基于上下文感知的智能合约要素提取方法

钱 肖[1],蒋忠元[1*],陶梅悦[1],刘柄呈[1],李任翔[1],高 胜[2],马建峰[1]
(1. 西安电子科技大学网络与信息安全学院,陕西西安,710126;2. 中央财经大学信息学院,北京,100081)

摘 要: 针对各行各业海量文本文档的智能合约化需求,提取文本关键数据要素是首要基础. 与传统命名实体识别(Named Entity Recognition,NER)相比,合约要素提取(Contract Element Extraction,CEE)技术旨在提取泛在较长、更多样、较冗余合约要素,然而目前面临着中文研究不足、对新颖大语言模型(Large Language Model,LLM)技术应用不够充分、对文本上下文关联特征感知不足等挑战. 本文首先提出了新颖的上下文语义感知动态填充方法(Context-sensitive Dynamic Padding Method,CDPM)、三重注意力层和要素边缘加权损失函数模块,在不增加硬件需求的前提下,为模型提供额外上下文语义信息,增强对上下文关联特征的感知能力,从而提升基于序列标注范式的CEE训练效率. 其次,融合上述模块和BERT(Bidirectional Encoder Representations from Transformers)嵌入模型构建了一种基于上下文感知的合约要素提取模型(Context-Aware Model for Contract Element Extraction,CAM-CEE),实现了面向智能合约化场景的高性能要素提取. 最后,在本文自主构建的数据集以及相关公开数据集上进行了大量实验. 结果表明,本文提出框架CAM-CEE在micro F_1、macro F_1等指标上的性能超越最佳基线模型,并具有高通用性.

关键词: 合约要素提取(CEE);自然语言处理(NLP);序列标注;信息提取
基金项目: 国家重点研发计划(No.2022YFB2701800)
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2025)04-1322-15
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20241038

A Context-Aware Approach for Smart Contract Element Extraction

QIAN Xiao[1], JIANG Zhong-yuan[1*], TAO Mei-yue[1], LIU Bing-cheng[1], LI Ren-xiang[1],
GAO Sheng[2], MA Jian-feng[1]
(1. *School of Cyber Engineering*, *Xidian University*, *Xi'an*, *Shaanxi* 710126, *China*;
2. *School of Information*, *Central University of Finance and Economics*, *Beijing* 100081, *China*)

Abstract: Extracting key data elements from text is the primary foundation for the intelligent contract conversion demand of massive text documents in various industries. Compared with traditional named entity recognition (NER), contract element extraction (CEE) aims to extract ubiquitous, lengthy, diverse, and redundant contract elements. However, it faces challenges such as limited research in Chinese, lack of application of novel large language model (LLM) techniques, and insufficient perception of contextual features in text. This article first proposes a novel context-sensitive dynamic padding method (CDPM), a triple attention layer, and an edge-weighted loss function. They provide additional context semantics without increasing hardware requirements, enhance the perception of context related features, and improve the efficiency of element extraction training under the sequential annotation paradigm; Secondly, a context-aware deep learning framework context-aware model for contract element extraction (CAM-CEE) was proposed by integrating the above modules with the bidirectional encoder representations from transformers (BERT) embedding model, achieving high-performance element extraction for smart contract scenarios; Finally, extensive experiments are conducted on the independently constructed and publicly available datasets in this article. The results indicate that the proposed framework CAM-CEE outperforms the best baseline model in metrics such as micro F_1 and macro F_1, and has high generality.

Key words: contract element extraction (CEE); nature language processing (NLP); sequence labeling; information extraction
Foundation Item(s): National Key Research and Development Program of China (No.2022YFB2701800)

收稿日期:2024-11-17;修回日期:2025-03-17;责任编辑:覃怀银
*通讯作者:蒋忠元

1　引言

随着经济的快速发展和信息化的大力推进,各行各业的长文本文档数量发生了爆炸性的增长. 由于自然语言文档充满模糊性、二义性和冗余的语义,基于人工的传统文档处理方式自动化程度低、人力需求大、时间成本高. 为了实现各行业的进一步智能化与自动化,需要将海量文档转换为基于区块链代码的智能合约,即实现文档智能合约化[1].

文档智能合约化的基本特征是实现文档自然语言数据要素与智能合约代码数据要素的等价,因此文档向智能合约转变的核心在于自然语言表达的文本要素向智能合约数据要素的转换,其首要基础则是从文档中提取出与智能合约数据要素对应的文本要素,即智能合约要素提取(Contract Element Extraction,CEE).

当前智能CEE的主流方法是CEE[2]和命名实体识别(Named Entity Recognition,NER). 其中,CEE用于长文本文档合约中的多种要素,既包括较长文本句也包括较短的名词. NER则用于从文本中提取多种预定义实体,主要针对短名词、单一文本句的提取进行优化,因此CEE相对更适合智能CEE场景. 在CEE和NER以外,大语言模型(Large Language Model,LLM)是自然语言处理(Natural Language Processing,NLP)的最新成果,其中,以GPT(Generative Pre-trained Transformer)[3]、DeepSeek[4]为代表的生成式LLM在许多任务中展现了强大的自然语言理解与生成能力. 然而,LLM基于通用语境语料训练,在处理文档智能合约化场景下专业性较强、上下文较特殊的文档时可能受通用语料影响,导致提取能力不稳定. 本文通过给定元素类型描述与样本内容测试了DeepSeek-V3的提取性能,结果其在两测试集上的提取准确率仅为35.33%和5.83%,远低于预期,表明其需要使用专业数据集进行训练以提升CEE能力. 同时,生成式LLM的训练与推理均依赖大规模计算集群,其硬件投入远超中小机构承受能力. 因此,目前生成式LLM不太适用于智能CEE任务,特别是在专业文档、低算力场景下,CEE方法在成本、效率与领域适配性上优势更为显著.

但当前CEE存在一些亟待解决的挑战. 首先,虽然生成式LLM本身难以适用于要素提取任务,但同样基于Transformer架构的较小判别式LLM技术,仍然可以成为解决方案的嵌入组件,并有效提升NLP任务(如NER)的性能,而现有CEE方法以机器学习和长短期记忆网络(Long Short-Term Memory,LSTM)、图神经网络(Graph Neural Network,GNN)、卷积神经网络(Convolutional Neural Network,CNN)等传统深度学习网络为主,仅有少数方法简单应用了LLM技术,在中文CEE领域,对LLM的应用更是缺乏. 其次,智能CEE需要对整篇文档进行提取,但输入整篇长文档将导致硬件需求过高,将文档分成句子输入又可能破坏完整语义,使提取退化至NER的单句提取模式. 再次,现有要素提取相关工作并未充分利用长文本文档的特征,例如连续文本句上下文. 最后,主流NER和CEE均采用边界标注或序列标注范式,但前者显存占用高,后者提取长要素的性能较低.

为解决上述挑战,本文提出了一种基于上下文感知的合约要素提取模型(Context-Aware Model for Contract Element Extraction,CAM-CEE). 与现有主流CEE方法相比,CAM-CEE首先基于BERT(Bidirectional Encoder Representations from Transformers)[5]模型实现文本的初步向量化,充分利用了LLM的强大嵌入能力. 其次,该方法融合新颖的上下文敏感的动态填充方法(Context-sensitive Dynamic Padding Method,CDPM),在不增加训练硬件需求和时间成本的情况下,提供了目标句前后的额外上下文信息,解决了输入整篇长文档/输入单文本句的两难问题,并增强了LLM的上下文特征提取能力. 再次,该方法引入了新颖的三重注意力层,从CDPM构建的前上下文-目标句-后上下文三元组中充分捕获不连续特征,提升上下文感知能力. 最后,本文提出基于要素边缘加权的交叉熵损失计算方法,进一步提升了CAM-CEE在序列标注范式下的要素提取性能. 广泛实验的结果表明,提出的CAM-CEE在实验数据集上优于所有最佳(State Of The Art,SOTA)基线,且可帮助多种嵌入模型提升性能,证明CAM-CEE能有效地提取细粒度、广范围的要素,且具备高度通用性.

2　相关工作

2.1　CEE方法

CEE方法可分为基于规则、传统机器学习、深度学习三种.

首先,针对少量形式相对固定的文档,例如建筑业合同,可使用基于词法、句法、语义等规则的方法识别风险与免责条款[6~8]. 基于规则的方法在组织形式类似的文本上提取准确性高、速度快,但在面对语言变异、文本不规范或预期外文本结构时,规则难以适用,导致提取方法失效,通用性、灵活性较低.

其次,基于传统机器学习的CEE方法主要通过特征工程和分类算法从文本中自动识别和提取关键要素,例如与知识图谱技术结合提取规范性条款和约束条件[9],或直接从文本中检索时间、成本等量化信息[10]. 实验表明在合约较多情况下机器学习效果优于基于规则方法,同时二者结合可以进一步提升性能[2]. 虽然对硬件需求较低,但传统机器学习模型往往层数少、结构简单,因此相关研究所提取的元素类别较少、

粒度较粗,存在局限性.

再次,基于深度学习的方法能够学习文本中的更多特征和语义关系,适配更多文档的提取需求,其性能明显优于机器学习分类器,且不同方法在基于不同粒度的评估中表现不一[11]. 中文CEE方法大多基于深度学习实现,例如,文献[12]提出了一种基于条件随机场的中文建筑文档提取方法;文献[13]提出了toi-CNN+LR(text-of-interest-Convolutional Neural Network+Logistic Regression),该框架结合文本信息提取和位置回归技术,主要使用卷积神经网络从保险合约中提取条款相关元素;文献[14]提出了Bi-FLEET(Bi-directional Feedback cLause Element relaTion network)框架,其结合了包括CNN、LSTM和GNN在内的多种网络,用于实现跨域CEE.

