基于双层语义映射的大语言模型辅助古汉语事件抽取半自动标注框架

卫聪聪, 李炜, 冯振冰, 邵艳秋*
北京语言大学，信息科学学院，
国家语言资源监测与研究平面媒体中心，
北京市海淀区学院路15号，100083
{weicongcong0214, zbfengblcu, yqshao163}@163.com, liweitj47@blcu.edu.cn

摘要

尽管自然语言处理技术（NLP）在现代语言事件抽取任务（EE）上已有较为成熟的解决方案，但针对古汉语事件抽取的研究却受限于标注数据匮乏和文本语义复杂等挑战。因而我们提出使用当前取得巨大成功的大语言模型（LLMs）来辅助人类标注员进行数据标注。为了应对LLMs在古汉语上存在的训练不足、语义理解能力欠缺的问题，我们提出了一种基于双层语义映射的LLMs辅助古汉语事件抽取半自动标注框架，利用古汉语的现代汉语译文，结合事件语义学理论及语义依存分析技术，为LLMs提供丰富的语义信息表示，从而进一步将语义依存关系逐步映射为具体的事件信息。经过人类标注员的审核反馈，有效克服了现有NLP工具和LLMs在古汉语事件抽取标注时的局限。实验结果表明，我们的方法不仅提高了古汉语事件抽取标注的准确性和效率，而且减少了对专业人员的依赖和人工标注工作量，为低资源语言标注实践提供了新的方法论，探索了大模型时代数据标注的新方向。

关键词： 事件抽取；大语言模型；语义依存分析；事件语义学

A Semi-automatic Annotation Framework for Event Extraction in Classical Chinese Assisted by Large Language Models Based on Dual-Layer Semantic Mapping

Congcong Wei, Wei Li, Zhenbing Feng, Yanqiu Shao*
Information Science School, Beijing Language and Culture University,
Language Resources Monitoring and Research Center,
15 Xueyuan Road, HaiDian District, Beijing, 100083
{weicongcong0214, zbfengblcu, yqshao163}@163.com, liweitj47@blcu.edu.cn

Abstract

While Natural Language Processing(NLP) techniques have advanced for modern languages, Classical Chinese Event Extraction (EE) research is limited by data scarcity and semantic complexity. Thus, we propose using successful Large Language Models (LLMs) to aid human annotators. To address LLMs' training shortcomings and semantic comprehension in Classical Chinese, we introduce a semi-automatic annotation framework assisted by LLMs Based on Dual-Layer Semantic Mapping. This framework uses Modern Chinese translations of Classical Chinese texts, combined with event semantics theory and semantic dependency analysis, to enhance the semantic representations for LLMs. It methodically converts these semantic dependencies into detailed event information. The process is iteratively refined by human annotators, thus effectively addressing the limitations of existing NLP tools and LLMs in annotating Classical Chinese events. Experimental evidence shows that our method improves

* 通讯作者 Corresponding Author

第二十三届中国计算语言学大会论文集，第588页-第599页，太原，中国，2024年7月25日至28日。
卷1：主会论文

Classical Chinese EE annotation's accuracy and efficiency, lowers reliance on experts, and introduces innovative annotation methods for low-resource languages, forging new avenues in data annotation during the LLM era.

Keywords: Event Extraction , Large Language Models , Semantic Dependency Analysis , Event Semantics

1 引言

事件抽取（EE）是自然语言处理（NLP）的一个重要任务，旨在从大规模非结构化文本中自动识别事件及其相关实体和属性(Chang et al., 2000)。事件抽取通常包括以下几个子任务：事件检测、事件关系分析和事件论元抽取等（如图1所示）。**事件检测**旨在确定文本中所描述的事件属于哪一种或多种事件类型，包括识别文本中的触发词，并判断这些词所属的事件类型(Wang et al., 2020)；**事件关系分析**则识别和理解文本中不同事件之间的关系，包括逻辑及时序关系等(Wang et al., 2022)；**事件论元抽取**是从文本中识别和抽取与特定事件相关的实体及角色(Ma et al., 2022)。

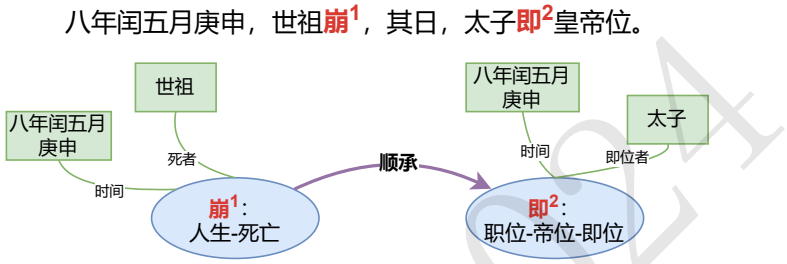

图 1: 事件抽取任务示意图。识别“崩”与“即”为触发词，对应“死亡”与“即位”事件类型，事件关系为顺承。论元抽取识别出时间（八月闰五月庚申）和主体（世祖、太子）。

针对古汉语的事件抽取是一个深具意义的研究领域，不仅有助于构建古汉语事件的本体知识，为深入挖掘和理解古汉语事件信息提供了理论框架，而且对推动事件知识图谱的构建(Guan et al., 2021)、智能问答系统的开发(Meng et al., 2021)、大语言模型（LLMs）知识评测(Yu et al., 2023)等多个领域起着重要作用。此外，研究古籍中的事件对于保护和传承历史文化遗产也至关重要，有助于更深刻地理解和传播历史文化。

然而，当前取得较大成功的事件抽取工具均依赖高质量、数量大的数据集资源。尽管NLP技术在现代语言处理上已有较成熟的解决方案(Li et al., 2021)，但由于古汉语的显著差异，其事件抽取任务无法直接利用现代汉语的数据资源，当前古汉语事件抽取标注面临着诸多挑战，数据集资源尤为匮乏。

