

基于语义增强的装备事件抽取方法

方睿¹ 崔良中¹ 方圆婧²

1 海军工程大学电子工程学院 武汉 430033

2 联勤保障部队 武汉 430000

(jsj_fr8836@163.com)

摘要 信息时代下,装备领域的数据量急剧增长,使得论证人员难以高效地从中获取关键信息,进而支持相应的数据分析和论证工作。针对装备领域事件抽取事件论元边界模糊的问题,提出了一种基于语义增强的装备事件抽取方法。该方法利用装备领域的专业术语和词汇信息,构建领域词向量,并设计能够兼容和整合不同粒度语义信息的模型结构,将装备领域词向量与预训练模型 ERNIE 生成的字符向量进行融合,将专业术语知识和通用语言理解能力相结合,实现更全面的语义信息捕捉,增强模型对装备领域文本语义的理解,从而提升模型对事件论元边界的识别能力。实验结果表明,该方法在装备领域数据集上取得了优于基线方法的 F1 值,相比 CK-BERT 模型 F1 值提升了 3.83%;在公开数据集 ACE2005 上进行的实验验证了其能有效提升装备领域事件要素抽取的性能。

关键词: 装备领域;事件抽取;语义增强;领域词向量;预训练模型

中图分类号 TP391

Equipment Event Extraction Method Based on Semantic Enhancement

FANG Rui¹, CUI Liangzhong¹ and FANG Yuanjing²

1 School of Electronic Engineering, Naval University of Engineering, Wuhan 430033, China

2 Joint Logistics Support Force, Wuhan 430000, China

Abstract In the information age, the volume of data in the equipment domain has surged dramatically, making it challenging for analysts to efficiently extract critical information to support relevant data analyses and arguments. To address the issue of ambiguous event argument boundaries in the extraction of events within the equipment sector, a semantic-enhanced event extraction method is proposed. This method utilizes specialized terminology and vocabulary information in the equipment domain to construct domain word vectors, and designs a model structure that can be compatible with and integrate semantic information of different granularity, fuses the equipment domain word vectors with character vectors generated by ERNIE of the pre-trained model, combines the knowledge of specialized terminology with the ability of general language comprehension, and realizes a more comprehensive capturing of semantic information that enhances the model's understanding of the textual semantics of the equipment domain, so as to improve the model's ability to recognize the boundaries of the event thesis elements. Experimental results demonstrate that, on the equipment domain dataset, the proposed method's F1 values outperforms baseline approaches, with an improvement of 3.83% compared to the CK-BERT model, and has been validated on the public dataset ACE2005, thereby effectively improving performance in the extraction of event elements in the equipment domain.

Keywords Equipment domain, Event extraction, Semantic enhancement, Domain-specific word vector, Pre-trained model

事件抽取作为信息抽取中的关键部分,旨在从文本中识别出描述事件的关键元素,如触发词、事件主体、事件客体等,它在信息检索^[1]、知识图谱构建^[2-3]和舆情分析^[4]等多个领域具有重要的应用价值。在装备领域中,尽管数据量庞大且富含价值,但由于缺乏有效的数据组织架构和直观的展示手段,这些数据难以被系统性地整合与呈现,阻碍了信息的充分利用和知识体系的形成,难以进行深层次的数据分析,也难以满足典型论证场景的应用需求。因此,探索装备数据中事件要素和类型的抽取愈发重要。

装备领域的特殊性在于其标注数据相对有限,而且文本

数据语境复杂,包含大量的专业术语和领域知识,上下文依赖性强,增加了事件抽取的难度。例如一词多义的问题,同一个词在不同的上下文中可能表示不同的事件类型或论元,如“导航”既可以作为名词导航装备,也可以理解为引导方向的动作;还有多词同义的问题,“苏-57”“苏-57 战斗机”和“Su-57”为同一型号战斗机的不同词语表述。若模型在处理时没有考虑到词汇在上下文中的特定信息,便会产生语义上的错误或缺失,影响事件抽取的准确性。

传统的事件抽取方法,如基于规则和模式匹配的方法,虽在特定领域和语言环境中表现良好,但其泛化能力有限,难以

基金项目:装备预先研究项目(30209040702)

This work was supported by the Equipment Advanced Research Project(30209040702).

通信作者:崔良中(szzll@163.com)

有效应对复杂多变的文本数据。现有的深度学习模型,尤其是神经网络架构,依赖于大规模、高质量的训练数据,导致基于数据驱动的方式在装备领域的应用中受到限制。随着BERT^[5],XLNET^[6],ERNIE^[7]等预训练语言模型的出现,基于深度学习的事件抽取方法取得了新的进展,为处理装备领域中复杂文本数据的理解与分析提供了更为有效的工具和方法^[8]。但预训练模型通常基于通用域语料进行训练,缺乏垂直领域的专业知识,因此模型对上下文的理解能力仍需进一步提升。

为此,研究者们尝试通过增强模型语义理解的方式进行改进。Hu等^[9]通过引入实体词典的方式,既缓解了小数据集上的性能问题,也减少了未登录词的影响;Zhao等^[10]采用将句子语义信息融入汉字上下文表示的方法,有效地增强了词典知识的应用,使得模型能够更加精细地感知和识别词语的边界;Zhuang等^[11]通过融合事件本体的语义信息来增强事件表示,并通过注意力机制增强关键知识的注入,有效缓解了知识缺失的问题。

上述研究通过引入实体词典以及融合句子语义信息和事件本体的语义信息,显著提升了模型的性能,由此说明了增强模型语义理解在信息抽取任务中的重要性。受此启发,本文针对装备领域文本语义复杂导致的论元边界模糊问题,提出了一种基于语义增强的装备事件抽取方法。该方法的核心是将领域词向量与预训练模型生成的字符向量进行融合,从而有效提升模型对复杂语境中事件论元边界的识别能力。其中,Soft-lexicon技术的应用,不仅增强了模型的灵活性和适应性,使其能够更好地捕捉装备领域文本中的专业术语和上下文信息,而且无需进行结构性改动,即可与现有的预训练模型等深度学习架构无缝集成,保持模型的高效性和可扩展性。

实验结果表明,本文提出的方法在装备领域事件抽取任务上取得了成效。与现有方法相比,本文方法在精确率、F1值等多个评估指标上均有提升。消融实验进一步验证了语义增强策略对性能提升的重要作用。

