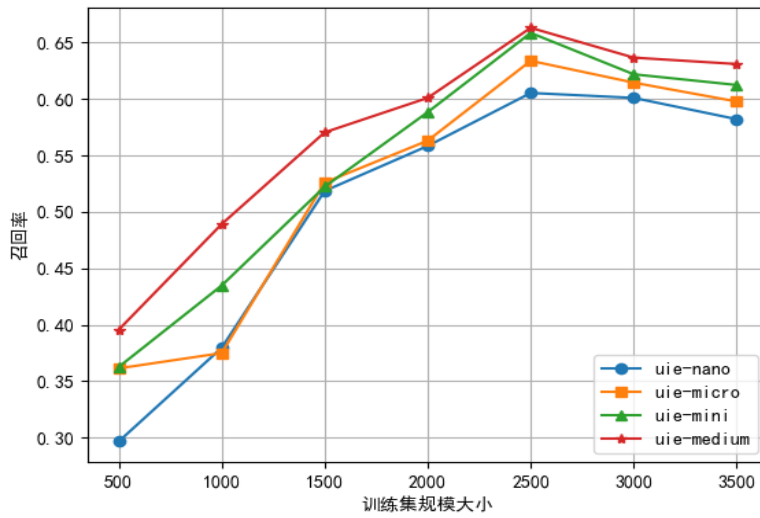


图 11 事件论元抽取 Recall 对比

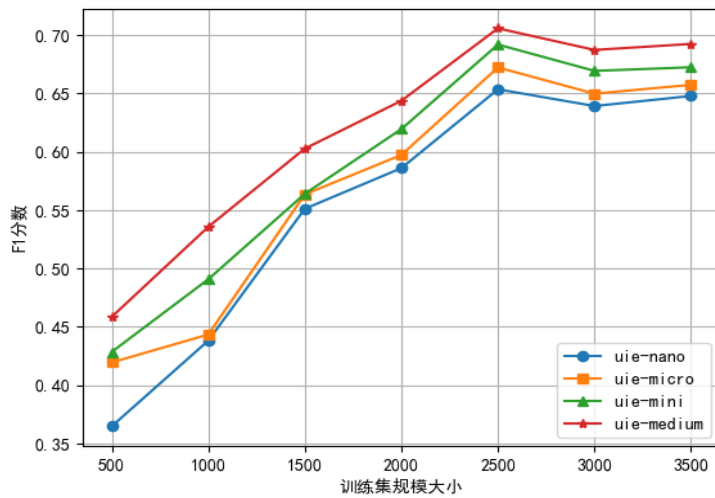


图 12 事件论元抽取 F1 值对比

回率分析、处理速度评估、领域适应性四个方面进行分析：

①整体性能对比：从实验结果来看，随着模型规模的增大，UIE 系列模型在古籍文本自动标注任务上的 F1 分数显著提升。具体而言，uie-medium 模型相较于 uie-nano 模型，在多个测试集上的 F1 分数平均提升了约 5%，表明更大规模的模型在复杂文本理解和标注任务中具有更强的能力。②精确率与召回率分析：在精确率方面，uie-medium 模型表现出更高的稳定性，尤其是在处理实体关系和事件论元抽取时，其精确率显著高于其他模型，主要原因是 uie-medium 模型的 hidden 层均大于其余模型，从而增强了其复杂特征提取和表示学习的能力。然而，uie-mini 模型在部分简单任务上也能达到较高的精确率，主要原因是其网络架构相较于 uie-nano 和 uie-micro 模型的显著扩展，具体体现在模型层数 (layers) 从 4 层增加至 6 层；但在复杂场景下有所下降，主要原因在于其相较于 uie-medium 模型隐藏层减少，限制了模型在复杂任务中深层特征的学习与抽象能力，从而影响了整体性能。在召回率方面，uie-medium 模型同样表现出优势，能够更全面地识别并标注出测试集中的实体关系和事件论元。然而，随着模型规模的增大，对计算资源的需求也相应增加，这在一定程度上限制了其在资源受限场景下的应用。③处理速度评估：在处理速度方面，uie-nano 模型凭借其较小的模型规模和较简单的结构，展现出更快的处理速度。而 uie-medium 模型虽然性能优越，但处理速度相对较慢。因此，在实际应用中需要根据具体需求选择合适的模型平衡点。④领域适应性探讨：