具体来说，现有的NLP工具多数基于现代汉语训练，对于古汉语的特殊语法和语义理解不足，尤其缺乏专门针对古汉语的开源语义分析工具。同时，由于古汉语的歧义性、模糊性以及独特的语法现象（如省略和词类活用），从古汉语文本中准确提取事件信息变得极为困难。自从以GPT-3.5、GPT-4等为代表的模型发布以来，NLP领域进入了一个大模型时代(Zhao et al., 2023)。因而有学者也开始尝试直接使用这些模型处理古汉语问题(Wang et al., 2023)或事件抽取任务(Gao et al., 2023)，但由于古汉语训练数据的缺乏，LLMs在理解古汉语的语义上存在局限，因而在事件抽取任务上也无法达到理想效果。

尽管直接使用LLMs处理古汉语存在局限，我们提出可以通过为其提供结构更加清晰的现代汉语文本和丰富的语义信息来提高LLMs在古汉语领域的语义解析能力。因而我们提出了一种基于双层语义映射的LLMs辅助半自动标注框架（如图2所示）。具体来说，双层语义映射是指，（1）**第一层语义映射**：将古汉语翻译为现代汉语，利用**语义依存分析**技术将现代汉语映射为丰富的语义信息表示，旨在解决古汉语语义复杂性的问题。（2）**第二层语义映射**：在第一层映射的基础上，结合**事件语义学理论**及**语义依存关系框架**建立事件语义解析框架并设计提示

词，进一步利用LLMs对这些已分析的文本进行事件抽取标注，将语义依存关系逐步映射为具体事件信息。最后，**半自动标注框架**则结合了LLMs的自动标注和人类标注员的审核反馈，既提高了效率，又保证了标注的准确性和可靠性。

为验证方法的有效性，我们进行了交叉验证标注实验。结果表明，我们的方法不仅提高了古汉语事件抽取标注的准确性和效率，而且减少了对专业人员的依赖和标注人员的工作量，为古汉语事件抽取标注提供了一种新的解决方案。总的来说，我们的贡献如下：

- 我们提出引入LLMs用于古汉语文本标注，并开发半自动标注框架，结合了LLMs的自动标注能力和人类标注员的审核反馈，显著提升了低资源语言标注的准确率与效率，探索了大模型时代的数据标注的新方向。
- 我们通过双层语义映射方法、结合语义依存分析技术和事件语义学理论，为LLMs提供了丰富的语义信息表示，有效缓解了其在古汉语上存在的训练不足、语义理解能力欠缺的问题，从而更深入地分析古汉语文本中的事件信息。
- 我们从标注准确率、标注时间等多个维度验证了所提方法的可行性与有效性，展现了我们方法在应对古汉语文本标注挑战方面的优势，为低资源语言的标注实践提供了新的理论框架，推动了古汉语文本处理的发展。

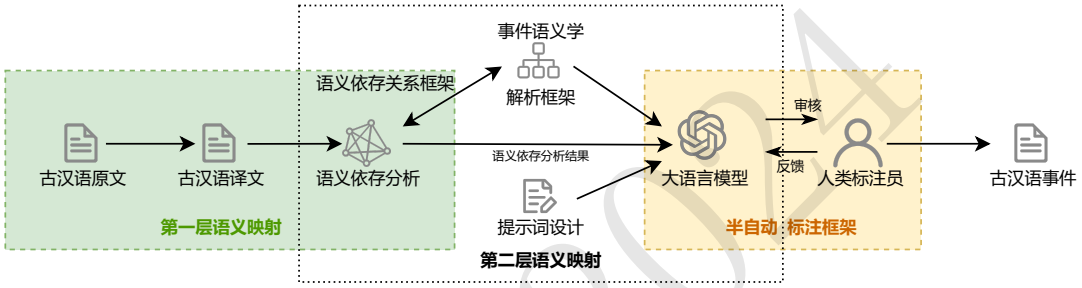

图 2: 基于双层语义映射的LLMs辅助古汉语事件抽取半自动标注框架示意图

2 相关工作

2.1 古汉语事件抽取

近年来，部分学者已经开始对古汉语事件抽取进行了初步的研究。这些研究大致可以分为基于模式匹配、机器学习和神经网络的方法。例如，李等(李章超et al., 2020)的研究为例，他们使用Fillmore的框架理论来研究《左传》中的战争事件，并通过模式匹配和条件随机场模型对相关事件进行识别和抽取。另外，刘等(刘忠宝et al., 2020)则在《史记》语料上，采用了BERT模型和LSTM-CRF模型进行历史事件和其组成元素的抽取。喻等(喻雪寒et al., 2021)学者研究了《左传》中的战争句，并利用RoBERTa-CRF模型进行事件抽取。更进一步，他们(喻雪寒et al., 2023)在基于神经网络的方法中融入了机器阅读理解模式，为事件抽取提供了新的思路，并在《左传》和《史记》上进行了实验。总之，这些方法在特定文本上有效且改进了古汉语事件抽取，但由于它们依赖于现代汉语优化模型，在处理古汉语的复杂性、特殊语法和语义时存在局限。

2.2 大模型事件抽取

目前，已有学者尝试探索LLMs用于事件抽取的可行性。Wei等(Wei et al., 2023)的研究从零样本信息抽取（IE）的角度出发，探讨了ChatGPT在未经特定训练情况下的信息提取效果。他们将零样本IE任务转化为一个包含两阶段框架的多轮问答问题（ChatIE）。在三个IE任务（实体-关系三元组提取、命名实体识别、事件抽取）的广泛评估中，ChatIE在六个数据集上的表现令人印象深刻。而Gao等(Gao et al., 2023)的工作探索了在长尾和复杂场景（即包含多个事件的文本）下ChatGPT用于事件抽取的效果，研究发现，ChatGPT在处理长尾和复杂场景（例如多事件文本）时的性能只有特定任务模型的一半。总之，虽然LLMs在事件抽取方面展现出潜力，但在处理特定类型的语言结构时仍存在显著挑战。

