CCL24-Eval 任务5系统报告：基于大小模型结合与半监督自训练方法的古文事件抽取

付薇薇，王士权，方瑞玉，李孟祥，何忠江，李永翔，宋双永
中国电信人工智能研究院
{fuweiwei, wangsq23, fangry, hezj, liyx25, songshy}@chinatelecom.cn
limengx@126.com

摘要

本文描述了队伍“TeleAI”在CCL2024古文历史事件类型抽取评测任务（CHED2024）中提交的参赛系统。该任务旨在自动识别出古代文本中的事件触发词与事件类型，其中事件类型判别被分为粗粒度和细粒度的事件类型判别两部分。为了提高古文历史事件类型抽取的性能，我们结合了大模型和小模型，并采用了半监督自训练的方法。在最终的评估中，我们在触发词识别任务得分0.763，粗粒度事件类型判别任务得分0.842，细粒度事件类型判别任务得分0.779，综合得分0.791，在所有单项任务和综合评分上均排名第一。

关键词： 事件抽取；半监督；自训练

System Report for CCL24-Eval Task 5: Ancient Chinese Text Event Extraction Based on Semi-Supervised Self-Training Method Combining Large and Small Models

Weiwei Fu, Shiquan Wang, Ruiyu Fang, Mengxiang Li, Zhongjiang He, Yongxiang Li , Shuangyong Song
Institute of Artificial Intelligence (TeleAI), China Telecom Corp Ltd
{fuweiwei, wangsq23, fangry, hezj, liyx25, songshy}@chinatelecom.cn
limengx@126.com

Abstract

This article describes the system submitted by the team ”TeleAI” for the CCL2024 Ancient Text Historical Event Type Extraction Evaluation Task (CHED2024). The task aims to automatically identify event trigger words and event types in ancient texts, with event type classification divided into coarse-grained and fine-grained parts. To improve the performance of ancient text historical event type extraction, we combined large and small models and adopted a semi-supervised self-training method. In the final evaluation, we scored 0.763 in the trigger word recognition task, 0.842 in the coarse-grained event type classification task, and 0.779 in the fine-grained event type classification task, with an overall score of 0.791. We ranked first in all individual tasks and the overall score.

Keywords: Event Extraction , Semi-Supervised , Self-Training

1 引言

古文事件抽取是从自然语言文本中识别和提取相关事件信息的过程，这是正确分析古汉语文本、进一步提升事件抽取技术及中国古代历史文本的数字化研究水平的重要步骤。然而，古代事件抽取任务缺乏公开的用于模型训练和评测的数据资源，制约了技术的进一步发展。除此之外，针对古代汉语的信息抽取任务还面临着古文句法语义复杂，使用范围小的复杂挑战。

为了进一步推动古代文本事件抽取任务的研究，研究者依托中国古文历史事件检测数据集（CHED2024）推出古文历史事件类型抽取评测任务，该评测任务共分为三个子任务，分别是触发词识别，识别出文本中触发词及其位置；粗粒度事件类型判别，识别出触发词所属的粗粒度事件类型；细粒度事件类型判别，识别出触发词所属的细粒度事件类型。具体如图1所示。

在本文中，我们使用了一种结合大小模型的半监督自训练方法，用于提升古文事件抽取任务的性能。具体来说，首先，我们利用比赛提供的CHED2024数据得到初始的模型；其次利用该模型在同领域的二十四史未标注古文数据上生成伪标签，利用标签一致性筛选得到的高质量的伪标签数据；最后利用伪标签数据与CHED2024数据进行训练得到最优结果。值得注意的是，在古文事件抽取任务中，由于古文事件抽取任务数据中存在标签不均衡的现象，我们利用大模型强大的泛化能力提高模型在稀缺标签数据上的准确率，弥补小模型在该类数据上的性能损失。

我们的系统在最终线上评测中触发词识别任务得分0.763，粗粒度事件类型判别任务得分0.842，细粒度事件类型判别任务得分0.779，综合得分0.791，排名第一。