最后,近年来,以BERT和GPT为代表的LLM是基于深度学习的NLP研究中最重要的进展之一. 然而,在CEE领域,仅有少数工作英文场景下简单使用LLM进行训练和测试. 例如,文献[15]在少量公开可用的英文合约上训练和测试了深度学习方法双向长短期记忆网络(Bidirectional Long Short-Term Memory, BiLSTM)和BERT,并对比了两者性能. 文献[16]构建了一个小规模的租赁合约数据集,并使用AlBERT在其上进行训练,提取实体类型和危险信号. 这些研究仅仅在数据集上训练和测试了BERT和AlBERT的性能,并未设计或添加任何组件来改进方法,也没有充分利用合约中的特征. 此外,LLM在中文CEE的应用尚未得到探索.

2.2 信息抽取与NER方法

虽然在CEE中应用较少,但近年来LLM在信息抽取,特别是与CEE相似的NER中广泛应用[17~19]. 在信息抽取领域,近年来,研究者常通过向基于LLM的嵌入模型和分类网络中提供更多特征来提升方法性能. 例如,文献[20]利用标记嵌入以及相对实体距离大幅提升了超参数关系提取性能. 联合多任务获取更多特征也是一大常用方法,例如,文献[21]对任务之间特定关系进行显式建模,提升关系抽取与实体识别的性能;文献[22]提出了一种联合命名NER和关系提取的方法,从研究论文中提取结构化知识.

具体到NER方法层面,从待提取文本中获取更多上下文特征表示往往能立竿见影地提升提取性能. 文献[23]将实体识别任务视为字与字之间关系的分类问题,改进了模型结构和特征表示. 文献[24]利用汉字的形状特征来提高医学文本中实体识别的准确性. 文献[25]通过整合以词汇特征为代表的多种元数据来提高模型的识别能力. NER方法主要关注简短和类型有限的名词实体,如人名、地名或组织名等,但CEE的目标更细粒度和多样化,既包括较长的逻辑文本,如合约条款句的原因、结果部分,又包括短名词,如甲、乙方. 同时,长文本文档中的文字组织方式,例如连续文本句等,与NER数据集中常见的单句独立文本有着根本的不同. 因此,尽管NER可被用作要素提取领域的参考解决方案[12,26]或基线模型[13,14],仍有必要专门进行智能CEE方法的研究.

3 问题定义与CAM-CEE

3.1 问题定义

本节给出CEE的问题定义. 给定一个文档文本,将其分解为一系列句子$\boldsymbol{S}=(\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_n)$,这里$\boldsymbol{s}_i$代表第$i$个句子,$n$则是该文档中句子的总数. 之所以输入句子而不是整个文档,是因为后者通常超过数千个令牌(token,即中文字符),导致显存需求过高,无法接受. 在中文场景下,每一个长度为m_i的句子$\boldsymbol{s}_i$都是一系列字符,或者说token,可表示为$\boldsymbol{s}_i=(t_{i,1},t_{i,2},\cdots,t_{i,m_i})$,这里$t_{i,j}$是$\boldsymbol{s}_i$中的第$j$个token.

然后,需要标记每个token属于哪个元素类别. 在token分类下,CEE模型扫描文档句中的每个token,并用细粒度子类标签对其进行分类$L^C=\{L_1,L_2,\cdots,L_C\}$,这里$C$代表分类类别的数量. token分类(即序列标注)的主要任务是将$\boldsymbol{s}_i$转换成一个分类的元组序列Y_i,这里$l_{i,j}$表示$t_{i,j}$的类别标签

$$Y_i=\left\{\left(t_{i,j},l_{i,j}\right)\right\}_{j=1}^{m_i},l_{i,j}\in L^C \tag{1}$$

3.2 模型框架

如图1所示,CAM-CEE的流程如下:输入整篇待提取文本文档后,首先,基于规则将多个连续和重复的空格、制表符或换行符转换为一个,并根据句末标点符号(包括"?""!""。"等)和换行符将文档文本分解为句子,在序列标注范式下,一个合同句子对应一个token序列和一个元素标签序列. 然后,每个句子及其标签序列都由CDPM填充,对齐长度并提供额外上下文. 其次,将文本和标签序列输入基于BERT的嵌入模型,并转化为嵌入张量. 接下来,双向门控循环单元(Bidirectional Gated Recurrent Unit, Bi-GRU)层和三重注意力层将分别提取嵌入张量的连续上下文关联信息和计算其不连续上下文关联信息. 最后,全连接(Full Connect, FC)层和Softmax层将张量转换为概率分布. 在非训练情况下,Argmax层将选择概率最高的类别作为每个token的标签,完成从token到标签的映射和要素提取.

3.3 上下文敏感的动态填充机制

在通常的填充方式中,基于预设的最大输入长度$N(N>m_i)$,对于每个长度为m_i的句子$\boldsymbol{s}_i$,$N-m_i$个token(其值为0,无任何特定含义)将被填充到$\boldsymbol{s}_i$中,从而对齐训练样本的长度. 这是深度学习中的一种常见方式.

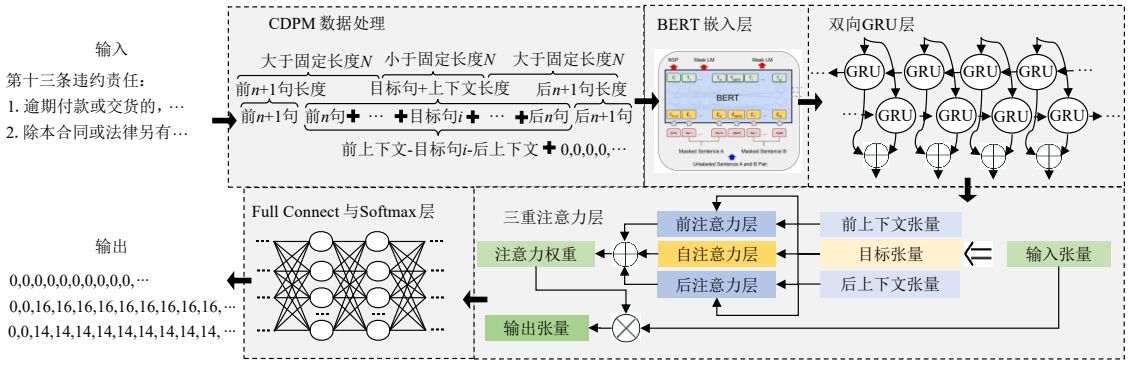

图1　CAM-CEE概览

然而,上述普通填充方法填入的内容无特定作用,且输入的单句文本包含较少上下文信息,缺失完整语义. 针对上述问题,本文提出了如图2所示的CDPM来填充尽可能多的上下文token,而不是通常的无意义填充. 具体来说,首先分别计算句子$\boldsymbol{s}_i$、$\boldsymbol{s}_{i+1}$和$\boldsymbol{s}_{i-1}$的长度m_i、m_{i+1}和m_{i-1}. 如果$m_{i-1}+m_i+m_{i+1}<N$,那么$\boldsymbol{s}_{i-1}$和$\boldsymbol{s}_{i+1}$的token将分别被填充到$\boldsymbol{s}_i$的前面和后面. 被填充的句子$\boldsymbol{s}_i'$将暂时表示为

$$\boldsymbol{s}_i'=(t_{i-1,1},t_{i-1,2},\cdots,t_{i-1,m_{i-1}},t_{i,1},t_{i,2},\cdots,t_{i,m_i},t_{i+1,1},t_{i+1,2},\cdots,t_{i+1,m_{i+1}}) \tag{2}$$

假如三个句子的长度之和仍然小于N,则句子$\boldsymbol{s}_{i\pm2},\boldsymbol{s}_{i\pm3},\cdots$也将以相同的方式计算长度,并交替添加至目标句的前后,直到新添加的句子抵达文档文本的开头或结尾,或者新填充的句子会使整个输入的长度超过N. 具体来说,假设作为填充上下文的前后句数目分别为η_1和η_2,文档中句子的总数为n,那么η_1和η_2必须满足

$$(i-\eta_1=0)\wedge(i+\eta_2=n)\wedge\left(\sum_{k=i-\eta_1}^{i+\eta_2}m_k<N\right) \tag{3}$$

上下文的边界和输入的边界将是$\boldsymbol{s}_{i-\eta_1}$的开头和$\boldsymbol{s}_{i+\eta_2}$的结尾. 因此,所填充上下文的边界取决于两个因素:目标句所在文档的边界以及设定的最大输入长度N. 对于前者,假如上下文范围超出了文档边界,在不同的数据存储形式下,可能导致其他文档的不相关内容被填入上下文等问题;对于后者,假如新填入的上下文使得整个输入的长度超过N,就需要将输入拆分,对目标句的训练失去益处;最后,基于句子粒度的填充使得上下文的边界总是句子开头和结尾,从而尽可能保证填充的上下文语义是通顺、完整的. 如果填充后的输入长度没有恰好达到N的限制,且无法再填充新的上下文,则仍将基于普通填充方法填充剩余的较少量token. CDPM填充的上下文将在过程中提供更多信息,帮助模型进一步提取待分类的token的特征,且几乎不会增加硬件需求与训练时间.

