基于词向量的自适应领域术语抽取方法

唐溪[1]，**蒋东辰**[1*]，**蒋翱远**[2]
1.北京林业大学信息学院，北京，100083
2.中原银行股份有限公司，郑州，450046
{tangxi,jiangdongchen}@bjfu.edu.cn jiangaoyuan@zybank.com.cn

摘要

术语分布呈现长尾特性。为了有效提取低频术语，本文提出了一种基于词向量的自适应术语抽取方法。该方法使用基于假设检验的统计方法，自适应地确定筛选阈值，通过逐步合并文本的强关联性字符串获得候选术语，避免了因固定阈值导致的低频术语遗漏问题；其后，本文基于掩码语言模型获得未登录候选术语的词向量，并通过融合词典知识的密度聚类算法获得候选术语归属的领域簇，将归属于目标领域簇的候选术语认定为领域术语。实验结果表明，我们的方法不仅在F值上优于对比方法，而且在不同体裁的文本中表现更为稳定。该方法能够全面有效地抽取出低频术语，实现领域术语的高质量提取。

关键词： 术语抽取；自适应；假设检验；词向量

An Adaptive Domain-Specific Terminology Extraction Approach Based on Word Embedding

Xi Tang[1],**Dongchen Jiang**[1*],**Aoyuan Jiang**[2]
1. School of Information Science and Technology,
Beijing Forestry University, Beijing, 100083, China
2.Zhong Yuan Bank Co.,Ltd, Zhengzhou, 450046, China
{tangxi,jiangdongchen}@bjfu.edu.cn jiangaoyuan@zybank.com.cn

Abstract

Terminology distribution shows a long-tail pattern. This study presents an adaptive term extraction method based on word embedding to effectively extract low-frequency terms. Using a hypothesis-testing statistical approach, the method adaptively sets filtering thresholds and acquires candidate terms by incrementally merging strongly related text strings, avoiding omission of low-frequency terms due to fixed thresholds. Word embeddings for out-of-vocabulary candidates are obtained through a masked language model, and a dictionary-integrated density clustering algorithm identifies domain clusters for these terms. Candidates within target domain clusters are recognized as domain-specific terms. Experimentally, our method outperforms competitors in F-score and maintains stability across diverse text genres. This approach effectively extracts low-frequency terms, ensuring high-quality domain-specific term extraction.

Keywords: Terminology extraction , Self-adaptation , Hypothesis testing , Word embedding

* 通讯作者

第二十二届中国计算语言学大会论文集，第186页-第195页，哈尔滨，中国，2023年8月3日至5日。

1 介绍

术语是特定领域内用以传达专业概念的约定俗成的语言符号。术语抽取作为自然语言处理领域的一项核心技术，能够从大规模语料库中自动提取领域术语，从而降低手工发现术语的人力成本。此外，术语抽取可应用于诸如文本分类、信息抽取、机器翻译等下游任务中，为其提供更为丰富的领域概念知识。例如，在机器翻译任务中，将领域术语作为先验知识引入，能够有效提升翻译质量(Michon et al., 2020; 游新冬et al., 2021)。因此，术语抽取在自然语言处理领域具有重要的应用价值。

研究发现，领域术语的分布遵循长尾分布(Williams et al., 2015)，低频术语在全部领域术语中所占比例极高。这意味着如果不能有效实现低频术语抽取，将导致大量术语无法被识别。在实际应用场景中，新兴术语和非规范术语作为术语标准化工作的重点(刘书剑and 彭道黎, 2011)，它们大都以低频形式出现。准确识别低频术语是及时发现新兴术语和规范术语使用的基础，能够有效辅助术语标准化工作。因此，关注低频术语识别具有重要意义，低频术语抽取质量直接影响术语抽取的全面性。

