2025 年 第 25 卷 第 18 期
2025, 25(18):07668-10

科 学 技 术 与 工 程
Science Technology and Engineering

ISSN 1671—1815
CN 11—4688/T

DOI:10.12404/j.issn.1671-1815.2404063

引用格式:王祎涵, 张思佳, 曹恒, 等. 融合 *N*-Gram 和多重注意力机制的能源领域新词发现方法[J]. 科学技术与工程, 2025, 25(18): 7668-7677.
Wang Yihan, Zhang Sijia, Cao Heng, et al. New word discovery method in the energy field combining *N*-Gram and multiple attention mechanism[J]. Science Technology and Engineering, 2025, 25(18): 7668-7677.

自动化技术、计算机技术

融合 *N*-Gram 和多重注意力机制的能源领域新词发现方法

王祎涵[1], 张思佳[1,2,3*], 曹恒[4], 刘珈宁[1], 张正龙[1]

(1. 大连海洋大学信息工程学院/辽宁省海洋信息技术重点实验室, 大连 116023;
2. 设施渔业教育部重点实验室(大连海洋大学), 大连 116023;
3. 大连市智慧渔业重点实验室, 大连 116023; 4. 中国科学院大连化学物理研究所, 大连 116023)

摘 要 随着能源行业的快速发展和技术革新,大量的专业术语和表达方式不断更新,新词不断涌现。然而,传统的新词发现方法通常依赖于词典或规则,且难以高效率地处理和更新大量的专业术语,特别是在快速变化的能源领域。因此,结合能源领域文本数据特性,提出了一种融合 *N*-Gram 和多重注意力机制的能源领域新词发现方法(new word discovery method in the energy field combining *N*-Gram and multiple attention mechanism,ENFM)。该方法首先利用 *N*-Gram 模型对能源领域的文本数据进行初步处理,通过统计和分析词频来生成新词候选列表。随后,引入融合多重注意力机制的 ERNIE-BiLSTM-CRF 模型,以进一步提升新词发现的准确性和效率。与传统的新词发现技术相比,在新词的准确识别和整体效率上均有显著提升,将其于能源领域政策文本数据集,准确率、召回率和 F_1 分别为 95.71%、95.56%、95.63%。实验结果表明,该方法能够准确地在能源领域的大量文本数据中识别新词,有效识别出能源领域特有的词汇和表达方式,显著提高了中文分词任务中对能源领域专业术语的识别能力。

关键词 能源领域; 新词发现; 预训练模型; *N*-Gram; 中文分词

中图法分类号 TP391.1; **文献标志码** A

New Word Discovery Method in the Energy Field Combining *N*-Gram and Multiple Attention Mechanism

WANG Yi-han[1], ZHANG Si-jia[1,2,3*], CAO Heng[4], LIU Jia-ning[1], ZHANG Zheng-long[1]

(1. College of Information Engineering, Dalian Ocean University/Liaoning Key Laboratory of Marine Information Technology, Dalian 116023, China;
2. Key Laboratory of Environment Controlled Aquaculture (Dalian Ocean University), Ministry of Education, Dalian 116023, China;
3. Dalian Key Laboratory of Smart Fisheries, Dalian 116023, China; 4. Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian 116023, China)

[**Abstract**] With the rapid development of the energy industry and technological innovation, a large number of professional terms and expressions are constantly updated, and new words continue to emerge. However, traditional neologism discovery methods often rely on dictionaries or rules, and it is difficult to efficiently process and update a large number of specialized terms, especially in the rapidly changing energy field. Therefore, combined with the characteristics of text data in the energy field, a new word discovery method in ENFM(energy field combining *N*-Gram and multiple attention mechanism) was proposed. Firstly, the *N*-Gram model was used to process the text data in the field of energy, and the candidate list of new words was generated by statistics and analysis of word frequency. Subsequently, the ERNIE-BiLSTM-CRF model integrating multiple attention mechanism was introduced to further improve the accuracy and efficiency of neologism discovery. Compared with the traditional neologism discovery technology, the accurate identification and overall efficiency of neologism have been significantly improved. The accuracy rate, recall rate and F_1 value of neologism in the data set of policy text in the energy field are 95.71%, 95.56% and 95.63%, respectively. The experimental results show that this method can accurately identify new words in a large number of text data in the field of energy, effectively identify the specific words and expressions in the field

收稿日期: 2024-05-31 **修订日期**: 2025-03-05

基金项目: 辽宁省教育厅高等学校基本科研项目面上项目(LJKMZ20221095);辽宁省重点研发计划(2023JH26/10200015);中国科学院战略性先导科技专项(A 类)(XDA21000000)

第一作者: 王祎涵(2000—),女,汉族,山东威海人,硕士研究生。研究方向:数据挖掘。E-mail:1305042661@qq.com。

* **通信作者**: 张思佳(1982—),女,汉族,吉林白城人,博士,副教授。研究方向:自然语言处理、知识图谱。E-mail:zhangsijia@dlou.edu.cn。

of energy, and significantly improve the recognition ability of professional terms in the field of energy in Chinese word segmentation tasks.
[**Keywords**]　energy field; new word discovery; pre-trained model; *N*-Gram; Chinese word segmentation

碳达峰、碳中和是党中央经过深思熟虑作出的重大战略决策,科学有序推进能源结构及相关工业体系从高碳向低碳、绿色发展,逐步构建“清洁低碳、安全高效”的新型能源体系,才能助力实现“双碳”目标,支撑中国经济社会高质量发展,其中科技创新必须发挥至关重要的引领作用,推动多能融合科技发展路径研究具有重要意义[1]。在这一背景下,如何有效地处理和分析大量的政策文本和文献资料已成为自然语言处理(natural language processing,NLP)研究的一个重要方向[2]。能源政策文本中蕴含丰富的能源技术数据,通常包含大量特定的术语与缩写,结构关系较复杂、个性化突出,这对新词发现提出了更高的要求。在能源领域中,新词发现不仅是智能化能源管理系统的基础任务,也是一个关键任务[3],且在能源技术创新路径分析和能源知识图谱构建等任务中发挥着至关重要的作用。

目前新词发现算法可分为 3 类:基于统计的方法、基于规则的方法和基于模型的方法[4]。基于统计的方法通过计算词语在文本中的出现频率和词语之间的互信息来识别新词,但这种方法可能会误将高频使用的专业术语或习惯用语错误识别为新词,尤其是在没有足够上下文信息的情况下。申兆媛等[5]提出一种面向特定领域语料特性的新词识别方法,以获得该领域的新词集合,将新词定义为现有领域词典内的未登录词。然而,该方法忽视了那些虽未收录于词典中但实际上符合新词定义的情况。