本文的主要贡献可以概括为以下3个方面。

(1)通过整合装备领域的专业术语和相关词汇信息,构建了装备领域词向量,增强了模型对专业术语的识别和理解能力,为后续的事件抽取任务奠定基础。

(2)设计了一种语义增强的事件抽取模型,该模型能够兼容并整合来自不同粒度的语义信息,实现领域词向量与预训练模型生成的字符向量的融合,提供更丰富的语义信息,帮助模型更好地捕捉装备领域文本的复杂语义关系,进而提升模型识别和抽取事件要素的准确性。

(3)在装备数据集和ACE2005数据集上进行实验评估,选取精确率(P)、召回率(R)、F1值(F1)作为评估指标,验证了模型的有效性。

1 相关工作

1.1 基于深度学习的方法

相比于传统的事件抽取方法,深度学习利用神经网络提升了模型的学习能力,可以更有效地利用大规模的数据;同时通过无监督学习的方式,可以更好地提取数据的原始特征,解决了机器学习中特征选择的难题,是目前主流的研究领域^[12]。

在卷积神经网络的应用上,Yu等^[13]利用知识-注意力和卷积神经网络(KA-CNN)进行多尺度事件因果关系提取。该模型设计了一种以单词嵌入作为输入,并以语料库的语义特征作为输出的多尺度CNN。实验结果表明,所提方法在提取多尺度事件因果关系方面取得了良好的效果。在循环神经网络方面,Wang等^[14]提出了一种将增强注意力机制集成到Tree-LSTM中的模型Child-Sum EATree-LSTM,其通过对相邻节点分配注意力权重,提高对事件触发词的检测。在MLEE和BioNLP数据集上的结果,展示了增强注意力机制在检测生物医学事件触发词方面的优势。在图神经网络的应用方面,Sun等^[15]提出了一种新的事件抽取模型Seq2EG,该模型将事件抽取转换为图解析问题,利用预训练的序列到序列模型将输入的句子转化为事件图,在ACE2005数据集上相比最佳基线模型F1值提升了3.4%。Guo等^[16]提出了DHG-EE模型,通过构建包含领域、事件类型和论元角色的多粒度异构图网络,将重叠事件的论元与对应的领域和事件类型节点分开,有效表示了重叠事件的复杂关联关系。

1.2 基于预训练模型的方法

预训练模型通过在大规模无标注文本上进行训练,学习到了丰富的语言知识,包括词汇的上下文表示、句法结构和语义关系等,可以作为事件抽取任务中特征提取的强有力工具。

Liu等^[17]提出了一种基于预训练语言模型BERT的会话式机器阅读理解框架,以解决以往机器阅读理解事件抽取模型存在的一些问题,如模型局限于单轮问答,缺乏问答之间的依赖关系,未充分利用句子中的实体信息等,从而提高了事件抽取的准确性和效率。Ding等^[18]针对BERT在中文事件抽取中面临的挑战,包括在没有自然分隔符的语言中触发词不匹配,以及句子中包含多个事件时固有的多角色噪声限制方法性能的问题,提出了一种基于掩码注意的转换器BERT(Mask-Attention-based BERT, MABERT)框架。MABERT能够更有效地处理中文语境下的触发词匹配问题,提升了在多事件句子中的事件提取准确性。

1.3 融合外部知识的语义增强方法

基于通用领域语料训练的大模型虽然包含丰富的常识知识,但缺乏垂直领域的专业知识,因此在处理特定领域的任务时表现出一定的局限性。为解决这一问题,研究者们开始探索将外部知识与预训练模型相结合的方法,以增强模型在特定领域的语义理解能力。

Pei等^[19]利用提示模版和微调等技术,通过少量数据构建了基于大语言模型的公安专业小样本知识抽取方法;Li等^[20]提出的基于Fin-BERT的金融事件抽取方法,通过在编码阶段融入领域先验知识,提高了模型在金融领域数据集上事件抽取的效果;PMC-LLaMA^[21]提出了一种针对生物医学文献的预训练语言模型,通过微调LLaMA模型并融入医疗领域知识,显著增强了模型在医疗领域的专业处理能力和理解能力;Li等^[22]提出了一种基于槽位语义增强的提示学习框架,该方法将论元角色的深层语义信息融入提示模版的槽位设计中,为模型在预测槽位内容时提供更为精准的论元类型约束,从而提升事件抽取的精确度;Yu等^[23]提出了一种结合预训练语言模型XLNET和图注意力网络GAT的句法信息增强型事件抽取模型,该模型利用XLNET进行文本的语义表示,并通过依存句法树中的依赖关系来增强信息流。

除此之外,ERNIE^[24],K-BERT^[25],CoLAKE^[26],KEP-LER^[27],JAKET^[28]和KG-Adapter^[29]等模型,都是在预训练模型中引入外部知识以增强其编码能力,实验结果表明,语义增强的方法能够有效提升下游任务的性能。

在装备事件抽取中,现有的深度学习模型,尤其是神经网络架构,依赖于大规模、高质量的训练数据。然而,装备领域数据的稀缺性,限制了模型通过大量数据学习复杂语义信息的能力。受到融合外部知识方法的启发,本文提出了一种基于语义增强的装备事件抽取方法。该方法将装备领域的专业知识和预训练模型的语义理解能力相结合,有效提升了模型对装备领域文本语义的理解能力,从而提高了事件抽取的准确性。

2 本文方法

2.1 任务描述

针对装备事件抽取任务,本文采用 BIO 的序列标注方法。BIO 标签方法是一种用于标注文本中实体或事件论元的常用方法,它将每个单词标记为 3 种状态。其中,“B-”表示该单词是某个实体或论元的开始;“I-”表示该单词是某个实体或论元的一部分,但不是开始;“O-”表示该单词不属于任何实体或论元。

本文方法首先进行触发词的抽取,通过识别文本中描述事件发生的核心动词或名词短语,为后续的事件论元识别奠定基础。随后,模型再根据触发词的上下文信息,识别与该事件相关的实体,如事件的主体、客体、时间、地点等,最终输出包含触发词和事件论元标签的序列。