现有研究在抽取领域术语方面已取得一定成果，然而针对低频术语的抽取仍存在较大挑战。低频术语在语料库中出现次数较少，使得诸如基于统计、机器学习或深度学习的方法在处理这类术语时很难获得理想的效果。针对低频术语抽取困难这一问题，部分学者关注低频术语的构词规律，通过语法规则特征匹配来抽取低频术语(俞琰and 赵乃, 2018; 李思良et al., 2018)。尽管这类方法在低频术语抽取方面具有一定的有效性，但方法所依赖的构词规律却呈现出显著的领域依赖性，使得领域迁移性相对较弱。

为了实现各领域低频术语的有效识别，本文提出一种基于词向量的自适应领域术语抽取方法。该方法首先采用假设检验的统计方法自适应地获取目标文本中候选术语的频率阈值，对低频候选术语筛选具有显著效果。在术语确认阶段，本文采用词向量技术，通过融合词典知识的密度聚类评估候选术语的领域相关性，并将具有领域语义的候选术语确定为最终领域术语。本方法能够有效抽取低频术语，实现准确度较优的术语抽取效果。

本文的组织结构如下：第二节介绍了研究的相关工作。第三节详细阐述了我们的术语抽取框架的细节。第四节展示了我们的实验设置和结果。第五节给出了研究结论。

2 相关工作

目前，研究者主要采用语法规则、统计计算、传统机器学习以及深度学习四类方法抽取术语。基于语言规则的方法从术语的构词规律出发，通过语法模板匹配得到术语。对于英语等拉丁语系语言，词性分析准确率相对较高，因此可通过匹配词性序列来抽取文本中的术语，甚至是术语变体(Kafando et al., 2021)。但中文的情况更为复杂，在词性分析前，通常需要先对语料分词，这一额外步骤可能会带来累积误差，从而增加了词性分析的难度。因此，基于语言规则的中文术语抽取方法更多使用领域常用词根和词缀辅助分析(孙水华et al., 2016; 李思良et al., 2018)。这种方法简单且有效，但其适应不于同领域和文本风格的能力有限；同时，规则匹配的形式较为僵化，这也限制了识别新兴术语的能力。

术语识别是一项普遍应用于各领域的任务，而统计方法为此提供了一种不针对特定领域的通用解决方案。基于统计的术语识别方法通常包括两个步骤。首先，从文本中找出能够表达独立概念的语义单元，作为候选术语。这一步中常借助互信息和邻接熵(刘伟童et al., 2019; 李贞贞et al., 2022)、对数似然比(王大亮et al., 2008)等统计指标，评估语义单元内部结合强度以及整体独立性，一旦统计指标超过预定的阈值，即可确认该语义单元为候选术语。在第二步中，则是要综合领域相关度对候选术语的置信度排名，确定前若干名为术语。这一步往往根据候选术语在目标领域与普遍场景的词频分布差异，利用C-Value(Frantzi et al., 2000)、TF-IDF(董洋溢et al., 2017)及其改进版本(俞琰et al., 2020; Kosa et al., 2020)等统计量，评估候选术语的领域相关度。然而，这类方法存在一些问题：由于统计量大多以词频作为核心计算因素，低频的术语的统计指标可能会偏低，甚至接近于非术语；同时，为了有效地排除噪声，这类方法的统