实验结果还揭示了 UIE 系列模型在古籍文本领域的一些适应性挑战。例如，在识别古语词汇和特定语境下的语义关系时，所有模型均表现出不同程度的不足。大语言模型在语义理解深度方面也优于 UIE 系列模型，这说明未来研究应更加注重模型的领域适应性训练，通过引入更多具有领域特色的训练数据来提升模型性能。

## 4 结语

本文探索了 UIE 系列模型在古籍文本自动标注中的实体关系抽取和事件论元抽取两类任务的有效性，通过一系列实验对比分析了不同 UIE 模型 (uie-medium、uie-mini、uie-micro、uie-nano) 在古籍文本自动标注任务中的性能表现。实验表明，当训练样本规模为 2500 条上下时，F1 值达到最优，实体关系抽取和事件论元抽取的 F1 值分别为 72.08% 和 70.57%，说明 UIE 模型能够在古籍文本的自动标注中展示出强大的适应能力。通过数据处理、模型训练与评估流程得出以下主要结论：①模型规模与性能关系：随着模型规模的增大 (从 uie-nano 到 uie-medium)，模型在古籍文本标注任务上的整体性能显著提升。这表明，在资源允许的情况下，采用更大规模的模型能够带来更好的标注效果。②领域适应性：未来研究应更加注重模型的领域适应性训练，通过引入更多的古籍文本数据进行模型微调，以提升模型在特定领域的性能。③标注质量评估：在实际应用中，需要根据具体需求选择合适的模型或采用模型融合策略，以达到最佳的标注效果。综上所述，本文不仅为古籍文本自动标注任务提供了有价

值的参考与启示,也为未来相关领域的深入探索奠定了坚实基础。

## 参考文献

- [1] LU Y, LIU Q, DAI D, et al. Unified structure generation for universal information extraction[C]// Proceedings of the Association for Computational Linguistics, 2022: 5755-5772.
- [2] 皮俊波,齐世雄,孙文多,等.基于UIE框架的电网故障处置预案实体和事件识别方法[J].中国电力,2023,56(12):138-146.
- [3] 朱杰,刘苏文,李军辉,等.基于UIE的情感可解释分析[J].中文信息学报,2023,37(11):152-157.
- [4] 李昌鏢,李蓓,段永恒.基于UIE的深圳市公共卫生事件应急管理知识图谱构建与应用[J].中国数字医学,2024,19(2):48-55.
- [5] 任安琪,柳林,王海龙,等.面向文本实体关系抽取研究综述[J].计算机科学与探索,2024,18(11):2848-2871.
- [6] 吴梦成,王东波,黄水清.古农书翻译与知识组织研究[J].中国农史,2024,43(2):52-64.
- [7] 崔斌,王东波,黄水清.基于“食货志”典籍文本的农作物空间分布特征研究[J].图书情报工作,2024,68(8):133-145.
- [8] 薛继伟,胡馨元,薛鹏杰.基于提示学习的篇章级事件论元抽取方法研究[J].计算机技术与发展,2024,34(6):126-130.
- [9] 王潞翔,陈艳平,黄辉,等.结合二维增强融合机制的事件论元抽取方法[J].计算机工程与应用,2025,61(10):111-119.
- [10] 于媛芳,张勇,左皓阳,等.基于角色信息引导的多轮事件论元抽取[J].北京大学学报(自然科学版),2023,59(1):83-91.
- [11] 刘忠宝,党建飞,张志剑.《史记》历史事件自动抽取与事理图谱构建研究[J].图书情报工作,2020,64(11):116-124.
- [12] 张琪,孔嘉,胡昊天,等.重建知识源流:将结构化知识自动溯源至史籍原文[J].情报学报,2024,43(4):406-415.
- [13] 刘忠宝,赵文娟.古籍信息处理回顾与展望[J].大学图书馆学报,2021,39(6):38-47.
- [14] SUN Y, WANG S, FENG S, et al. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation[EB/OL]. (2021-07-15) [2023-05-16]. <https://doi.org/10.48550/arXiv.2107.02137>.
- [15] ETHANYT. GuwenBERT[EB/OL]. (2020-01-01) [2024-02-21]. <https://github.com/ethan-yt/guwenbert>.
- [16] 喻雪寒,何琳,徐健.基于RoBERTa-CRF的古文历史事件抽取方法研究[J].数据分析与知识发现,2021,5(7):26-35.
- [17] 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.

(责任编辑:徐红姣)