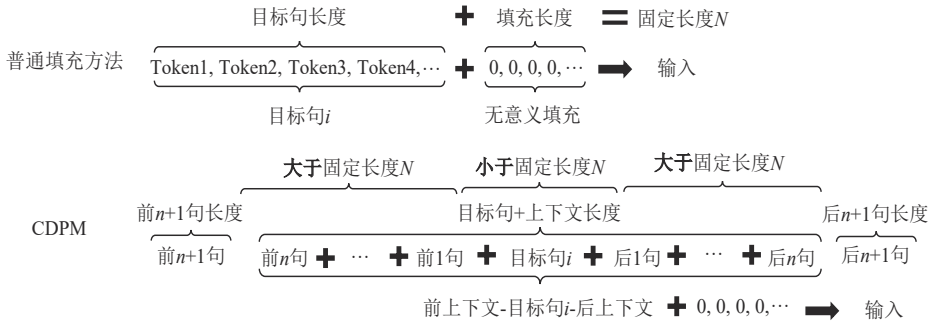

图2　普通填充方法与CDPM的对比

一般来说,CDPM填充后的单个输入由三部分组成:前上下文、目标句和后上下文. 假如输入是$\boldsymbol{s}_i$,不妨用$\boldsymbol{s}_i^{\text{pad}}=\{t_{i,j}^{\text{pad}}\}_{j=1}^{N}$来代表CDPM的输出. 由于$\boldsymbol{s}_i$的token位于$\boldsymbol{s}_i^{\text{pad}}$的中间,为简便起见,设$z=\sum_{k=i-\eta_1}^{i-1}m_k$,则可以用$z+1$来代表$t_{i,1}$($\boldsymbol{s}_i$的第一个token)在$\boldsymbol{s}_i^{\text{pad}}$中的位置. 那么前上下文可表示为$\{t_{i,j}^{\text{pad}}\}_{j=1}^{z}$,目标句$\boldsymbol{s}_i$则表示为$\{t_{i,j}^{\text{pad}}\}_{j=z+1}^{z+m_i}$,后上下文表示为$\{t_{i,j}^{\text{pad}}\}_{j=z+m_i+1}^{N}$.

3.4　基于BERT的嵌入层

在嵌入过程中,输入句子$\boldsymbol{s}_i^{\text{pad}}$后,BERT模型输出的

最后一个隐藏状态被用作嵌入张量. 句子的嵌入实际上是由token嵌入组成的序列,可表示为$\boldsymbol{s}_i^{\text{emb}}=\{t_{i,j}^{\text{emb}}\}_{j=1}^{N}$.

3.5 分类层

3.5.1 Bi-GRU层

接下来,Bi-GRU层从输入的嵌入张量中提取文本中常见的连续相关性信息. 例如,行为类别的元素总是出现在条件元素之后. 假设$h_{i,j}^{\rightarrow}$和$h_{i,j}^{\leftarrow}$分别代表每个token的前向和后向GRU隐藏状态,则Bi-GRU的输出将如下计算:

$$h_{i,j}^{\rightarrow}=\text{GRU}^{\rightarrow}\left(h_{i,j-1}^{\rightarrow},t_{i,j}^{\text{emb}}\right) \tag{4}$$

$$h_{i,j}^{\leftarrow}=\text{GRU}^{\leftarrow}\left(h_{i,j+1}^{\leftarrow},t_{i,j}^{\text{emb}}\right) \tag{5}$$

$$\boldsymbol{s}_i^{\text{gru}}=\left\{\left(h_{i,j}^{\rightarrow}+h_{i,j}^{\leftarrow}\right)/2\right\}_{j=1}^{N} \tag{6}$$

3.5.2 三重注意力层

如3.3节所述,输入文本由三部分组成,其中目标句需以较离散的方式从前后上下文两部分提取关联关系. 例如,合约中的标的表格被转换为文本后,表头"名称\t金额\t金额\n"将位于前上下文,表内容行"抑尘车辆\t1\t562000元\n"则是目标句子. 那么目标文本中总价元素的分类("56 200元")可能与前上下文的特定部分("金额")相关联. 由于它们不直接相邻、不连续,Bi-GRU难以提取两者间的关联关系,难以利用CDPM填充的上下文句子的所有特征.

为了解决上述问题,本文提出了三重注意力层. 与常见的多头注意力方法[5]不同,如图3所示,本文使用三个注意力层,包括前注意力、自注意力和后注意力,分别处理输入的前上下文、目标句和后上下文三个部分. 例如,假如三重注意力层堆叠在Bi-GRU层之上,输入则为$\boldsymbol{s}_i^{\text{gru}}=\{t_{i,j}^{\text{gru}}\}_{j=1}^{N}$,其中,目标句的开头和结尾token分别为$t_{i,z+1}^{\text{gru}}$和$t_{i,z+m_i}^{\text{gru}}$. 然后,前注意力层使用$\{t_{i,j}^{\text{gru}}\}_{j=1}^{z}$作为注意力的value和key输入,而自注意力层使用$\{t_{i,j}^{\text{gru}}\}_{j=z+1}^{z+m_i}$,后注意力层使用$\{t_{i,j}^{\text{gru}}\}_{j=z+m_i+1}^{N}$. 同时,所有三个注意力部分使用$\{t_{i,j}^{\text{gru}}\}_{j=z+1}^{z+m_i}$作为注意力的query输入$Q_i$.

然后,对于前注意力层,翻转key输入的首z个值,让key值输入变为$\boldsymbol{K}_i^{\text{pre}}=\left(t_{i,z}^{\text{gru}},t_{i,z-1}^{\text{gru}},\cdots,t_{i,1}^{\text{gru}},0,\cdots,0\right)$. 对于后注意力层,也用类似方式翻转最后$N-z-m_i$个值,让key输入变为$\boldsymbol{K}_i^{\text{next}}=(0,\cdots,0,t_{i,N}^{\text{gru}},t_{i,N-1}^{\text{gru}},\cdots,t_{i,z+m_i+1}^{\text{gru}})$. 其原因为,对于前上下文和后上下文中的token,相比"token自身在前/后上下文的位置","到目标的距离"更有可能反映它与目标文本之间的关联. 当然,目标句本身不需要翻转,故其输入为$\boldsymbol{K}_i^{\text{self}}=\boldsymbol{Q}_i$. 最后,对于每个注意力层,分别由$\varepsilon\in\{\text{pre},\text{self},\text{next}\}$代表,其未加权的value输入等于key输入$\boldsymbol{V}_i^{\varepsilon}=\boldsymbol{K}_i^{\varepsilon}$.

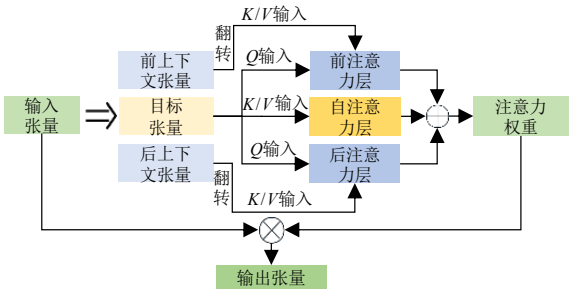

图3 三重注意力的工作模式

接下来,设置$\boldsymbol{W}_{\boldsymbol{Q}_i}$、$\boldsymbol{W}_{\boldsymbol{V}_i^{\text{pre}}}$、$\boldsymbol{W}_{\boldsymbol{V}_i^{\text{self}}}$和$\boldsymbol{W}_{\boldsymbol{V}_i^{\text{next}}}$四个权重矩阵,并使用Glorot[27]标准函数进行初始化. 它们会在训练期间分别基于$\boldsymbol{Q}_i$、$\boldsymbol{V}_i^{\text{pre}}$、$\boldsymbol{V}_i^{\text{self}}$和$\boldsymbol{V}_i^{\text{next}}$进行权重更新,从而使三重注意力层参与深度学习网络的训练过程. 以Luong-style[28]进行注意力计算,则每个输入部分的注意力权重$\boldsymbol{A}_i^{\varepsilon}$如下:

$$\sigma\left(\boldsymbol{Q}_i,\boldsymbol{K}_i^{\varepsilon}\right)=\boldsymbol{Q}_i\cdot\boldsymbol{W}_{\boldsymbol{Q}_i}\cdot\boldsymbol{K}_i^{\varepsilon\,\text{T}} \tag{7}$$

$$\boldsymbol{A}_i^{\varepsilon}=\text{softmax}\left(\sigma\left(\boldsymbol{Q}_i,\boldsymbol{K}_i^{\varepsilon}\right)\right)\cdot\left(\boldsymbol{V}_i^{\varepsilon}\cdot\boldsymbol{W}_{\boldsymbol{V}_i^{\varepsilon}}\right) \tag{8}$$

以$a_{i,j}$代表每个位置的注意力权重,连接$\boldsymbol{A}_i^{\text{pre}}$、$\boldsymbol{A}_i^{\text{next}}$和$\boldsymbol{A}_i^{\text{self}}$得到$\boldsymbol{A}_i^{\text{all}}=\{a_{i,j}\}_{j=1}^{N}$. 则三重注意力层的输出为

$$\boldsymbol{s}_i^{\text{att}}=\boldsymbol{s}_i^{\text{gru}}*\boldsymbol{A}_i^{\text{all}} \tag{9}$$

3.5.3 最终分类层

最后,最终分类层由FC层和Softmax层组成,计算$\boldsymbol{s}_i^{\text{att}}$为概率分布序列$\widehat{P}_i$:

$$\widehat{P}_i=\text{softmax}\left(\text{FC}\left(\boldsymbol{s}_i^{\text{att}}\right)\right) \tag{10}$$

其中,$\widehat{P}_i=\left\{\widehat{p_{i,j}^{L_1}},\widehat{p_{i,j}^{L_2}},\cdots,\widehat{p_{i,j}^{L_C}}\right\}_{j=1}^{N}$,$\widehat{p_{i,j}^{L_k}}$代表$t_{i,j}^{\text{pad}}$属于$L_k,k\in\{1,2,\cdots,C\}$的预测可能性.