基金项目：北京市优秀人才培养资助项目(2018000020124G063)；

计量往往会主观设定一个较高的阈值，这可能导致低频术语被遗漏(蒋婷, 2021)。因此，现有的基于统计的术语抽取方法，在识别低频术语上仍面临较大的挑战。

传统机器学习和深度学习方法均将术语抽取视作序列标注任务。在用于序列标注的机器学习模型中，条件随机场最为经典，已被用于多个领域的术语抽取任务(木合亚提·尼亚孜别克et al., 2016; 黄菡et al., 2019)。不同于最大熵模型或隐马尔科夫模型，条件随机场直接对标签序列的联合分布建模，从而允许引入各种复杂的特征，这在前两者中是较难实现的。然而，机器学习模型的性能通常受限于人工特征选择的准确性。近年来，深度学习的发展为这一问题提供了解决方案。神经网络模型，如门控循环单元(Kucza et al., 2018)、双向长短期记忆网络(吴俊et al., 2020)、图卷积网络(任秋彤et al., 2021)等，可以自动提取上下文特征；将这些特征应用于条件随机场，会进一步提升术语抽取的效果。但无论是特征工程方法还是深度学习方法，都需要一定规模的标注语料协助模型训练。这就需要专业人员参与数据标注，标注成本相对较高。另外，由于低频术语在训练语料中的出现次数较少，模型可能无法从有限的样本中学习到低频术语的特征知识，从而导致模型对低频术语的识别能力不足

由此可见，虽然现有的各种术语抽取方法从不同角度深入研究了术语抽取任务，但对于低频术语的抽取仍然存在不同的局限。因此，我们提出了一种术语抽取方法，该方法融合了基于假设检验的统计方法和词向量技术，能够有效地抽取低频术语。

3 方法

术语是指在特定领域或学科内，用于表示专业概念的一组约定性的语言符号。这些符号可以是单个词语、短语或缩略词，其目的是为了在学术研究、技术开发和行业应用等场景下，确保概念的精确表达与交流。本文参照术语的使用场景，认为领域术语应具有以下特点：

(1) 单元性：构成一个术语的字符串在文本中作为一个独立单元使用，单元内部呈现出强关联性，以术语“钢筋混凝土”为例，“钢筋”与“混凝土”在此共同参与构成了一个完整的术语，单独识别“钢筋”或“混凝土”是不符合单元性要求的；

(2) 领域性：术语是特定领域内的专门用语，术语所表达的语义内容应归属于其所属的领域；

(3) 专业认可性：术语在特定领域内具有权威性，术语的使用需要得到专业人士的认可，从而确保领域内沟通的准确、高效。

在上述标准中，专业认可性必须由领域专业人士确定，目前难以用机器替代；但单元性和领域性确是在一定程度上可由计算机完成。因此，针对术语的低频特性，我们基于术语的单元性和领域性设计了一种无监督领域术语抽取方法。整体流程如图1所示。该方法分别针对单元性、领域性要求，设计候选术语单元性识别、候选术语领域性筛选两个模块。候选术语单元性识别模块依靠文本自身信息，识别文本中语义完整、使用独立的字符串作为候选术语；候选术语领域性筛选模块评估候选术语与所属领域的领域相关性，将归属于特定领域的候选术语认定为领域术语。

术语自动抽取能够显著减少专家在术语整理和归纳过程中的工作量。通过将术语抽取技术融入专家工作，能够进一步优化术语标准化流程的效率与准确性。

3.1 候选术语单元性识别

参照单元性定义，本文将“钢筋混凝土”这样具有单元性的字符串称为一个语义单元，一个语义单元可能由一个词或多个词构成。本步骤将文本划分为若干语义单元，作为候选术语。

候选术语识别与中文新词识别任务有共同之处：术语是具有领域性的语义单元，新词是未登录的语义单元，两者都需要满足单元性要求。在新词识别中，识别语义单元的方法包括互信息、信息熵和假设检验等。前两者需人为设定阈值，达到阈值的候选项视为语义单元。然而，阈值设定通常具有主观性，其高低影响筛选效果。过低的阈值无法筛除干扰项，而过高的阈值可能导致低频词汇被误筛。基于假设检验的方法根据给定文本和语料库的频率信息自适应地为不同频率字符串设置不同的阈值，消除了主观设定阈值的局限性。Jiang等人(2022)提供了一个基于假设检验的语义单元识别方法，对语义单元尤其是低频语义单元具有良好的识别效果。

Jiang等人指出语义单元内汉字或词语所具备的两个关键特征：非偶然相邻性和强关联性，并设计了相应的假设检验方法用于判别两个特征。具体来说，针对语义单元内部任意相邻字符

