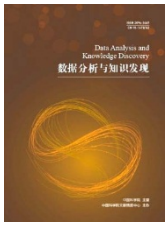

数据分析与知识发现
Data Analysis and Knowledge Discovery
ISSN 2096-3467,CN 10-1478/G2

《数据分析与知识发现》网络首发论文

题目：基于 RoBERTa-CRF 的古文历史事件抽取方法研究
作者：喻雪寒，何琳，徐健
网络首发日期：2021-03-26
引用格式：喻雪寒，何琳，徐健．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.
https://kns.cnki.net/kcms/detail/10.1478.g2.20210325.1127.002.html

网络首发时间：2021-03-26 17:29:10
网络首发地址：https://kns.cnki.net/kcms/detail/10.1478.g2.20210325.1127.002.html

基于 RoBERTa-CRF 的古文历史事件抽取方法研究[1]

喻雪寒 何琳[2] 徐健
（南京农业大学信息管理学院 南京 210095）

摘要：
[目的]为有效抽取典籍中蕴含的事件信息，本文构建了面向典籍的事件抽取框架，并采用 RoBERTa-CRF 模型实现事件类型、论元角色和论元的抽取。
[方法]选择《左传》的战争句作为实验数据，建立事件类型和论元角色的分类模板。基于 RoBERTa-CRF 模型先用多层 transformer 提取语料特征，再结合前后文序列标签学习相关性约束，由输出的标记序列识别论元并对其抽取。
[结果]对比 guwenBERT-LSTM、BERT-LSTM、RoBERTa-LSTM、BERT-CRF、RoBERTa-CRF 等五种模型在数据集上的事件抽取实验结果，RoBERTa-CRF 的精确度为 87.6%、召回率为 77.2%、F1 值达到 82.1%，验证了 RoBERTa-CRF 模型的有效性和可操作性。
[局限]现阶段使用的数据集规模较小，无法使主题类别更均衡化。
[结论]本文构建的 RoBERTa-CRF 模型，提升了面向《左传》战争句的事件抽取效果。
关键词：RoBERTa；CRF；事件抽取；古文
分类号：TP391，G255
DOI: 10.11925/infotech.2021.0094

Research on Event Extraction from Ancient Books based on RoBERTa-CRF

Yu Xuehan, He Lin, Xu Jian
(College of Information Management, Nanjing Agricultural University, Nanjing 210095, China)

Abstract:
[Objective] In order to extract the event information contained in ancient books effectively, this paper constructs an event extraction framework for ancient books, and uses the RoBERTa-CRF model to extract the event type, argument role and argument.
[Methods] The war sentences in Zuozhuan are selected as the experimental data to establish the classification templates of event types and argument roles. Based on the RoBERTa-CRF model, the multi-layer transformer is used to extract the features of the corpus, and then combined with the sequence tags to learn the correlation constraints, the argument is identified by the output tag sequence.
[Results] Comparing the experimental results of five models of guwenBERT-LSTM, Bert-LSTM, RoBERTa-LSTM, Bert-CRF and RoBERTa-CRF on dataset, the validity and operability of the model were verified.
[Limitations] The scale of the data set used at this stage is small and cannot make the topic categories more balanced.
[Conclusions] The RoBERTa-CRF model constructed in this paper improves the effect of event extraction for war sentences in Zuozhuan.
Keywords: RoBERTa; CRF; event extraction; Ancient Chinese Language

---

[1]本文系南京农业大学中央高校基本科研业务费“面向典籍文本的触发动词语义体系构建研究”（SKCX2020006）及中国博士后面上基金“基于实体语义关联的学术观点知识表示与服务研究”（2020M681652）的研究成果之一。
[2]通讯作者:何琳，ORCID:0000-0002-4207-3588，E-mail:helin@njau.edu.cn。

1 引言

近年来伴随数字人文研究的兴起，面向人文研究的“数据基础设施”建设正在成为共同的呼声。数字人文基础设施是一种支持数字人文研究活动的“研究基础设施(Research Infrastructure)”，包括数字化的文献资源、数据库、工具平台、支持知识生产和信息交流的网络空间等[1]。长期以来，图书情报等记忆机构保存了大量历史文献，将这些历史文献进行细粒度的内容生产和组织构成了支撑数字人文研究的重要基础。随着信息检索技术和自然语言处理技术的不断发展，文本分词及命名实体抽取得到了相对有效的解决，在此基础上进行特定类型的关系抽取是能够实现深层次文本挖掘的重要手段。而事件抽取作为信息检索中的基础工作，是通过使用预定义的事件模板从文本中发现和提取所需的特定类型的事件，借助抽取触发词、识别事件论元达到对文本细粒度的揭示，在数字人文“数据基础设施”构建中发挥着必不可少的作用。

中国文化历史悠久、绵延五千年而长盛不衰，典籍成为中华文化源远流长的有力见证。从典籍中定位、挖掘和归纳信息是构建数据基础设施的重要基石，然而由于典籍数据集体量不大，分类后容易导致类别不均，古文较白话文而言语句言简意赅，句长大多较短，造成事件论元集中分布，因而面向典籍的事件抽取存在不小的难度。在机器学习时代，典籍抽取主要依赖于模板匹配的方式，如利用模式匹配法进行战争句识别，选择条件随机场模型对命名实体进行识别和抽取[2]；利用正则表达式等从已经去除官名、地名的墓志铭中抽取有关亲属信息等[3]。而在深度学习到来之后，神经网络自动学习特征表示的特点给典籍抽取带来了新的机遇，与人工构建的离散特征不同，神经网络通过提供连续的向量表示，帮助挖掘词语之间的潜在关联。刘忠宝等在 BERT 模型和 LSTM-CRF 模型的基础上，提出面向《史记》的历史事件及其组成元素抽取方法，并基于此构建《史记》事理图谱[4]。

