# 语义增强的多策略政策术语抽取系统①

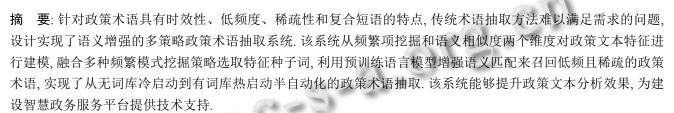
曹秀娟<sup>1,2</sup>, 马志柔<sup>2</sup>, 朱 涛<sup>3</sup>, 张庆文<sup>3</sup>, 杨 燕<sup>2</sup>, 叶 丹<sup>2</sup>

¹(广西大学 计算机与电子信息学院, 南宁 530004)

2(中国科学院 软件研究所 软件工程技术研究开发中心, 北京 100190)

3(政和科技股份有限公司, 济南 250000)

通信作者: 马志柔, E-mail: mazhirou@otcaix.iscas.ac.cn



关键词: 术语抽取; 多策略; 语义增强; 低频度; 词库构建

引用格式: 曹秀娟,马志柔,朱涛,张庆文,杨燕,叶丹.语义增强的多策略政策术语抽取系统,计算机系统应用,2022,31(9):152-158. http://www.c-sa.org.cn/1003-3254/8693.html

# Semantic Enhanced Multi-strategy Policy Term Extraction System

CAO Xiu-Juan<sup>1,2</sup>, MA Zhi-Rou<sup>2</sup>, ZHU Tao<sup>3</sup>, ZHANG Qing-Wen<sup>3</sup>, YANG Yan<sup>2</sup>, YE Dan<sup>2</sup>

<sup>1</sup>(School of Computer, Electronics and Information, Guangxi University, Nanning 530004, China)

<sup>2</sup>(Technology Center of Software Engineering, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China)

<sup>3</sup>(Zhenghe Technology Co. Ltd., Jinan 250000, China)

Abstract: Policy terms are characterized by timeliness, low frequency, sparsity, and compound phrases. To address the difficulty of traditional term extraction methods in meeting demands, we design and implement a semantic enhanced multi-strategy system of policy term extraction. The system models the features of policy texts from the two dimensions of frequent item mining and semantic similarity. Feature seed words are selected by integrating multiple frequent pattern mining strategies. Low-frequency and sparse policy terms are recalled by pre-training the language model and enhancing semantic matching. Transforming from a cold start without a thesaurus to a hot start with a thesaurus, the system achieves semi-automatic extraction of policy terms. The proposed system can improve the effect of policy text analysis and provide technical support for the construction of a smart government service platform.

Key words: term extraction; multi-strategy; semantic enhancement; low frequency; thesaurus construction

政策文本是用来记录政策活动而产生的过程性文 件,是政策服务研究的重要载体和依据,包括通知、公 告、意见、批复等公文类别. 目前, 政府与企业之间在 政策服务上存在着一定的壁垒,一方面企业无法及时 解读相关政策,不能及时享受政府补贴;另一方面,政 府无法及时了解政策发布的受益面及其所发挥的作用, 而政策文本分析在政策解读、政企协同、企业决策和 成果转化等政策服务方面具有非常重要的现实意义. 由于政策术语新词的大量出现, 使得政策领域的分词 不准确, 严重影响了对政策文本的解读[1], 政策术语抽

① 基金项目: 国家自然科学基金 (61802381)

收稿时间: 2021-12-21; 修改时间: 2022-01-24; 采用时间: 2022-01-30; csa 在线出版时间: 2022-06-16

152 系统建设 System Construction



取成为了解决这一难题的当务之急. 政策术语具有时 效性、低频度、稀疏性和复合短语的特点,难以用频 繁模式和序列标注的方法直接抽取,多由领域专家手 工抽取.

为了实现半自动化的政策术语抽取,本文设计了 语义增强的多策略政策术语抽取系统,该系统融合频 数、自由度、凝固度等多种策略,获得包含政策结构 信息的术语新词; 并利用预训练语言模型增强语义相 似度匹配来召回包含政策语义信息的术语新词,结合 两者信息来生成政策术语词库并可对其迭代更新,切 实解决了人工抽取政策术语的困难.