2.2 整体框架

本文提出的模型框架整体分为 5 个部分,如图 1 所示。

输入层:模型的输入是一段文本序列。

嵌入层:由字符级别的嵌入表示和词汇级别的语义增强嵌入表示两部分组成。给定一个输入文本序列,通过 ERNIE 模型获得字符级别的嵌入表示,通过 Soft-lexicon 方法将词汇级别的语义信息融入到字符表示中,以增强模型对文本语义的捕捉能力。

融合层:将来自 ERNIE 的字向量嵌入和 Soft-lexicon 的词向量嵌入进行拼接,整合不同粒度的语义信息,为模型提供更丰富的输入表示。

编码层:通过门控循环单元提供对输入序列的双向上下文理解,捕捉长距离依赖关系。

CRF 层:CRF 层通过学习标签间的转移概率来增强标注的准确性,最终输出预测序列结果,对应相应的事件元素。

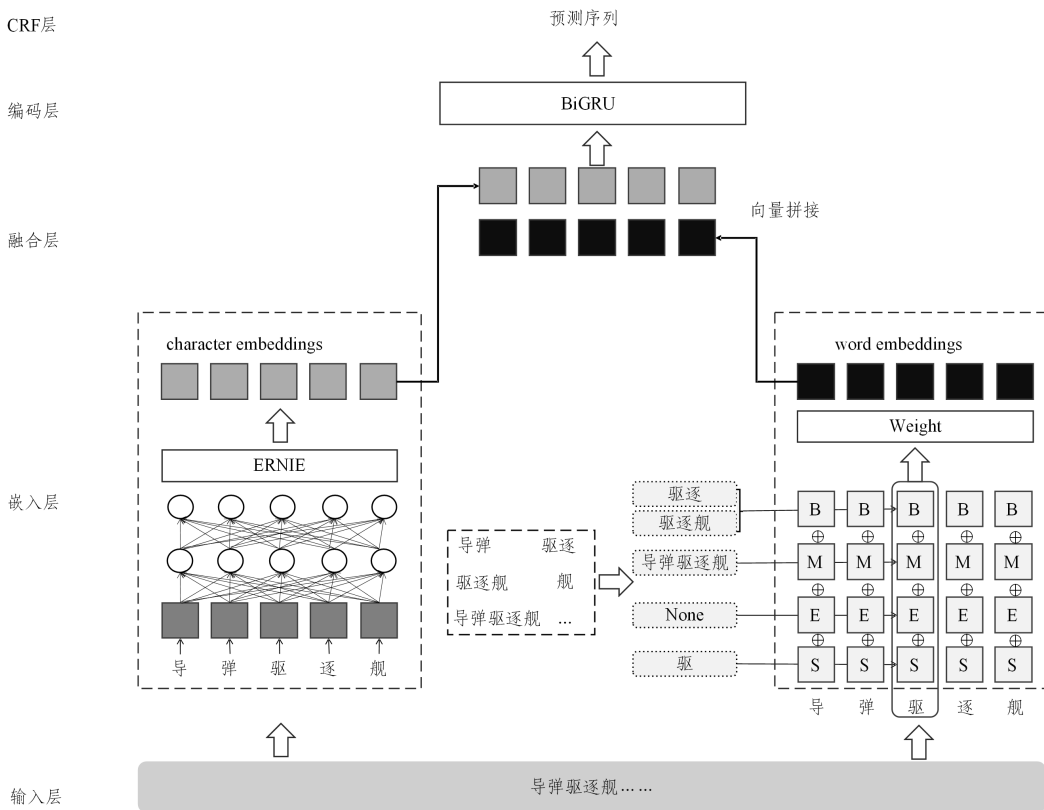


图 1 模型结构图

Fig. 1 Model structure

2.3 领域词向量构建

在进行装备领域词向量的构建时,选取 Word2vec 模型^[30],而没有选取 BERT 预训练模型进行词向量训练。原因在于 BERT 模型在训练词向量时,采用 WordPiece 分词方法,将词语分割成子词(subword)进行训练,子词无法完整地表达词语的语义信息,而 Word2vec 模型则以词汇为基本单元。鉴于装备领域涉及的专业术语具有特定的组合性,采用以词汇为单元的 Word2vec 模型更贴合本文的研究需求。同时,

装备领域词向量构建的目标是捕捉专业术语的语义特征,而 BERT 模型的预训练目标更侧重于语言的通用表示,需要针对下游任务进行微调,与特定任务的目标不完全匹配,因此选择 Word2vec 模型构建装备领域词向量。

通过利用公开的与装备数据相关的百度百科、新闻报道等数据,构建装备领域语料库;然后对语料库进行清洗和去噪,去除无关信息和停用词,并识别出装备领域的专业术语和常用词汇;再使用 Word2vec 模型对清洗后的语料库进行训

练,得到能够反映装备领域语义特征的词向量;最后,通过相似度评估等方法评估词向量的质量。

2.4 嵌入层

嵌入层由字符级别的嵌入表示和词汇级别的语义增强嵌入表示两部分组成。

给定一个输入文本序列,模型首先通过 ERNIE 模型获得字符级别的嵌入表示。ERNIE 模型采用多层 Transformer 架构,能够学习到字符层面的信息,并通过分析上下文的语义来深入理解知识。这种字符级别的嵌入表示为模型提供了基本的文本表示能力。

为了增强模型对装备领域专业术语的理解能力,模型通过 Soft-lexicon 方法将词汇级别的语义信息融入到字符表示中。Soft-lexicon 方法的核心在于将领域词汇信息以软向量的形式融入到网络编码中,提升模型对装备领域文本的语义理解能力,从而提高对装备领域事件抽取中论元边界的识别能力。

(1) ERNIE

ERNIE 模型是百度推出的一种预训练模型。与 BERT 主要基于字符级别的训练方式不同,ERNIE 采用了一种更为细化的 Mask 策略,通过设计字符级、实体级和短语级 3 个层级的 Mask 机制,不仅能够学习到字符层面的信息,还能够通过分析上下文的语义来深入理解知识。同时,ERNIE 的训练过程还融入了丰富的多源数据,显著增强了模型对语言的理解和表示能力,从而在处理中文自然语言处理任务时效果更佳。