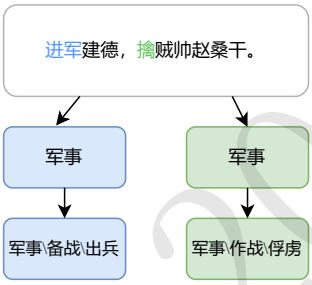

Figure 1: CHED2024任务示例

2 相关工作

事件抽取的主要目标是从文本数据中获取事件信息，事件作为一种特定的信息形式，是指某一特定时间、某一特定地点发生的某一特定事件，涉及一个或多个参与者，通常可以用状态的变化来描述(Doddington et al., 2004)。事件抽取任务旨在将此类事件信息从非结构化的纯文本中提取成结构化的形式，这种形式主要描述现实世界中发生的事件的信息。事件抽取可以作为许多任务的前提，比如信息检索、推荐系统、智能问答、知识图谱构建等，这些任务都依赖于事件抽取任务从文本中提取出的结构化事件信息。事件抽取的主要方法可以分为四种，分别是基于分类任务的事件抽取方法、基于阅读理解任务的事件抽取方法、基于序列标注任务的事件抽取方法和基于序列到结构生成任务的事件抽取方法。

基于分类任务的事件抽取方法通常将事件抽取任务转化为多类别分类问题，该方法将每个文本片段或句子分类为预定义的事件类型(Mekala and Shang, 2020; Guo et al., 2021)。这要求预先定义一组事件类别，并为每个类别构建相应的训练样本。在基于分类的任务中，触发器识别是对一个词是否为触发器进行分类，在确定事件类型后，再将句子中的实体分类为预定义的论元角色。该方法的特点是简单直观，易于实现和解释。它适用于事件类型较少且确定的情况，且对每个事件类型的分类效果较好。然而，它忽略了事件内部结构的细节。

基于机器阅读理解任务的事件抽取方法通过利用机器阅读理解模型来从文本中抽取事件信息(Guo et al., 2020)。该方法要求模型能够理解文本的上下文，定位事件的触发词、事件论元和其他关键信息，并将其结构化表示。在基于机器阅读理解的时间抽取任务中，模型首先需要确定文本所属的事件类型，然后根据事件类型确定要提取的事件论元。首先根据触发词分类确认文本的事件类型，然后为每个论元设计一个问题模式。最后，基于机器阅读理解的事件提取

方法将设计好的问题模式逐个应用于提取模型，根据输入文本获取答案，每个答案对应一个事件论元。基于机器阅读理解任务的事件抽取方法具有捕捉事件上下文信息的优势，能够提供更丰富的事件抽取结果(Chen et al., 2021)。然而，该方法对模型的理解和推理能力要求较高，需要模型具备对文本上下文进行推理和抽象的能力。

基于序列标注任务的事件抽取方法将文本中的每个词标注为事件的不同组成部分，如触发词和论元，常用的序列标注方法包括命名实体识别和关系抽取(Gui et al., 2020)。序列标注任务是基于词级别的多分类任务，旨在基于词级别事件类型进行直接匹配和提取事件论元。它主要涉及两个核心任务：事件类别的识别和分类，以及事件论元的提取。基于序列标注的事件抽取方法简单、快速，能够实现事件类型与事件参数的匹配，而无需使用额外的特征。在基于序列标注的任务中，触发词识别是将文本中一个单词标记为触发词的过程。通过序列标注方法，可以将目标事件从文本中标注出来，使其适用于事件提取任务。对于给定的文本事件类型，通过序列标记模型对事件论元进行标记。序列标注模型输出的序列对文本中的所有单词进行了标注。虽然基于序列标注任务的事件抽取方法能够对文本进行细粒度的标注，捕捉事件的具体信息，但它需要高质量的标注数据和对序列建模的技术支持。