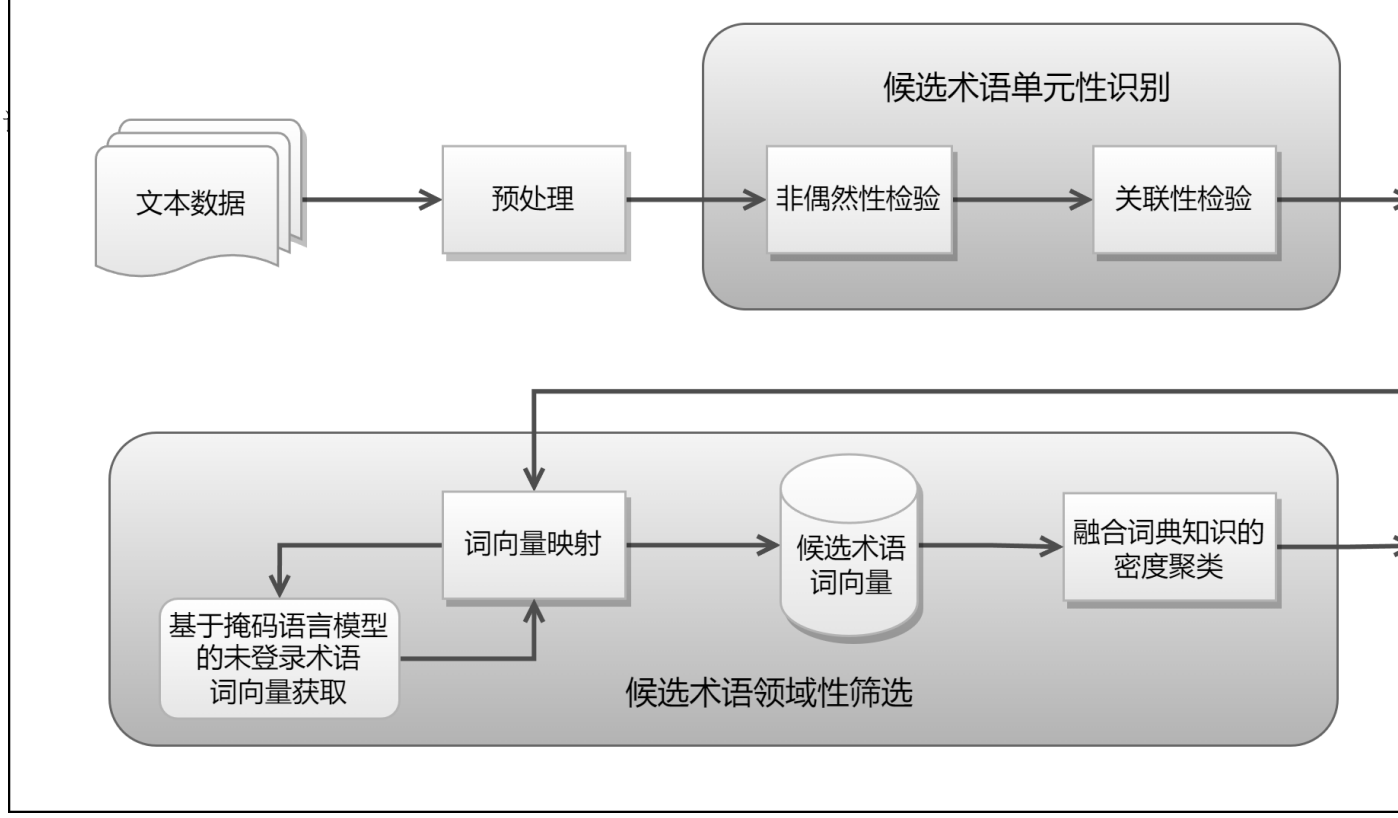

图 1: 领域术语抽取流程

之间的非偶然相邻特征，可以采用非偶然性检验评估相邻字符共同出现的显著性；考虑到构成语义单元的字符串作为一个整体在文本中表现出的强关联性时，可以使用关联性检验来判断相邻语义单元结合的紧密程度。