在非训练情况下,argmax函数会选择最大可能的分类$\widehat{l_{i,j}}\in L^C$(其中$z<j\leqslant z+m_i$)作为式(1)中$t_{i,j-z}$的$l_{i,j-z}$,并完成序列标注任务,实现要素提取:

$$\widehat{l_{i,j}}=\text{argmax}\left(\left(\widehat{p_{i,j}^{L_1}},\widehat{p_{i,j}^{L_2}},\cdots,\widehat{p_{i,j}^{L_C}}\right)\right) \tag{11}$$

3.6 训练

在序列标注范式下,一般基于交叉熵损失函数进行权重更新即模型训练. 训练时,由$\left(p_{i,j-z}^{L_1},p_{i,j-z}^{L_2},\cdots,p_{i,j-z}^{L_C}\right)$代表真值标签的热编码$l_{i,j-z}$,则交叉熵损失可被计算为

$$\text{CELoss}(i)=-\sum_{j=z+1}^{z+m_i}\frac{1}{C}\sum_{k=1}^{C}p_{i,j-z}^{L_k}\ln\left(\widehat{p_{i,j}^{L_k}}\right) \tag{12}$$

序列标注任务旨在正确分类尽可能多的token. 然而,在要素提取中,相对于token级的评估,元素级的评估可能更为重要. 对于后者,只有当元素中的每一个token都被正确识别时,对元素的预测才被认为是正确的. 因此,当使用序列标注进行要素提取时,正确预测的标注的位置可能比它们的数量更重要. 一方面,在中

文CEE中,元素的边界相对难以分类,在序列标注中,体现为真值元素边界的token难以分类. 如图4所示,图中的不同颜色的文本代表不同类型的元素,下划线标识每个元素的边缘token. 另一方面,如果错误判断真实元素中间的某个token,该token将会成为预测元素序列的边界;同时,该真实元素将被识别为两个甚至更多的错误元素,说明预测元素的边界的错误判断会严重影响要素提取. 因此,在训练中重视真值元素和预测元素的边界token的分类将减少由错误序列标注导致的错误元素提取,让模型实现更有效的学习.

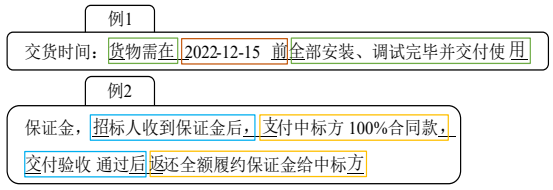

图4　元素边界示例

同时,加权损失函数和基于任务定制的损失函数已被广泛应用于深度学习训练[29,30]. 基于上述考虑,本文提出了基于真值和预测要素边缘的加权交叉熵损失函数. 具体来说,对于每一个的真值标签序列$\{l_{i,j}\}_{j=z+1}^{z+m_i}$和其预测标签序列$\{\widehat{l_{i,j}}\}_{j=z+1}^{z+m_i}$,如果存在任意$i,j$对使得

$$\left(l_{i,j}\neq l_{i,j+1}\right)\vee\left(l_{i,j}\neq l_{i,j-1}\right)\vee\left(\widehat{l_{i,j}}\neq\widehat{l_{i,j+1}}\right)\vee\left(\widehat{l_{i,j}}\neq\widehat{l_{i,j-1}}\right) \tag{13}$$

即某一token的预测标签或真实标签不同于其任一邻居token,则它是真实或预测元素的边界,那么该标签计算的损失将加倍. 基于此,边缘加权损失函数可如下计算:

$$\mathrm{TokenLoss}(i,j)=\begin{cases}\dfrac{2}{C}\displaystyle\sum_{k=1}^{C}p_{i,j-z}^{L_k}\ln\left(\widehat{p_{i,j}^{L_k}}\right), & \text{如式(13)为真}\\ \dfrac{1}{C}\displaystyle\sum_{k=1}^{C}p_{i,j-z}^{L_k}\ln\left(\widehat{p_{i,j}^{L_k}}\right), & \text{其他}\end{cases} \tag{14}$$

$$\mathrm{WeightedLoss}(i)=-\sum_{j=z+1}^{z+m_i}\mathrm{TokenLoss}(i,j) \tag{15}$$

4　实验与结果分析

4.1　实验设置

4.1.1　基本设定

本文的实验主要包括五个部分:(1)有效性实验,比较所提出的框架CAM-CEE和SOTA方法的要素提取性能,以验证CAM-CEE的有效性;(2)消融实验,逐步将组件添加到基本模型上并对其进行评估,验证每个组件的效用;(3)通用性实验,将提出的框架应用于其他的类BERT嵌入模型,从而验证CAM-CEE的通用性;(4)分析不同元素上的性能差异,即详细分析每种元素的性能及其原因;(5)案例分析,在完整合约文档上进行面向实用场景的要素提取案例演示,基于错误抽取实例分析进一步优化的可能方向.

实验中分别列出并评价了使用边缘加权损失函数和普通交叉熵损失函数训练的CAM-CEE框架的结果,从而更全面地评估框架性能. 在实验结果中,CAM-CEE(w)代表使用边缘加权损失函数训练的CAM-CEE,CAM-CEE(n)代表使用普通交叉熵函数训练的CAM-CEE.

4.1.2　实验数据集

(1)数据集1

当前公开可用的中文要素提取数据集极少[13,14],且被过度清洗,导致数据集无法反映真实场景下的智能合约化需求. 因此,本文面向智能合约化要素提取场景,构建了首个中文买卖合约数据集CPCD.

首先,本文从上海、重庆和湖北等多个地方政府的官方网站①上,以合法合规的方式采集了64 314份真实有效的政府采购合同. 接下来,这些合同被自动转换为文本,并由三名经过系统培训的研究生进行人工标注. 最后,本文基于规则自动将带标注的合约分解为数据样本,以便模型在可接受的硬件条件下输入待提取数据. 至此,本文在对原始文档文本进行最小程度修改的情况下构建了数据集,在除人工标注以外的流程中都实现了自动化,使数据集反映了智能合约场景下的自动化要素提取需求.

构建的数据集共包含9个大类和22个子类元素,总计122 687个数据样本. 表1列出了每个类别的数据样本数. 这里的无标签(No Label,NL)是指该token不属于任何元素,相当于NER中的O(Other,其他),不在所统计的9个大类元素中. 根据合约的数量,本文以7∶1.5∶1.5的比例将数据集分为训练集、测试集和开发集,确保每个集都包含所有类别. 然而,包含智能合约所需元素的文本在整个合约中非常稀疏,以至于近84.6%的句子不包含任何元素. 这些句子被称为only-NL数据. 将它们全部删除并不能很好地反映智能合约化场景,但保留它们将影响评估结果,因为所有被测试的方法都难以处理如此不平衡的数据. 因此,本文随机抽取5%的only-NL数据并将其留在集合中,形成具有23 582个样本的Part-NL数据集(训练集16 406个,开发集3 807个,测试集3 369个),并将其用作实验中的数据集1.

(2)数据集2

除了数据集1,本文也采用目前可下载的公开数据

①上海市政府采购网 http://www.ccgp-shanghai.gov.cn/site/home
重庆市政府采购网 https://www.ccgp-chongqing.gov.cn/
湖北省政府采购网 https://ccgp-hubei.gov.cn/

表1 数据集元素类别及样本数

类别		样本数量				
大类	小类	简称	训练集	测试集	验证集	总和
NL	NL	NL	84 997	18 420	19 270	122 687
甲方	甲方	PtA	721	147	154	1 022
乙方	乙方	PtB	772	154	168	1 094
签订时间	签订时间	Tm	494	107	105	706
支付	条件	PmC	1 095	235	246	1 576
	后置条件	PmP	105	26	25	156
	行为	PmB	1 048	248	243	1 539
发货	条件	AtC	224	43	42	309
	后置条件	AtP	98	24	23	145
	行为	AtB	325	67	60	452
验收	条件	DlC	185	36	41	262
	后置条件	DlP	69	9	21	99
	行为	DlB	299	57	70	426
终止违约条款	条件	TdC	1 310	320	270	1 900
	合同终止	TdT	1 228	289	250	1 767
	结果	TdR	1 437	330	275	2 042
非终止违约条款	条件	NdC	1 588	301	366	2 255
	后置条件	NdP	491	99	96	686
	结果	NdR	1904	371	422	2 697
标的	名称	SmN	1 413	224	351	1 988
	数量	SmQ	1 289	216	297	1 802
	单价	SmU	1 115	188	291	1 594
	总价	SmT	1 720	297	400	2 417
总计			84 997	18 420	19 270	122 687

集ETIP①[13]，其包含150份经人工标注的短保险合同，包含7种条款要素．具体来说，本文使用了由数据集作者划分的训练集和测试集作为数据集2，其大小分别为2 709和2 699条数据样本．由于CAM-CEE和大部分基线模型不是为嵌套NER设计的，本文将数据集中的嵌套元素归一化为具有最小范围的类型．这实际上增加了需提取的元素总数，使在数据集2上的提取更具挑战性．