ERNIE 采用多层 Transformer 架构作为其核心编码器。对于一个由 k 个字符组成的句子 $s = \{c_1, c_2, \dots, c_k\}$,模型会分别添加唯一的标记 [CLS] 和 [SEP] 来表示句子的开始和结束。句子中的每个字符通过字符嵌入、段嵌入和位置嵌入的相加被表示为嵌入向量,即 $E_{w_i} = E_{\text{token}} + E_{\text{seg}} + E_{\text{pos}}$ 。然后,句子被转换成向量序列 $\text{embedding} = \{E_{c_1}, E_{c_2}, \dots, E_{c_k}\}$,并输入到双向 Transformer 中以提取特征。Transformer 的自注意力机制捕获每个标记的上下文信息,最终生成富含语义特征的序列向量 $X = \{x_1, x_2, \dots, x_k\}$ 。

(2) Soft-lexicon

Soft-lexicon 是一种简单而有效的将词汇信息纳入字符表示的方法^[31],如图 2 所示。它无需设计复杂的序列建模结构,仅需对字符表示层进行细微调整,即可引入词典信息。在输入表示层,每个字符被映射到一个高维向量,并构造出软词典特征,进而将其与每个字符的表示相融合。这种方法有效弥补了仅基于字符信息的模型在利用词汇信息方面的不足,而且在保持较高推理速度的同时,性能也更好,能够更容易地与预训练模型相结合。具体过程如下:

对于输入文本 $s = (c_1, c_2, \dots, c_k)$,在处理每个字符 c_i 时,将其与词典中的词语进行匹配,并将匹配到词语的字符分类到相应的集合中。字符 c_i 构成的 4 个集合如式(1)所示。

$$\begin{cases} B(c_i) = \{w_{i,n}, \forall w_{i,n} \in L, i < n \leq k\} \\ M(c_i) = \{w_{m,n}, \forall w_{m,n} \in L, 1 \leq m < i < n \leq k\} \\ E(c_i) = \{w_{m,i}, \forall w_{m,i} \in L, 1 \leq m < i\} \\ S(c_i) = \{c_i, \exists c_i \in L\} \end{cases} \quad (1)$$

其中, L 表示词典,即构建的领域词向量文件中包含的词汇,

$w_{m,n}$ 表示以 c_m 开始,以 c_n 结束的词典中的词; $B(c_i)$ 代表以 c_i 开始的词语集合, $M(c_i)$ 代表 c_i 为中间字的词语集合, $E(c_i)$ 代表以 c_i 结尾的词语集合, $S(c_i)$ 代表只包含 c_i 单个字符的集合。如果在词典中没有找到符合 $\{B, M, E, S\}$ 结构的词语,则该集合为空集。

在得到字符词集后,通过统计词集中每个词的频率,将词集压缩成固定维度的向量,并直接拼接到每个字符中,如式(2)、式(3)所示。通过对输入文本序列的字向量和词向量融合,字符与其对应的词汇信息可以一一对应地融合。

$$v^s(S) = \frac{4}{Z} \sum_{w \in S} z(w) e^{w(w)} \quad (2)$$

$$\begin{cases} e^s(B, M, E, S) = \{v^s(B), v^s(M), v^s(E), v^s(S)\} \\ x^c \leftarrow \{x^c, e^s(B, M, E, S)\} \end{cases} \quad (3)$$

其中, S 代表一个单词集合, $v^s(S)$ 代表单词集合 S 的权重, $z(w)$ 代表单词 w 在统计词集中出现的频率, $e^{w(w)}$ 代表单词 w 在词向量查找表中对应的向量。

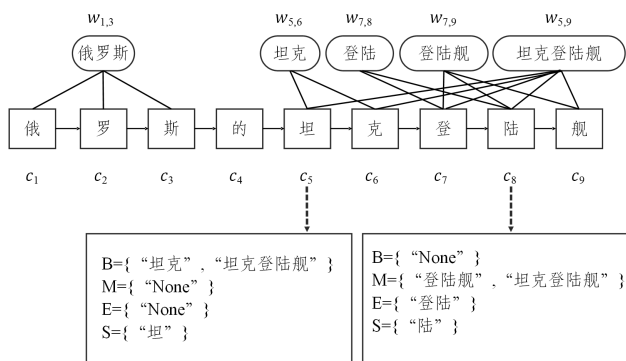


图 2 Soft-lexicon 方法图

Fig. 2 Diagram of Soft-lexicon method

2.5 融合层

融合层将字符级别的嵌入表示和词汇级别的嵌入表示结合起来,以充分利用两种不同粒度的语义信息。

输入文本序列通过 ERNIE 模型获得字符级别的嵌入表示,通过 Soft-lexicon 方法调整和优化词汇级别的嵌入表示。对于输入文本序列 $s = (c_1, c_2, \dots, c_k)$,ERNIE 模型将每个字符 c_i 转换为一个嵌入向量 $e_{char,i}$;同样,Soft-lexicon 方法将每个字符转换为一个结合了词集特征的嵌入向量 $e_{word,i}$ 。融合后的 E_{fusion} 可以表示为:

$$E_{\text{fusion},i} = e_{char,i} + e_{word,i} \quad (4)$$

其中,“+”表示向量的拼接操作,为模型提供了更全面和丰富的输入特征。

2.6 编码层

Bi-GRU 编码层由前向 GRU 和后向 GRU 组成。对于前向 GRU,其在第 i 个时间步的隐藏状态 $h_{f,i}$ 可以表示为:

$$h_{f,i} = GRU_f(e_i, h_{f,i-1}) \quad (5)$$

其中, GRU_f 表示前向 GRU 函数, e_i 是当前时间步的输入向量, $h_{f,i-1}$ 是前一个时间步的隐藏状态。

对于后向 GRU,其在第 i 个时间步的隐藏状态 $h_{b,i}$ 可以表示为:

$$h_{b,i} = GRU_b(e_i, h_{b,i+1}) \quad (6)$$

其中, GRU_b 表示后向 GRU 函数, e_i 是当前时间步的输入向量, $h_{b,i+1}$ 是后一个时间步(实际上是前一个时间步在逆序序

列中的位置)的隐藏状态。

为了充分利用双向信息,Bi-GRU将前向GRU和后向GRU在每个时间步的隐藏状态进行拼接,得到第*i*个时间步的输出向量 h_i ,可以表示为:

$$h_i = h_{f,i} + h_{b,i} \quad (7)$$

其中,“+”表示向量的拼接操作。

经过Bi-GRU编码层处理后,得到了一组输出向量序列 $X = (h_1, h_2, \dots, h_k)$,其中*k*是输入序列的总长度。输出向量既包含了字符级别的信息,也包含了词汇级别的信息,并且由于双向GRU的使用,还捕捉了文本的上下文信息。

