面向数字经济政策文本的新词发现方法研究

舒春萍

广东财经大学，广东 广州 510320

摘　　要： 随着数字经济的快速发展，数字经济政策文本中涌现出大量新兴词汇，准确识别这些新词汇对于政策分析至关重要。传统的新词发现方法存在一定局限性，难以应对大规模文本和快速变化的技术词汇。本文提出了一种结合统计特征、规则基础和语义分析的无监督新词发现方法，旨在提升数字经济政策文本中新词识别的准确性。通过 N-gram 建模、词性规则、改进互信息（PMI）和左右邻接熵等特征，并结合 BERT 语义验证，本方法能够有效提高新词提取的精度和召回率。实验结果表明，所提方法在准确性和效率上优于传统方法，为数字经济政策的文本分析提供了一种新的技术工具，并为政策制定者提供了更加精准的分析手段。

关 键 词： 数字经济政策；新词发现；N-gram；无监督方法；统计特征

Research on Neologism Discovery Methods for Digital Economy Policy Texts

Shu Chunping

Guangdong University of Finance and Economics, Guangzhou, Guangdong 510320

Abstract: With the rapid development of digital economy, a large number of new words have emerged in the policy texts of digital economy. Accurately identifying these new words is very important for policy analysis. Traditional neologism discovery methods have some limitations, which are difficult to cope with large-scale texts and rapidly changing technical vocabulary. This paper proposes an unsupervised neologism discovery method that combines statistical features, rule basis and semantic analysis to improve the accuracy of neologism recognition in digital economy policy texts. By means of N-gram modeling, part-of-speech rules, improved mutual information (PMI), left and right adjacency entropy, and BERT semantic validation, this method can effectively improve the accuracy and recall rate of new words. The experimental results show that the proposed method is superior to the traditional method in accuracy and efficiency, and provides a new technical tool for the text analysis of digital economy policies and a more accurate analysis means for policy makers.

Keywords: digital economy policy; new word discovery; N-gram; unsupervised method; statistical characteristics

引言

随着数字经济的快速发展，数字经济政策在国家经济发展中的地位愈加重要，政策内容和技术词汇也在不断变化与更新。特别是在数字技术不断进步的背景下，数字经济政策文本中涌现出大量新兴词汇，反映了政策趋势和技术进步。因此，准确有效地识别这些新词汇，成为政策分析中的一个重要问题。

传统文本分析方法依赖人工标注或规则驱动，虽然能在一定程度上发现新词，但随着文本量增加和领域变化，现有方法逐渐显现局限性。如何通过创新算法提高新词发现的准确性和效率，成为亟须解决的关键问题。

本文提出一种无监督新词发现方法，结合统计特征、规则基础和语义分析的优势，旨在提升数字经济政策文本中新词识别的准确性。通过验证集和完整数据集的实验验证了方法的有效性。

一、新词发现方法概述

新词发现是文本挖掘领域的一个重要任务，尤其是在分析数字经济政策等专业领域的文本时，其准确性直接影响到后续的分析结果。现有的新词发现方法大致可以分为有监督和无监督两大类。

有监督新词发现方法

有监督的方法通常需要大量的人工标注数据，通过机器学习模型进行训练。常见的有监督方法包括支持向量机（SVM）[1]、隐性马尔可夫模型（HMM）和条件随机场（CRF）等。[2] 这些方法依赖人工标注语料库，以训练分类器识别文本中的新词。虽然有监督方法在精度上表现较好，但由于标注成本高、数据要求严格，因此在大规模自动化的新词发现任务中，应用受限。[3]

无监督新词发现方法

无监督方法不依赖人工标注数据，能够自动从大规模文本中提取新词，因此在实际应用中更具优势。无监督方法的研究主要集中在以下几类：

（1）基于规则的方法[4]：如 W.Ma 等（2003）提出的归并算法，通过引入构词规则自动检测新词。这类方法简单易用，但对文本的适应性差，容易受到规则的限制。

（2）基于统计的方法：统计方法主要依赖信息熵、互信息、改进互信息等统计量，通过计算词语共现关系来发现潜在的新词。例如，Pecina 等（2006）采用点互信息（PMI）算法，通过词汇相关度来识别新词。统计方法的优势在于无需人工干预，能够高效处理大规模数据，但它往往忽略了语义信息，导致对复杂词组的判断较为粗糙。[5]