## 1 相关工作

随着大数据和人工智能时代的到来, 自动术语抽 取技术作为实现领域术语抽取系统的关键技术, 受到 了广泛的关注和研究. 解决自动术语抽取的主流方法 主要有3大类:基于语言学方法、基于统计学方法、 基于深度学习方法.

#### 1.1 基于语言学方法的术语抽取

基于语言学方法的术语抽取根据领域术语的语言 特征规则,或与词典中的术语相匹配.首先将文本进行 分词和词性标注,然后对比分词结果和词法规则,匹配 一致的内容为候选术语. 研究者主要通过对行业领域 术语的构词模式进行分析,实现不同领域的术语抽取. 曾浩等人[2] 制定了 4 条扩展规则并结合统计特征进行 术语抽取. 赵志滨等人[3] 运用句法分析和词向量技术 对新词发现进行研究, 在护肤品论坛的真实文本数据 集上取得了较好的效果. Kafando 等人[4] 结合统计特征 和语言学定性定量规则分析,利用 BioTex 工具抽取生 物医学领域组合术语. 基于语言学方法的术语抽取需 要领域专家的知识背景进行支撑及维护, 无法完成领 域迁移.

## 1.2 基于统计学方法的术语抽取

基于统计学方法的术语抽取主要采用 N-Gram 统 计语言模型建模,结合扩展统计特征对术语进行抽取. 常见的统计特征主要有词频数 (TF)、凝固度 (PMI)、 自由度 (DF) 和 C-value 等. 目前应用统计学方法进行 术语抽取具有较多工作. Chen 等人[5] 为有效地确定专 利领域新词的边界, 引入二元词的双向条件概率信息, 提取专利领域长词. 王煜等人[6] 利用改进的频繁模式 树算法,结合 DF、PMI 和时间特征,对网络新闻热点 新词进行了有效识别. Li 等人[7] 改进 PMI 并结合 DF 特征自动抽取未登录词. 陈先来等人[8] 采用融入逻辑 回归的凝固度模型提取新词,有效地提高了电子病历 文本数据分词准确率. 基于统计学方法的术语抽取能 抽取到高频且高质量的术语, 无法抽取低频且稀疏的 术语.

## 1.3 基于深度学习方法的术语抽取

随着机器学习尤其是深度学习的发展,推动术语 抽取研究产生了各类模型和方法的领域应用. Chen 等 人<sup>[9]</sup> 采用统计特征提取候选术语, 利用 CNN 模型生成 消费品缺陷领域词典. 基于术语语义相关性的思想, 张 一帆等人[10] 使用 TextRank 抽取领域种子词典, 而后计 算候选术语与种子集的余弦相似度进行术语抽取. Qian 等人[11] 使用包含词语信息的 Word2Vec 词向量 对 N-Gram 频繁字符串候选词组进行剪枝, 无监督地 进行术语抽取,但其并未考虑中文词语的一词多义问 题. 张乐等人[12] 提出将汉字笔画知识和知网中的义原 知识引入 Word2Vec 词向量训练, 从而获得多语义词 向量, 但其针对社交媒体领域. 近年来, 预训练语言模 型 BERT 提出后, 在术语抽取上得到了广泛应用, Choi 等人[13] 将统计特征 TF-IDF 与 FastText 和 BERT 模型 结合,实现了韩文语料的自动术语抽取.

上述研究表明,单一的方法均无法达到最佳的术 语抽取效果,基于统计学方法抽取的候选术语仍需进 行停用词过滤和对应领域的语言规则过滤,基于深度 学习的方法需要海量的标注数据来训练模型,对分布 稀疏的政策术语来说, 难以达到抽取效果. 因此, 本文 考虑引入预训练语言模型来增强语义,并融合多策略 频繁模式来提高政策术语抽取效果, 实现政策术语的 半自动化抽取.

## 2 关键技术研究

政策术语抽取系统的半自动化实现, 其关键技术 是如何利用人工智能和自然语言处理技术,尝试将自 动术语抽取与语义知识相结合, 高效地构建政策领域 术语词典,有效提升政策术语抽取的效果.