4.1.3 基线模型

据本文作者所知，目前在中文CEE领域仅有TOI-CNN[13]可供开源测试，然该模型是为数据集2定制的，使其难以在数据集1上进行测试．但原始论文中给出了TOI-NN在数据集2的每种元素上的提取性能．因此，在分析不同元素性能差异的实验中，比较了TOI-NN和CAM-CEE(w)在数据集2上的性能．

在有效性研究中，由于中文CEE基线存在上述问题，本文选择了中文NER领域五个最新SOTA模型作为基线，包括MECT4CNER[25]、W2NER[23]、Graph4CNER[31]、NFLAT4CNER[32]和LWICNER[33]．

为保证公平，除了LWICNER之外，训练的批次(batch)大小和轮次(epoch)均一致设置．LWICNER例外的原因是，其源代码中表明该模型暂时仅在batch大小设置为1时适用．其他超参数则根据原始论文设置．在测试MECT4CNER时，由于其论文中使用的radials词典存在版权问题，本文使用了作者在开源代码中推荐的词典．在通用性研究中，本文使用最流行的数个类BERT模型作为基准模型，包括RoBERTa[34]、AlBERT[35]和DistilBERT[36]．

4.1.4 参数与模型设定

本文使用AdamWeightDecay优化器训练模型，学习率设置为2×10^{-5}，衰减率设置为0.01．在每个实验中，训练epoch数设置为20，batch大小设置为8，除了前面提到的LWICNER．对于提出的CAM-CEE，根据大多数句子的长度，数据集1的最大输入长度N设置为256，数据集2设置为128．嵌入模型基于huggingface的Transformers框架使用②，所有模型都在同一计算机上使用单个显存为24 GB的GTX3090Ti GPU进行训练．

在通用性实验中，预训练嵌入模型的权重来自huggingface，并使用Transformers初始化．本文使用“xlm-roberta-base”[37]作为RoBERTa模型，其由2.5 TB经过滤的多语言CommonCrawl数据预训练；将“uer/albert-base-chinese-cluecorpussmall”作为AlBERT模型，其基于py-uer[38]在CLUECorpusSmall[39]数据集上预训练；“distilbert-base-multilingual-cased”作为DistilBERT模型，其在Wikipedia语料库上预训练．

4.1.5 评估指标

本文使用四个指标来评估模型的要素提取性能．第一个是token级别的Accuracy：

$$\text{Accuracy}=\frac{\text{判断正确token数}}{\text{token总数}} \tag{16}$$

即通过将正确识别的token数量除以token总数获得．其他指标均为元素级的F_1-score，因为存在不包含任何元素的文本且元素分类较多，无法简单地使用Accuracy来计算元素级指标．F_1函数中的第一个是micro F_1，通过对所有元素进行F_1-score计算得到，即

$$\text{Precision}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{17}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{18}$$

$$F_1=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{19}$$

其中，TP表示所有真实存在且被正确提取的元素；FP表示实际不存在却被提取出的元素；FN表示实际存在

①https://github.com/ETIP-team/ETIP-Project/

②https://huggingface.co/

却未被提取出的元素. 第二个是macro F_1,通过单独计算每种类型的元素 F_1再取平均得到,即

$$\text{macro}F_1=\frac{\sum_{i=1}^{C}F_{1i}}{C} \tag{20}$$

其中,F_{1i}表示单独对L_i类元素执行F_1-score计算,反映了所有类别元素的平均性能,无论其数量多少. 除了常用的micro F_1和macro F_1外,本文还引入minimum F_1,从而以更细粒度的方式评估要素提取性能. 具体来说,单独计算每种类型的元素 F_1后,取所有元素分类中最小的F_1为minimum F_1,即

$$\text{minimum}F_1=\min\left(F_{11},F_{12},\cdots,F_{1C}\right) \tag{21}$$

从而使该指标反映了模型在最难提取的元素上的性能.

4.2　有效性实验

4.2.1　CAM-CEE(w)的有效性

如表2所示,当使用边缘加权损失在数据集1上进行训练时,所提出的CAM-CEE(w)在每个指标上都优于所有基线模型. 其Accuracy、micro、macro和minimum F_1分别比最佳基线高5.62、2.60、0.85和13.73个百分点. 其中,CAM-CEE(w)在minimum F_1上的优势尤为突出,这表明CAM-CEE能比SOTA基线模型更好地处理最难的元素. 总体而言,结果证明了CAM-CEE在智能合约化场景中的有效性. 在数据集2上,尽管CAM-CEE(w)在minimum F_1中没有达到最佳水平,但在大多数指标上仍优于所有基线方法. 具体来说,CAM-CEE框架的Accuracy、micro F_1和macro F_1分别比最佳基线高4.91、0.93和0.56个百分点,表明CAM-CEE架构在过度清洗的数据集中也极具竞争力.

表2　有效性实验结果

数据集	模型	Accuracy/%	micro F_1/%	macro F_1/%	minimum F_1/%	每轮训练时间/s	显存占用/GB
数据集1	MECT	86.25	74.95	70.10	38.62	52.00	16.540
	W2NER	/*	77.45	74.92	0	113.00	20.770
	Graph	85.44	70.68	66.87	41.90	88.00	1.040
	NFLAT	77.96	59.31	51.92	4.88	24.00	5.930
	LWICNER	81.20	70.08	68.50	35.90	376.00	1.120
	CAM-CEE(n)	91.74	78.45	73.20	49.12	206.00	8.960
	CAM-CEE(w)	91.87	80.05	75.77	55.63	205.00	8.960
数据集2	MECT	93.15	82.15	82.78	68.87	12.00	5.270
	W2NER	/*	93.65	93.09	85.45	16.00	7.830
	Graph	93.72	83.44	82.51	69.32	12.72	0.772
	NFLAT	91.86	78.25	67.10	26.41	6.70	2.310
	CAM-CEE(n)	98.71	94.39	93.67	78.26	19.00	4.960
	CAM-CEE(w)	98.63	94.58	93.65	78.26	19.30	4.960

注:*只有W2NER将填充token计入Accuracy计算中,与其他所有方法不同,故未计入.

4.2.2　CAM-CEE的有效性

如表2所示,即使统一使用普通的交叉熵损失函数训练,CAM-CEE框架仍然表现最佳. 在micro F_1和Accuracy方面,CAM-CEE(n)在两个数据集上都仍然优于所有基线. 具体来说,CAM-CEE(n)的micro F_1和Accuracy在数据集1上比最佳基线高1.0个百分点和5.49个百分点,在数据集2上比最佳基线高0.74个百分点和4.99个百分点.

对比剩余的指标,macro F_1和minimum F_1,CAM-CEE(n)至少在一个数据集上仍取得最佳性能,同时在另一个数据集中排名第二. 具体来说,数据集1上CAM-CEE(n)的minimum F_1比最佳基线高7.22个百分点,数据集2上的macro F_1比最佳基线高0.58个百分点. 实验结果表明,CAM-CEE即使在常规的训练方式下,也总体优于基线方法.

4.2.3　基线模型间的比较

在基线模型中,W2NER在两个数据集的大多数指标上都优于其他基线,但在数据集1上,其minimum F_1非常低. 这表明,尽管提取能力较佳,但在包含多种目标元素的复杂情况中,W2NER很难提取其中某些元素,这在智能合约化场景中可能是致命的,因为某种较棘手的要素可能完全无法被正确识别. 相对而言,Graph4NER、LWICNER和MECT4CNER的性能较中庸,但在所有指标上都没有致命缺陷.

4.2.4　训练效率

此外,本文还记录并列出了每个模型每轮平均训练时间和GPU显存使用情况. 虽然CAM-CEE的训练时间和显存使用不是最突出的,但它们均在可接受范围内,且表明CAM-CEE可在低价消费级显卡上进行训练. 与W2NER和MECT4CNER等性能最佳的基线相

比,CAM-CEE在显存占用方面甚至具有一定的优势. 考虑到智能合约化场景下的硬件条件和性能要求,10 GB以下的显存使用以及强大的性能使CAM-CEE具备竞争力.

4.2.5 生成式LLM测试

最后,本文在DeepSeek-V3上进行了实验,测试生成式LLM的提取性能. 具体而言,对于测试集中的每条样本,本文将所有要素类别(包括无要素文本)的描述、数据样本、要素提取要求与返回格式要求组合为提示(prompt)发送给DeepSeek-V3,并接收返回结果,从而测试生成式LLM的要素提取性能. 假如返回结果出现格式明显错误或请求超时现象,实验程序将丢弃错误结果并向DeepSeek重新发送请求,从而尽可能保证测试的可靠性. 最终,DeepSeek-V3在两个数据集上的token级Accuracy分别为35.33%和5.83%,显著低于所有基线模型;同时,每次发送提取请求-接收生成结果的时间花费在10~30余秒,效率较低且波动性强,表明生成式LLM在专业性较强的要素提取任务中可能表现不佳.

4.3 消融实验

如表3所示,显然,仅由BERT和全连接层组成的基本模型在两个数据集上性能较低. 特别是在元素级指标和数据集1上,基本BERT模型的性能显著低于大多数基线. 使用所提出的CDPM代替普通的填充方式可显著提高数据集1上BERT的性能,每个元素级指标均提升了超过4.79个百分点,而token级Accuracy则提高了2.6个百分点以上. 在数据集2上,micro F_1和macro F_1仍增加2个百分点以上,而minimum F_1下降. 其原因为,数据集2中最大的挑战来自条款元素之间的相似性,而CDPM带来的丰富上下文使得线性层难以处理小部分相似性强但样本数量少的条款. 总体而言,CDPM提高了模型性能,特别是在智能合约化场景中.