2.3 事件语义学研究

事件语义学是语义学的一个重要理论分支，核心是将事件作为理解句子意义的基本元素，强调事件在句子中的角色。Davidson(Davidson, 1967)首次在行动句（action sentences）中引入“事件”论元，认为动词不仅关系到主语和宾语，还关联到一个表示实际动作的“事件”。随后，事件语义学的研究领域不断扩展。如研究者们尝试区分不同类型的事件（如状态、过程和结果），并探讨它们是如何通过语言结构被表达和理解的(Maienborn, 2011)。事件的因果关系，即事件之间的相互作用和影响，也是研究的重点之一。Talmy(Talmy, 2000)对此进行了深入探讨，提出了初始，持续等因果关系概念，强调了事件语义在理解语言中的动态过程的重要性。随着计算语言学的发展，事件语义学在NLP和语言资源构建中也受到关注。例如，有学者利用事件语义学分析日语和汉语中的存在句，建立它们的事件结构模型，以此来处理语言信息(熊苇渡, 2020)。此外，一些学者还利用事件语义学理论，探索如何计算复杂句子中事件的语义距离，进而帮助构建语言知识词典(马腾and 詹卫东, 2014)。总之，事件语义学的发展不仅加深了我们对语言语义的理解，而且对事件抽取的框架提供了理论基础。

3 方法

我们使用图3来简要阐述我们的方法流程。图中主要分为四大部分，输入为二十四史语料，首先进行数据准备和预处理工作，其次构建事件语义解析框架，对古汉语译文进行语义依存分析（第一层语义映射），并设计提示词，再使用LLMs标注分析（第二层语义映射），然后人类进行审核映射（半自动标注框架），最终输出为二十四史事件抽取语料。

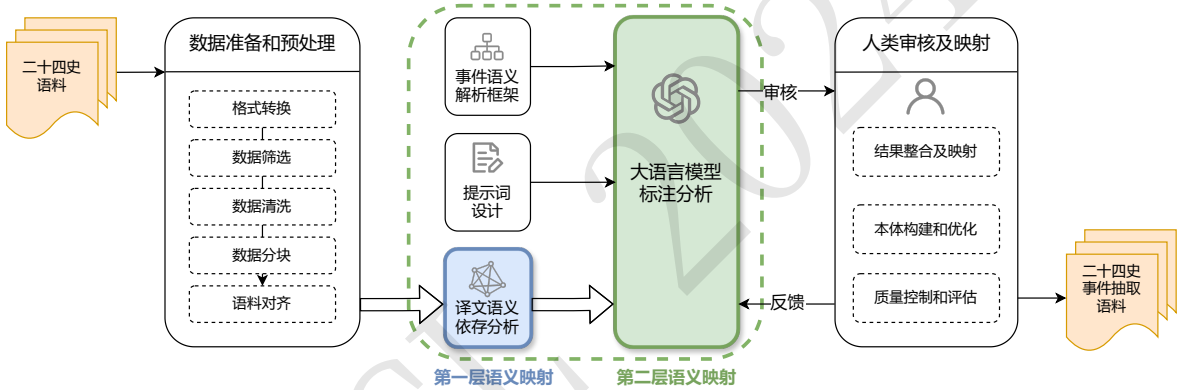

图 3: 方法流程图。展示从数据预处理到二十四史事件抽取语料的生成过程，其中蓝色为第一层语义映射，将译文映射为丰富语义信息；绿色（含虚线部分）为第二层语义映射，将语义信息进一步映射为事件信息。

3.1 数据准备和预处理

我们选择《二十四史》作为主要研究对象，这套权威文献详细记录了从黄帝时期（约公元前2550年）至明朝崇祯十七年（1644年）的中国历史，展现了不同朝代、各色人物的历史纪实。其中的“本纪”和“列传”章节，作为各史书的核心内容，是我们研究的重点。为确保语料的完整性和质量，我们采用了文白对照的《二十四史》精华版（共12册）[0]作为语料来源，并将对其进行预处理。如图4所示为数据预处理示例。具体来说：

- **格式转换**：将数据源格式转换为便于标注的格式。互联网公开的电子书籍多为equb等格式，我们需要将其转换为txt或json文本格式，以便进行标注。同时，为了更好地检索和理解文档，我们会在数据中加入文档ID和篇章信息。

- **数据筛选**：筛选出我们的目标语料。主要是以人物或篇章为检索单位，筛选出原文和对应译文数据。

[0]来源网址: http://www.happydot.top/9151.html

- **数据清洗**：删去无关数据内容，如文本中的多余符号等。这一步骤使用正则化表达式进行规则匹配，清洗数据中的一些噪音，以降低错误率。
- **数据分块**：将大批量文本数据切分为更小批次，方便处理，存储及排查错误。对于过长的文本，我们会在保持上下文连贯性的前提下进行分块，以提高效率。
- **语料对齐**：将古汉语原文与现代汉语译文进行句子级别的对齐，为语义映射打下基础。由于译文通常比原文长，这一步骤有助于确保分析结果能准确地对应到原文。


```
{
      "doc_id": "1史记卷八十一列传第二十一廉颇蔺相如列传",
      "sentence_id": "2",
      "total_sentences": "235",
      "original_text": "赵惠文王十六年，廉颇为赵将伐齐，大破之，取阳晋，拜为上卿，以勇气闻于诸侯。",
      "translation": "赵惠文王十六年，廉颇率领赵军攻打齐国，大败齐军，攻占阳晋，以军功官拜上卿，他也就以勇敢无畏而闻名于诸侯各国。"
}
```