（3）基于规则与统计结合的方法：如 Xu Sun 等（2012）提出的结合词特征和 CRFs 的方法，利用统计信息和规则特征的结合提高新词检测的准确性。这种方法在提升新词识别精度的同时，也能够适应特定领域的特点。[6]

（4）基于语义分析的方法：如 Chen 等（2018）提出了一种基于词嵌入和共现分析的新词发现方法，结合了 Word2Vec 和共现信息分析，捕捉词汇间的相似性进而识别新词。近年来，深度学习方法，如 BERT 模型，已被广泛应用于新词发现任务。BERT 能够通过上下文信息提供更精确的语义表示，从而提高新词的准确识别。[7]

尽管现有的无监督新词发现方法在一定程度上取得了较好的效果，但它们在数字经济政策文本分析中的应用仍存在一定局限性。传统的基于规则和统计的方法对新兴技术词汇的适应性差，难以捕捉到数字经济政策文本中的新兴技术词汇，尤其是在政策语言快速变化的背景下；此外，大多数现有方法，特别是基于统计和规则的方法，通常只关注词汇的共现频率和局部模式，忽略了新词的语义完整性，容易导致部分新词的遗漏。针对现有方法的局限性，本文提出了一种结合统计、规则和语义特征的无监督新词发现方法。该方法不仅能提高新词识别的准确性，还能更好地适应数字经济政策文本的特点。[8]

方法

在具体实验中，本文采用了以下技术流程进行新词发现：

（1）N-gram 建模

N-gram 是指由 N 个连续的单词或字符组成的序列。在 N-gram 模型中，整个文本被分割成连续的 N 个项，通常是单词或字符。在本实验中，通过计算 N-gram 的频率，筛选出满足特定频率阈值的候选词组作为新词候选项。

（2）基于规则的词性筛选

为了提高新词发现的准确性，本实验参考了张越等的词性规则，对候选词组进行了基于规则的筛选。具体而言，实验首先设定了特定的词性组合规则，用于筛选短词组（如2-gram、3-gram 和4-gram）的结构，具体如表3-1所示。此外，本实验也尝试了对较长词组（如5-gram 及以上）进行筛选，但发现长词组的语义完整性较差，实用性不强。因此，最终实验仅输出了2-gram、3-gram 和4-gram 的结果。这种基于规则的筛选方式，减少了无关短语的出现，提高了新词发现的精准度。

表3-1 词性筛选规则

词串数	词性组合规则
2-gram	n.+n.、v.+n.、v.+v.、n.+v.、a.+n.
3-gram	n.+n.+n.、v.+n.+n.、v.+n.+n.、a.+n.+n.
4-gram	n.+n.+n.+n.

（3）凝固度（PMI）

凝固度是衡量 N-gram 内字与字或词与字之间紧密程度的指标，通常使用互信息（PMI）来计算。PMI 的计算公式（3-1）如下所示。对于一个二元的 N-gram（如 w1 和 w2），其 PMI 值为：

$$PMI(w_1,w_2)=\log\frac{p(w_1,w_2)}{p(w_1)\times p(w_2)} \qquad 式（3-1）$$

其中，$p(w_1,w_2)$ 表示 w_1 和 w_2 同时出现的概率，而 $p(w_1)$、$p(w_2)$ 分别表示 w_1 和 w_2 单独出现的概率。当 PMI 值较高时，表示 w_1 和 w_2 之间具有较强的相关性，即该 N-gram 可能组成一个有意义的新词。本实验对二元、三元和四元的 N-gram 分别计算 PMI 值，以评估词组内部的凝固度。