本文将在已有研究[2][3][4]的基础上，将典籍中蕴含的主题事件抽取看作序列标注任务，结合典籍文本事件抽取语言精练、简短的特点，构建既能继承 transfomrer 优势，又能考虑句子局部特征的深度学习模型 RoBERTa-CRF，有效克服数据集较小、古文句短造成事件论元集中分布等问题。实验结果证明，本文提出的模型在典籍的事件抽取上取得了较好的效果。

2 相关研究

早期的事件抽取主要采用了基于模式匹配的方法，它首先构造一些特定的事件模板，然后通过模式匹配从文本中抽取单个论元的事件。最早的模式匹配系统可以追溯到 1993 年 Riloff 等人开发的用于抽取限定域的恐怖事件 AutoSlog[5]，之后受 AutoSlog 系统的启发，许多基于模式的事件抽取被研发出来用于不同领域，包括生物医学[5]、金融[6]等。Cohen 等利用 OpenDMAP 语义解析器通过生物医学本体分析提取模板，为生物医学概念及其属性提供各种高质量的本体模板[6]。Arendarenko 和 Kakkonen 开发了一个基于本体的事件抽取系统 BEECON，从在线新闻中提取商业知识[7]。

基于模式匹配的方法因为是由具备专业知识的专家手工构建，所以生成的事件模板定义明确、质量较高，故而在特定领域应用中往往表现优异。相对的，该类模板的缺点是需要大量的人工标注，耗时耗力，且存在移植性差的问题，因此基于模式匹配的事件抽取方法比较适合应用在特定的领域。

随后发展的机器学习技术，本文将分为传统的机器学习和深度学习两种方式加以介绍。传统的机器学习首先需要从文本中提取特征作为分类模型的输入，常用的文本特征可分为词汇、句法和语义特征，陈慧炜定义了破案、抓获和报案三种事件类型，利用手工标注的词形、词形、实体和事件特征，以此辅助 CRF 模型进行事件类型和论元的识别[8]；赵文娟等通过对语义框架和语法知识的介绍，提出了基于句法依存分析的角色填充思路和技术，以“灾难场景”框架下的“森林火灾”事件为例，用最大熵算法对填充过程进行了说明，例证了方法的有效性[9]。

传统的机器学习方法并不完美，特征工程是其面临的主要挑战。尽管词汇、句法、语义等多种特征可以作为分类器的输入，但它们的构建也需要语言知识和领域专长，限制了分类模型的应用性和适应性。此外，这些特征往往以独热向量的形式表示，这不仅造成了数据稀疏问题，还使训练时的特征选择变得更加复杂。

近年来，随着深度学习技术在多个领域的成功应用，基于神经网络的自然语言处理任务也相继涌现。在人工神经网络中，最底层的原始数据以一个非常简单的低维向量作为输入，每一层可以学习上一层的输入并将其转换为更混合抽象的表示，然后输入到下一层，直到最高层的输出用于分类。与传统的机器学习技术相比，深度学习可以大大降低特征工程的难度。

当前基于机器学习的事件抽取大都依据 ACE2005 评测会议的标准，将事件抽取分成四个子任务：触发词识别、事件类型分类、论元识别和论元角色分类。通常将前两个任务合并，称为事件识别；将后两个任务合并，称为事件论元角色抽取。四个子任务可以管道模式或联合抽取方式执行，经典的管道式事件抽取模型包括 Chen 于 2015 年提出的 DMCNN（动态多池化卷积神经网络），通过一个动态多池层同时提取词汇层和句子层特征来评估句子的每个部分[10]；联合抽取的典型案则有 Sha 等设计的 DBRNN（依赖桥循环神经网络），该模型在两个 RNN 神经元的基础上增加了单词之间的依赖桥信息，为每种依赖关系分配一个权重，提升论元角色的分类效果[11]；Duan 等人设计出的 DLRNN 模型（文档级循环神经网络），则通过使用分布式向量进行文档表示来提取跨句子甚至跨文档的线索[12]。

管道抽取与联合抽取相比，一方面容易造成级联错误，即将上游事件识别的错误传播到下游的论元角色分类中；另一方面，下游分类器无法影响上游分类器的决策，单独的触发词检测或者事件论元识别考虑不到触发词-论元之间的关系，这将直接导致上下文信息的丢失。相反，联合抽取则会通过对事件识别和论元角色分类两个阶段的联合建模，来解决了管道抽取存在的常见问题，进而提升模型的整体性能。基于上述分析，本文采用联合抽取的方式，结合 RoBERTa-CRF 模型，对《左传》事件及其论元抽取展开深入研究。

3 研究方法

3.1 研究框架

图 1 给出了《左传》事件抽取框架图，该框架主要分为三个部分：数据预处理、模型训练、模型预测及性能评估。

数据预处理部分首先根据《左传》战争句的内容特点构造事件类型，再依据各种事件类型涉及到的论元建立论元角色分类，由生成的事件类型及论元角色模板标注数据、提炼出其中的具体论元。