然而，这种自适应获得阈值的形式会为高频常见词设置较高的阈值，导致一些高频常见词不能通过非偶然性检验，进而影响了对包含高频常见词的语义单元的识别。不同于新词，许多术语含有高频常见词，如计算机领域术语“支持向量机”中的“支持”。为适应术语特点，本文将分词信息整合至Jiang等人的方法中，从而提高对包含高频常见词的语义单元的识别效果。具体而言，本文首先使用分词工具分割文本。现有的分词工具能够高效准确地识别高频常见词，因此我们无需对文本中全部相邻汉字对进行非偶然检验；相反，我们只需对相邻分词结果的邻接汉字对进行非偶然性检验，既提高了判断效率，又消除了非偶然性检验识别高频常见词的局限性。这样的融合策略能够更有效地识别高频常见词，进一步优化语义单元识别。

具体的，本文首先使用PKUSEG(Luo et al., 2019)对文本分词，分词结果存储为一个词语序列。对序列中相邻词语$[P,Q]$, 使用$< l_P, f_Q >$ 表示$[P,Q]$ 之间的相邻汉字对, 其中, l_P 表示词P 中的最后一个汉字, f_Q 表示词Q 中的第一个汉字。对$< l_P, f_Q >$ 进行非偶然性测试。

非偶然性检验的原假设为: 文本中任何相邻汉字对的频率都与它的一般频率特征一致。如果相邻汉字对的实际频率明显高于它的一般频率, 就可以判断该相邻汉字对为非偶然性相邻。

假设文本中任意两个汉字相邻出现的概率服从泊松分布, 则可以根据相邻汉字对的频率估计参数λ :

$$\lambda = N_T \times p = N_T \times \frac{n_{i,j}}{N} \tag{1}$$

式中$n_{i,j}$ 为相邻汉字对$< c_i, c_j >$ 在语料库中出现的频次，N 表示所有相邻汉字对出现的总频次，N_T 表示文本中相邻汉字对个数。可以通过以下公式计算$< c_i, c_j >$ 出现n 次的累计概率:

$$F_c\left(c_i, c_j, n\right) = \frac{\sum_{k=1}^{n} \frac{e^{-\lambda}\lambda^k}{k!}}{1 - e^{-\lambda}} \tag{2}$$

给定显著性水平α_p, N_a 表示$< c_i, c_j >$ 在文本中的实际出现次数, 若$F_c\left(c_i, c_j, N_a\right) > 1 - \alpha_p$, 则可以推断$< c_i, c_j >$ 是因非偶然性因素相邻的汉字对。

当$< l_P, f_Q >$ 满足非偶然性检验时, 才对相邻词语$[P,Q]$ 进行关联性检验。关联性检验的原假设为: 对于任意相邻词语$[A,B]$, A 、B之间的关联性不足以构成语义单元。

给定文本T, a, b, c, d 分别表示相邻语言单元$[A,B], [\bar{A},B], [A,\bar{B}], [\bar{A},\bar{B}]$ 出现在文本T 中的频

次, 用$[\bar{A}, B]$ 表示词B 的前一位置不是词A 的情况。构建统计量$Q^2_{A,B}$：

$$Q^2_{A,B} = \frac{(a+b+c+d) \times (ad-bc)^2}{(a+b) \times (c+d) \times (a+c) \times (b+d)} \tag{3}$$

给定显著性水平, 如果统计量$Q^2_{A,B}$ 的值在拒绝域中, 应拒绝原假设。这说明A 、B呈现强关联, 通过关联性检验。

若$[P, Q]$ 通过非偶然性检验、关联性检验，则意味着P 、Q可以共同构成语义单元或语义单元的一部分，应当合并。通过多轮迭代，合并所有能组成语义单元的词语，至文本中无可合并的词语为止。这时将当前所有获得的语义单元结果作为候选术语。