图 4: 数据预处理后的古汉语语料JSON格式示例。展示文档编号（doc_id）、句子编号（sentence_id）、总句数（total_sentence）、原文（original_text）及对应翻译（translation）。

3.2 事件语义解析框架构建

在探索事件抽取和语义依存分析的联系部分，我们结合了事件语义学理论和语义依存关系框架[1]，旨在将语义依存分析标签映射到事件语义标签上，以构建一个全面事件语义解析框架。

事件语义学认为“事件”是一种具有确定时空属性并能够整合参与者功能的特定实体；强调事件的时空特性和可感知性，即事件作为世界中的实体，在空间和时间中有其特定的位置，并且在实现方式上可能有所不同（多样化）。例如，句子“七月丙寅，始皇死于沙丘平台。”中，此事件发生在特定时间（七月丙寅）和地点（沙丘平台），其中始皇为参与者，其死亡为事件核心。此事件是历史上真实发生的，可以被感知。进一步，死亡方式可能有多种，如自然死亡、疾病或他杀等，这表明同一事件（死亡）可以不同方式发生（如图5b所示）。

而**语义依存分析**旨在理解和分析语句中词语之间的语义关系。语义依存分析规范认为事件是以一个核心谓词为中枢论元，一个或多个相关的论元为周边的语义组合所反映的客观现实。例如，上例句子的语义依存分析中，核心谓词“死”连接多个相关论元。其中，始皇是当事，意味着他是事件的主体；“七月丙寅”代表时间，“沙丘平台”代表地点，共同构建了这一历史事件的背景（如图5a所示）。

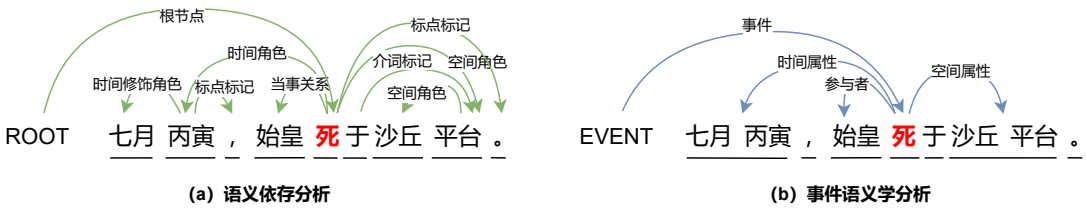

图 5: 语义依存分析（a）与事件语义学分析（b）对比分析

通过比较，可以发现两者都视事件为具有时空属性的客观现实，包含核心元素和相关参与者，且均具有可感知性和多样性。我们可以认为语义依存关系的部分框架和事件语义解析框架

[1]https://csdp-doc.readthedocs.io/zh-cn/latest/

相关，且存在映射关系（如图6所示）。事件抽取着眼于文本中的实体行为，涉及特定词语触发的事件结构。这一结构包括：触发词、参与者、情境、情态及事件关系。具体来看：

- **触发词（trigger）**：事件发生的核心词，常为动词或名词，对应语义依存分析的核心谓词。
- **参与者（participant）**：事件参与的主体和客体，对应语义依存分析中的主体角色和客体角色。
- **情境（situation）**：事件发生的时间、地点等背景要素，对应语义依存分析的情境角色。
- **情态（modal）**：人对事件的主观态度，如是否真实发生，对应语义依存分析中的否定和情态标记。
- **事件关系（event_relations）**：事件间的联系，对应语义依存分析的事件关系。

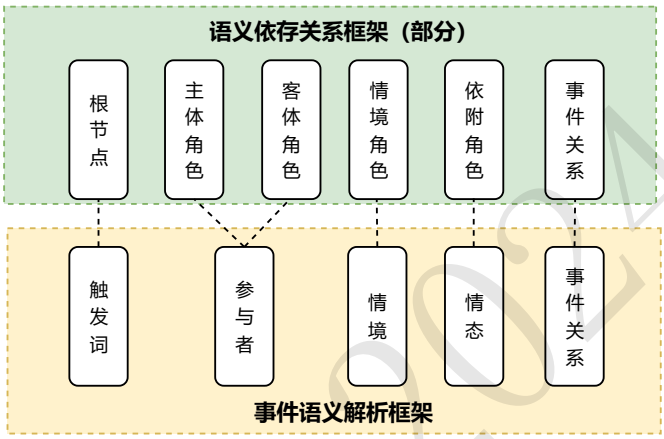

图 6: 语义依存关系框架（部分）与事件语义映射关系示意图。展示语义依存关系框架中各角色与事件语义解析框架中相应元素的对应关系。

3.3 第一层语义映射：译文语义依存分析

由于古汉语语义的复杂性，目前缺乏专门针对古汉语文本的语义依存分析工具，且现代NLP工具主要基于现代语言语料训练，对古汉语的特殊语法和语义处理能力有限。相反，现代汉语清晰的语法结构有助于NLP工具更准确地进行语义依存分析，从而提高事件信息提取的准确性，具有可行性。

因而我们使用语义依存分析工具[2]对古汉语译文进行语义依存分析，作为双层语义映射方法的第一层映射。语义依存分析是理解和抽取事件信息的关键，对译文进行语义依存分析对于事件抽取不仅是必要的，而且是提供深度语义理解的关键步骤。它能够有效地辅助抽取古汉语文本中的复杂事件关系，准确识别触发词和解析事件结构，为构建精确的事件本体和理解事件动态提供坚实基础。具体来说:

- **确定事件触发词**：事件的触发词通常是动词或名词，它们表明了发生了什么动作或事件。语义依存分析能够识别句子中的核心动词和名词，而这些词往往是事件的触发词。如句子“八年闰五月庚申日，孝武帝驾崩，当天，太子业登皇帝位。”中，“驾崩”和“登”就是两个事件的触发词。
- **解析事件论元**：语义依存分析中，每个词语被赋予特定的语义角色，而这些语义角色中就已包含了时间、空间、主体、客体等事件发生的要素。如在上述例句中，“孝武帝”和“太子业”是两个事件的主体。

[2]https://hanlp.hankcs.com/

- **揭示事件间的关系**：语义依存分析不仅识别词语之间的表层句法关系，还深入探究它们之间的语义关系，通过分析语义角色和关系，能够揭示事件之间的逻辑关系，如因果、顺承或并列等。如在上句中，“太子业登皇帝位”紧接着“孝武帝驾崩”发生，形成了一个顺承关系。

3.4 第二层语义映射：大语言模型标注分析

LLMs的通用性在于其能将多种任务转化为文本生成。通过设计简洁的提示词或提供少量示例，它能执行众多NLP任务，无需为特定任务构建专门的数据集或模型。利用这一特性，我们提出使用LLMs进行数据初步标注。尽管LLMs在直接处理古汉语方面存在局限，但我们通过提供结构更加清晰的现代汉语文本和丰富的语义信息，既能发挥LLMs高效的语义解析能力，又能规避其在处理古文时的不足，提高标注效率。

因而，在第一层语义映射及事件语义解析框架的基础上，我们使用LLMs对已进行语义依存分析的现代汉语文本进行事件抽取标注，将语义依存关系映射为具体事件信息（如图8所示）。这一过程包括两个阶段：提示词设计和模型标注。

- **提示词设计**：根据“角色+ 任务（详情）+ 目标+ 格式/内容要求+ 示例”的模板创建提示词（如图7所示）。“角色”界定了LLMs的角色，“任务”说明了具体的任务背景和细节，“目标”是模型应输出的结果，而“格式/内容”要求则规定了输出的具体形式。这个过程是迭代的，通过反复测试和评估模型输出，不断调整提示词以获取稳定的结果。

- **思维链及少样本学习**：为了提高LLMs分析的准确性，我们还采用了思维链（CoT）(Wei et al., 2022)技术和少样本学习方法(Brown et al., 2020)。利用思维链技术，提示词模拟了人类处理事件语义信息的逻辑顺序，这一过程逐步分析和提取语义依存关系框架中的关键成分。首先识别触发词以确定事件的核心，接着将其主体及客体角色对应到事件参与者上，然后通过提取时间和地点信息，确定事件的具体情境，处理情态词以捕捉事件的主观态度，并构建事件之间的逻辑和因果关系。整个映射过程确保了事件信息的准确性和完整性。最后，少样本学习则通过展示具体的标签映射和输出格式示例，使模型迅速适应并执行新任务。

- **模型标注**：我们利用gpt-4-1106-preview模型作为基础，通过API进行自动化标注。这包括将译文的语义依存分析结果输入模型，并将这些关系映射到事件分析上，即把语义依存关系转换为具体的事件信息。

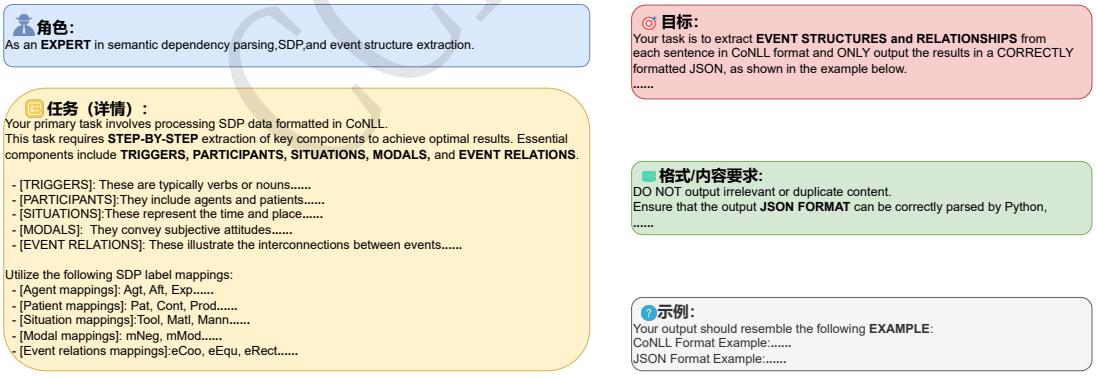

图 7: 输入给LLMs的提示词示意图。展示提示词的各部分，“角色+ 任务（详情）+ 目标+ 格式/内容要求+ 示例”，以及思维链、少样本学习方法。

3.5 人类审核及映射

尽管LLMs在许多方面表现出色，但在处理复杂和特定的语言任务时仍可能产生错误。而人类标注员可以识别并纠正这些错误，更进一步提高标注结果的准确性。因而在半自动标注框架

```
{
    "translation": "三年，蒙骜攻打韩国，夺取了十三个城邑。",
    "events": [
    {"event_id": "E1","trigger": "攻打","participant": [{"Agent": {"Agt": "蒙骜"}},
{"Patient": {"Datv": "韩国"}}],"situation": [{"Time": "三年"}],"modal": []},
    {"event_id": "E2","trigger": "夺取","participant": [{"Agent": {"Agt": "蒙骜"}},
{"Patient": {"Pat": "城邑"}}],"situation": [{"Time": "了"},{"Quan": "十三个"}],"modal": []}
    ],
    "event_relations": [{"source_event": "E1","target_event": "E2","relation_type":
"eResu"}]
}
```


图 8: LLMs输出的JSON格式展示。呈现古汉语事件的结构化信息，含译文（translation）、事件触发词（trigger）、参与者角色（participant）、情境（situation）及事件关系（event_relation）