基于规则的方法通过规则从文本中匹配和筛选出新词[6]。Zhang 等[7]提出了一种无监督领域新词发现方法,通过 LDA 模型分析文本主题,从而辅助识别与主题相关的潜在新词。但是基于规则的方法依赖于规则库,导致一些新词被遗漏。祝钰莹等[8]提出了一种基于信息熵-切分概率模型的中文新词发现方法。虽然该方法能够有效处理较短文本或具有高频出现的新词,但对长文本的处理能力有限,且依赖于语料库质量。

基于模型的方法通过深度学习模型来自动学习和识别新词,利用神经网络的强大建模能力,将大规模有标签的数据集进行训练,从而提高学习效率。耿骞等[9]通过利用 Word2vec 算法进行词向量嵌入以捕捉词语之间的上下文语义关系,然后通过 BiLSTM-Attention-CRF 算法从序列数据中识别出新词结构,最后使用 *K*-means 算法进行聚类,以提取出特定领域的关键新词集合,但在特定领域的长文本中,该方法可能难以充分理解某些语境下的新词含义。刘凡平等[10]利用 BERT 模型对句意的较强理解能力,将新词识别任务转化为分类任务来完成新词的识别,在开放性领域的数据集上拥有更高的精准率和 F_1,但对于特定语境下产生的新词可能存在一定的偏差。

在能源领域,张一帆等[11]提出一种基于条件随机场(conditional random field,CRF)和词向量的能源政策新词发现方法,该方法融合了种子词典与条件随机场,并通过词向量筛选实现新词挖掘。但该方法依赖于种子词典的质量和完整性,如果种子词典不够全面会导致新词发现的准确率受限。因此,需要一种更加高效、准确的新词发现方法,以适应能源领域的快速发展和变化。

综上所述,现有的新词发现方法常常不正确地拆分词组,导致重要的词结构信息丢失,进而影响到词语的真正含义。这些问题不仅降低新词发现技术在能源领域中的有效性和可靠性,而且会影响技术和政策的准确实施。针对上述问题,现提出一种融合 *N*-Gram 和多重注意力机制的能源领域新词发现方法(new word discovery method in the energy field combining *N*-Gram and multiple attention mechanism,ENFM),通过结合无监督和深度学习的方法,利用大规模的专业语料库进行验证和集成,提高能源领域专业术语词汇的识别准确性和效率。

为克服传统新词发现方法在处理单词级别信息的局限性,现引入 *N*-Gram 模型。通过多个连续词语的联合特征捕捉局部上下文信息,避免分词错误带来的信息丢失。为提升新词识别中对复杂上下文理解,引入多重注意力机制。模型从多个维度深入挖掘文本新词,动态调整对词语及其上下文的关注,增强对全局信息的掌握,尤其在能源领域专业文献中,能够精准识别潜在新词。为解决能源领域数据中复杂语义依赖和特征提取不足的问题,结合 *N*-Gram 与融合多重注意力机制的 ERNIE-BiLSTM-CRF 模型。*N*-Gram 模型揭示数据中潜在结构特征,为模型提供更为丰富的输入特征,而 ERNIE-BiLSTM-CRF 模型借助 ERNIE 语义理解与生成能力,同时通过双向 LSTM 网络和 CRF 的结合,有效捕捉长距离的依赖和全局上下文信息。

1　DLOU-NW 数据集构建

1.1　数据预处理

数据采集是构建能源技术词条语料库的第一

步。实验所用数据从科学技术部、国家能源局等官方网站获取相关政策文件,主要包括《科技支撑碳达峰碳中和实施方案(2022—2030 年)》《江苏省科技支撑碳达峰碳中和实施方案》等 18 份政策。

由于不同来源的数据存在格式差异,需要将原文本数据格式统一转换为. txt 标准化格式,以便于后续的整合和分析。数据处理流程如图 1 所示。图 1采用的预处理步骤如下。

(1)对于采集到的原始非结构化文本,首先去除文本中的停用词,并移除图表等非文本元素,之后将所有数据转换成标准的. txt 格式,得到纯文本数据。

(2)构建政策文本段落化自动切分工具,通过对政策文件进行智能分析,将步骤(1)得到的文本数据按照预定规则切分为多个简洁明了的段落,形成一次数据语料库。

(3)将步骤(2)得到语料库分成发布机构、发布时间、总体要求、重点任务、保障措施等模块。经过分析语料特点,采用人机协同的方式将重点任务模块作为核心语段进行识别抽取,再进行人工筛选,得到核心语段文本数据集。

(4)使用正则表达式来匹配标点符号,并将步骤(3)得到的文本数据分割成句子,得到二次文本数据集,形成能源技术词条语料库。

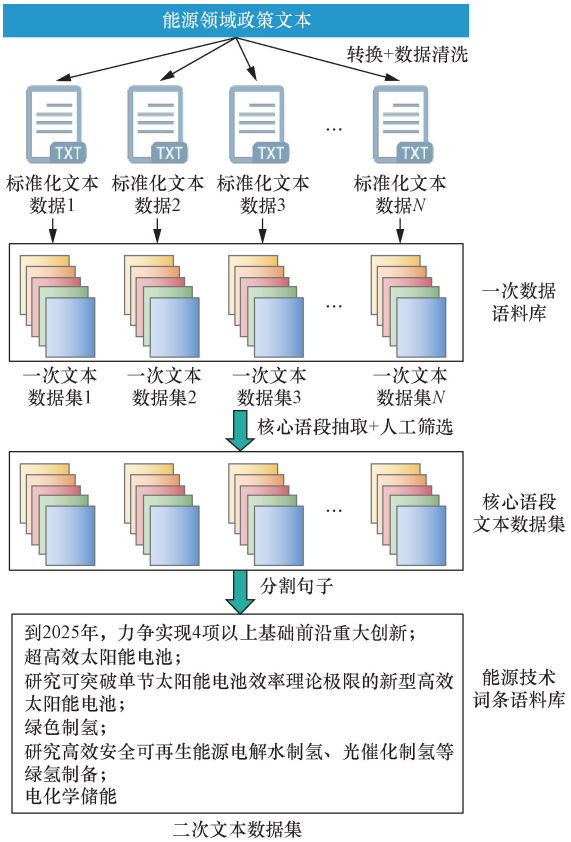


图 1 数据处理流程

Fig. 1 Data processing flow

1.2 数据标注流程

数据标注是指给原始数据(如图像、视频、文本等)添加标签的过程,经过这一步骤处理后的数据被称为训练数据。这些标签在机器学习模型中发挥着重要作用,它们使模型能够在遇到未曾接触过的数据时,依然能够精准识别数据中的核心内容[12]。数据标注流程如图 2 所示。