3.2 候选术语领域性筛选

术语的领域性是其另一个基本特征。由于候选术语是基于统计方法识别的，尚未判断其语义，导致筛选结果可能包含一些与领域无关的词汇。为解决这一问题，本文将进一步评估候选术语语义与所属领域语义的相关性，从而筛除领域外词汇，仅保留与领域相关的候选术语。为实现候选术语的领域性评估，本文采用基于词向量的聚类方法，这种方法相较于传统的TF-IDF方法具有优势，无需领域专用语料库即可执行。

词语的语义内涵可以通过词向量表示。词向量是基于大量语料训练得到的，常见的词向量模型如Word2Vec(Mikolov et al., 2013)能够有效表示训练语料中高频词汇。然而，针对领域术语的低频特性，Word2Vec模型存在一定的局限性。对于那些在训练语料中出现频次极低或根本未出现的词汇，Word2Vec无法生成相应的词向量，这类词汇很可能是新兴术语、非规范术语等低频术语，这一局限性使得Word2Vec在处理这类低频术语时面临诸多挑战。鉴于此，我们根据词向量的可用性，将候选术语分为已登录和未登录两类，使用基于Word2Vec模型的词向量数据集(Song et al., 2018)获取已登录候选术语的词向量。对于未登录候选术语，本文提出一种基于掩码语言模型的方法构建其向量表示，以克服低频特性对词向量表示的负面影响。

3.2.1 基于掩码语言模型的未登录候选术语词向量获取

掩码语言模型最初作为预训练任务在2018年提出，广泛应用于Bert(Devlin et al., 2019)、RoBERTa(Liu et al., 2019)等预训练模型。其思想方法为：按一定策略遮盖文本中的单词，让预训练模型根据遮盖位置的上下文信息，预测最适合填入遮盖位置的单词。具体而言，给定一个包含N个词的文本$\{x_1, x_2, x_3, \ldots, x_N\}$, 使用特殊标记[MASK] 遮盖其中第$j$ 个单词x_j, 掩码语言模型建模:

$$p\left(x_j \mid x_1, \ldots, x_{j-1}, [MASK], x_{j+1}, \ldots, x_N\right) \tag{4}$$

掩码语言模型在各类完形填空式任务中展现出良好的适应性，可直接应用于句法分析(Wu et al., 2021)、实体类型推断(Dai et al., 2021)等任务。本文利用掩码语言模型获取与未登录候选术语语义相近的替代词，并使用这些替代词的词向量构建未登录候选术语的词向量。具体来说，本文采用WoBERT预训练模型(Su, 2020)预测替代词。WoBERT是一种以词为训练单位的预训练模型，其训练任务基于掩码语言模型，因此更适合处理中文文本。

被遮盖候选术语	模型输入	模型输出结果
南荣油茶	攸县油茶、小果油茶、[MASK]、尾叶山茶等4个物种，鲜出籽率最高，达50%～60%以上。	大红袍, 油茶, 山茶
双苗砧嫁接	[MASK]效果好，当年嫁接苗高可达1.59cm，叶片数可达10片以上，它可以应用于快速培育嫁接大苗方面	嫁接, 移栽, 扦插
连续清查报告	根据《第9次全国森林资源[MASK]》，目前我国的国有林面积8274万公顷、集体林面积3874万公顷、个人所有林9673万公顷。	普查, 规划, 清查

表 1: WoBert生成替代词示例

为获得未登录候选术语的词向量，需要在WoBERT后添加softmax层，以便获取WoBERT对每一个替代词的预测概率。假设未登录候选术语w 在文本中出现M 次, 取出其所有出现位置的段落集合$S=\{S_1,S_2,\ldots,S_M\}$, 遮盖未登录候选术语后送入WoBERT，将所有位置的预测结果按预测概率由高至低排序，保留前top-K个替代词$\{w_i \mid i=1,2,3,\ldots,k\}$ 及其预测概率$\{p_i \mid i=1,2,3,\ldots,k\}$ 。由于替代词结果可能来自不同位置预测结果, 因此对预测概率归一化:

$$p'_i=\frac{e^{p_i}}{\sum_{j=1}^{K}e^{p_j}} \tag{5}$$

未登录候选术语的静态词向量$vecW$ 按如下公式计算得到:

$$vecW=\sum_{i=1}^{K}VEC\left(w_i\right)*p'_i \tag{6}$$

式中VEC (w_i)为词w_i 在词向量集合中映射得到的词向量，若VEC (w_i) 不存在, 则置为零。

3.2.2 融合词典知识的密度聚类

词向量聚类方法可以将语义相近的同领域术语映射到空间距离上的聚集，使得语义相近的同领域术语聚集在同一簇中，从而实现领域性筛选。由于术语抽取可能面临一个或多个学科分支主题的混合文本，这使得术语簇的形状难以确定。此外，候选术语中的与领域无关的词汇可能产生一定的噪声。因此，本方法采用具有抗噪声能力并能发现任意形状簇的密度聚类算法划分候选术语的语义类别。

为了在文本中的术语分布稀疏且数量较少的情况下，确保由候选术语映射的词向量在空间中能够规模化地聚集，本文从领域词典中收集已有的领域术语。将这些术语映射为词向量后，将它们与候选术语一起参与密度聚类。

本文将余弦相似度作为密度聚类中的距离指标，余弦相似度通过计算两个向量夹角的余弦来衡量两个词向量的语义相似度。词向量a,b 的余弦相似度记作$\mathrm{sim}(a,b)$ 。在此基础上，给出密度聚类中邻近集合的定义：给定向量v, 它的邻近集合$N_{\mathrm{eps}}\left(v\right)$ 定义为以v 为中心, 与v 的余弦相似度大于等于相似度阈值Eps 的向量集合, 即:

$$N_{eps}(v)=\{q \mid \mathrm{sim}(v,q)\geq \mathrm{Eps}\} \tag{7}$$

对于词向量v，给定一个密度阈值$minPts$，词向量v的邻近集合如果满足:

$$\left|N_{\mathrm{eps}}\left(v\right)\right|\geq \mathrm{minPts} \tag{8}$$

则称v 为在相似度阈值Eps，密度阈值$minPts$条件下的高密度点；自身不是高密度点但属于某个高密度点邻域集合的对象称为边界点；其余为噪声点。根据密度聚类的原理，从空间中任意一个高密度点Q开始，通过将Q附近的高密度点和边界点加入到Q所属的簇中，从而扩大簇的规模。通过不断地连接邻近集合中的高密度点，簇的方向得以拓展，最终实现对空间内词向量的聚类。在获取候选术语的聚类结果后，计算各个聚类簇与已有领域术语的交集。交集规模最大的簇被认定为领域簇，而属于领域簇的候选术语则被认定为领域术语。

4 实验

4.1 实验数据

在本节中，我们将测试所提方法的术语抽取能力，尤其是低频术语抽取能力，并与现有方法对比抽取效果。

由于尚无公开的中文术语抽取数据集，本文以林业领域为例手工构建实验文本。具体的，我们从国家林业和草原局政府网收集2022年4月至7月发布的林业碳中和主题新闻25篇，并选取湖南省常德市林业科学研究所编撰的《油茶优质高产栽培技术》一书的第四章“油茶良种繁育技术”，第八章“油茶科研应用及发展趋势”作为实验文本。实验文本主题涵盖了林业碳中和、良种繁育、林业科研成果三类不同林业主题，体裁包括林业新闻、技术类书籍，能够考察方法在不同场景的适用性。

本文采用词典与人工结合的方式标注实验文本中的林业术语，作为实验标准集。具体而言，如果一个词包含在《林学名词》、《中国林业辞典》、《中国树木志》《草业大辞典》林业辞典中，本文将其标注为林业术语。在此基础上，我们另外邀请五位林学专业研究人员以人工的方式标注出未在辞典内的林业术语，若一个词得到过半人数认可，则认定为林业术语。