表3 消融实验结果

数据集	模型	Accuracy/%	micro F_1/%	macro F_1/%	minimum F_1/%	每轮训练时间/s	显存占用/GB
数据集1	BERT	89.17	69.92	65.37	33.40	187.0	8.77
	↑* +CDPM	91.83	75.87	70.34	38.19	191.0	8.77
	↑+ Bi-GRU	91.45	77.36	70.20	40.66	210.0	8.96
	↑+三重注意力(CAM-CEE(n))	91.74	78.45	73.20	49.12	206.0	8.96
	CAM-CEE(w)	91.87	80.05	75.77	55.63	205.0	8.96
数据集2	BERT	97.15	87.26	79.43	35.54	17.0	4.77
	↑*+CDPM	98.00	90.39	81.78	17.03	17.0	4.77
	↑+ Bi-GRU	98.11	89.43	92.83	78.25	19.0	4.96
	↑+三重注意力(CAM-CEE(n))	98.71	94.39	93.67	78.26	19.0	4.96
	CAM-CEE(w)	98.63	94.58	93.65	78.26	19.3	4.96

注:* ↑+表示在上一行的模型基础上添加.

其次,在数据集1上,Bi-GRU在micro F_1和minimum F_1上进一步提高了1.49个百分点和2.47个百分点,但在Accuracy和macro F_1上略微降低了0.38个百分点和0.14个百分点. 在数据集2上,Bi-GRU提高了Accuracy、minimum F_1和macro F_1. 考虑到性能的改进比降低显著得多,仅添加Bi-GRU层可在整体上提升性能,但仍有改进空间.

再次,与之前的模型相比,添加三重注意力机制,让模型成为所提出的CAM-CEE则可提高所有指标的性能. 在数据集1上,就Accuracy而言,模型恢复到BERT+CPDM的水平,差异小于0.1个百分点. 在三个元素级别的指标上,CAM-CEE实现了目前为止最高的性能,与之前模型的最佳性能相比,分别提高了1.09、3.00和8.46个百分点. 在数据集2上,CAM-CEE在每个指标上都取得了目前为止最好的结果. 这些结果证明了三重注意力机制的有效性,当其与Bi-GRU层结合时,可显著提高要素提取性能.

最后,要素边缘加权训练使CAM-CEE在智能合约化要素提取场景中表现更好,数据集1上的每个指标分别提升0.13、1.60、2.57和6.51个百分点. 这表明,在使用序列标注任务实现特征提取时,更多地关注真值或预测要素的边界可提高性能. 然而,在数据集2上,边缘加权函数训练的表现与普通的交叉熵函数训练非常接近,每个指标的差异均低于0.2个百分点. 这表明边缘加权函数的性能增益在两种场景下是不同的,在过度清洗的数据集上,它对CAM-CEE的影响较小.

同时,CAM-CEE的显存内存使用量为8.96 GB,仅比基本模型高出0.19 GB,并允许其在消费级GPU上运行. 两个数据集的训练时间分别平均增加约19 s和2 s,差距微小. 实验结果表明,CAM-CEE的模块组合对智能CEE而言是最优的.

4.4　通用性实验

4.4.1　数据集1实验结果

如表4所示,对于其他类BERT嵌入模型,所提出的框架也显著提高了它们的要素提取性能. 特别是在数据集1上,每种嵌入模型在每个指标上都有性能提升. 最显著的是micro F_1和macro F_1,在使用普通交叉熵函数训练时提升分别为3.9~8.0个百分点,在使用边缘加权函数情况下甚至更高(8~11个百分点). 使用普通交叉熵函数训练时,minimum F_1和Accuracy的提升相对不高,但仍然显著. 这些结果表明,所提出的框架可增强智能合约化要素提取场景中其他LLM的要素提取能力,具备在多种嵌入模型上应用的通用性.

表4　通用性实验结果

数据集	模型	Accuracy/%	micro F_1/%	macro F_1/%	minimum F_1/%	每轮训练时间/s	显存占用/GB
数据集1	RoBERTa基线	87.30	71.76	64.66	36.71	255.0	8.77
	融合CAM-CEE(n)	91.00	78.68	73.79	40.94	278.0	8.96
	融合CAM-CEE(w)	91.30	80.33	75.10	33.99	275.0	8.96
	AlBERT基线	87.16	64.03	61.46	27.05	155.0	2.77
	融合CAM-CEE(n)	90.52	72.02	65.45	31.82	175.0	2.96
	融合CAM-CEE(w)	90.95	75.41	69.94	29.12	174.0	2.96
	Distil基线	87.06	66.13	61.41	25.19	128.0	8.77
	融合CAM-CEE(n)	90.06	73.89	68.03	30.22	150.0	8.96
	融合CAM-CEE(w)	90.04	74.37	69.41	42.16	150.0	8.96
数据集2	RoBERTa基线	97.30	85.49	87.21	72.08	28.3	4.77
	融合CAM-CEE(n)	97.98	91.68	91.51	76.13	30.3	4.96
	融合CAM-CEE(w)	98.36	93.42	94.93	88.65	30.3	4.96
	AlBERT基线	97.35	85.31	86.08	76.18	12.2	1.77
	融合CAM-CEE(n)	98.34	90.34	89.62	71.15	14.2	1.96
	融合CAM-CEE(w)	98.26	90.66	90.37	78.26	14.3	1.96
	Distil基线	98.32	91.92	89.64	66.27	14.3	4.77
	融合CAM-CEE(n)	98.62	94.01	93.48	78.26	16.2	4.96
	融合CAM-CEE(w)	98.70	94.13	92.83	73.91	16.2	4.96

在边缘加权训练情况下,token级Accuracy的提高相对轻微,因为所提出的框架已经提供了强大的序列标注能力. 更重要的是,对于RoBERTa和AlBERT来说,使用边缘加权函数训练并没有提高他们的minimum F_1,这表明针对不同嵌入模型,边缘加权损失函数仍有改进空间. 考虑到RoBERTa模型比BERT更大,AlBERT则比BERT更小,可能根据模型的大小调整损失加权乘数将有助于处理智能合约化场景中最具挑战性的元素. 然而,边缘加权函数仍然实现了大多数指标的提升,表明其对其他模型的要素提取仍然有益.

4.4.2　数据集2实验结果

如表4所示,相比之下,类BERT的模型在数据集2上仍然从所提出框架上获得了性能提升,但不如数据集1上显著. micro F_1和macro F_1的提升范围为2~6个百分点,Accuracy则提高不到1.1个百分点. 与数据集1相比,数据集2上minimum F_1的性能提升颇具差异. 具体而言,对于RoBERTa和AlBERT来说,边缘加权训练只提高了数据集2上的minimum F_1性能. 对于DistilBERT来说,情况则相反. 显然,就小部分指标而言,模型和训练方法在两个数据集上的性能影响可能存在差异. 因为这两个数据集中反映了不同的场景——智能合约化要素提取场景和数据过度清洗场景. 同时,它们还包含不同类型的元素——粒度细、范围广的各种元素与仅包含粗粒度保险条款元素. 数据集和元素导致的具体性能差异及其原因将在下一节中进行分析.

4.5　不同元素的性能分析

4.5.1　数据集1中元素的性能

由于预定义元素的总类别数高达22个,本节具体分析CAM-CEE在各种元素上的性能. 如图5所示,在数据集1上,绝大多数(22个中的17个)类别的元素F_1超过60%,大多数(12个)类别甚至超过75%. 这些结果表明,在细粒度标准下,CAM-CEE仍然能很好地提取各种元素. 同时,不同类别的元素之间存在显著的性能差异. 例如,终止违约条款(TdC,TdT,TdR)的F_1比验收条款(AtC,AtP,AtB)的F_1更高.

为了找出不同类别之间性能差异背后的原因,本文将训练集中各种元素的数量放在图5中,并将其与F_1进行比较. 本文发现从支付条款(PmC)到终止违约条款结果(TdR),F_1和元素计数的变化趋势非常相似,而形式固定的元素,例如甲方(PtA)和乙方(PtB),尽管数

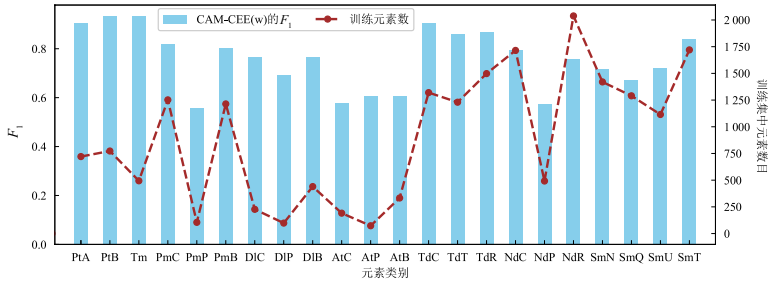

图5 数据集1各类元素的提取性能与元素数目

量并不显著，但F_1成绩最高．这些事实表明，数据集1的性能同时受到训练数据的计数和元素的形式影响．

4.5.2 数据集2中元素的性能

如图6所示，CAM-CEE(w)在数据集2的7大类别中的6个中实现了F_1≥90%，表现优于基线模型(TOI-CNN+LR)．同时，基线模型在不同元素上的性能差异与各种元素的数量趋势更为一致．对于CAM-CEE(w)，其在数据集2上的性能不平衡程度较轻，与元素数量的变化较不一致．