中，LLMs首先进行初步标注，随后人类标注员对结果进行审核反馈。通过对照古今对齐语料，标注结果可准确映射回古汉语原文，有效弥补了模型在处理古汉语特有问题时的不足。

此外，基于LLMs的初步标注结果，人类标注员可以进一步进行**本体构建和优化**。结合语义依存分析提供的深层语义知识和LLMs的语义理解能力，我们能有效抽取古汉语文本中复杂事件关系，为构建精确事件本体和理解事件动态提供坚实基础，帮助人类标注员发现一些新的事件类型和事件关系，从而进一步优化本体知识（如图9所示）。

如在句子“契丹寇澶州，（安重霸）以临阵忸怩，为景延广所诛。”中，传统观点未将“忸怩”这类情感词视为事件类型。然而，通过分析结果，我们发现“忸怩”对事件发展有重要作用，呈现契丹军队攻击澶州与安重霸临阵畏缩间的转折关系，以及导致其被景延广处决的因果关系。因此，我们的方法能挖掘古汉语中特定事件类型，尤其是影响事件发展的人物情感词汇。

图 9: 我们方法在发现新事件类型和关系中的应用示例。图中句子分析揭示了传统未视为事件类型的情感词“忸怩”在事件发展中重要性，显示了古汉语文本的特色。

4 实验

4.1 设置

本实验旨在验证我们提出的方法对古汉语事件抽取标注任务的准确率、标注效率及标注难度的影响。实验设计包括**两个对照组**，一组在原文上直接进行标注（原文标注组），另一组则应用我们方法进行标注（辅助标注组）。具体实验任务涉及对古汉语原文句子进行触发词识别、事件论元和事件关系标注，标注难度依次递增，标签选取基于语义依存关系。

为了全面评估我们方法的效果，我们采取了**多维度**的测量方法。评价指标包括准确率（Precision）、召回率（Recall）和F1值（F1 Score），同时考量标注时间。首先，我们计算对照组在三个任务上的准确率、召回率和F1值，以全面评估标注员标注效果。这些指标计算的基准（baseline）是来自多位专家的综合标注结果。其次，我们统计了对照组的平均标注时间。具体操作中，我们在限定时间内观察对照组的标注句子数量，以此来计算平均标注时间。各指标的具体概念和计算方法如下：

- **准确率(Precision)**：指正确标注的正例数量与所有标注为正例的数量之比。准确率度量了标注人员在标注触发词、事件论元和事件关系时的准确性。

$$\text{准确率（Precision）} = \frac{\text{正确标注的正例数目}}{\text{所有标注为正例的数目}} \tag{1}$$

- **召回率(Recall)**：指正确标注的正例数量与所有应该标注为正例的数量之比。召回率度量了标注人员从文本中正确标注事件元素的能力，召回率高意味着标注人员漏标的情况较少。

$$\text{召回率（Recall）} = \frac{\text{正确标注的正例数目}}{\text{所有应该标注为正例的数目}} \tag{2}$$

- **F1值(F1 Score)**：指准确率和召回率的调和平均数。F1值被用来综合评价标注人员在识别和标注文本中事件元素（如触发词、事件论元和事件关系）的效果。

$$\text{F1值（F1 Score）} = 2 \times \frac{\text{准确率} \times \text{召回率}}{\text{准确率} + \text{召回率}} \tag{3}$$

- **平均标注时间**：是标注一定数量句子所需时间的平均值，用于衡量标注效率和难度。这反映了标注人员在给定时间内完成标注任务的速度。

$$\text{平均标注时间（分钟/句）} = \frac{\text{标注所用总时间（分钟）}}{\text{标注的句子数量}} \tag{4}$$

同时，为了保证实验的一致性和合理性，我们从二十四史的63篇文档中随机筛选了320句古汉语原文作为标注实验语料。这些语料在句子总长度和标签数量上进行了严格控制，以确保句子难度的均匀分布。**标注人员的选择**也是实验设计的关键环节。我们选取了具备一定古汉语基础知识的人员，并将标注分为**试标注**（20句）和**正式标注**（300句）两个阶段（如图11所示）。试标注环节的目的在于确定标注人员的知识水平，以确保参与实验的标注人员的水平相当。正式标注环节则分为**两轮**进行，每轮各150句，同一标注人员需采用不同对照组方式对句子进行多次**交叉标注**。此外，我们制定了标注规范，并对标注人员进行了初步培训，从而减少了人为因素对实验结果的潜在偏差，增强了实验结果的可靠性。

4.2 结果及分析

在本次实验中，我们对提出的基于双层语义映射的LLMs辅助古汉语事件抽取半自动标注框架进行了全面的实验评估。在表1和图10中我们给出了实验结果，可以看到我们的方法在提高古汉语事件抽取标注的准确性、完整性和效率方面都具有**显著优势**。特别是在事件关系标注这一复杂任务上，提升效果尤为突出，显示了其在处理古汉语文本中复杂语义信息时的高效性和准确性。

具体而言，在触发词识别任务中，我们的方法相较于直接标注方法，尤其在召回率方面表现尤其优异，这表明我们的方法能够更全面地识别文本中的关键信息，减少了遗漏重要触发词的情况。而对于事件论元及事件关系标注任务，我们的方法同样显示出了更高的准确率和召回率，尤其是事件关系标注的提升最为显著，F1值几乎提升了一倍，这证明了我们方法在理解和分析复杂事件关系方面的有效性。

同时，图10显示，**我们的方法显著减少了平均标注时间，提高了标注效率**。这一点在两位标注员的数据上都得到了验证，其中标注员B的效率提升百分比高于标注员A，这可能与个体差异有关，但整体趋势一致表明我们的方法能有效缩短标注所需时间，提升标注效率。此外，平均标注时间也从侧面反映出**标注任务的难度**，一般情况下我们认为标注时间少则意味着标注任务较为简单。