2.7 CRF层

CRF通过建模标签间的依赖关系来提高序列标注任务的准确性。

在序列标注任务中,给定一个输入序列,经过编码层得到的输出序列为 $X = \{h_1, h_2, \dots, h_i, \dots, h_k\}$, h_i 表示序列中第*i*个位置的隐藏状态向量,CRF层的目标是预测一个最优的标签序列 $y = \{y_1, y_2, \dots, y_k\}$ 。

对于给定的标签序列*y*,其得分函数的计算公式为:

$$S(h, y) = \sum_{i=1}^k P_{i, y_i} + \sum_{i=1}^k R_{y_i, y_{i+1}} \quad (8)$$

其中,*k*是序列总长度; P_{i, y_i} 表示序列中第*i*个序列输出标签为 y_i 的概率值; $R_{y_i, y_{i+1}}$ 表示字符标签 y_i 到字符标签 y_{i+1} 的转移概率。

然后使用Softmax函数进行归一化,计算标签序列*y*的概率值:

$$P(y|h) = \frac{\exp[S(h, y)]}{\sum_{y' \in Y(h)} \exp[S(h, y')]} \quad (9)$$

其中, $Y(h)$ 表示所有可能的标签序列集合。

损失函数采用最大似然估计的形式,即最小化负对数似然损失:

$$Loss = -\log P(y|h) = -\log \frac{\exp[S(h, y)]}{\sum_{y' \in Y(h)} \exp[S(h, y')]} \quad (10)$$

CRF模型能够学习每个标签之间转移的概率,并利用这些概率来预测每个位置上的标签。通过最大化标签序列的概率,CRF模型能够自动学习出最佳的标签序列。

3 实验结果与分析

3.1 数据集

本文基于CCKS 2022开源装备数据事件要素抽取技术评测任务的数据,自行整理构建了装备领域事件抽取数据集进行模型的训练、验证和测试,同时也在公开数据集ACE2005上进行测试。数据集的基本信息如表1所列。

表1 数据集信息

Table 1 Dataset details

数据集	类型	训练集	验证集	测试集
Ours	句子	1600	200	200
	字符	173600	16500	16000
ACE2005	句子	6000	720	730
	字符	260000	30000	33000

本文数据集主要来自于开源网站上军事领域的新闻文本数据,共包含了6种事件类型以及发起者、承受者、使用器械、时间、地点5类事件要素。原始的CCKS 2022中数据格式是JSON形式,本文针对序列标注任务的需要,通过编写脚本对

原始数据进行了相应处理,将其转换成BIO形式的标注数据用以进行模型的训练、验证和测试。

ACE2005是语言数据联盟(LDC)发布的由实体、关系和事件注释组成的各种类型的数据^[32],包括英语、阿拉伯语和中文数据。本文主要使用中文事件抽取数据,共包含了生活、移动、事务、商业、冲突、交流、个人职业、司法相关8类事件类别以及33种子类别。

3.2 实验环境

本文的实验环境设置如表2所列。

表2 实验环境

Table 2 Experiment environment

实验环境	配置
操作系统	Windows 10
CPU	Intel(R) Core(TM) i7-8700K
内存	16 GB
显卡	NVIDIA GeForce RTX 3090
显存	24 GB
开发环境	Python 3.9
开发框架	Pytorch 1.13.0+cu116

3.3 实验参数

实验的具体参数设置如表3所列。

表3 超参数设置

Table 3 Hyper-parameters setting

参数	数值
char embedding size	50
gaz embedding size	50
hidden dim	200
gaz dropout	0.5
learning rate	0.015
learning ratedecay	0.05
dropout	0.5
epoch	50

3.4 实验结果与分析

本文在装备数据集和ACE2005数据集上开展实验,选取LSTM-CRF, Lattice LSTM, LR-CNN, FLAT, BERT-BiLSTM-CRF, ALBERT-BiLSTM-CRF, RoBERTa-BiLSTM-CRF, CK-BERT和LERT几种模型作为基线与本文模型进行对比,选取精确率(P)、召回率(R)和F1值(F1)作为评价指标。表4和表5列出了不同数据集上的对比实验结果,表6列出了装备数据集上各事件要素的抽取结果,图3给出了装备数据集上本文模型的训练损失值变化情况,图4给出了装备数据集上各模型的指标变化情况。

表4 装备数据集实验结果

Table 4 Results of equipment datasets

模型	P	R	F1
LSTM-CRF	62.50	49.68	55.36
Lattice LSTM	64.86	65.96	65.41
LR-CNN	68.55	66.81	67.67
FLAT	71.01	59.96	65.02
BERT-BiLSTM-CRF	77.11	67.65	72.07
ALBERT-BiLSTM-CRF	76.44	64.48	69.95
RoBERTa-BiLSTM-CRF	71.75	66.60	69.08
CK-BERT	80.85	68.71	74.29
LERT	79.66	68.71	73.78
Ours	80.95	75.48	78.12

表5 ACE2005 数据集实验结果

Table 5 Results on ACE2005

(%)

模型	P	R	F1
LSTM-CRF	42.14	33.23	37.16
Lattice LSTM	42.74	38.74	40.64
LR-CNN	43.71	35.68	39.29
FLAT	45.61	33.28	38.48
BERT-BiLSTM-CRF	52.39	48.70	50.48
ALBERT-BiLSTM-CRF	46.55	38.28	42.02
RoBERTa-BiLSTM-CRF	52.88	40.74	46.02
CK-BERT	58.53	47.71	52.57
LERT	47.81	58.21	52.50
Ours	60.22	50.08	54.68

表6 各事件要素结果

Table 6 Results of each event element

(%)

事件要素	P	R	F1
触发词	79.90	72.60	76.08
发起者	84.02	85.79	84.90
承受者	72.00	62.07	66.67
使用器械	72.22	61.90	66.67
时间	89.13	83.67	86.32
地点	81.93	70.10	75.56