模型训练部分把事件类型、事件论元和论元角色表示为三元组的形式，需要说明的是，本文不考虑对触发词的识别，而是将抽取任务转化为序列标注工作。在语料标注后，训练数据将通过预训练模型 RoBERTa，利用内置的哈工大 LTP 作为分词工具，对组成同一个词的汉字进行向量化表示。最后借助 CRF 层结合前后文序列标签学习相关性约束，输出最终的标记序列。

模型预测及性能评估部分选用精确率、召回率和 F1 值三个评测标准，评估系统标注出正确的论元数量，从而考察模型的性能价值。

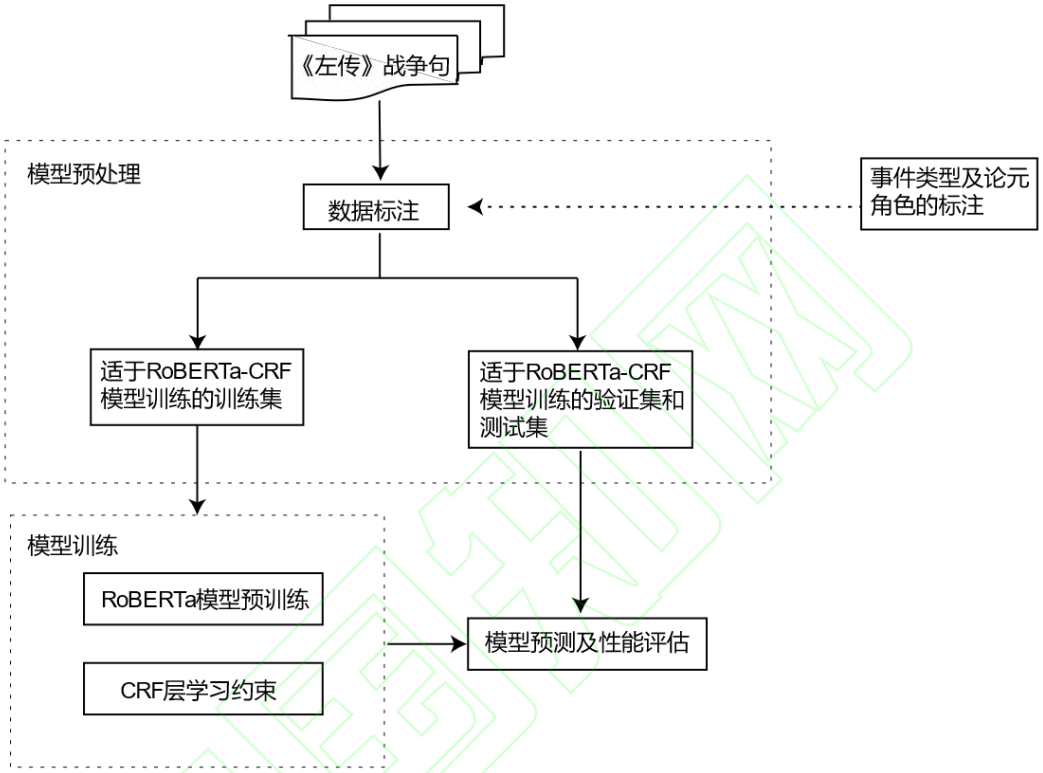

图 1　基于 RoBERTa-CRF 的古文历史事件抽取

Fig.1 Event Extraction Framework From Chinese Ancient Books based on RoBERTa-CRF

3.2　事件类型及论元角色的建立

《左传》记录了春秋时期诸侯争霸的历史，依照鲁国十二公的继位顺序对春秋战争有较为详尽的记述，邓勇依据《十三经注疏》[13]、《春秋左传正义》[14]、《左氏兵法》[15]、《中国历代战争年表》[16]对《左传》中的战争做了完全统计，撰写了《春秋左传战争表》[17] 。本文将在邓勇所注的战争句基础上，依据战争目的进行对比和归纳，将战争划分为三大类型：征战类、戕杀类、救援类。由统计 结果可知：征战类数据最多，救援类次之，戕杀类最少。

（1）征战类

征战类包括交战双方的讨伐、攻打、包围、追击、缴获、驱逐等动作，描述的主要是国家之间有组织的暴力斗争，一次战争一般由同样的交战方在同一地点进行。征战类的论元角色较为丰富，包含时间（战争发生的季节或月份）、进攻方（战争发起方）、防守方（受攻方）、战争原因、战争地点（战争发生地）、战利品（缴获的城池、土地、财物及人质）、助战方（由进攻方带领）等。以定公八年晋国侵郑事件为例：

原句是“秋，晉士鞅會成桓公侵鄭，圍蟲牢，報鄭侵周”，该事件发生在秋

天（时间），晋国的士鞅（进攻方）为了报复定公六年郑国攻打周国（战争原因），联合成桓公（助战方）侵入郑国（防守方），包围虫牢（战争地点）。

（2）戕杀类

戕杀类的事件触发词统计后主要是“殺”和“弒”，该类区分度非常鲜明，即出现了进攻方和受害者。由于是战争中的杀戮，因此不同于普通的刺客暗杀事件，戕杀类特指存在于某场进攻或反叛集团争斗中的死亡事件，论元角色包含时间、进攻方、受害人（受到袭击的个人）、战争原因、战争地点、战利品、助战方。以昭公五年鲁国内战举例：