基于序列到结构生成的事件提取方法采用端到端的方式从文本中提取事件信息，它将所有任务统一建模为一个模型，并直接生成事件的触发词和事件论元，这种方法通常使用编码器-解码器结构(Lu et al., 2021)。在基于序列到结构生成的任务中，模型通过统一建模和预测不同的标签，可以实现对事件抽取的端到端处理。这种方法的特点是直接生成事件抽取结果，无需预定义的类别或标签。这种方法能够捕捉文本中复杂的事件结构和上下文信息，并提供结构化的事件抽取结果。然而，基于端到端生成任务的事件抽取方法对大量的训练数据和计算资源要求较高。

在本文中，我们结合大语言模型(Wang et al., 2024)与小语言模型(Liu et al., 2023)来完成评测任务，具体来说，在小模型上我们使用GRTE(Ren et al., 2021)，该方法通过一个迭代的模型来提升对全局特征(Global Feature)的学习，进而提高模型性能；在大模型上我们使用XunziALLM，该模型经过古文领域的持续训练在古文任务上具有较好的表现，可以准确剖析古籍文本的复杂性，进一步提高模型在古文事件抽取任务上的效果。

3 方法

在本文中，我们采取了一种半监督自训练学习的方法，通过融合大小模型的优势，显著提升了古文事件抽取的性能。半监督自训练学习方法如算法1所示，下面我们将在各个小节详细介绍每个模块的具体细节。

Algorithm 1: 半监督自训练学习方法

Input: 有标签数据D_l, 无标签数据D_u, 初始模型M
Output: 训练好的模型M'

用有标签数据D_l 初始化模型M;
$D'_u \leftarrow D_u$;
while 未达到停止准则 **do**
  在$D_l \cup D'_u$ 上训练模型M;
  使用模型M 为无标签数据D'_u 生成伪标签;
  基于一致性检验筛选高质量的伪标签数据;
  $D'_u \leftarrow$ 选取的高质量伪标签数据;
在$D_l \cup D'_u$ 上训练最终模型M';
return M';

3.1 模型初始化

在古文事件抽取任务中，由于古文事件抽取任务数据中存在标签不均衡的现象，我们利用大模型强大的泛化能力提高稀缺样本的准确率，弥补小模型在该类数据上的性能损失。具体来说，我们使用CHED2024训练集初始化大模型与小模型，其中小模型部分经过对比我们选择在古文事件抽取中表现最好的GuwenBERT作为基座模型，我们将事件抽取任务转换为关系抽取任务，使用GRTE方法通过对全局特征的学习提高从古文文本中抽取事件触发词与细粒度事件

类型的准确率。大模型部分我们选择经过古籍数据微调的Xunzi-Baichuan-7B作为基座模型，构造直接从原始文本中获取事件触发词与细粒度事件类型的指令。

3.2 伪标签生成

为了获取更多与古文事件抽取相关的高质量训练数据，我们首先从网络上获取了开源的《二十四史》原文，其次通过规则方式筛选出17w无标签数据，再次利用初始化的大小模型对无标签数据进行预测，最后利用标签一致性检验策略从中过滤筛选出4w条高质量伪标签数据。

标签一致性检验主要通过比较大模型和小模型对于同一条数据预测标签是否一致来判断该标签是否可靠，同时为了缓解训练数据中类别不均衡问题，我们提高了大模型预测结果中稀有类别标签数据的采纳比例。通过一致性检验可以筛选出高质量的伪标签，减少低质量标签对模型训练的负面影响，利用多次预测的标签一致性能够捕捉更可靠的样本信息，从而提升模型的泛化能力，通过有效过滤掉噪声数据和错误标签可以确保模型训练数据的纯净度。