根据表4和表5的实验结果,从以下3个方面进行实验分析。

(1) 模型层面

LSTM-CRF 和 Lattice LSTM 都是基于序列标注的方法,其中 LSTM 仅基于字符信息进行解码,而 Lattice LSTM 使用字词级别的表示进行解码。从实验结果来看,添加字词级别的信息后效果更好,F1 值提升了 10.05%。

LR-CNN 是一种基于卷积神经网络的方法,而序列标注任务通常需要捕捉序列中的长期依赖关系,CNN 主要是基于局部感受野进行特征提取,虽然可以通过堆叠卷积层来扩大感受野,但相比于循环神经网络(RNN)或 Transformer 等模型,其在捕捉长距离依赖方面的能力仍然有限。在装备领域事件元素抽取中,事件元素之间的关系可能跨越较长的序列,文本数据通常具有特定的术语和表达方式,这要求模型具有强大的特征表示能力,而 CNN 在特征提取时主要关注局部特征,这使得 CNN 在处理此类任务时效果不佳。

FLAT 是一种基于 Transformer 的方法。装备领域事件元素抽取是一个涉及复杂语义关系和细粒度分类的任务,但是标注数据稀缺,FLAT 模型在训练数据不足时无法学习到足够的特征来表示这些复杂的语义关系,从而导致性能不佳。

BERT-BiLSTM-CRF, ALBERT-BiLSTM-CRF, RoBERTa-BiLSTM-CRF 三者均是利用预训练语言模型进行事件抽取,可以看出,预训练模型的引入明显提升了抽取的效果,在各项指标上均优于前几种方法。

CK-BERT 融合了语言学知识和事实知识,提升了模型对文本深层次语义的理解能力;LERT 通过多任务预训练和语言知识增强生成了包含丰富语言特征的文本表示。两种模型都是知识增强型的预训练模型,相较于 BERT(ALBERT, RoBERTa)-BiLSTM-CRF 3 种模型取得了更好的效果,但预训练模型主要学习的是通用语言知识,所以对特定领域的专业术语和领域知识的理解程度有限,效果差于本文所提出的方法。

本文提出的方法融合了预训练语言模型的深层语义理解能力和领域词汇信息的语义增强效应,有效提升了模型对装备领域文本的语义理解能力,从而在事件元素识别和抽取方面相较于基线方法取得了较好的结果。

(2) 事件要素层面

“时间”的识别效果较好,F1 值分别为 86.32%;“地点”的 F1 值为 75.56%。“时间”在文本中具有较强的确定性,易于识别;而“地点”的形式则更为复杂,因此识别效果相对较差。

“发起者”的识别效果要优于“承受者”,原因在于“发起者”通常描述为明确的实体或行为主体,使得模型能够较容易地识别;而“承受者”的表述相对更加多样,且有时与“发起者”混淆,在某些情况下,“承受者”可能是隐含的或未明确提及的,如“目标”“敌机”等类似通用的表达,导致模型难以准确区分。

“使用器械”的识别效果相对较差,F1 值为 61.90。“使用器械”通常为特定的装备专业术语,并且类型多样,导致识别的难度增大。

触发词的识别效果较好,因为触发词通常是描述事件的核心动词或名词短语,在文本中以较为明显的形式出现,因此相对容易识别。但是在数据标注过程中,对于触发词的定义和标注标准可能存在不一致,也会导致模型学习到的特征不准确,从而影响识别效果。

(3) 指标层面

实验结果显示,大部分模型尽管在精确率方面表现良好,但召回率却相对较低。如在装备数据集上,本文模型的精确率为 80.95%,召回率为 75.48%。这说明模型在训练过程中更多关注于识别置信度高的论元,而对潜在事件要素的识别不够敏感。具体是因为装备数据集的数量相对有限,模型对某些事件类型的训练不足或特征提取不充分,缺乏足够的样本来学习完整的论元识别模式。

在具体的事件要素中,“承受者”“使用器械”“地点”这 3 类要素的精确率和召回率相差较大。一方面,由于其表达方式更复杂多样,模型无法完全覆盖;另一方面,在数据分布上也会存在不平衡的现象,“使用器械”等要素的标注数据较其他要素少,因此模型对其学习能力也较弱。

总的来说,本文的方法在装备领域事件要素抽取任务上取得了一定的效果。同时,从表 5 可以看出,在 ACE2005 数据集上,本文方法也取得了优于对比方法的效果,说明其具有一定的泛化性。但是由于装备领域特点以及数据标注质量等因素的影响,本文方法还存在一定的提升空间,后续可以通过对这些因素进行深入分析和相应的优化,进一步提升事件要素抽取的整体性能。

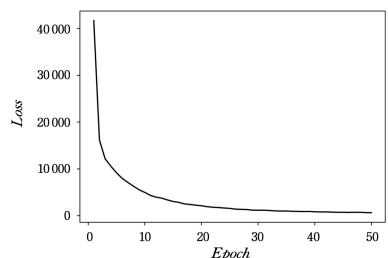


图3 训练损失值

Fig. 3 Train loss

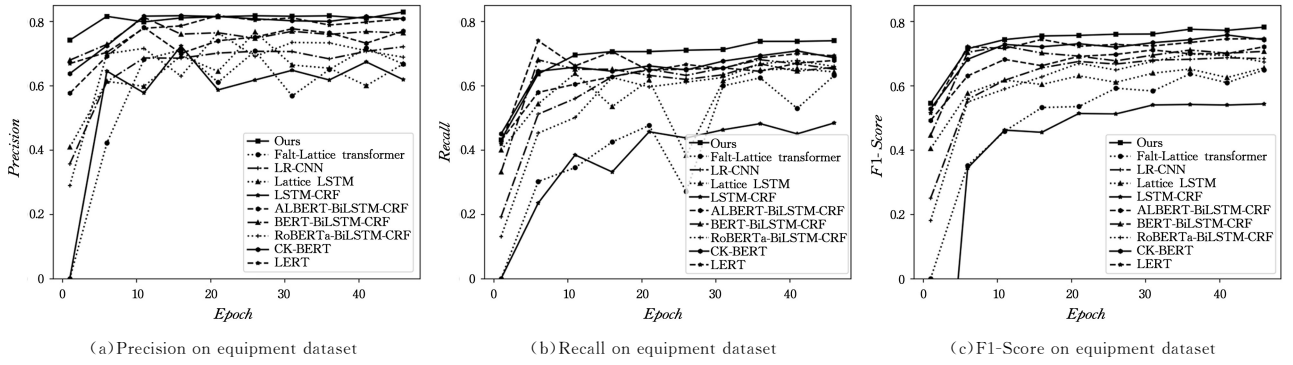


图 4 模型结果对比

Fig. 4 Model results comparison

3.5 消融实验

为了验证本文提出的基于语义增强方法的有效性,即领域词向量融合和预训练模型在装备事件抽取中的作用,进行消融实验。其中,“BiGRU-CRF”作为基线模型,即不采用任何语义增强策略;“ours”代表本文方法;“+ERNIE”代表只使用预训练模型的语义增强;“+ Soft-lexicon”代表仅引入领域词向量的语义增强。实验结果如表 7 所列。