“(春)，魯南遺使國人助豎牛，攻仲壬於大庫之庭，殺之。豎牛取東鄙三十邑于南遺”指的是春天（时间），鲁国的南遗（助战方）派遣国人帮助竖牛（进攻方）在府库的庭院里（战争地点）攻打并杀死仲壬（受害人），竖牛取得了东部的三十个城邑（战利品），将其送给了南遗。

（3）救援类

救援类指战争发生后，集团、国家等力量参与实施解救行动，以减轻人员伤亡和财产损失为目标的过程，救援类是救护和援驰行为，其特点是产生了援军和被救方。论元角色定义包含时间、援军（施救方）、被救方、战争原因、战争地点、敌军（袭击被救方的军队）、助战方。以昭公六年吴楚房钟之役为例：

“(秋)，吳救徐，楚令尹子蕩帥師伐吳，吳人敗諸房鐘”发生于秋天（时间），吴国人（援军）救援徐国（被救方），楚国令尹子荡（敌军）率领军队进攻吴国，吴国人在房钟（战争地点）击败了令尹子荡的军队。

总体而言，征战、戕杀和救援三大类共有的论元角色包括：时间、战争原因、战争地点、助战方。征战类由于涉及到交战双方，因此又包含“进攻方”、“防守方”；戕杀类的最大特点是出现了受害方，因此增添“受害人”一项作为论元角色；救援类指对交战双方的一方进行支援，因而产生了“援军”、“被救方”、“敌军”等角色。具体的事件类型和论元角色设置如表 1 所示：

表 1　《左传》战争句事件类型及论元角色

Table1 Event Types and Argument Roles of War Sentences in Zuozhuan

触发词	事件类型	论元角色
伐、敗、入、取、侵、討、圍、滅、戰、追、克、降、襲、執、攻、獲、門、徼、軍、逐	征战	时间、进攻方、防守方、战争原因、战争地点、战利品、助战方、参与人物
殺、弒	戕杀	时间、进攻方、受害人、战争原因、战争地点、战利品、助战方
救、援	救援	时间、援军、被救方、战争原因、战争地点、敌军、助战方

3.3 事件抽取的模型

《左传》战争句的抽取难点在于数据集较小，三大事件类型分类不平衡，事件论元在同一句集中分布等。针对这类数据的特点，本文选取了 RoBERTa 模型作为事件抽取的模型。

（1）BERT 模型

目前深度学习的预处理阶段大多使用谷歌开发的 BERT 语言编码模型[18]，该模型利用双向 Transformer 对大规模无标注语料进行训练，进而获取包含丰富语义信息的编码表示。关于 BERT 模型，文献中有诸多介绍，本文不再赘述。

（2）RoBERTa 模型

RoBERTa 模型[19]不仅继承了 BERT 模型的优势，将输入的句子表示为字向量、句向量、位置向量三者之和，而且在模型结构和数据层面上对 BERT 模型进行了改进，用更大的单次训练样本数和更多的数据训练模型；移除了 NSP（next sentence prediction）目标函数，用更长的序列长度训练；在预训练阶段采用了中文 wwm（Whole Word Masking，全词遮掩）技术，使用了哈工大 LTP 作为分词工具，对组成同一个词的汉字全部进行遮掩。

（3）guwenBERT 模型

BERT 模型和 RoBERTa 模型的共同点在于训练数据都为维基百科数据，运用的是现代汉语的字词逻辑，而由北京理工大学阎覃开发的 guwenBERT[20]模型则改为基于殆知阁古文文献语料训练，包含 17 亿字古文，共 15,694 本古文书籍，其中有古文诗词、小说、四书五经等诸多种类古文文本。文献语料中的所有繁体字均经过简体转换处理，结合现代汉语 RoBERTa 权重和大量古文语料，将现代汉语的部分语言特征向古代汉语迁移。

（4）三种模型的比较

BERT、RoBERTa、guwenBERT 三种模型的部分区别见表 2。

表 2　三种预训练模型的区别比较

Table2 Comparison of three Pre-training Models

预训练模型	BERT	RoBERTa	guwenBERT
本文调用的模型名	BERT-Base, Chinese	RoBERTa-wwm-ext, Chinese	ethanyt/guwenbert-base
训练数据	中文维基百科	中文维基百科	殆知阁古文文献
字形	简体中文、繁体中文	简体中文、繁体中文	简体中文
句子切分粒度	以字为粒度	以词为粒度	以字为粒度
词表大小	21128	21128	23292
支持框架	Pytorch、TensorFlow	Pytorch、TensorFlow	Pytorch
是否采用 NSP 函数	是	否	否
是否选用 wwm 技术	否	是	是

在深度学习流行起来之前，常见的序列标注问题的解决方案都是借助 HMM（Hidden Markov Model，隐马尔可夫）模型、最大熵模型、CRF 模型等，尤其是 CRF，它是解决序列标注问题的主流方法。故本文选择在 RoBERTa 模型后接上 CRF 层，由 RoBERTa 模型输出句子中每个字和符号到对应的实体标签的规律。CRF 层会采用 BIO 的标注方式，学习相邻实体标签之间的转移规则从而预测标签序列。

（5）RoBERTa-CRF 模型的处理流程

图 2 给出了 RoBERTa-CRF 模型的整体结构，利用 RoBERTa-CRF 模型对《左传》战争句进行事件抽取的基本流程具体如下：先利用 RoBERTa 模型将标注语料转换为相应的向量化表示$\boldsymbol{E_i}(\boldsymbol{i}=\boldsymbol{1},\boldsymbol{2},\cdots,\boldsymbol{n})$，接着用多层 Transformer 对语料提取特征后生成特征向量$\boldsymbol{T_i}(\boldsymbol{i}=\boldsymbol{1},\boldsymbol{2},\cdots,\boldsymbol{n})$，最后引入 CRF 模型学习约束，根据约束规则分析事件论元之间的语义关系，预测论元的标签序列。