将 DLOU-NW 数据集按照训练集:测试集:验证集为 8:1:1 的比例划分,并进行标注,具体标签类别如表 1 所示。

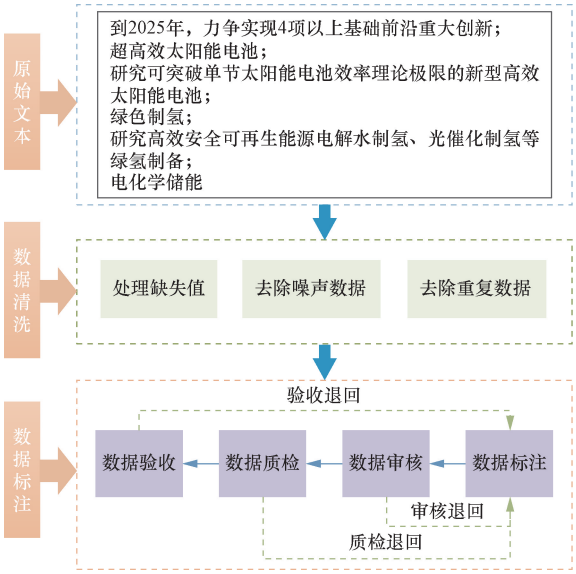


图 2 数据标注流程

Fig. 2 Data annotation process

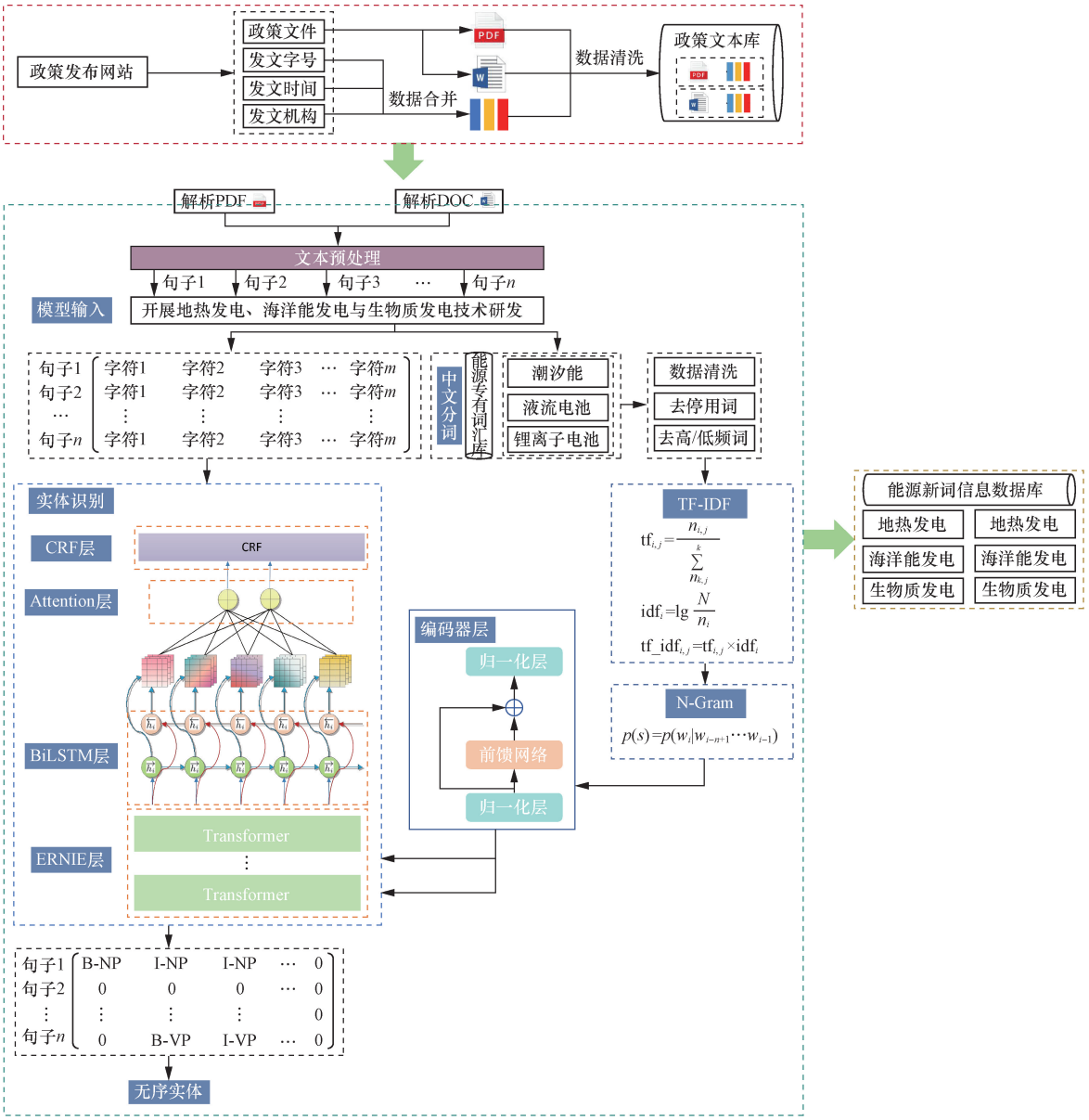
表 1 DLOU-NW 数据集具体标签类别

Table 1 Specific label categories of the DLOU-NW dataset

序号	具体标签类别	符号表示
1	修饰词短语	PP
2	名词短语	NP
3	动词短语	VP
4	动名词短语	VNP

2 ENFM 模型

由于中文缺乏明确的词边界,与基于空格分词的英文相比,直接采用原始的、针对英文设计的预训练模型在中文上往往不能达到最佳效果[13],这一问题在处理包含众多专业术语的能源文本时尤为突出。为了解决现有分词工具通常训练于通用语料库,缺少对能源领域特定术语和新兴词汇的认知的问题,提出了融合 *N*-Gram 和多重注意力机制的能源领域新词发现模型 ENFM,该模型通过融合能源领域的专业词典和使用专门训练的注意力机制,能够更好地识别和理解领域内的新词和术语。模型结构如图 3 所示。



$\mathrm{tf}_{i,j}$ 为词 i 在文档 j 中的频率占比;$n_{i,j}$ 为词 i 在文档 j 中的出现次数;$\sum_{n_{k,j}}^{k}$ 为文档 j 中所有词的总出现次数;idf_i 为逆文档频率;N 为文档总数量;n_i 为包含词 i 的文档数量;$\mathrm{tf\text{-}idf}_i$ 为词 i 在文档 j 中的最终权重;w_i 为当前待预测的词;$p(s)$ 为在给定前 $n-1$ 个词时,当前词 w_i 的条件概率