表 7 消融实验结果

Table 7 Results of ablation experiments

模型	P	R	F1
ours	80.95	75.48	78.12
+Soft-lexicon	75.00	66.60	70.55
+ERNIE	77.96	69.56	73.52
BiGRU-CRF	66.83	57.51	61.82

从表 7 中可以看出,仅引入领域词向量(+Soft-lexicon)后,模型的 F1 值相较于基线模型提升了 8.73%,表明领域词向量的融合显著增强了模型对事件元素边界的识别能力。当仅使用预训练模型(+ERNIE)时,F1 值提升了 11.7%,证实了预训练模型在提供丰富的语义表示方面的优势。领域词向量的融合专注于领域专业术语的理解,而预训练模型则提供

更全面的语义表示,当两者结合使用时,模型在装备领域的语义和上下文信息理解方面表现更佳,从而进一步提升了事件抽取的性能。

结果表明,语义增强策略,无论是单独使用还是与预训练模型结合使用,都能有效地提升装备事件抽取的准确性。特别是两者结合使用时,可以最大程度地发挥语义增强的优势,实现对装备领域文本深层次语义的精准捕捉和理解。

3.6 案例分析

图 5 展示了来自开源网站的新闻文本示例,通过应用本文提出的事件抽取方法,得到如图 6 所示的抽取结果,进一步验证了本文方法的有效性和实用性。

Text1:
“林肯”号航空母舰参与支援南方守望行动 (OperationSouthernWatch), 监视伊拉克南方的禁飞区。
Text2:
2022年3月24日,乌克兰军队发挥了圆点U弹道导弹高精度的优势,直接命中了正在卸货的萨拉托夫号 (BKD-10) 1171型 (短吻鳄) 大型登陆舰,库尼科夫号和新切尔卡斯科号在受损之后迅速撤离。
Text3:
杜威号驱逐舰利用ODIN激光拦截器应对敌方无人机及其它精确制导武器对舰艇的进攻。

图 5 新闻文本
Fig. 5 News texts

Text1: 触发词: “监视” 发起者: ““林肯”号航空母舰” 承受者: 无 使用器械: 无 时间: 无 地点: “伊拉克南方的禁飞区” 事件类型: 侦查事件	Text2: 触发词: “命中” 发起者: “圆点U弹道导弹” 承受者: “萨拉托夫号 (BKD-10) 1171型 (短吻鳄) 大型登陆舰” 使用器械: 无 时间: “2022年3月24日” 地点: 无 事件类型: 攻击事件	Text3: 触发词: “拦截” 发起者: “杜威号驱逐舰” 承受者: “敌方无人机” 使用器械: “ODIN激光拦截器” 时间: 无 地点: 无 事件类型: 防守事件
--	--	---

图 6 抽取结果

Fig. 6 Extraction results

从图 6 的抽取结果来看,本文的方法针对触发词和事件论元的抽取均取得了一定的效果,但也反映出本方法目前仍存在的一些问题。比如,对于“时间”这类特定元素的抽取准确率较高,而涉及到论元多样性时,抽取的结果并非十分完整,如 Text3 中“承受者”——“精确制导武器”没有抽取出的内容。出现这种问题可能有以下几个原因:(1)模型虽然结合了字符和词汇级别的特征表示,但是词典的覆盖范围有限,仍然存在未登录词,影响模型对词汇的准确识别,进而影响论元抽取的准确性;(2)模型以简单拼接的方式进行特征的

融合,难以充分捕捉词汇之间的语义关系和上下文信息,导致模型在处理复杂多样的论元时表现不佳;(3)模型的训练和评估策略也会对事件抽取的效果产生显著影响。目前,大多数抽取模型的评估是以论元为单位的,即只要论元及其关联的事件被正确识别,该论元就被认为是抽取正确的。但这种评估方法有其局限性,因为它只关注单个论元的准确性,而忽略了事件的整体结构和其他论元抽取的正确性。这种评估方式无法全面反映模型在实际应用中的性能,尤其是在处理复杂事件时,单个论元的正确识别并不能保证整个事件抽取的准

确性。因此,为提升事件抽取的准确性和完整性,还需对模型的训练与评价策略进行深入研究,以实现更全面的评估。

结束语 本文研究了基于语义增强的装备领域事件抽取方法。针对装备领域存在大量专业术语导致的论元边界模糊的问题,通过融合领域词向量的语义信息和ERNIE预训练模型的语言理解能力,在装备数据事件抽取上取得了一定的效果。实验结果表明,语义增强策略有效地提高了模型对事件论元的识别精度。然而,由于数据集质量和评估策略的局限性等问题,该模型仍需进一步改进和完善。

未来的研究方向:一是要继续优化数据集,提高标注质量和多样性,构建高质量的装备事件抽取数据集;二是引入更丰富的领域知识和语义信息,增加有助于模型识别的文本特征,进一步提升模型对相关要素的识别效果;三是设计更全面的事件抽取评估指标,例如基于事件整体结构的评估指标,以更准确地评估模型在实际应用中的性能。