此外，**不同任务本身的复杂度导致了各评价指标之间的显著差异**。具体来说，触发词识别任务的评价指标相对较高，因为它主要涉及到在文本中识别显而易见的动词或名词，较为简单。相反，事件论元识别和事件关系标注任务的评价指标较低，这一差异归因于这两项任务的复杂性。事件论元识别任务要求识别与特定事件相关的多种元素，这些元素可能包括古汉语中特有的人名、地名和官职等实体。这要求标注人员拥有对相关文化背景的深入了解。而对于事件关系标注任务则要求标注人员不仅理解文本中明确表达的信息，还要对古汉语中隐含的复杂语义关系有深刻的理解。因此，该任务的评价指标较低反映了其在认知和分析层面的高难度。

表 1: 对照组在不同标注任务上的效果对比

对照组	任务	标注员A			标注员B		
		准确率	召回率	F1值	准确率	召回率	F1值
原文标注组	触发词识别	78.95	77.74	78.34	65.48	81.18	72.49
	事件论元标注	46.42	38.80	42.27	58.22	53.11	55.55
	事件关系标注	26.35	25.36	25.85	26.28	28.72	27.45
辅助标注组	触发词识别	77.08	85.61	**81.12**	76.74	80.49	**78.57**
	事件论元标注	58.92	57.05	**57.97**	58.93	55.85	**57.35**
	事件关系标注	44.90	46.81	**45.83**	37.61	38.86	**38.23**

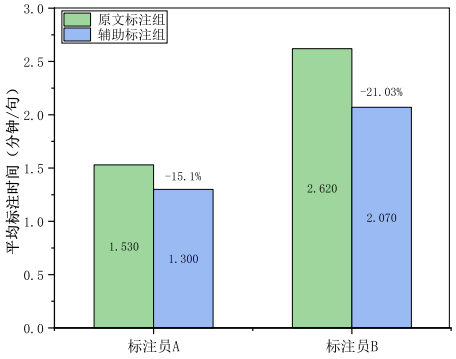

图 10: 对照组在标注时间上的对比

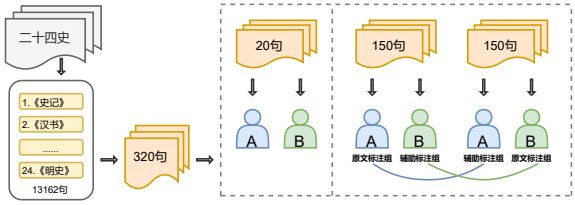

图 11: 实验标注阶段示意图。如图标注分为试标注（20句）和正式标注（300句）两个阶段，其中正式标注分为两轮（150句）进行，标注人员AB需采用不同对照组方式对句子进行交叉标注。

5 讨论

5.1 优点

我们的方法引入了一种基于双层语义映射的LLMs辅助半自动标注框架，专门针对古汉语事件抽取的数据标注。这一方法的核心在于融合了NLP工具和LLMs的优势，并结合人类标注员的审核反馈，有效缓解了古汉语文本的语义复杂性问题。具体来说：

（1）**通过提供丰富语义帮助LLMs进行分析，提高准确性**：古汉语文本的复杂性一直是NLP的一大挑战。我们的方法不仅保留了古文的丰富语义，还通过精确和可靠的分析，降低了文本的复杂度。同时，借助现代汉语译文的语义依存分析结果，我们的方法能够减少歧义，减少误标的可能性。

（2）**借助自动化工具和LLMs，提升标注效率与质量**：通过利用自动化工具和LLMs快速生成初步标注数据，结合人类专家的审核和反馈，我们的方法显著提高了标注的效率和质量。这种双管齐下的策略确保了数据标注的准确性和可靠性。

（3）**减少专业背景依赖和标注工作量**：传统的NLP工具在处理古汉语文本时面临诸多挑战，尤其是语义层面的复杂性和专业知识的需求。我们的方法通过双层语义映射，不仅减少了对背景知识的限制，还减少了工作量，使得标注过程更加容易和普遍适用。

5.2 局限

虽然我们的方法在提升古汉语事件抽取准确性等方面取得了效果，但仍存在一些局限。首先，**译文的选择直接影响语义依存分析的准确性**。依赖质量不高的译文进行事件抽取标注，可能导致标注结果质量不高。为此，我们强化数据预处理操作，选用出版书籍语料作为高质量译文。同时，我们对古汉语原文和现代汉语译文进行了对齐。

语义依存分析局限于句子内，缺乏上下文信息。由于我们分析的是古汉语历史文本篇章，而语义依存分析通常局限于单个句子内部，可能无法完全捕捉跨句子或跨篇章的事件关系。面对这一局限，我们在数据预处理阶段中加入上下文索引信息（如文档id和篇章位置），每一个句子都能明确索引到具体的文档中的上下文信息，为标注人员提供了参考。

LLMs的标注结果的质量受限于提示词设计。如果提示词设计不当，模型可能无法生成准确的标注结果。此外，由于LLMs在处理复杂输入时可能出现“幻觉（hallucination）(Li et al.,

2023)”现象，即产生与实际情况不符或逻辑混乱的回答。针对这一局限，我们在提示词设计阶段使用数据进行反复测试，观察评估LLMs的输出结果，不断调试提示词以便获得稳定的预期输出结果。同时使用了一些提示技术（如思维链(Wei et al., 2022)和少样本学习(Brown et al., 2020)）。

总而言之，对于我们方法的局限性，一个解决方法是结合**半自动标注框架**，其中LLMs提供初步标注，而人类标注员负责审核反馈，这种方法结合了LLMs的效率和人类标注员的深入理解，可以最大化事件抽取的准确性和可靠性。