3.3 训练策略

为了更好的利用数量较少的高质量CHED2024数据，我们在训练过程中首先使用伪标签数据作为训练集，CHED2024数据作为验证集，当模型在伪标签数据上达到稳定状态时，切换到高质量的CHED2024数据。先使用伪标签数据再使用真实标签数据训练策略可以有效提高模型性能。先使用伪标签数据进行训练，有助于提高模型的泛化能力和对数据的理解；后用真实标签数据进行训练可以使模型更快地收敛并获得更好的性能。同时这种方法还能够降低对有标签数据的依赖，可以最大程度地利用有限的有标签数据资源。

除此之外，我们还在伪标签数据与CHED2024数据中通过上下文增强策略提高模型对于稀缺样本预测的准确率，具体来说，对于稀缺样本，我们从训练集中将其与前后若干条数据进行拼接，然后将其加入训练数据。

4 实验

为了提高古文事件抽取任务上模型的性能表现，我们针对CHED2024数据集进行了若干实验，值得注意的是，本小节实验结果均基于CHED2024验证集中细粒度事件类型判别任务计算得到。

4.1 基座模型对比

为了寻找到最适合处理古籍领域文本的基座模型，我们首先在CHED2024数据集上分别尝试了若干开源预训练语言模型，实验结果如表1示。

Model	BERT	RoBERTa-base	SIKU-BERT-base	SIKU-RoBERTa-base	BTfhBER	Guwen-NER-RoBERTa-base
CHED2024_dev	0.7032	0.7287	0.7735	0.798	0.793	**0.828**

Table 1: CHED2024数据集上开源预训练语言模型的实验结果

表1展示了在CHED2024验证集上在GRTE结构下，不同开源预训练语言模型的性能表现，其中的F1得分为该开源预训练语言模型的最高Micro-f1得分。从实验结果可以看出Guwen-NER-RoBERTa-base在CHED2024数据集上表现最佳，因此在本次评测中我们选择该模型作为后续实验的基座模型。

4.2 GRTE VS GPLINKER

我们在CHED2024数据集对GRTE方法与GPLINKER方法上进行对比，利用5折交叉训练的方式在训练集上进行训练，在验证集上的的实验结果如表2所示：

我们将事件抽取的任务改为关系抽取的任务形式，表2展示了GRTE方法与GPLINKER方法在CHED2024数据集上微调后的实验结果，实验结果表明基于GRTE方法的模型性能优于GPLINKER方法，在Micro-f1上获得了4.6%的相对提升，我们在后续的实验中选用GRTE作为抽取方法。

Method	GRTE	GPLINKER
Fold-0-Micro-f1	0.765	0.768
Fold-1-Micro-f1	0.789	0.774
Fold-2-Micro-f1	0.824	0.704
Fold-3-Micro-f1	0.847	0.827
Fold-4-Micro-f1	0.802	0.798
Average	0.828	0.791

Table 2: CHED2024数据集上GRTE与GPLINKER的实验结果

4.3 CHED2024实验结果

为了利用大量的无标注数据，我们基于半监督自训练的方式获取了高质量的伪标签数据，并通过上下文增强的方式提高稀缺样本的准确率，为了证明伪标签数据与上下文增强策略的有效性，我们基于伪标签数据与CHED2024数据进行了实验，实验结果如表3所示。

Metric	Micro-f1	Macro-f1
CHED2024_dev	0.828	0.723
CHED2024_dev+Context	0.831	0.752
CHED2024_dev+Pseudo	0.904	0.768

Table 3: CHED2024验证集上模型微调的实验结果

表3展示了CHED2024验证集上Guwen-NER-RoBERTa-base+GRTE方法的实验结果，其中CHED2024_dev表示只使用CHED2024训练集与验证集，CHED2024_dev+Context表示在CHED2024数据集上利用稀缺样本的上下文对其进行数据增强，CHED2024_dev+Pseudo表示使用CHED2024与通过半监督自训练方法获取的高质量伪标签数据。基于上下文对数据集稀缺样本进行增强，具体来说，对于稀缺样本，我们从训练集中将其与前后若干条数据进行拼接，然后将其加入训练数据。实验结果表明加入上下文增强后的稀缺样本数据可以在Micro-f1上获得0.3%的相对提升，在Macro-f1上获得4.0%的相对提升。我们还利用半监督自训练的方式从大量无标注数据中获取高质量的伪标签数据提高模型最终的性能。在对于伪标签数据的利用上，我们先利用伪标签数据作为训练集，CHED2024作为验证集，当模型在伪标签数据上达到稳定状态时，切换到CHED2024数据集。实验结果表明该方法在验证集上Micro-f1获得了9.2%的相对提升，在Macro-f1上获得了5.2%的相对提升。