图 3　ENFM 模型框架图

Fig. 3　ENFM model framework diagram

ENFM 模型是由以下三部分组成。

(1)第一部分是基于 *N*-Gram 模型的编码器(NGE),使用 *N*-Gram 模型生成候选词汇列表,分析文本中的连续字符序列来识别潜在的新词。通过设置合适的 N 值,可以控制模型对上下文的敏感度,从而捕捉到不同长度的词汇模式。

(2)第二部分通过将融合多重注意力机制的 ERNIE-BiLSTM-CRF 模型(EBAC)的编码器层与 NGE 得到的特征拼接,形成一个综合的特征表示。其中 EBAC 模型提供的深度语义特征与 NGE 的局部上下文特征进行线性组合,增强了特征间的互补性,使模型更有效地捕捉文本中的细粒度信息。融合后的特征向量输入到 BiLSTM 层,优化了来自不同来源的特征集成,并且能深入捕获上下文相关语义特征。同时,NGE 弥补了 ERNIE 模型在处理未知或罕见词汇时的不足,提升了模型处理复杂文本的能力,有助于提高新词发现的准确性与鲁棒性。

(3)第三部分是 EBAC 模型,其通过在模型之间融合多重注意力机制来提取文本中的局部特征,不仅增强了模型对文本细节的敏感性,也提高了训

练过程中的并行处理能力,从而优化了整体的运算效率。合并后的特征用于训练一个 CRF 层,该层最终输出最优的标签序列,完成新词发现任务。

2.1 基于 *N*-Gram 的编码器(NGE)

在自然语言处理领域,*N*-Gram 是一种易于理解的统计模型,其核心概念是采用大小为 *N* 的滑动窗口对数据中的连续字节序列进行处理,生成一系列长度为 *N* 的字节片段,通常是词或字符[14]。每个这样的片段被称为 gram。随后,对这些 gram 的出现次数进行统计,并根据预定阈值进行筛选,从而得到关键 gram 列表,该列表构成了文本的向量特征空间,每个 gram 对应一个特征向量的维度。*N*-Gram 模型通过捕捉这些序列来获取文本的局部上下文信息,有助于理解和预测语言中的模式和结构。

为了解决过多自由参数的问题,引入了马尔科夫假设。假设句子 U 是有词序列 $W_1, W_2, \cdots, W_n$ 组成,文本中的每个词 W_i 和前面 $N-1$ 个词有关,而与更前面的词无关。这种假设被称为 $N-1$ 阶马尔科夫假设,对应的语言模型称为 *N* 元模型。*N* 元模型的具体表达式如表 2 所示。

利用 *N*-Gram 模型获取能源政策文本中连续出现的 *N* 个词(*N* 的取值为 1,2,3,4,5),将获取的 *N* 个词作为一个 *N* 元词串,*N* 元词串具体举例如表 3 所示。

表 2 *N* 元模型表达式

Table 2 *N*-gram model expression

模型	参数形式
一元模型	$P(W_i)$
二元模型	$P(W_i \mid W_{i-1})$
三元模型	$P(W_i \mid W_{i-2}W_{i-1})$
⋮	⋮
N 元模型	$P(W_i \mid W_{i-n+1}\cdots W_{i-1})$

表 3 *N* 元词串具体举例

Table 3 Specific examples of *N*-gram sequences

文本	近零排放的煤制清洁燃料和化学品技术
分词结果	近零排放的 煤制 清洁燃料 和化学品技术
$N=1$	(近零排放的)(煤制)(清洁燃料)(和)(化学品)(技术)
$N=2$	(近零排放的煤制)(煤制清洁燃料)(清洁燃料和)(和化学品)(化学品技术)
$N=3$	(近零排放的煤制清洁燃料)(煤制清洁燃料和)(清洁燃料和化学品)(和化学品技术)
$N=4$	(近零排放的煤制清洁燃料和)(煤制清洁燃料和化学品)(清洁燃料和化学品技术)
$N=5$	(近零排放的煤制清洁燃料和化学品)(煤制清洁燃料和化学品技术)

通过将文本划分为 *N*-Gram 词元,构建词汇表并统计各个 *N*-Gram 词元出现的次数,进而通过特征计算的方式将文本表示成维度固定的向量,并用于文本的新词发现和其他各项任务中。

在基于 *N*-Gram 的编码器部分,根据现有的 DLOU-NW 数据集构造词汇表,并统计 *N*-Gram 的频度。在词汇表和 *N*-Gram 频度的基础上,使用滑动窗口的方式,从文本中提取出所有的 *N*-Gram。在模型训练阶段,将单字序列和带 *N*-Gram 标记的文本数据作为模型的输入。之后将 EBAC 模型的编码器层与 NGE 得到的特征通过拼接形成一个综合的特征表示,每个来自 EBAC 的编码器输出向量与相应的 NGE 输出向量并排排列,然后通过一个连接操作将它们线性组合成一个单一的长向量,实现多尺度特征提取,可以提高模型在不同领域的适应性,特别是在能源领域,这种方法可以更好地捕捉领域特有的词汇和表达,有助于发现更多潜在的新词,从而提高新词发现的效果。

2.2 融合多重注意力机制的 ERNIE-BiLSTM-CRF 的新词发现模型(EBAC)

传统词嵌入方法 Word2Vec 未考虑词的位置信息,在词嵌入过程中,由于词在不同位置中表达的含义可能不同,易出现一词多义的问题[15]。该模型结合了 ERNIE 模型、BiLSTM 模型和 CRF 模型,并引入了多重注意力机制以提高模型的准确性和灵敏度。

2.2.1 ERNIE 层

ERNIE 是在 BERT 模型的基础上进行改进,采用 Transformer 作为基本结构,选用 ERNIE 模型作为词嵌入过程,是因为 ERNIE 模型在预训练阶段融合了实体等知识信息[16],因此它在理解和表示词汇时,能够更好地捕捉到词汇的语义和上下文信息,这有助于在复杂文本中识别出新词。

在 ERNIE 层,选择使用 ERNIE2.0 模型而非最新的 ERNIE3.0 模型,主要是因为 ERNIE2.0 模型表现出了更好的性能。虽然 ERNIE3.0 提供了知识增强和多任务学习等优势,但由于能源领域的文本数据往往包含大量的专业术语和复杂的表达方式,且在通用语料中并不常见,因此,ERNIE2.0 模型基于大量的中文语料进行预训练,能够更好地捕捉中文文本的语义信息,可以更有效地识别能源领域的新词和专有名词。

与之前的模型不同,ERNIE 模型通过对海量数据中的实体概念等先验语义知识进行建模。例如,在处理海[mask]能和气[mask]堆这类词汇时,尽管 BERT 模型可以通过字的搭配推测掩码字的信息,

却未能显式地对如海洋能、气冷堆等语义概念单元及其语义关系进行建模。与此相反,ERNIE 使用的掩码语言模型不仅覆盖单个字或词,还包括完整的词语和命名实体。这种方法不仅遮盖后能预测整体,还使得语言模型能够训练出更好的全局信息,从而学习到更为深入的先验知识。