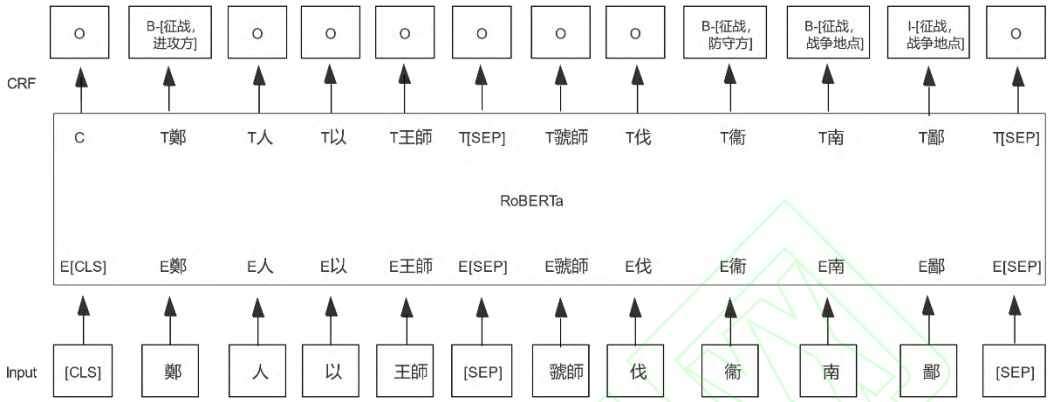

图 2　RoBERTa-CRF 模型的整体结构

Fig.2 Structure of RoBERTa-CRF Model

4　实验与结果分析

4.1　数据来源

本文的实验数据来自于邓勇的《春秋左传战争表》[17]，该表认为若交战双方休战后，隔一段时间再战，如果已记录战争发生的时间明确不同、或能够根据上下文推测战争发生在不同季节，则即使交战双方不变，该表仍然将其视作两次战争。经统计，该表共记录了 783 次战役。本文在此基础上定义了 3 种事件类型，22 个论元角色，一条语句有可能包含多个事件类型，标注后“征战”类型占比较多，有 696 条；“戕杀”类包含 23 条；“救援”类占 69 条。训练集、验证集、测试集按照 8:1:1 的比例进行划分。

4.2　数据预处理

根据 BERT、RoBERTa、guwenBERT 三种预训练模型支持的字形（具体可见表 2）以及初步随机测试结果，在 BERT、RoBERTa 两种模型上使用简繁体中文对训练结果没有产生较大差异，而在 guwenBERT 模型上，简体中文结果明显优于繁体中文。故本文在前两种模型上使用繁体中文，在 guwenBERT 模型中采用简体中文。

事件抽取包括两类任务：一类是事件识别，另一类是事件论元角色抽取，本文将第一类任务中的事件类型识别与第二类任务中的论元角色及论元抽取视作三元组的联合抽取任务。以语料“為報衛伐鄭，(冬)，鄭人以王師、虢師伐衛南鄙”为例进行事件标注，第一个论元“衛伐鄭”，它的事件类型是“征战”，论元角色对应为“战争原因”，标注结果如图 2 所示：

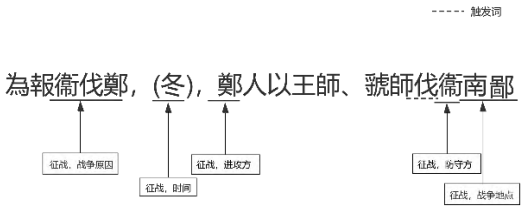

图 3　事件标注样例

Fig.3 Example of Event Annotation

4.3　测评指标及评价方法

本实验中，选用的实验测评指标主要包含精确率(Precision)、召回率(Recall)和 F1 值，其计算公式为：

$$精确率 = \frac{系统标注正确的论元数量}{系统标出的论元数量} \quad (1)$$

$$召回率 = \frac{系统标注正确的论元数量}{测试集中出现的论元数量} \quad (2)$$

$$F1\ 值 = \frac{2\times精确率\times召回率}{精确率+召回率} \quad (3)$$

本次事件抽取中“系统标注正确的论元数量”是指在事件类型和论元角色分类正确的情况下，标注正确的论元数量。

本文从模型、事件类型、论元角色层面共设计了三种评价方法：第一就模型角度考量，为了验证提出的 RoBERTa-CRF 模型的有效性，笔者挑选了谷歌的 BERT 模型和北京理工大学的 guwenBERT 模型作为预训练方法作为比较。又鉴于随着深度学习的发展，LSTM 在序列建模上表现出的强大性，它可以捕捉长远上下文，并且具备神经网络独有的拟合非线性特点。笔者因而增加了 BERT 和 LSTM 模型，共计设计了 5 组对比实验，分别是 guwenBERT-LSTM、BERT-LSTM、RoBERTa-LSTM、BERT-CRF、RoBERTa-CRF 模型；第二从事件类型角度，将在第一种模型评价结果的基础上，挑选出最优模型，考察其在三个事件类型上的指标效果；第三从论元角色出发，由于征战类的语料数量最为丰富，故将征战类的论元角色结合第一种评价方法中的最优模型进行测评，分析模型在具体语料预测上的优劣差异。