通常政策文本术语抽取示例如表1所示.

由表 1 可知, 政策术语有如下的特点: 1) 复合短 语: 由多个词语嵌套、复合、派生组成的固定短语; 2) 词性分布: 多为名词性短语或动名词性短语; 3) 长度 分布: 长度分布于 4 至 15 字词之间; 4) 低频度: 出现的

System Construction 系统建设 153



频次普遍不高; 5) 时效性: 政策术语随着时间的推移会 不断更新.

针对低频且稀疏的政策术语抽取难的问题, 本文

提出了一种零样本语义增强的多策略政策术语抽取方 法来实现系统,包括多策略频繁模式抽取算法和语义 增强抽取算法。

表 1 政策文本术语抽取示例

编号	原始文本	政策术语	术语频次
1	······扎实推进商务领域" <b>双招双引</b> "工作,奋力推动全市商务经济加快·····	双招/双引	7
2	······企业顺利入围山东省 <b>准独角兽企业</b> 公示名单. 进一步完善······	准/独角兽/企业	1
3	······提高技能人才培养质量, 决定组织实施山东省 <b>高技能人才培养特色载体</b> 建设工程······	高技能/人才培养/特色/载体	2
4	······切实加强对济南 <b>新旧动能转换</b> 起步区建设的组织领导, 落实······	新旧/动能/转换	7
5	······关键核心零部件、新材料首批次应用、 <b>首版次高端软件</b> 保险补偿······	首/版次/高端/软件	4

## 2.1 多策略频繁模式抽取算法

肖仰华等人[14] 指出衡量一个术语的质量, 主要考 虑 4 个方面: 高频率、一致性、信息量和完整性. 高频 率主要指术语应该在给定文档中出现足够频繁;一致 性是指术语和不同词之间的搭配是否合理或是否常见; 信息量主要考虑术语传达的信息, 其应当表达一定的 主题或者概念; 完整性主要指术语在特定上下文中是 一个完整的语义单元. 凝固度衡量文本片段中字与字 之间的紧密程度,即术语的一致性;自由度衡量一个文 本片段左右两侧字符组合的丰富度,即术语的完整性; C-value 衡量候选短语质量即术语的信息量, 通过有效 校正父子嵌套短语重复统计带来的频次估计的偏差, 提取多词嵌套的长政策术语.

为了抽取政策文本中内部凝结紧且外部组合自由 度高的政策术语,设计了一种多策略频繁模式抽取算 法. 该算法以 N-Gram 统计语言模型为基础, 采用综合 词频、自由度、凝固度和 C-value 特征各自优势的指 标 FPDC 来衡量术语, 结合停用词和常用词前后缀搭 配规则过滤术语. 算法步骤如下:

Step 1. 文本预处理. 对文本进行预处理, 删除政策 文本中的邮箱、电话号码、手机号码、日期、网址等, 置换标点符号为空格.

Step 2. 候选短语生成. 基于 N-Gram 统计语言模 型对文本语料进行统计, 过滤词长阈值以下的文本片 段,得到候选文本片段.

Step 3. 术语质量评分. 首先对各候选文本片段计 算词频 tf、凝固度 pmi、自由度 df 和 C-value 值 cval, 然后对各特征进行 Sigmoid 函数归一化, 最后融合各 特征值计算指标 FPDC, 初始化为每个特征平均分配 权重, 考虑到政策领域多词嵌套的中心词, 对词频进行 了 0.15 的惩罚, 对 C-value 进行了 0.15 的奖励, 如式 (1) 所示. 根据阈值筛选, 得到候选政策术语.

$$FPDC(c_1 \cdots c_n) = 0.1tf(c_1 \cdots c_n) + 0.25pmi(c_1 \cdots c_n) + 0.25df(c_1 \cdots c_n) + 0.4cval(c_1 \cdots c_n)$$

$$\tag{1}$$

其中, $c_1 \cdots c_n$ 表示多个字构成的候选文本片段.

Step 4. 语言规则过滤. 对候选政策术语进行停用 词过滤和常用词作为前后缀的语言学规则过滤.