2.2.2　BiLSTM 层

BiLSTM 模型是长短期记忆网络(long short-term memory,LSTM)的一个变体,用于更好地捕捉序列数据中的双向依赖关系。LSTM 是一种特殊类型的循环神经网络(recurrent neural network,RNN),能够学习序列数据中长距离的依赖关系[17]。LSTM 通过引入 3 个门控机制(遗忘门、输入门和输出门)解决了传统 RNN 在处理长序列时的梯度消失或梯度爆炸问题。3 个门控机制的公式为

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_{\mathrm{f}}[h_{t-1},x_t] + \boldsymbol{b}_{\mathrm{f}}) \tag{1}$$

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_{\mathrm{i}}[h_{t-1},x_t] + \boldsymbol{b}_{\mathrm{i}}) \tag{2}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_{\mathrm{o}}[h_{t-1},x_t] + \boldsymbol{b}_{\mathrm{o}}) \tag{3}$$

式中:σ 为 sigmoid 激活函数;$\boldsymbol{f}_t$为遗忘门的激活向量;$\boldsymbol{W}_{\mathrm{f}}$为遗忘门的权重矩阵;$\boldsymbol{b}_{\mathrm{f}}$为遗忘门的偏置项;$[h_{t-1},x_t]$为将 h_{t-1}和 x_t拼接起来;$\boldsymbol{i}_t$为输入门的激活向量;$\boldsymbol{W}_{\mathrm{i}}$为权重矩阵;$\boldsymbol{b}_{\mathrm{i}}$为偏置项;$\boldsymbol{o}_t$为输出门的激活向量;$\boldsymbol{W}_{\mathrm{o}}$为输出门的权重矩阵;$\boldsymbol{b}_{\mathrm{o}}$为偏置项。

传统的 RNN 和 LSTM 通常只能从前到后处理序列,这限制了模型对未来信息的利用。BiLSTM 通过将两个独立的 LSTM 层并置在一起,主要由两个阶段构成:首先是前向传播阶段,在这一阶段中,训练序列被送入前向 LSTM 网络,通过前向传播过程计算出前向的特征信息;接着是后向传播阶段,此时,序列输入到后向 LSTM 网络,并通过后向传播过程计算出后向的特征信息,最后将前后两向的特征信息结合,形成最终的隐藏层状态。这样就汇总了双向的语义特征,使得模型能够更好地理解句子中每个词的上下文,有助于解决语言中的歧义和依赖性问题。

2.2.3　CRF 层

CRF 层作为输出层,对标记序列进行解码,从而通过联合建模整个序列,确保了最终识别结果的合理性和准确性。其评分函数包含转移函数、发射函数两部分,转移函数旨在表示相邻标签间的依赖关系,而发射函数则表示当前样本在某一特定状态下的概率。CRF 的评分函数公式为

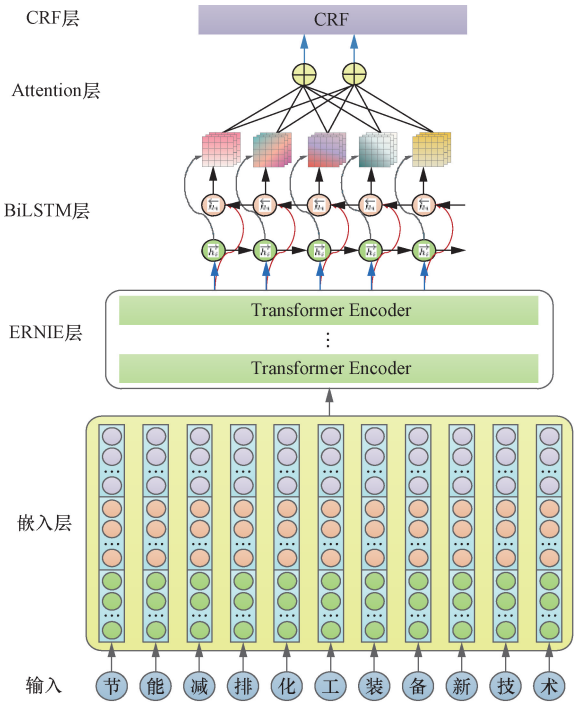
$$\mathrm{Score}(x,y) = \sum_{i=1}^{n}[S_{y_i}^{i}(x) + t_{y_i,y_{(i+1)}}] \tag{4}$$

式(4)中:y 为标签序列;x 为观察序列;s 为各个标签位置的评分;t 为转移评分;$S_{y_i}^{i}(x)$ 为发射函数;$t_{y_i,y_{(i+1)}}$ 为转移函数。

2.2.4　多重注意力机制层

注意力机制会根据输入序列中不同位置的相关性,动态分配权重。多重注意力机制可以学习到输入数据的多种表示,这些表示被拼接并转换后,能够提供更丰富的特征组合,增强模型的表达能力,这使得模型在处理文本时,能够根据上下文信息自适应地调整对不同词汇的关注度,进而更好地识别出新词。在 BiLSTM 层之后、CRF 层之前融入多重注意力机制,通过多角度的信息捕获和增强的模型表达能力,提高了模型处理复杂任务的能力,同时保持了较好的可解释性,这些优势使得它在新词发现等领域得到了广泛的应用。融合多重注意力机制的 ERNIE-BiLSTM-CRF 新词发现模型如图 4 所示。

能源政策文件通常包含大量专业术语、长距离依赖以及复杂的语义结构,这使得传统新词发现方法难以充分捕捉深层语义信息。针对上述问题,将 *N*-Gram 模型与融合多重注意力机制的 ERNIE-BiLSTM-CRF 模型相结合,在处理能源领域长文本和专业术语时能充分利用 *N*-Gram 在捕捉局部上下文信



Transformer Encoder 为 Transformer 编码器;h_t为在时间步 t 的隐藏状态,由前向和后向 LSTM 的隐藏状态拼接而成

图 4　融合多重注意力机制的 ERNIE-BiLSTM-CRF 新词发现模型

Fig. 4　ERNIE-BiLSTM-CRF new word discovery model with integrated multiple attention mechanism

息上的优势,如低碳技术、绿色能源等复合词的识别。通过将 ERNIE 与 BiLSTM 相结合,能够引入语境和领域知识,从而更好地理解长距离的依赖关系和复杂语义结构。BiLSTM 的双向上下文建模加强了对序列信息的全面理解,而 CRF 则进一步优化了序列标签之间的全局依赖,确保了最终标注的准确性和一致性。