4.4　实验环境与参数设置

本文设计了 5 组对比实验：guwenBERT-LSTM、BERT-LSTM、RoBERTa-LSTM、BERT-CRF、RoBERTa-CRF 模型，其中前三种为 pytorch 框架下的管道抽取模型，对典籍抽取的方式都是先对事件类型进行识别，再进一步对论元角色抽取；后两种为 tensorflow 框架训练的联合抽取模型，同时抽取事件类型和论元角色。本实验使用了 GeForce RTX 2080 Ti 的 GPU 进行加速，中文 BERT 模型为谷歌提供的 bert-base-chinese，中文 RoBERTa 模型为哈工大讯飞研究室发布的 RoBERTa-wwm-ext，guwenBERT 模型为北京理工大学开发的 ethanyt/guwenbert-base。实验中使用的软件版本为 Python 3.7.6，Pytorch 1.3.1，Tensorflow-gpu 2.0.0，Keras 2.3.1，详细的参数设置如表 3 所示：

表 3　实验模型参数设置

Table3 Model Parameters Setting Details

参数名	参数值
序列长度(maxlen)	128
迭代次数(epochs)	45
每批训练大小(batch_size)	32
学习率(learning_rate)	0.00002
CRF 层的学习率(crf_lr_multiplier)	100

4.5　实验结果评估与分析

（1）不同模型的抽取效果评价

将本文的 RoBERTa-CRF 模型与 guwenBERT-LSTM 、BERT-LSTM、RoBERTa-LSTM、BERT-CRF 三种模型进行比较，五种模型分别记为 d、a、b、c、d，实验结果如表 4 所示：

表 4　不同模型的抽取性能比较

Table4 Comparison of Extraction Performance of Different Models

模型	精确率	召回率	F1 值
(a)guwenBERT - LSTM	68.3%	45.7%	54.7%
(b)BERT - LSTM	73.4%	64.6%	68.7%
(c)RoBERTa - LSTM	77.2%	66.2%	71.3%
(d)BERT - CRF	85.0%	74.9%	79.7%
(e)RoBERTa - CRF	87.6%	77.2%	82.1%

从表 4 可以看出，本文使用的模型 e 在《左传》战争句得到了相对不错的结果，笔者自三方面进行分析：

e 较之 d，c 较之 b、a 的 F1 值分别提升了 2.4%、2.6%、16.6%，说明 RoBERTa 语言模型相比于 BERT 和 guwenBERT 语言模型在典籍的事件抽取上效果有所提升。RoBERTa 预处理模型取消了下一句预测的任务，延长训练步数，用中文维基百科进行训练，预训练阶段使用哈工大 LTP 分词，并创新性采用全词遮掩的方式，一系列的改进措施促使 RoBERTa 语言模型又被誉为“一个调参达到最优的 BERT 模型”，因此它相较 BERT 能取得更优异的结果；而 guwenBERT 语言模型虽然采用了古文语料进行训练，但受殆知阁文献内容限制，部分语料未经分句、没有标点符号，尤其是史藏中的传记、正史等最能反映本文《左传》战争句语句特点的文本，该类文本由于缺少标点训练时将产生大量歧义，因此目前虽然存在古文的预训练模型，然而依旧由于文本数据不够成熟，其训练效果依旧不尽如人意。

d、e 比 b、c 在精确率、召回率、F1 值 3 个指标上都有明显的提升，说明 CRF 模型比 LSTM 模型在《左传》战争句数据集上表现优异。LSTM 模型在序列建模上非常强大，可以捕捉长远的上下文信息，除此之外 LSTM 还具备神经网络拟合非线性的能力，然而 LSTM 模型的缺陷是无法对不同时刻的输出产生

约束，假设这些输出之间存在较强的依赖关系，LSTM 模型的性能将因此受到限制；而 CRF 模型虽然不能像 LSTM 模型考虑上下文信息，它更多是通过特征模板去扫描句子局部特征的线性加权特征，同时 CRF 模型计算的是联合概率，优化的最终目标是整个序列。总体而言，在数据规模较小时，CRF 的效果优于 LSTM；从场景而言，如果需识别的任务不需要太依赖长久的信息，此时 CRF 模型也将更适合。而《左传》战争句满足上述两个条件，一是本实验的数据集较小，仅 783 条语料；其二是每一条语料由于是古文，长度仅在 20 至 50 字之前，属于短句，没有较长的上下文信息，因而 CRF 模型在典籍的事件抽取表现更优。

由上一指标比较结果，参考 4.4 节实验环境里曾经提及的“a、b、c 为管道抽取模型，d、e 为联合抽取模型”不难想到造成该指标结果的第二个原因，即管道模型常会带有的级联错误，管道模型将事件抽取分为事件类型识别和论元角色抽取两个子任务进行，第二个子任务的开始是在第一个子任务结束后，因而事件类型识别的效果很容易影响到论元角色的分类。相比之下，联合抽取由于是同时进行两个子任务，在本实验中是共同抽取“事件类型”、“论元”、“论元角色”，因而可以避免管道种常见的级联问题。

（2）不同事件类型的抽取效果评价

由于 RoBERTa-CRF 模型在 4.5.1 节中表现最优，因此笔者选用该模型分别抽取“征战”、“戕杀”、“救援”三个事件类型，比较其实验结果，其结果如表 5 所示：