（4）自由度（左右邻接熵）

自由度用于衡量 N-gram 是否具备丰富的左右搭配，即其独立成词的潜力。本文使用邻接熵来衡量自由度，通过计算 N-gram 左、右两侧邻字的分布熵值，反映出候选词组左右搭配的丰富性。左右邻接熵的计算公式（3-2）（3-3）如下所示。

$$H_L(X)=\sum_{w_l\in D_l}P(w_l|X)\times\log_2 P(w_l|X) \qquad 式（3-2）$$

$$H_R(X)=\sum_{w_r\in D_r}P(w_r|X)\times\log_2 P(w_r|X) \qquad 式（3-3）$$

其中，其中，$H_L(X)$ 和 $H_R(X)$ 分别是片段 X 的左右边界信息熵，D_l 和 D_r 分别是片段 X 的左右邻字集合。$P(w_l|X)$ 为片段 X 出现时左侧邻字为 w_l 的条件概率，同理，$P(w_r|X)$ 为片段 X 的出现时右侧邻字为 w_r 的条件概率。较高的邻接熵值表明该 N-gram 在文本中具有多样的搭配，较可能构成独立的新词。

（5）BERT 语义相似度

BERT（Bidirectional Encoder Representations from Transformers）是一种基于深度学习的语言模型，能够通过上下文信息输出融合了全文语义的词向量。本文使用 BERT 模型（Bert-base, Chinese）生成文本中各个词的向量表示，并通过余弦相似度衡量候选新词与上下文的语义一致性，以确保提取的新词在语义上具有完整性和一致性。相似度公式（3-4）如下所示：

$$Cosine\ Similarity=\frac{A\cdot B}{A\times B} \qquad 式（3-4）$$

BERT 模型在本实验中的作用主要体现在为候选词组提供语义层面的验证，结合 N-gram 模型的统计信息，实现语义和统计的双重筛选。

二、实验设计与结果

（一）数据收集与处理

实验首先从中国政府网收集选取了226篇国家层面发布的数字经济政策文件，并对文本数据进行了预处理，主要步骤包括去除标点符号、去除多余的空格与换行符，并应用自定义的停用词表过滤噪声词。为了确保实验方法的通用性与科学性，本文从收集的数字经济政策文本中随机抽取500条句子，手动标注新词，构建测试集。测试集覆盖政策和法律文本中的关键术语和技术短语，确保标注内容具有多样性和代表性。[9]

（二）评价指标

为了量化评估新词发现方法的性能，本文采用了精确率、召

回率和F1值三个常用指标，具体定义如下：

（1）精准率（Precision）：用于衡量模型预测的新词中有多少是真正的新词，定义如下：

$$Precision = \frac{|T \cap P|}{|P|} \quad \text{式（4-1）}$$

其中，T 表示测试集中标注的新词集合，P 表示模型预测的新词集合，$|T \cap P|$ 表示正确预测的新词数量，即真实新词与预测新词的交集。

（2）召回率（Recall）：衡量测试集中标注的新词中有多少被模型成功预测，定义如下：

$$Recall = \frac{|T \cap P|}{|T|} \quad \text{式（4-2）}$$

其中，分母 $|T|$ 表示测试集中标注的新词总数，分子与精准率公式中的分子一致。

（3）F1值：是精准率与召回率的加权调和平均值，用于综合衡量模型的预测性能，定义如下：

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad \text{式（4-3）}$$

其中，F1值的范围为[0,1]，值越高表示模型在精准率与召回率之间取得了更好的平衡。

（三）参数调优

在测试集上，本文通过遍历不同的参数组合，寻找F1值最高的阈值配置。设置词组出现频率的下限为5，剔除出现次数少于5的候选词组，以减少低频无意义短语的干扰；设定PMI阈值为2.0，确保词组内字符间的关联性；通过左右邻接熵过滤掉边界搭配单一的候选词组，阈值设定为1.0；利用BERT模型筛选语义完整的词组，最终确定相似度阈值为0.6。

最终，模型在测试集上的精确率为0.9216，召回率为0.6714，F1值为0.7768，表明参数配置具有较高的合理性与适应性。

消融实验

为了评估每个筛选特征在新词发现方法中的具体贡献，本文设计了消融实验，依次去除特定的筛选特征并观察实验结果的变化，包括去除BERT语义特征、词性规则、改进互信息（PMI）和左右邻接熵。实验结果如表4-1所示。