Step 5. 结果排序输出. 按照 FPDC 指标由高到低 排序,输出政策术语抽取结果.

#### 2.2 语义增强抽取算法

在零样本无监督挖掘情况下,多策略算法可以抽 取到大量频繁、高质量的政策术语,但针对低频、稀 疏的政策术语抽取效果仍不够好. 引入预训练语言模 型来增强政策领域术语语义特征匹配, 在多策略算法 的基础上,设计了语义增强抽取算法来召回低频术语 新词. 语义增强抽取算法流程如下所示:

Step 1. 候选术语生成. 将现有词库中的政策术语 ngrams dict 和文本语料特征词集合 ngrams fw 特征词 计算归一化的 C-value 指标, 更新父子嵌套类型术语 的 FPDC 值, 将其作为 Jieba 分词的自定义词典, 对原 始语料重新分词,过滤不符合词长和词语频数要求的 文本片段作为候选术语.

Step 2. 语义向量生成. 从 ngrams fw 特征词中选取 FPDC排序前20%的特征词作为种子词,采用RoBERTa 预训练语言模型[15] 对候选术语和种子词语义特征向量 化, 得出每个候选术语和种子词的语义特征向量表示.

Step 3. 语义相似度计算. 从每个种子词出发, 计算 每个种子词和所有候选术语的语义向量的归一化欧式 距离相似度. 欧氏距离计算结果受到向量长度以及向 量维度的影响,取值范围不固定,采用 L2-norm 对候选 术语和种子词的语义特征向量标准化. 假设X是n维

154 系统建设 System Construction

的语义特征向量 $X = (x_1, x_2, x_3, \dots, x_n)$ , 则向量 X 的 L2 标准化公式如下:

$$||X||_2 = \sqrt{\sum_{i=1}^n x_i^2}$$
 (2)

向量 X 和向量 Y 的归一化欧式距离计算公式如下:

$$dist_{\text{edu}}(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i \times y_i)^2}$$
 (3)

$$sim = \frac{1}{1 + dist_{\text{edu}}(X, Y)} \tag{4}$$

Step 4. 语义特征相似度匹配. 遍历每个特征种子词, 找到与每个特征词相似度最大的候选术语, 当相似度大于设定阈值时认为该候选术语与种子词相似, 将候选术语加入结果术语集合; 考虑到候选术语之间的连通性, 对相似度阈值进行指数衰减法来将词与词分开. 设定最小相似度阈值为 *MinSim*, 阈值将随着词连通个数增大, 指数衰减法公式如下:

 $MinSim = MinSim + (1 - MinSim) \times (1 - e^{-\alpha \times idx})$  (5) 其中,  $\alpha$ 为衰减因子, idx表示种子词的序号.

Step 5. 结果排序输出. 通过每个特征种子词与候选术语的语义特征相似度匹配, 得到相似度匹配结果, 根据相似度由高到低排序, 输出最终的政策术语抽取结果, 并对词库进行了更新.

## 3 系统设计与实现

#### 3.1 系统架构设计

为了解决人工抽取政策术语的问题,本文设计了一套语义增强的多策略政策术语抽取系统.系统的组织架构如图 1 所示,分为数据层、模型层、服务层和应用层.

#### (1) 数据层

数据层包括系统中模型使用的停用词库、噪声词 规则库和政策术语词库.

停用词库用于过滤术语抽取结果中的垃圾串,即如果候选术语中的任意一个子串包含在停用词库中,则丢弃该候选术语.该词库初始化为通用的停用词库.

噪声词规则库用于过滤前后缀为常用词的候选术语.该规则库中初始化为常与政策术语作为前后缀进行搭配的模式,如"采用#""提供#""#与"和"#如下"等,"#"与常用词结合的位置代表该常用词作为候选政策

术语的前缀或者后缀.

政策术语词库用于保存政策术语抽取结果. 词库中包含政策术语、术语频次、术语词性、术语类别等信息. 系统提供了对于词库的增、删、改、查和词库统计信息可视化. 词库初始化为空, 通过设定或调整特征指标 FPDC 阈值, 由系统从候选术语列表中批量增加或更新术语词库.