Task	**Rank**	**Exact**	**Subset**	**Score**
Trigger Word Recognition	1	0.636	0.889	0.763
Task	**Rank**	**Macro-f1**	**Micro-f1**	**Score**
Coarse-grained Event Type Classification	1	0.839	0.845	0.842
Fine-grained Event Type Classification	1	0.743	0.814	0.779

Table 4: CHED2024测试集上模型的实验结果

表4展示了我们队伍在最终测试集上的实验结果，其中触发词识别任务得分为0.763，排名第一；粗粒度事件类型判别任务得分为0.842，排名第一；细粒度事件类型判别任务得分为0.779，排名第一，综合得分0.791，排名第一，具体计算过程如公式1所示:

$$\text{Total\_Score} = \text{Task\_1} \times 0.4 + \text{Task\_2} \times 0.3 + \text{Task\_3} \times 0.3 \quad (1)$$

5 总结与展望

在本次CCL2024古文事件抽取评测中，我们利用大小模型结合与半监督自训练的方式从开源的无标注数据中获取到高质量的伪标签数据，然后依靠CHED2024伪标签数据提高模型在古文事件抽取上的表现，我们“TeleAI”队伍提交的系统在三项任务中均取得第一名的成绩，综合得分为0.791，排名第一。然而，本次比赛由于时间因素我们未对伪标签数据进行多轮迭代，未来可以通过多轮迭代的方式持续提高伪标签数据的质量，进一步提高模型性能。

参考文献

Shaowei Chen, Yu Wang, Jie Liu, and Yuelin Wang. 2021. Bidirectional machine reading comprehension for aspect sentiment triplet extraction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12666–12674.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Lrec*, volume 2, pages 837–840. Lisbon.

Tao Gui, Jiacheng Ye, Qi Zhang, Zhengyan Li, Zichu Fei, Yeyun Gong, and Xuan-Jing Huang. 2020. Uncertainty-aware label refinement for sequence labeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2316–2326.

Shaoru Guo, Ru Li, Hongye Tan, Xiaoli Li, Yong Guan, Hongyan Zhao, and Yueping Zhang. 2020. A frame-based sentence representation for machine reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 891–896.

Biyang Guo, Songqiao Han, Xiao Han, Hailiang Huang, and Ting Lu. 2021. Label confusion learning to enhance text classification models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12929–12936.

Shixuan Liu, Chen Peng, Chao Wang, Xiangyan Chen, and Shuangyong Song. 2023. icsberts: Optimizing pre-trained language models in intelligent customer service. *Procedia Computer Science*, 222:127–136.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2event: Controllable sequence-to-structure generation for end-to-end event extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2795–2806.

Dheeraj Mekala and Jingbo Shang. 2020. Contextualized weak supervision for text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 323–333.

Feiliang Ren, Longhui Zhang, Shujuan Yin, Xiaofeng Zhao, Shilei Liu, Bochao Li, and Yaduo Liu. 2021. A novel global feature-oriented relational triple extraction model based on table filling. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2646–2656.

Zihan Wang, Xinzhang Liu, Shixuan Liu, Yitong Yao, Yuyao Huang, Zhongjiang He, Xuelong Li, Yongxiang Li, Zhonghao Che, Zhaoxi Zhang, et al. 2024. Telechat technical report. *arXiv preprint arXiv:2401.03804*.