表4-1 消融实验

	P	R	F1
全特征	0.9216	0.6714	0.776852
去除BERT	0.8942	0.62	0.732273
去除词性	0.8415	0.616	0.711306
去除PMI	0.9174	0.5714	0.704194
去除熵值	0.785	0.672	0.724118

结果显示：去除BERT语义特征后，F1值从0.7769降至0.7323，主要体现在召回率（Recall）的下降。这表明BERT的语义信息在捕捉上下文中起到了重要作用。此外，语义特征还弥补了统计和规则特征对复杂语义关系捕捉不足的局限性；去除词性规则后，F1值进一步下降至0.7113，导致精确度下降，模型对噪声词组的识别能力减弱。去除PMI后，F1值下降到0.7042，同时召回率显著降低（从0.6714降至0.5714），使模型难以区分语义独立性强的词组和随机组合的低频词组；去除左右邻接熵后，F1值降至0.7241，精确率下降更为明显（从0.9216降至0.7850），导致许多上下文依赖性较高的词组被错误保留，降低了模型的精确度。

从以上实验结果可以看出，BERT语义特征、词性规则、改进互信息（PMI）和左右邻接熵各自在新词发现中发挥了不同的作用。在这些特征的共同作用下，最终实现了精确率和召回率的平衡。

对比实验

为了验证所提方法的有效性，本文与两种主流的新词发现方法进行了对比实验，分别是基于改进互信息和左右邻接熵的方法（MMI+Entropy）[17]以及基于词嵌入和频繁n元组挖掘的WEBM方法[10]，具体结果如表4-2所示。

表4-2对比实验

	P	R	F1
论文方法	0.9216	0.6714	0.776852
MMI+Entropy	0.762	0.6897	0.72405
WEBM	0.8087	0.6435	0.716704

结果表明，本文方法在P、R和F1值方面均优于对比方法，体现了多特征融合方法在新词发现中的显著优势。

三、结论

本文提出了一种结合统计特征、规则基础和语义分析的无监督新词发现方法，旨在提升数字经济政策文本中新词识别的准确性。通过结合N-gram建模、词性规则、改进互信息（PMI）和左右邻接熵等特征，并引入BERT语义验证，本文方法有效提高了新词提取的精度和召回率。实验结果表明，所提方法在精确度、召回率和F1值上优于传统方法，尤其在识别低频词和新兴技术术语方面表现突出。尽管如此，未来研究可以通过优化词向量模型和引入更多领域特定数据进一步提升方法的适应性和准确性。

参考文献

[1] 师博，方嘉辉. 数字经济赋能中国式新型工业化的理论内涵、实践取向与政策体系［J］. 人文杂志. 2023,(1).

[2] 唐超，陈颖淇，胡宜挺. 我国数字素养教育政策的演进脉络与结构特征［J］. 图书馆论坛. 2023,43(11).

[3] 曹树金，曹茹烨. 基于研究主题和引文分析的信息资源管理学科发展探究［J］. 信息资源管理学报. 2023,(2).

[4] 陈光，钟方媛，明翠琴，等. 地方政府创新政策工具偏好测量
——基于四川省政策文本的分析［J］. 科技进步与对策. 2023,40(2).

[5] 曹玲静，张志强. 政策信息学视角下政策文本量化方法研究进展［J］. 图书与情报. 2022,42(6).

[6] 雷浩伟，廖秀健. 省级政府大数据发展应用政策的规制导向与执行优化研究——基于政策文本的分析［J］. 公共管理与政策评论. 2022,(2).

[7] 李川川，刘刚. 发达经济体数字经济发展战略及对中国的启示［J］. 当代经济管理. 2022,44(4).

[8] 潘琳，徐鸣. 我国社区治理领域政策分析与评价研究——基于”过程—工具—内容”三维分析框架［J］. 理论学刊. 2022,(6).

[9] 雷鸿竹，王谦. 中国地方政府数字经济政策文本的量化研究［J］. 技术经济与管理研究. 2022,(5).

[10] 张清慧，陈谊，武彩霞. 基于词表示模型的领域文献数据可视分析方法［J］. 图学学报. 2022,43(4).