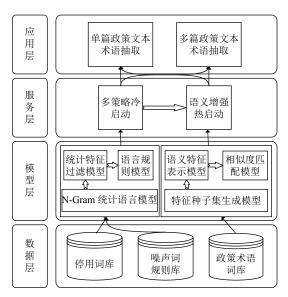


图 1 系统架构图

#### (2) 模型层

模型层是术语抽取系统所使用的核心模型,为多策略冷启动服务和语义增强热启动服务提供模型支持,包括 N-Gram 统计语言模型、统计特征过滤模型、语言规则模型、特征种子集生成模型、语义特征表示模型和相似度匹配模型.以下对各个模型的作用进行简要介绍.

N-Gram 统计语言模型为初始文本片段生成模型.模型对文本语料进行长度为1到n的滑动窗口操作,形成长度为1到n的字符片段序列,按给定的词长阈值过滤字符片段序列,得到候选文本片段集合.

统计特征过滤模型接收 N-Gram 模型的输出,对 候选文本片段进行 TF、PMI、DF、C-value 特征的统 计,计算术语特征融合指标 FPDC,按设定阈值过滤, 输出高于阈值的候选政策术语.

语言规则模型对候选政策术语进行噪声过滤,分为停用词库过滤和噪声词规则库过滤,输出去噪后的候选政策术语.

特征种子集生成模型主要生成语料的政策术语特征种子集. 模型根据候选术语和已有政策术语词库的

FPDC 值计算 C-value 进行更新, 选取 FPDC 值排序 前 20% 的候选术语, 输出为语料特征种子集.

语义特征表示模型主要生成候选术语和特征种子 词的语义特征表示. 模型对所有候选术语和特征种子 词利用中文预训练语言模型生成相应的语义特征向量, 并对语义特征向量进行 L2 标准化.

相似度匹配模型主要利用候选术语和特征种子词 的相似度挖掘低频且稀疏的政策术语. 模型遍历语料 特征种子集中的每个特征种子词, 计算所有候选术语 与该词的语义向量的归一化欧式距离相似度, 根据指 数衰减的相似度阈值进行连通性匹配,输出最终抽取 的政策术语结果.

#### (3) 服务层

针对零样本的术语抽取需求,提供了多策略冷启 动服务和语义增强热启动服务,即分别集成了多策略 频繁模式算法和语义增强的多策略术语抽取算法,为 两种算法提供 RESTful API 访问接口.

多策略冷启动服务提供无词库支持的多策略政策 术语抽取服务,模型使用第2.1节介绍的算法.通过设 定术语 TF 阈值、术语长度阈值、术语 PMI 阈值、术 语 DF 阈值、C-value 阈值以及是否进行语言规则过 滤, 先利用 N-Gram 统计语言模型从政策文本中抽取 候选文本片段,接着基于统计特征过滤模型和语言规 则模型进行候选文本片段分析与过滤, 最后排序输出 冷启动抽取结果.

语义增强热启动服务提供有词库支持的语义增强 政策术语抽取服务,模型使用第2.2节介绍的算法.冷 启动服务得到的抽取结果存在一定的不足,一方面抽 取术语中带有噪声词汇,一方面遗漏了低频数据.在冷 启动术语抽取结果的基础上, 先利用特征种子集生成 模型得到语料特征种子集,接着依次使用语义特征表 示模型和相似度匹配模型去除已抽取噪声词和召回未 登录低频词, 最后排序输出热启动抽取结果.

## (4) 应用层

应用层提供零样本条件下的交互式政策术语抽取 构建词库的功能,按照术语抽取的使用场景不同,分为 单篇政策文本术语抽取和多篇政策文本术语抽取两个 场景,提供政策术语词库的维护管理,包括增加、删 除、修改、查询等交互功能,以及统计可视化功能.

在单篇政策文本术语抽取场景下,用户可设定和 调整政策术语抽取参数 (术语 TF 阈值、术语长度阈 值、术语 PMI 阈值、术语 DF 阈值、C-value 阈值以 及是否进行语言规则过滤、是否加入当前词库和相似 度阈值) 实现从无词库冷启动到有词库热启动半自动 化的政策术语抽取.