参 考 文 献

- [1] AFYOUNI I, KHAN A, AL AGHBARI Z. Deep-Eware: spatio-temporal social event detection using a hybrid learning model[J]. *Journal of Big Data*, 2022, 9(1).
- [2] CHANG C, TANG Y, LONG Y, et al. Multi-Information Pre-processing Event Extraction With BiLSTM-CRF Attention for Academic Knowledge Graph Construction[J]. *IEEE Transactions on Computational Social Systems*, 2023, 10(5): 2713-2724.
- [3] KNEZ T, ZITNIK S. Event-Centric Temporal Knowledge Graph Construction: A Survey[J]. *Mathematics*, 2023, 11(23).
- [4] ZHANG Q C, WEI S W, LI Z H, et al. Combining NSP and NER for public opinion event extraction model[J]. *Frontiers in Physics*, 2022, 10.
- [5] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. ACL, 2019: 4171-4186.
- [6] YANG Z, DAI Z, YANG Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding[J]. *Advances in Neural Information Processing Systems*, 2019, 32.
- [7] SUN Y, WANG S, LI Y, et al. Ernie 2.0: A continual pre-training framework for language understanding[C]//Proceedings of the AAAI Conference on Artificial Intelligence. AAAI, 2020: 8968-8975.
- [8] WANG G D. Research on event analysis technology for military texts[D]. Harbin: Harbin Institute of Technology, 2022.
- [9] HU X B, YU X Q, LI S M, et al. Chinese Named Entity Recognition Based on Knowledge Enhancement[J]. *Computer Engineering*, 2021, 47(11): 84-92.
- [10] ZHAO Z Y, ZHU J J, ZHANG Y X, et al. Chinese named entity recognition based on enhancing lexicon knowledge integration utilizing character context information[J]. *Journal of Sichuan University(Natural Science Edition)*, 2024, 61(4): 110-118.
- [11] ZHUANG L, FEI H, HU P. Knowledge-enhanced event relation extraction via event ontology prompt[J]. *Information Fusion*, 2023, 100: 101919.
- [12] LI Q, LI J, SHENG J, et al. A Survey on Deep Learning Event Extraction: Approaches and Applications[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2024, 35(5): 6301-6321.
- [13] YU X, WANG X, LUO X, et al. Multi-scale event causality extraction via simultaneous knowledge-attention and convolutional neural network[J]. *Expert Systems*, 2022, 39(5).
- [14] WANG L, CAO H, YUAN L, et al. Child-Sum EATree-LSTMs, enhanced attentive Child-Sum Tree-LSTMs for biomedical event extraction[J]. *Bmc Bioinformatics*, 2023, 24(1).
- [15] SUN H, ZHOU J, KONG L, et al. Seq2EG: a novel and effective event graph parsing approach for event extraction[J]. *Knowledge and Information Systems*, 2023, 65(10): 4273-4294.
- [16] GUO X Y, MA B, AIBIBULA A, et al. Dynamic Heterogeneous Graph Enhanced Cascade Decoding Event Extraction Model[J]. *Computer Engineering*, 2024: 1-11.
- [17] LIU L P, ZHOU X, CHEN J J, et al. Event Extraction Method Based on Conversational Machine Reading Comprehension Model[J]. *Computer Science*, 2023, 50(2): 275-284.
- [18] DING L, CHEN X, WEI J, et al. Mabert: mask-attention-based Bert for Chinese event extraction[J]. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2023, 22(7): 1-21.
- [19] PEI B S, LI X, JIANG Z T, et al. Research on Public Security Professional Small Sample Knowledge Extraction Method Based on Large Language Model[J]. *Journal of Frontiers of Computer Science and Technology*, 2024, 18(10): 2630-2642.
- [20] LI Y, GENG C Y, YANG D. Fin-BERT-Based Event Extraction Method for Chinese Financial Domain[J]. *Computer Engineering and Applications*, 2024, 60(14): 123-132.
- [21] WU C, ZHANG X, ZHANG Y, et al. PMC-LLaMA: Further Finetuning LLaMA on Medical Papers[J]. *arXiv:2304.14454*, 2023.
- [22] LI H P, MA B, YANG Y T, et al. Document-level Event Extraction Method Based on Slot Semantic Enhanced Prompt Learning[J]. *Computer Engineering*, 2023, 49(9): 23-31.
- [23] YU C M, DENG B, TAN L Y, et al. Syntax-Enhanced Event Extraction Model Based on XLNET and GAT[J]. *Data Analysis and Knowledge Discovery*, 2024, 8(4): 26-38.
- [24] ZHANG Z, HAN X, LIU Z, et al. ERNIE: Enhanced Language Representation with Informative Entities[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. ACL, 2019: 1441-1451.
- [25] LIU W, ZHOU P, ZHAO Z, et al. K-bert: Enabling language representation with knowledge graph[C]//Proceedings of the AAAI Conference on Artificial Intelligence. AAAI, 2020: 2901-2908.
- [26] SUN T, SHAO Y, QIU X, et al. CoLAKE: Contextualized Language and Knowledge Embedding[C]//Proceedings of the 28th International Conference on Computational Linguistics. ACL, 2020: 3660-3670.
- [27] WANG X, GAO T, ZHU Z, et al. KEPLER: A unified model for

knowledge embedding and pre-trained language representation [J]. Transactions of the Association for Computational Linguistics, 2021, 9: 176-194.

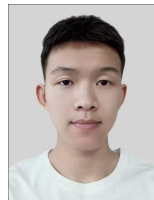
[28] YU D, ZHU C, YANG Y, et al. Jaket: Joint pre-training of knowledge graph and language understanding [C] // Proceedings of the AAAI Conference on Artificial Intelligence. AAAI, 2022: 11630-11638.

[29] TIAN S, LUO Y, XU T, et al. KG-Adapter: Enabling Knowledge Graph Integration in Large Language Models through Parameter-Efficient Fine-Tuning [C] // Findings of the Association for Computational Linguistics ACL 2024. ACL, 2024: 3813-3828.

[30] CHURCH K W. Word2Vec [J]. Natural Language Engineering, 2017, 23(1): 155-162.

[31] MA R, PENG M, ZHANG Q, et al. Simplify the Usage of Lexicon in Chinese NER [C] // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020: 5951-5960.

[32] CUNHA L F, SILVANO P, CAMPOS R, et al. ACE-2005-PT: Corpus for Event Extraction in Portuguese [C] // Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024. Association for Computing Machinery, 2024.



FANG Rui, born in 2000, postgraduate. His main research interests include natural language processing, knowledge graph, etc.



CUI Liangzhong, born in 1979, Ph.D., associate professor. His main research interests include knowledge graph, data mining and analysis, etc.