表 5　不同事件类型的论元抽取性能比较

Table5 Comparison of Argument Extraction Performance of Different Event Types

事件类型	精确率	召回率	F1 值
战争-征战	87.1%	76.9%	81.7%
战争-戕杀	80.0%	50.0%	61.5%
战争-救援	96.6%	93.3%	94.9%

由表 5 可以得到，三类事件中 F1 值从高到低分别是：救援、征战、戕杀。在三个事件类型中，救援类的 F1 值高达 94.9%，通过分析救援类语料发现论元分布集中在触发词两侧，例如“(秋)，齊宋魯救鄭，楚師夜遁”，触发词为“救”，触发词的左右两端即为“援军（齊宋魯）”和“被救方（鄭）”，易于 CRF 模型采取就近原则识别附近的论元；与此相对的，戕杀类的 F1 值最低，仅为 61.5%，分析发现触发词后会出现用代词指代“受害人”的情况，如句子“(春)，魯南遺使國人助豎牛，攻仲壬於大庫之庭，殺之”，该句中触发词为句末的“殺”，“殺”后紧跟的“之”为代词，回溯整句可以发现“进攻方”为“豎牛”，受害人是“仲壬”，都位于句中位置，对于该类复杂句，没有大量语料作为基础训练，将很难提升模型的 F1 值；而征战类由于占比较大，拥有 696 条语料，虽然句型复杂，但学习到的特征却比占比最小的戕杀类语料全面，故征战类最终的 F1 值为 81.7%，在三个事件类型中更接近上一小节中模型 e 的表现。

（3）不同论元角色的抽取效果评价

在上一节的基础上，笔者进一步对征战类的 8 种论元角色进行抽取比较，实验参数不变，实验结果如表 6 所示：

表 6　不同论元角色的抽取性能比较

Table6 Comparison of Extraction Performance of Different Argument Roles

事件类型	论元角色	精确率	召回率	F1 值
战争-征战	时间	98.5%	100.0%	99.3%
	进攻方	88.3%	75.7%	81.5%
	防守方	92.9%	76.5%	83.9%
	战争原因	94.4%	73.9%	82.9%
	战争地点	88.0%	84.6%	86.3%
	战利品	71.4%	55.5%	62.5%
	助战方	66.7%	40.0%	50.0%
	参与人物	25.0%	20.0%	22.2%

28 个论元角色中，“时间”F1 值遥遥领先，其次 “战争地点”、“防守方”、“战争原因”、“进攻方”4 种论元角色的 F1 值都位于 80%以上，而 “参与人物”最低，F1 值仅有 22.2%。

首先，论元角色“时间”的 F1 值高达 99.3%，由于战争句中提及到季节、月份的词语一般都在首句，如“夏，齊侯伐魯北鄙，圍成”，开头第一个字“夏”就标明了战争发生的季节；“秋八月，晉與鄭衛戰于鐵，鄭師敗績”中的“秋八月”注明了战争的月份。

而“战争地点”、“防守方”、“战争原因”、“进攻方”4 种论元角色 F1 值相对也较高，原因主要因为涉及到的论元都分布在触发词前后不远的位置上，出现指代词作干扰的概率较低。例如“秦為報令狐之役，冬，秦伯伐晉，取羈馬，戰于河曲，交綏”，触发词为“伐”，触发词左端出现“令狐之役”、“秦”分别对应“战争原因”、“进攻方”，右端的“晉”、“羈馬”、“河曲，交綏”对应“防守方”、“战利品”、“战争地点”，古文语言精练、大多数句子较为短小，因此 RoBERTa-CRF 模型在这 4 种论元角色上适应良好。

5 结语

本文利用深度模型算法实现了典籍事件抽取的研究，探讨了典籍事件类型及论元角色的构建方法，针对典籍事件抽取中存在数据集不大、古文句长较短、论元分布集中等问题，提出了 RoBERTa-CRF 模型以提取战争句中事件类型、论元角色和论元，该模型用哈工大 LTP 作为分词工具，对组成同一个词的汉字全部进行遮掩，最后利用 CRF 模型完成序列预测。RoBERTa 模型虽然是在现代汉语的基础上训练，运用到古文语料里存在迁移偏差，但经过比较用殆知阁古籍文献作为训练依据的 guwenBERT 模型后，笔者认为就目前而言，RoBERTa 模型依旧是古文训练的首选。通过实验验证，在 guwenBERT-LSTM、BERT-LSTM、RoBERTa-LSTM、BERT-CRF、RoBERTa-CRF 五种模型中，数据集最终在 RoBERTa-CRF 模型上的有效性和可操作性最好。

但由于现阶段使用的数据集规模较小，无法使主题类别更均衡化，实现模型的高准确率。对事件进行论元角色标注选取的角色数量较多，部分角色对应的语料却相对较少，也影响了模型的准确率。因此，在尽可能多的提取事件内容的前提下，扩展标注数据集和权衡论元角色以提高模型的准确率是下一步的研究方向。

参考文献

[1] 夏翠娟.面向人文研究的“数据基础设施”建设——试论图书馆学对数字人文的方法论贡献[J].中国图书馆学报,2020,46(03):24-37. (Xia Cuijuan. The Construction of “Data Infrastructure” for Humanities Research: The Methodological Contribution of Library Science to Digital Humanities[J]. Journal of Library Science in China, 2020,46(03):24-37.)