在多篇政策文本术语抽取场景下, 与单篇政策文 本术语抽取不同之处在于,抽取时不仅要考虑候选政 策术语在单篇语料中的局部特征,而且还需考虑其在 多篇语料中的全局统计特征,实现对某类政策文本的 全局政策术语抽取.

系统整体流程如图 2 所示

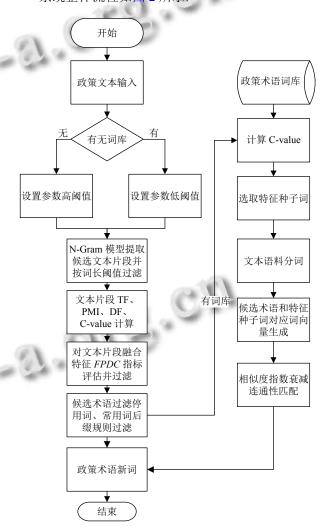


图 2 语义增强的多策略政策术语抽取流程图

#### 3.2 系统实现与展示

系统实现采用 Python 语言作为程序开发语言, 选 用具有强扩展性和兼容性的 Flask 框架作为 Web 服务 框架,以 Keras 框架作为快速加载预训练语言模型的 深度学习框架. 系统展示如图 3 所示.

156 系统建设 System Construction

系统包括政策术语词库统计、政策术语词库管理 和政策术语抽取 3 大功能模块. 系统首页为政策术语 词库统计模块,包括政策术语词库中政策术语总数、 政策术语长度分布、政策术语类型分布、政策术语词 性分布、政策术语频数分布. 政策术语词库管理模块 提供了对政策术语词库的增、删、改、查. 政策术语 抽取模块, 分为单篇政策文本术语抽取和多篇政策文 本术语抽取两部分.



图 3 系统界面效果图

# 4 应用与结果分析

本系统在某公司政务通平台进行术语抽取应用验 证,选取数据集为1942篇来自各省、直辖市或以上行 政级别政府单位所公布的政策文本, 由业务人员判断 抽取的术语是否有用. 抽取效果评价指标如下:

## (1) 术语抽取准确率

$$P = \frac{\text{抽取正确的术语数}}{\text{抽取的术语总数}} \times 100\%$$
 (6)

#### (2) 术语抽取召回率

$$R = \frac{\text{抽取的正确的术语数}}{\text{文本包含的术语总数}} \times 100\% \tag{7}$$

(3) F1 值

$$F1 = \frac{2 \times P \times R}{P + R} \tag{8}$$

## 4.1 系统方法可行性分析

为了说明系统抽取方法的必要性和可行性,设计 了消融实验探究各个特定模块对抽取结果的影响,得 出了如表 2 所示的实验结果.

由表 2 可知语义增强的多策略算法取得了最好的 政策术语抽取效果,移除了语义增强、凝固度、自由 度、规则过滤和 C-value 特征中的任一策略都使得政 策术语抽取效果变差.

表 2 1940 篇政策文本术语抽取效果 (%)

算法	P	R	<i>F</i> 1
语义增强的多策略算法	75.6	78.0	76.8
移除语义增强的多策略算法	68.3	71.5	69.9
移除凝固度的多策略算法	62.6	65.1	63.8
移除自由度的多策略算法	53.3	58.3	55.7
移除规则过滤的多策略算法	53.0	60.9	56.7
移除C-value的多策略算法	50.3	59.4	54.5

## 4.2 系统结果有效性分析

为了说明系统抽取结果的可用性和有效性,对验 证数据集抽取的 3 436 条术语进行统计分析, 词库中的 低频长词占比为55%,通过普通的术语抽取方法难以 抽取得到. 系统抽取的政策术语示例如表 3 所示.