3 实验

3.1 实验环境及参数配置

本文所使用的数据集为 DLOU-NW 数据集,计算机算力为 RTX 4090,操作系统为 Windows 10(64 bit),内存 16 GB,硬盘大小 1 000 GB,CPU 核数六核,实验参数:批处理(batch_size)为 16,学习率(learning rate)为 0.001,失活率(dropout)为 0.5,最大序列长度(max_seq_len)为 128。

3.2 实验结果分析

为评价能源领域新词发现的效果,采用准确率(precision,P)、召回率(recall,R)和 F_1(F_1-score)常用的 3 个模型评价指标,对模型的性能进行评价。

在能源领域新词发现任务中,实验使用的各个模型分别对新词发现识别效果的 3 个指标的计算公式为

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{5}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{6}$$

$$F_1 = \frac{2RP}{P + R} \tag{7}$$

式中:TP 为正确的匹配数目;FP 为误报、没有的、匹配不正确的数目;FN 为漏报、没有找到正确匹配的数目;TN 为正确的非匹配数目。

本文提出的 ENFM 算法在 DLOU-NW 数据集上基于不同的取值进行了实验评估。为验证 N-Gram 模型中不同的 N 值设置对新词发现准确率的影响,本文设置 N 值分别为 1 ~ 5 开展验证实验,实验结果如表 4 和图 5 所示。

表 4 不同 N-Gram 取值的新词发现方法评价指标

Table 4 Evaluation metrics for new word discovery methods with different N-Gram values

数据集	N	准确率/%	召回率/%	F_1/%
DLOU-NW 数据集	1	90.49	89.82	90.15
	2	**95.71**	**95.56**	**95.63**
	3	92.67	92.31	92.49
	4	84.97	85.90	85.40
	5	83.70	83.52	83.61

注:加粗数值为每列的最佳结果。

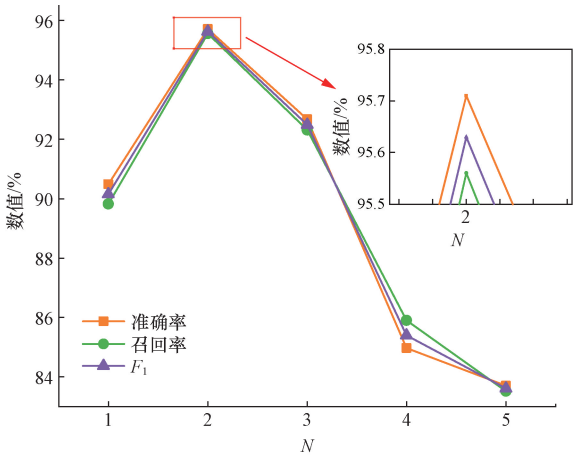


图 5 不同 N-Gram 取值的新词发现方法评价指标

Fig. 5 Evaluation metrics for new word discovery methods with different N-Gram values

理论上来说,N 越大,模型预测的准确率越高。但是模型训练依赖的语料库是有限的,N 越大,模型统计出的数据就越稀疏,反而会影响性能。从图 5 以及表 4 印证了这一想法,适当增大 N 可以提升准确率,但是当数据集的大小不足以支撑过大的 N 时,模型的准确率反而有所下降。实验结果表明,当 $N=2$ 时,准确率、召回率和 F_1 均高于其他情况。在最终实验中,本文方法选择 $N=2$ 与其他算法进行对比实验。基于以上结果可知,当 $N=2$ 时,本文方法能够相对取得更好的准确率,具有更好的实用性和应用价值。

某些情况下,领域内的词汇构成有其特定的规律,在能源领域,专业术语往往由两个词组合而成,如能源效率、碳排放等。使用 $N=2$ 的取值可以更好地捕捉这样的双词组合,从而识别出领域内的新词或新概念,并且相较于其他 N,取 $N=2$ 时又更少受到数据稀疏性的影响。这意味着在有限的数据集上,该数值能够更有效地统计和学习词汇模式,从而优化新词发现的效果。

3.3 消融实验

在消融实验的训练过程中,模型的各个参数均保持不变。选择 ERNIE-BiLSTM-CRF 模型、EBAC 模型、N-Gram 模型在 DLOU-NW 数据集上作为 ENFM模型的消融实验。ENFM 模型的消融实验结果如表 5 所示。

表 5 消融实验结果

Table 5 Ablation experiment results

模型	准确率/%	召回率/%	F_1/%
ERNIE-BiLSTM-CRF	92.92	94.08	93.49
EBAC	93.82	94.62	94.22
ENFM;2gram	95.71	95.56	95.63

(1)ERNIE-BiLSTM-CRF 模型:ERNIE-BiLSTM-CRF 模型结合了 ERNIE、BiLSTM 和 CRF 的优点,使得模型能够同时考虑文本的全局和局部信息。虽然模型的准确率、召回率和 F_1 都比较高,但由于能源领域的多样性和复杂性,模型可能对某些新词不够敏感。

(2)EBAC 模型:引入多重注意力机制后,EBAC 模型与 ERNIE-BiLSTM-CRF 模型相比,准确率、召回率和 F_1 分别提高了 0.9%、0.54%、0.73%。这表明多重注意力机制有效地增强了模型对文本中关键信息的捕捉能力,尤其是在提取与新词相关的上下文特征时更为有效。

(3)*N*-Gram 模型:独立的统计模型,在前文已经设置 N 为 1~5 进行实验,用来验证 *N*-Gram 方法在不同长度下的表现,证明当 $N=2$ 时,新词发现效果更好。使用 2-gram 作为特征的 ENFM 模型在所有性能指标上进一步提升,达到了最高的准确率、召回率和 F_1。这显示出 *N*-Gram 模型在处理新词发现任务时的强大能力,尤其是在捕捉和利用词语序列信息方面的优势。在新词发现的任务中,2-gram 能够帮助模型捕捉到词汇之间的紧密联系,特别是在词汇创新和术语演变快速的能源领域。

3.4　对比实验

为进一步验证实验结果,将现有方法作为参照进行比较,与本文提出的新词发现方法进行对比实验。

(1)对比方法Ⅰ:基于互信息和左右邻接熵的新词发现模型。依赖于统计量来确定词汇是否作为新词。该方法简单高效,但缺乏对语义和上下文的深入理解。

(2)对比方法Ⅱ:BERT 模型。使用 Transformer 的编码器部分作为其基本结构,通过两个自监督任务来学习语言表示利用 Transformer 的双向性,能够同时考虑一个词的前后文信息,从而更准确地捕捉词的上下文语义。

(3)对比方法Ⅲ:BERT-BiLSTM-CRF 模型。结合了 BERT 预训练模型、BiLSTM 和 CRF,在处理中文时,能够有效地捕捉上下文信息和语义信息,提高序列标注的准确性和鲁棒性。