[2] 李章超,李忠凯,何琳.《左传》战争事件抽取技术研究[J].图书情报工作,2020,64(07):20-29.(Li Zhangchao, Li Zhongkai, He Lin. Study on the Extraction Method of War Events in Zuozhuan[J]. Library and Information Service,2020,64(07):20-29.)

[3] 陈佩辉.人文数据库建设中人文学者何为——以《全宋文》墓志铭亲属信息提取为例[J].图书馆论坛,2019,39(05):17-23.(Chen Peihui. What Humanities Scholars Can Do in the Construction of Humanities Databases——Taking the Extraction of Kinship Data from Epitaphs in Quansongwen for Example[J]. Library Forum,2019,39(05):17-23.)

[4] 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.(Liu Zhongbao, Dang Jianfei, Zhang Zhijian. Research on Automatic Extraction of Historical Events and Construction of Event Graph Based on Historical Records[J]. Library and Information Service, 2020,64(11):116-124.)

[5] Ellen Riloff. Automatically constructing a dictionary for information extraction tasks[C]. Proceedings of the Eleventh National Conference on Artificial Intelligence. 1993:811–816.

[6] Cohen K B, Verspoor K, Johnson H L, et al. High-precision biological event extraction with a concept recognizer[C]. Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task. Association for Computational Linguistics, 2009:50-58.

[7] Arendarenko E, Kakkonen T. Ontology-Based Information and Event Extraction for Business Intelligence[C]. International Conference on Artificial Intelligence: Methodology. Springer Berlin Heidelberg, 2012:89-102.

[8] 陈慧炜. 刑事案件文本信息抽取研究[D]. 南京师范大学,2011.(Chen Huiwei. Research on text information extraction of criminal cases[D].Nanjing Normal University,2011.)

[9] 赵文娟,刘忠宝,王永芳.基于句法依存分析的事件角色填充研究[J].情报科学,2017,35(07):65-69.(Zhao Wenjuan, Liu Zhongbao, Wang Yongfang. Research on Event Role Annotation Based on Syntactic Dependency Analysis[J]. Information Science,2017,35(07):65-69.)

[10] Chen Y, Xu L, Liu K, et al. Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks[C]. The 53rd Annual Meeting of the Association for Computational Linguistics(ACL2015), 2015:167-176.

[11] Sha L，Qian F，Chang B，Sui Z F. Jointly Extracting Event Triggers and Arguments by Dependency-Bridge RNN and Tensor-Based Argument Interaction[C]. Association for the Advancement of Artificial Intelligence，2018:5916-5923.

[12] Duan S, He R, Zhao W. Exploiting document level information to improve event detection via recurrent neural networks[C]. Proceedings of the Eighth International Joint Conference on Natural Language Processing, 2017:352–361.

[13] 阮元.十三经注疏[M].中华书局,1980.(Ruan Yuan. The Confuscian Bible[M]. China Publishing House,1980.)

[14] 李学勤.春秋左传正义[M].北京大学出版社,1999.(Li Xueqin. The Standard of Chunqiu Zuozhuan[M]. Peking University Press,1999.)

[15] 朱宝庆.左氏兵法[M].陕西人民出版社,1991.(Zhu Baoqing. Zuo's art of war[M]. Shanxi People's Publishing House,1991.)

[16] 中国军事史编写组.中国历代战争年表[M].解放军出版社,2003.(Compilation group of Chinese Military

History. Chronology of Chinese Wars[M]. People's Liberation Army Press,2003.)
[17] 邓勇(邓曦泽).王霸：正义与秩序——从春秋战争到普遍正义[D].武汉大学,2007:270-295.(Deng Yong. Wang-Ba:Justice and Order——From Wars in Spring-Autumn Period To Universal Justice[D].Wuhan University,2007:270-295.)
[18] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. 2018.
[19] Cui Y M, Che W X, Liu T, et al. Pre-Training with Whole Word Masking for Chinese BERT[J]. The 2020 Conference on Empirical Methods in Natural Language Processing, 2020.
[20]阎覃. GuwenBERT:古文预训练语言模型（古文 BERT）[EB/OL]. [2020-11-22]. https://github.com/Ethan-yt/guwenbert.(Yan Tan. GuwenBERT:a Pre-trained Language Model for Classical Chinese (Literary Chinese) [EB/OL]. [2020-11-22]. https://github.com/Ethan-yt/guwenbert.)

(通讯作者:何琳，ORCID:0000-0002-4207-3588，E-mail:helin@njau.edu.cn。）

作者简介：

喻雪寒：女，南京农业大学信息管理学院硕士研究生，研究方向为文本挖掘。
何琳：女，南京农业大学信息管理学院教授，博士生导师，主要研究方向为信息组织与信息检索。
徐健：男，南京农业大学信息管理学院博士后，主要研究方向为文本挖掘。

作者贡献声明：

喻雪寒：负责进行实验，论文起草与最终版本修订；
何琳：提出研究思路，设计研究方案，修改论文；
徐健：分析数据，提供论文修改意见。

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：

支撑数据由作者自存储，E-mail:2019114010@njau.edu.cn。
[1] 喻雪寒.左传战争数据.xlsx.
[2] 喻雪寒.预训练-LSTM.rar.管道抽取的代码和模型训练文件.
[3] 喻雪寒.预训练-CRF.rar.联合抽取的代码，模型、事件类型、论元角色的训练文件.
[4] 喻雪寒.实验结果.rar.模型、事件类型、论元角色的训练结果截图.