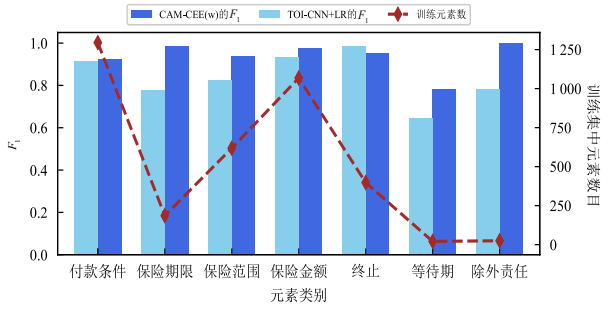

图6 数据集2各类元素的提取性能与元素数目

其原因为，首先，数据集2在构建中进行了数据的过度清洗，文本中的冗余语义被去除，导致CAM-CEE只需要相对较少的训练数据即可提取大多数特征；其次，数据集2中的要素仅包含保险合同中的不同条款，与数据集1中的要素相比，它们相对更相似，因此CAM-CEE在各元素上的较小性能不平衡源于不同目标元素特征之间的较小差异．总而言之，在数据集2中，对于性能更强的模型，文本的语义而不是样本的数量成为了不同元素性能差异的主要来源；相对而言，数据集1所代表的真实智能合约化场景包含更多冗余语义，为要素提取带来了更多挑战．

4.6 案例分析

本文使用CAM-CEE(w)、BERT基线方法和DeepSeek-V3生成式LLM在两篇真实法律合约上进行了要素提取拟真测试，从而评估各方法在真实应用场景下的效果．合约A包含2 852字，长度适中，代表大多数合约场景；合约B是数据集中最长的已标注合约，其字数超过41 000且包含大量标的条款信息，代表最具挑战性的提取场景．由于每篇合约要素数目较多，本文针对部分典型提取案例进行演示和分析，如图7所示，其中彩色部分文本表示模型识别出的不同要素．

首先，案例1是一句典型的非终止违约条款，包含条件和结果两种逻辑子句．在提取过程中，CAM-CEE(w)充分利用了上下文关联信息，成功提取总计四个逻辑子句；而基线方法在提取最后一个非终止违约结果时，上下文关联能力的不足导致无法正确区分要素所属的条款类型或逻辑子类，将大部分文本错误识别为了终止违约结果(条款类型错误)和非终止违约条件(逻辑子类错误)，导致提取失败．

其次，案例2反映了复杂上下文场景．具体而言，照片类型合约向文本的转换产生了大量无法通过自动化手段区分与去除的异常换行符(加粗转义字符“\n”)，将原文划分为了多个自然段．针对较复杂的跨段落上下文关联，CAM-CEE(w)仍然准确识别了四种标的信息．而基线方法在提取标的名称时失败，将名称的一部分和文本行编号错误识别为了标的总价，且误将半括号识别为标的数量．

再次，案例3则展示了仍待解决的挑战．CAM-CEE(w)在识别标的数量时未能正确排除半括号，该半括号是量词的一部分．基线模型除了在标的数量上出现同样错误以外，还未能识别标的名称．其原因在于，一方面，案例3位于标的内容中的第146行，与标的表头内容的字符距离远远超过了实验中设定的N，因此在填充时表头并未作为其上下文填入，导致CAM-CEE(w)无法从表头获取上下文关联信息，关联能力可能退化至基线模型水平；另一方面，几乎没有训练数据同时满足“标的信息与表头跨度未超过N”与“标的数量包含量词信息”两个条件，导致模型难以在训练中学习到与该案例类似的上下文关联，无法复刻针对标的名称的成功识别．这表明当上下文关联跨度过长(超过最大输入长度N)且训练集中缺少类似数据时，CAM-CEE(w)的上下文关联能力和提取性能可能退化至基线模型水平．因此，一方面，对于长度极长、信息量大、上下文关联跨度过长的

合约场景,有必要提出对应的长距离上下文关联方法;另一方面,需要构建更大规模的合约数据集与语料库,使模型得到更充分的训练.

最后,对于生成式LLM,其在案例2和案例3中受益于较长的输入长度和丰富的通用语料,成功识别了四种标的要素,但也将大量无关内容误判为标的数量和名称,其原因可能为,生成式LLM基于通用语料训练,因而在处理异常符号较多、结构较特殊的文本内容时易出错. 在案例1中,LLM虽然未出现误判问题,但在提取四个条款要素时均出现提取不完整现象. 特别是对于非终止违约条件要素,生成式LLM在提取中遗漏了"因乙方原因"关键信息,导致提取的元素语义出现关键差异. 上述现象表明,生成式LLM在专业化文档的要素提取过程中性能较不稳定,可能出现遗漏或误判现象,具体取决于上下文语境和文本包含的要素类型.

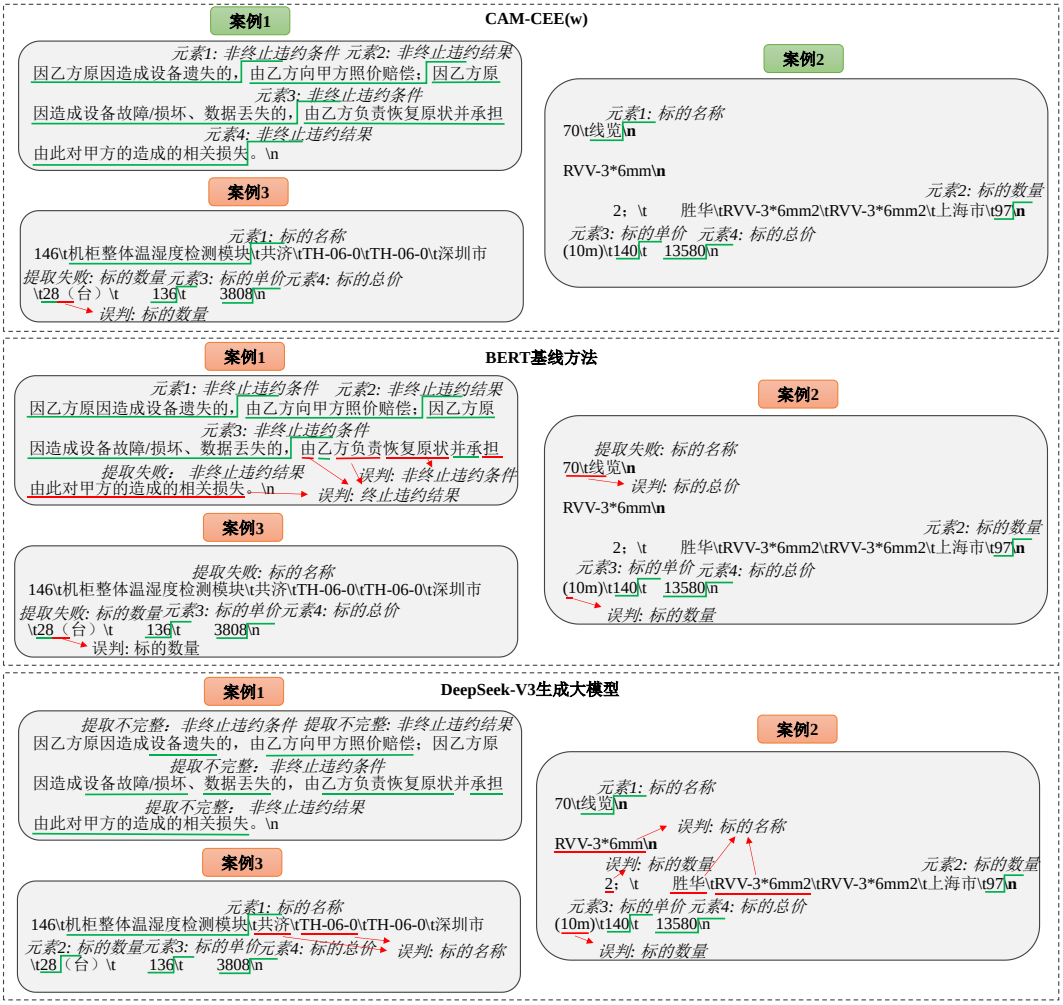

图7　要素提取案例

5　结论

本文针对智能合约化场景中文本文档要素提取的需求和挑战,提出了新颖的上下文语义感知的动态填充方法(CDPM)、三重注意力层和要素边缘加权损失函数模块,在不增加训练成本的前提下提供了额外上下文信息,增强了上下文感知能力,并提升了序列标注范式下的要素提取能力. 基于上述模块和BERT嵌入模型,提出了要素提取最优解模型CAM-CEE. 在模型实验过程中,本文发现可用中文要素提取数据集较少、质量较低,因此从合同文档入手,收集了64 314份真实合同,构建了首个中文买卖CEE数据集,该数据集包含多达122 687条数据和22类元素. 大量实验表明,在所实验的数据集中,所提出CAM-CEE的要素提取性能超越了所有被测试的基线模型,且本文提出的CAM-CEE能提高其他嵌入模型的性能,具备高度通用性. 未来,拟从多个维度扩展数据集,如规模、粒度以及元素和文档的类别;在方法方面,拟尝试进一步解决元素样本不平

衡的问题,实现更长跨度的上下文关联,并应用新技术从而更有效地提取被智能合约所需的元素.

参考文献

[1] FANG P C, ZOU Z H, XIAO X S, et al. iSyn: Semi-automated smart contract synthesis from legal financial agreements[C]//Proceedings of the 32nd ACM SIGSOFT International Symposium on Software Testing and Analysis. New York: ACM, 2023: 727-739.

[2] CHALKIDIS I, ANDROUTSOPOULOS I, MICHOS A. Extracting contract elements[C]//Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law. New York: ACM, 2017: 19-28.