表 3 政策术语抽取结果示例

	per series and perfect as a	
序号	正确术语	术语词频
1	智慧/长途/货运	2
2	失业/保险/技术/技能/提升/补贴	3
3	核酸/检测/阴性/证明	5
4	工业/强基/一条龙/应用/推进/单位	6
5	农业/供给/侧/结构性/改革	7

## 5 结束语

本文介绍了语义增强的多策略政策术语抽取系统

System Construction 系统建设 157

的设计与实现. 该系统针对政策术语的时效性、低频度和复合短语等特点,设计了一种基于统计学方法和语言学方法的多策略冷启动算法,并在冷启动得到政策术语词库后,利用预训练语言模型语义增强方式召回低频且稀疏的政策术语,提供交互式页面对词库进行了循环更新,实现了半自动化的政策术语抽取,有助于政务企业对政策内容的智能解读,提升企业政策精准推送服务效果.

## 参考文献

- 1 Wang H, Wang B, Zou MY, *et al.* New cyber word discovery using Chinese word segmentation. Proceedings of the IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). Chengdu: IEEE, 2019. 970–975. [doi: 10.1109/ITNEC.2019.8729065]
- 2 曾浩, 詹恩奇, 郑建彬, 等. 基于扩展规则与统计特征的未登录词识别. 计算机应用研究, 2019, 36(9): 2704-2707, 2711. [doi: 10.19734/j.issn.1001-3695.2018.02.0140]
- 3 赵志滨, 石玉鑫, 李斌阳. 基于句法分析与词向量的领域新词发现方法. 计算机科学, 2019, 46(6): 29-34. [doi: 10.11896/j.issn.1002-137X.2019.06.003]
- 4 Kafando R, Decoupes R, Valentin S, *et al.* ITEXT-BIO: Intelligent term extraction for biomedical analysis. Health Information Science and Systems, 2021, 9(1): 29. [doi: 10.1007/s13755-021-00156-6]
- 5 Chen MJ, Xie ZP, Chen XQ, et al. Novel bidirectional aggregation degree feature extraction method for patent new word discovery. Journal of Computer Applications, 2020, 40(3): 631–637. [doi: 10.11772/j.issn.1001-9081.2019071193]
- 6 王煜, 徐建民. 用于网络新闻热点识别的热点新词发现. 计算机应用, 2020, 40(12): 3513-3519. [doi: 10.11772/j.issn. 1001-9081.2020040549]

- 7 Li P, Guang YX, Qiao TL. Research on Chinese new word recognition method. Proceedings of the 4th International Conference on Electronic Information Technology and Computer Engineering. Xiamen: ACM, 2020. 703–707. [doi: 10.1145/3443467.3443839]
- 8 陈先来, 韩超鹏, 安莹, 等. 基于互信息和逻辑回归的新词 发现. 数据分析与知识发现, 2019, 3(8): 105-113. [doi: 10. 11925/infotech.2096-3467.2018.1445]
- 9 Chen P, Lv XQ, Sun N, et al. Building phrase dictionary for defective products with convolutional neural network. Data Analysis and Knowledge Discovery, 2020, 4(11): 112–120. [doi: 10.11925/infotech.2096-3467.2020.0214.]
- 10 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. 南京理工大学学报, 2021, 45(1): 37-45. [doi: 10.14177/j.cnki.32-1397n.2021.45.01.004]
- 11 Qian Y, Du Y, Deng XW, et al. Detecting new Chinese words from massive domain texts with word embedding. Journal of Information Science, 2019, 45(2): 196–211. [doi: 10.1177/0165551518786676]
- 12 张乐, 冷基栋, 吕学强, 等. MWEC: 一种基于多语义词向量的中文新词发现方法. 数据分析与知识发现, 2022, 6(1): 113-121. [doi: 10.11925/infotech.2096-3467.2021.0684]
- 13 Choi KH, Na SH. FastText and BERT for automatic term extraction. Annual Conference on Human and Language Technology. Human and Language Technology, 2021: 612–616.
- 14 肖仰华, 徐波, 林欣, 等. 知识图谱: 概念与技术. 北京: 电子工业出版社, 2020.
- 15 Liu YH, Ott M, Goyal N, et al. RoBERTa: A robustly optimized BERT pretraining approach. arXiv: 1907.11692, 2019.

(校对责编: 孙君艳)