6 结论与未来工作

我们通过使用基于双层语义映射的LLMs辅助半自动标注框架，有效地解决了古汉语事件抽取标注中的诸多挑战。我们首先将古汉语文本翻译为现代汉语，结合事件语义学理论和语义依存分析技术将译文映射为丰富的语义信息来解决古汉语语义复杂性的问题，进而利用LLMs进行深入分析，从而提高了对低资源语言标注的准确性。此外，我们结合了LLMs的自动标注能力与人类标注员的审核反馈，提高了标注效率及结果的可靠性。我们的方法在解决古汉语文本的复杂性等挑战方面表现出色，为低资源语言处理提供了新的方法论，探索了LLM时代数据标注的新方向。未来，我们将继续从数据质量和提示词设计方面优化我们的方法，并将其应用于更广泛的应用场景如历史文献的数字化分析等，以进一步推动数字人文学科和NLP领域的发展。

致谢

本成果受国家自然科学基金项目（62306045，61872402），北京语言大学校级项目（中央高校基本科研业务费专项资金）（18ZDJ03），北京语言大学梧桐创新平台项目（21PT04）基金资助；北京语言大学研究生创新基金（中央高校基本科研业务费专项资金）项目成果（24YCX073）。

参考文献

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Heng-Hsou Chang, Yau-Hwang Kuo, and Jang-Pong Hsu. 2000. An event-driven and ontology-based approach for the delivery and information extraction of e-mails. In *2000 International Symposium on Multimedia Software Engineering, ISMSE 2000, Taipei, Taiwan, December 11-13, 2000*, pages 103–109. IEEE Computer Society.

Donald Davidson. 1967. The logical form of action sentences. In Nicholas Rescher, editor, *The Logic of Decision and Action*, pages 81–95. University of Pittsburgh Press.

Jun Gao, Huan Zhao, Changlong Yu, and Ruifeng Xu. 2023. Exploring the feasibility of chatgpt for event extraction. *CoRR*, abs/2303.03836.

Saiping Guan, Xueqi Cheng, Long Bai, Fujun Zhang, Zixuan Li, Yutao Zeng, Xiaolong Jin, and Jiafeng Guo. 2021. What is event knowledge graph: A survey. *CoRR*, abs/2112.15280.

Qian Li, Jianxin Li, Jiawei Sheng, Shiyao Cui, Jia Wu, Yiming Hei, Hao Peng, Shu Guo, Lihong Wang, Amin Beheshti, and Philip S. Yu. 2021. A compact survey on event extraction: Approaches and applications.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6449–6464. Association for Computational Linguistics.

Yubo Ma, Zehao Wang, Yixin Cao, Mukai Li, Meiqi Chen, Kun Wang, and Jing Shao. 2022. Prompt for extraction? PAIE: prompting argument interaction for event argument extraction. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 6759–6774. Association for Computational Linguistics.

Claudia Maienborn. 2011. 34. event semantics.

Fanqi Meng, Wenhui Wang, and Jingdong Wang. 2021. Visual analysis of the research status of intelligent question answering system. In *2021 13th International Conference on Computational Intelligence and Communication Networks (CICN)*, pages 145–149.

Leonard Talmy. 2000. Toward a cognitive semantics, vol. 1: Concept structuring systems.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A massive general domain event detection dataset. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 1652–1671. Association for Computational Linguistics.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. MAVEN-ERE: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 926–941. Association for Computational Linguistics.

Zhaoji Wang, Shirui Zhang, Xuetao Zhang, and Renfen Hu. 2023. 古汉语通假字资源库的构建及应用研究(the construction and application of an Ancient Chinese language resource on tongjiazi). In Maosong Sun, Bing Qin, Xipeng Qiu, Jing Jiang, and Xianpei Han, editors, *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 535–546, Harbin, China, August. Chinese Information Processing Society of China.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2023. Zero-shot information extraction via chatting with chatgpt. *CoRR*, abs/2302.10205.

Jifan Yu, Xiaozhi Wang, Shangqing Tu, Shulin Cao, Daniel Zhang-li, Xin Lv, Hao Peng, Zijun Yao, Xiaohan Zhang, Hanming Li, Chunyang Li, Zheyuan Zhang, Yushi Bai, Yantao Liu, Amy Xin, Nianyi Lin, Kaifeng Yun, Linlu Gong, Jianhui Chen, Zhili Wu, Yunjia Qi, Weikai Li, Yong Guan, Kaisheng Zeng, Ji Qi, Hailong Jin, Jinxin Liu, Yu Gu, Yuan Yao, Ning Ding, Lei Hou, Zhiyuan Liu, Bin Xu, Jie Tang, and Juanzi Li. 2023. Kola: Carefully benchmarking world knowledge of large language models. *CoRR*, abs/2306.09296.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Z. Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jianyun Nie, and Ji rong Wen. 2023. A survey of large language models. *ArXiv*, abs/2303.18223.

刘忠宝, 党建飞, and 张志剑. 2020. 《史记》历史事件自动抽取与事理图谱构建研究. 图书情报工作.

喻雪寒, 何琳, and 徐健. 2021. 基于roberta-crf的古文历史事件抽取方法研究. 数据分析与知识发现, 5(7):26–35, 7.

喻雪寒, 何琳, and 王献琪. 2023. 基于机器阅读理解的古文事件抽取研究. 情报学报, 42(3):316–326, 3.

李章超, 李忠凯, and 何琳. 2020. 《左传》战争事件抽取技术研究. 图书情报工作, 64(7):20–29, 4.

熊苇渡. 2020. 面向语言信息处理的存现句事件语义学类型研究. 面向语言信息处理的存现句事件语义学类型研究.

马腾and 詹卫东. 2014. 基于事件语义距离的v1-v2述结式判别研究2014年2月27日. 计算机工程与应用.