[3] RADFORD A, NARASIMHAN K. Improving language understanding by generative pre-training[C]//Proceedings of the 2018 Conference on Neural Information Processing Systems. Montreal: Curran Associates, 2018.

[4] BI D-A X, CHEN D, CHEN G, et al. DeepSeek LLM: Scaling open-source language models with longtermism[J/OL]. (2024-01-05)[2025-03-17]. https://api.semanticscholar.org/CorpusID:266818336.

[5] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: Curran Associates Inc, 2017: 6000-6010.

[6] LEE J, YI J S, SON J. Development of automatic-extraction model of poisonous clauses in international construction contracts using rule-based NLP[J]. Journal of Computing in Civil Engineering, 2019, 33(3): 04019003.

[7] KIM Y, LEE J, LEE E B, et al. Application of natural language processing (NLP) and text-mining of big-data to engineering-procurement-construction (EPC) bid and contract documents[C]//2020 6th Conference on Data Science and Machine Learning Applications (CDMA). Piscataway: IEEE, 2020: 123-128.

[8] PADHY J, JAGANNATHAN M, KUMAR DELHI V S. Application of natural language processing to automatically identify exculpatory clauses in construction contracts[J]. Journal of Legal Affairs and Dispute Resolution in Engineering and Construction, 2021, 13(4): 04521035.

[9] GAO X, SINGH M P. Extracting normative relationships from business contracts[C]//International conference on Autonomous Agents and Multi-Agent Systems. Paris: International Foundation for Autonomous Agents and Multiagent Systems, 2014: 101-108.

[10] JAFARI P, AL HATTAB M, MOHAMED E, et al. Automated extraction and time-cost prediction of contractual reporting requirements in construction using natural language processing and simulation[J]. Applied Sciences, 2021, 11(13): 6188.

[11] ILIAS C, ION A. A deep learning approach to contract element extraction[M]//Legal Knowledge and Information Systems. Awsterdam: IOS Press, 2017: 155-164.

[12] ZHANG Q Q, XUE C, SU X, et al. Named entity recognition for Chinese construction documents based on conditional random field[J]. Frontiers of Engineering Management, 2023, 10(2): 237-249.

[13] ZHANG K, SUN L, JI F L. A TOI based CNN with location regression for insurance contract analysis[C]//2019 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE, 2019: 1-8.

[14] WANG Z H, SONG H Y, REN Z C, et al. Cross-domain contract element extraction with a bi-directional feedback clause-element relation network[C]//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2021: 1003-1012.

[15] AEJAS B, BOURAS A, BELHI A, et al. Smart contracts implementation based on bidirectional encoder representations from transformers[M]//Product Lifecycle Management, Green and Blue Technologies to Support Smart and Sustainable Organizations. Cham: Springer International Publishing, 2022: 293-304.

[16] LEIVADITI S, ROSSI J, KANOULAS E. A benchmark for lease contract review[EB/OL]. (2020-10-20)[2025-03-17]. https://arxiv.org/abs/2010.10386v1.

[17] GARCÍA-BARRAGÁN Á, CALATAYUD A G, PRIETO-SANTAMARÍA L, et al. Step-forward structuring disease phenotypic entities with LLMs for disease understanding[C]//2024 IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS). Piscataway: IEEE, 2024: 213-218.

[18] XU D R, CHEN W, PENG W J, et al. Large language models for generative information extraction: A survey[J]. Frontiers of Computer Science, 2024, 18(6): 186357.

[19] GOPALAKRISHNAN S, GARBAYO L, ZADROZNY W. Causality extraction from medical text using large language models (LLMs)[J]. Information, 2025, 16(1); 13.

[20] SAIER T, OHTA M, ASAKURA T, et al. HyperPIE: Hy-

perparameter information extraction from scientific publications[M]//Advances in Information Retrieval. Cham: Springer Nature Switzerland, 2024: 254-269.

[21] 刘小明. 任务协作表示增强的要素及关系联合抽取模型[J]. 电子学报, 2024, 52(6): 1955-1962.
LIU X M. Task collaboration representation enhanced model for element and relation joint extraction[J]. Acta Electronica Sinica, 2024, 52(6): 1955-1962. (in Chinese)

[22] DAGDELEN J, DUNN A, LEE S, et al. Structured information extraction from scientific text with large language models[J]. Nature Communications, 2024, 15, 1418. .

[23] LI J Y, FEI H, LIU J, et al. Unified named entity recognition as word-word relation classification[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(10): 10965-10973.

[24] 孟伟伦. 基于字形特征的中文医学命名实体识别方法[J]. 电子学报, 2024, 52(6): 1945-1954.
MENG W L. Chinese medical named entity recognition method based on glyph features[J]. Acta Electronica Sinica, 2024, 52(6): 1945-1954. (in Chinese)

[25] WU S, SONG X N, FENG Z H. MECT: Multi-metadata embedding based cross-transformer for Chinese named entity recognition[C]//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). Stroudsburg: PA, 2021: 1529-1539.

[26] AEJAS B, BELHI A, BOURAS A. Toward an NLP approach for transforming paper contracts into smart contracts[M]//Intelligent Sustainable Systems. Singapore: Springer Nature Singapore, 2023: 751-759.

[27] GLOROT X, BENGIO Y. Understanding the difficulty of training deep feedforward neural networks[J]. Journal of Machine Learning Research, 2010, 9: 249-256.

[28] LUONG T, PHAM H, MANNING C D. Effective approaches to attention-based neural machine translation[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Lisbon: Association for Computational Linguistics, 2015: 1412-1421.

[29] ALTURAYEIF N, LUQMAN H. Fine-grained sentiment analysis of Arabic COVID-19 tweets using BERT-based transformers and dynamically weighted loss function[J]. Applied Sciences, 2021, 11(22): 10694.

[30] LENG Z, TAN M, LIU C, et al. PolyLoss: A polynomial expansion perspective of classification loss functions[C]//Proceedings of the 10th International Conference on Learning Representations. Virtual Event: OpenReview.net, 2022: 25-29.

[31] SUI D B, CHEN Y B, LIU K, et al. Leverage lexical knowledge for Chinese named entity recognition via collaborative graph network[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg: Association for Computational Linguistics, 2019: 3828-3838.

[32] WU S, SONG X N, FENG Z H, et al. NFLAT: Non-flat-lattice transformer for Chinese named entity recognition[EB/OL]. (2020-05-12) [2025-03-17]. http://dx. doi. org/10.48550/ARXIV.2205.05832.

[33] WU W J, ZHANG C Y, NIU S Z, et al. Unify the usage of lexicon in Chinese named entity recognition[M]//Database Systems for Advanced Applications. Cham: Springer Nature Switzerland, 2023: 665-681.

[34] LIU Y H, OTT M, GOYAL N, et al. RoBERTa: A robustly optimized BERT pretraining approach[EB/OL]. (2019-07-26)[2025-03-17]. https://arxiv.org/abs/1907.11692v1.

[35] LAN Z Z, CHEN M D, GOODMAN S, et al. ALBERT: A lite BERT for self-supervised learning of language representations[EB/OL]. (2020-02-09) [2025-03-17]. https://arxiv.org/abs/1909.11942v6.

[36] SANH V, DEBUT L, CHAUMOND J, et al. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter[EB/OL]. (2019-10-02) [2025-03-17]. http://arxiv.org/abs/1910.01108.

[37] CONNEAU A, KHANDELWAL K, GOYAL N, et al. Unsupervised cross-lingual representation learning at scale[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2020: 8440-8551.

[38] ZHAO Z, CHEN H, ZHANG J B, et al. UER: An open-source toolkit for pre-training models[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations. Stroudsburg: Association for Computational Linguistics, 2019: 241-246.

[39] XU L, ZHANG X W, DONG Q Q. CLUECorpus2020: A large-scale Chinese corpus for pre-training language model[EB/OL]. (2020-05-05) [2025-03-17]. https://arxiv. org/abs/2003.01355v2.

作者简介

钱 肖 男,1999年出生于浙江省湖州市.现为西安电子科技大学网络与信息安全学院硕士研究生.主要研究方向为自然语言处理.
E-mail: xqian@stu.xidian.edu.cn

蒋忠元 男,1988年出生于陕西省榆林市.现为西安电子科技大学网络与信息安全学院教授、博士生导师.主要研究方向为复杂(无人)网络系统安全、大数据与AI安全、空天地一体化网络技术、法律自然语言处理.
E-mail: zyjiang@xidian.edu.cn

陶梅悦 女,2001年出生于安徽省芜湖市.现为西安电子科技大学网络与信息安全学院硕士研究生.主要研究方向为软件可靠性、异常检测.
E-mail: mytao@xidian.stu.edu.cn

刘柄呈 男,2001年出生于甘肃省庆阳市.现为西安电子科技大学网络与信息安全学院硕士研究生.主要研究方向为自然语言处理.
E-mail: bingchengliu@stu.xidian.edu.cn

李任翔 男,2001年出生于河南省漯河市.现为西安电子科技大学网络与信息安全学院硕士研究生.主要研究方向为自然语言处理和智能合约.
E-mail: renxiangli@stu.xidian.edu.cn

高 胜 男,1987年出生于湖北省黄冈市.现为中央财经大学信息学院教授、博士生导师.主要研究方向为区块链、智能合约、联邦计算和大语言模型安全.
E-mail: sgao@cufe.edu.cn

马建峰 男,1963年出生于陕西省西安市.现为西安电子科技大学网络与信息安全学院教授、博士生导师.主要研究方向为网络安全、系统安全、数据安全和无人机安全.中国电子学会会员编号:E190004733F.
E-mail: jfma@mail.xidian.edu.cn