(4)对比方法Ⅳ:DeBERTa 模型。通过引入解耦合的注意力机制和遮蔽位置预测等新技术,改进了 BERT 模型,从而改善了新词发现的准确性和召回率,提高了其在 NLP 任务上的性能和泛化能力。

(5)对比方法Ⅴ:GPT-2 模型。选择使用 GPT-2 在于其资源和成本、可访问性和易用性等方面。GPT-2 采用多层 Transformer 作为基础结构,它训练的数据跨越多个领域,而且模型的性能受到使用的评估标准和数据集选择的影响,能源领域的专业术语和概念需要特定的语义理解能力。

各模型结果对应的准确率、召回率和 F_1 如表 6 和图 6 所示。

表 6　对比实验结果

Table 6　Comparison experiment results

序号	模型	准确率/%	召回率/%	F_1/%
Ⅰ	基于互信息和左右熵	91.67	84.61	87.9
Ⅱ	BERT	88.47	90.34	89.39
Ⅲ	BERT-BiLSTM-CRF	90.18	91.58	90.87
Ⅳ	DeBERTa	92.55	92.91	92.73
Ⅴ	GPT-2	94.7	94.78	94.74
Ⅵ	ENFM	95.71	95.56	95.63

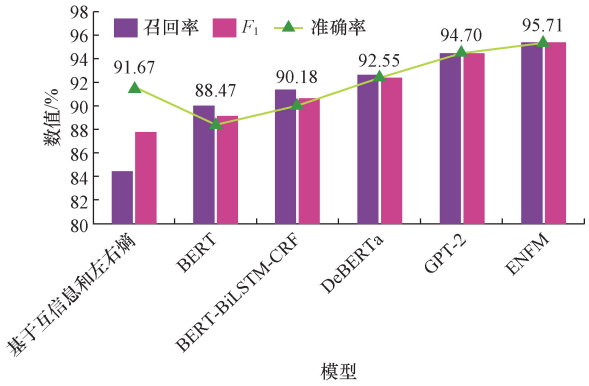


图 6　对比实验结果

Fig. 6　Comparison experiment results

图 6 显示,ENFM 模型在 DLOU-NW 数据集上达到了相对稳定的实验结果,表明模型在能源领域新词发现任务中的效果较好。

(1)基于互信息和左右熵的模型(模型Ⅰ):此模型显示了较高的准确率(91.67%)但较低的召回率(84.61%),导致 F_1 为 87.90%。这表明该模型在识别某些常见的能源领域实体时表现较好,但由于缺乏深度语义理解,可能在处理复杂或罕见的实体时表现不佳,导致召回率较低。

(2)BERT 模型(模型Ⅱ):作为一个通用的预训练模型,BERT 在处理各种文本任务时都表现出色。但在能源领域,它可能缺乏针对该领域的特定知识,也存在对长文本处理困难、依赖预训练数据集以及学习能力有限等缺点,因此在某些专业术语或概念的识别上准确率较低。

(3)BERT-BiLSTM-CRF 模型(模型Ⅲ):结合了 BERT 的语义理解能力和 BiLSTM-CRF 的序列标注能力,3 个主要指标上均优于 BERT 模型。

(4)DeBERTa 模型(模型Ⅳ):3 个主要指标上均

优于前3种对比模型。这显示出DeBERTa模型的均衡性能,有效提升了召回率,减少了遗漏新词的情况。但其对新词的敏感性可能仍受限于训练数据的代表性和多样性。如果训练数据中缺乏对新兴词汇或变化快速的语言用法的覆盖,模型可能不会有效识别真正的新词。

(5)GPT-2模型(模型Ⅴ):所有指标上都优于前4种模型。GPT-2作为一个生成式预训练模型,展示了其在处理新词发现任务时的优异能力,尤其是在理解和生成文本的上下文关联性方面,但在面对高度专业化的领域时缺乏足够的深度。

(6)ENFM模型(模型Ⅵ):与其他5种方法相比,模型的准确率分别提高了4.04、7.24、5.53、3.16、1.01个百分点,召回率分别提高了10.95、5.22、3.98、2.65、0.78个百分点,F_1分别提高了7.73、6.24、4.76、2.9、0.89个百分点。这表明ENFM模型通过融合有效的特征和机制,进一步提高了模型的性能,尤其是在处理非结构化文本数据时。模型在各测评指标值上都取得了一定的提升,达到了相对更优的效果。

能源政策文本部分新词实例如表7所示。从表7可知,本文方法应用于能源领域可以较好改善新词发现的效果,对不同的类别划分清晰明确、各概念的定义界定更为精确,分析能源政策文本部分新词实例结果如下。

例句"研发大规模可再生能源并网及电网安全高效运行技术",该句子的结构为包含连接词"及""和""与"等的并列结构,连接词左右两端描述了两个并列的技术目标,强调了研发工作的两个重要方向,表示研发的目标是解决"大规模可再生能源并网"和"电网安全高效运行"这两个问题。

例句"开展地热发电、海洋能发电与生物质发电技术研发",使用了"、"和"与"作为连接词,将"地热发电""海洋能发电"和"生物质发电"3个短语并列起来,表示将针对这3种发电技术进行研发。这3个并列的短语都是能源领域的新词,分别指代利用地热、海洋能和生物质资源进行发电的技术,展示了能源技术研发的广度。

表7 能源政策文本部分新词实例

Table 7 Examples of new words in the text of energy policy documents

序号	例句	新词
1	研发大规模可再生能源并网	大规模可再生能源并网
2	及电网安全高效运行技术	电网安全高效运行
3	开展地热发电、海洋能发电与生物质发电技术研发	地热发电
4		海洋能发电
5		生物质发电
6	研究富氢冶炼工艺技术、低	富氢冶炼工艺
7	碳清洁生产技术	低碳清洁生产

例句"研究富氢冶炼工艺技术、低碳清洁生产技术",该句子的结构为"、"作为连接词的并列结构,"、"两端分别表示了研究的两种不同类型的技术。这两个短语都是能源领域的新词,体现了现代工业在节能减排和可持续发展方面的趋势。

4 结论

提出了一种融合*N*-Gram和多重注意力机制的能源领域新词发现模型ENFM,在能源领域新词发现任务中展现出显著效果。

(1)该模型不仅高效识别新词,还提升了识别的精确度和效率,相比于传统技术,ENFM模型在新词准确识别和边界确定上展现出更高的性能,能解决传统方法在处理专业领域文本数据时面临的困难,特别是在面对能源领域术语的高度专业化和多样性。

(2)实验结果表明,ENFM模型在多个评估指标上均优于其他模型,这一成果不仅验证了ENFM模型的有效性,也为能源领域的文本分析和后续的能源技术词条抽取提供了新的解决方案。

(3)然而,ENFM模型在处理极端复杂或未知领域的文本时可能面临挑战。未来的工作将探索更多模型融合策略,以增强模型对非结构化文本的处理能力,并在不同数据集和领域验证其实用性和可扩展性。

参 考 文 献

[1] 蔡睿, 朱汉雄, 李婉君, 等. "双碳"目标下能源科技的多能融合发展路径研究[J]. 中国科学院院刊, 2022, 37(4): 502-510.
Cai Rui, Zhu Hanxiong, Li Wanjun, et al. Development path of energy science and technology under "Dual Carbon" goals: perspective of multi-energy system integration [J]. Bulletin of Chinese Academy of Sciences, 2022, 37(4): 502-510.

[2] 许雅玺, 孟天宇, 王欣, 等. 融合领域词典嵌入的航空不安全事件命名实体识别[J]. 科学技术与工程, 2024, 24(8): 3284-3290.
Xu Yaxi, Meng Tianyu, Wang Xin, et al. Named entity recognition of aviation unsafe events embedded with fusion domain dictionary [J]. Science Technology and Engineering, 2024, 24(8): 3284-3290.

[3] 刘清民, 王芳, 黄梅银. 我国人工智能政策新词发现与演化研究——一个多特征融合的算法[J]. 现代情报, 2024, 44(6): 18-32, 58.
Liu Qingmin, Wang Fang, Huang Meiyin. Discovery and evolution of new words in Chinese artificial intelligence policies: a multi-feature fusion algorithm[J]. Journal of Modern Information, 2024, 44(6): 18-32, 58.

[4] 王巍洁, 任慧玲, 李晓瑛, 等. 融合汉字多语义与文本统计特征的中文医学新词发现研究[J]. 图书情报工作, 2024, 68

(6): 119-128.
Wang Weijie, Ren Huiling, Li Xiaoying, et al. Chinese medical new word detection by chinese character's multi-semantic word vector and statistical text features [J]. Library and Information Service, 2024, 68 (6): 119-128.

[5] 申兆媛, 巢翌, 李晓龙, 等. 针对特定领域的新词发现方法研究[J]. 计算机仿真, 2022, 39(6): 269-273, 335.
Shen Zhaoyuan, Chao Yi, Li Xiaolong, et al. Research on new word discovery methods for specific domains [J]. Computer Simulation, 2022, 39(6): 269-273, 335.

[6] 汪琳, 王昊, 李晓敏, 等. 融合学习扩展的非遗陶瓷工艺领域术语库构建及应用[J]. 图书馆论坛, 2024, 44(2): 66-78.
Wang Lin, Wang Hao, Li Xiaomin, et al. Thesaurus development and application in the field of intangible cultural heritage ceramics incorporated with learning extension [J]. Library Tribune, 2024, 44(2): 66-78.

[7] Zhang C, Zhao S, He Y. An integrated method of the future capacity and RUL prediction for lithiumion battery pack[J]. IEEE Transactions on Vehicular Technology, 2021, 71(3): 2601-2613.

[8] 祝钰莹, 郭燕, 万亿兆, 等. 基于信息熵-切分概率模型的新词发现方法[J]. 计算机科学, 2023, 50(7): 221-228.
Zhu Yuying, Guo Yan, Wan Yizhao, et al. New word discovery method based on information entropy-segmentation probability model [J]. Computer Science, 2023, 50(7): 221-228.

[9] 耿骞, 邓斯予, 靳健. 融合词语义表示和新词发现的领域本体演化——以产品评论数据为例[J]. 图书情报工作, 2021, 65(8): 85-96.
Geng Qian, Deng Siyu, Jin Jian. Domain ontology evolution combining word semantic representation and new word discovery: a case study of product review data [J]. Library and Information Service, 2021, 65(8): 85-96.

[10] 刘凡平, 陈慧, 沈振雷, 等. 基于 BERT 的开放领域中文新词发现研究[J]. 计算机应用与软件, 2023, 40(6): 173-180.
Liu Fanping, Chen Hui, Shen Zhenlei, et al. Research on Chinese new word discovery in open domains based on BERT [J]. Computer Applications and Software, 2023, 40(6): 173-180.

[11] 张一帆, 张军莲, 汪鸣泉, 等. 基于条件随机场和词向量的能源政策领域新词发现[J]. 南京理工大学学报, 2021, 45(1): 37-45.
Zhang Yifan, Zhang Junlian, Wang Mingquan, et al. New word discovery in the field of energy policy based on conditional random fields and word vectors [J]. Journal of Nanjing University of Science and Technology, 2021, 45(1): 37-45.

[12] 鲁静. 中国特色社会主义文化的多模态融合传播策略研究[J]. 河南社会科学, 2023, 31(8): 117-124.
Lu Jing. Research on the multimodal integration communication strategy of socialism with Chinese characteristics culture [J]. Henan Social Sciences, 2023, 31(8): 117-124.

[13] 蒋丽媛, 吴亚东, 王书航, 等. 融合笔画特征的命名实体识别方法[J]. 科学技术与工程, 2023, 23(17): 7436-7443.
Jiang Liyuan, Wu Yadong, Wang Shuhang, et al. Named entity recognition method incorporating stroke features [J]. Science Technology and Engineering, 2023, 23(17): 7436-7443.

[14] Mattiev J, Salaev U, Kavsek B. Word game modeling using character-level N-Gram and statistics [J]. Mathematics, 2023, 11(6): 1380-1388.

[15] 刘巨升, 于红, 杨惠宁, 等. 基于多核卷积神经网络(BERT + Multi-CNN + CRF)的水产医学嵌套命名实体识别[J]. 大连海洋大学学报, 2022, 37(3): 524-530.
Liu Jusheng, Yu Hong, Yang Huining, et al. Recognition of nested named entities in aquature medicine based on multi-kernel convolution (BERT + Multi-CNN + CRF) [J]. Journal of Dalian Ocean University, 2022, 37 (3): 524-530.

[16] 裴炳森, 李欣, 胡凯茜, 等. 基于知识增强预训练模型的司法文本摘要生成[J]. 科学技术与工程, 2024, 24(20): 8587-8597.
Pei Bingsen, Li Xin, Hu Kaixi, et al. Judicial text summarization based on knowledge-enhanced pretrained language models [J]. Science Technology and Engineering, 2024, 24(20): 8587-8597.

[17] 郝宽公, 董兵, 吴悦, 等. 基于 BERT-Bi-LSTM-CRF 模型的机场类中文航行通告要素实体识别[J]. 科学技术与工程, 2024, 24(10): 4182-4188.
Hao Kuangong, Dong Bing, Wu Yue, et al. Airport class based on BERT-BILSTM-CRF model chinese navigation notice element entity recognition [J]. Science Technology and Engineering, 2024, 24(10): 4182-4188.