预处理工作包括在去除实验数据中的图表内容与特殊字符。预处理后的文本数据如下：

主题文本	字数	术语数
林业碳中和	17,172	174
良种繁育	16,212	263
林业科研成果	15,943	208

表 2: 实验文本数据

4.2 实验结果与分析

实验采用Python3.7 编程语言和PyTorch 1.12 深度学习框架、scikit-learn0.24机器学习框架。候选术语单元性识别的显著性水平α_p 按建议固定在1×10^{-8} 。林业碳中和、良种繁育、林业科研成果三个实验数据集对应的密度聚类参数[Eps, minPts]分别设置为[0.75, 7]、[0.75, 5]、[0.75, 7]。

在实验评价标准上，本文选用准确率、召回率和综合评价指标F值评估实验结果。准确率反映抽取结果的正确性，召回率反映抽取结果的全面性，F值基于两者给出抽取质量的综合性评价。

为了验证本文提出的方法在领域术语抽取上的有效性，我们选取了中文开源术语抽取工具Termolato作为对比方法。Termolator 是一种结合了语法规则和统计的领域术语识别方法。它采用分词和词性匹配识别候选术语，并使用一般性语料库和领域语料库计算候选术语的相关统计指标，以统计指标评估候选术语的领域相关程度。在对比实验中，文本采用推荐配置运行Termolator，领域语料库选用《油茶优质高产栽培技术》除第四、第八章外的内容，一般性语料库则选用中文维基百科语料库*。

值得注意的是，可以通过设置词频阈值来调整Termolator的抽取效果。由于本文方法能够抽取最低词频为1的术语，为了在同等条件下对比，我们将Termolator的词频阈值也设置为1。在林业碳汇、良种繁育、林业科研成果等三个数据集上实验，结果见表3。

数据集	方法	准确率	召回率	F值
林业碳中和	本文方法	73.30	74.14	73.71
	Termolator	8.81	37.36	14.27
良种繁育	本文方法	76.30	78.33	77.30
	Termolator	16.90	56.27	25.99
林业科研成果	本文方法	73.10	78.40	75.63
	Termolator	12.81	59.61	21.11

表 3: 林业数据集上的实验结果

实验结果显示，本文方法在这三个数据集上的F值分别为73.71%、77.30%、75.63%；相较于Termolator方法，本文方法在准确率、召回率和F值上均实现了显著提升。同时，本文方法在不同体裁数据集上的F值表现稳定，方法有效性受体裁影响小。Termolator会误将部分与领域有间接关联、但并非术语的词也标记为术语，如“塑料薄膜”、“塑料棚”等。经分析，我们认为：此类词在领域语料库和通用语料库中的分布差异显著，在统计上表现出和术语相似的特征，这导致基于统计的Termolator方法产生了混淆。本文方法除使用统计特性，还会利用词向量从语义角度判断候选术语的类别归属，这能有效避免此类错误。另一方面，Termolator采用的词性规则更倾向于识别名词词组作为候选术语，这使得该方法对“皮下枝接”和“炼苗”等带有动词特性的术语产生了遗漏；而本文方法不受词性限制，具有更高的召回率。

*https://dumps.wikimedia.org/zhwiki/

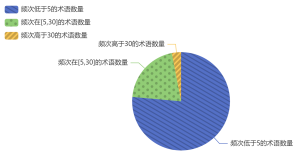

图 2: 良种繁育数据集术语词频分布

为了比较两种方法在低频术语上的抽取效果，本文首先统计良种繁育数据集的词频分布，结果见图2。统计结果显示，数据集中包含263个林业术语，其中出现频次低于5的低频术语有201个，占全部术语的76.43%。同时，所有术语的出现总频次为1246次，而低频术语出现频次仅为360次，占比28.89%。由此可见，良种繁育数据集中的术语分布具有明显的长尾特征，因此低频术语的抽取效果将对领域术语的整体抽取至关重要。

我们将本文方法与词频阈值分别设置为1、3的Termolator方法分别在良种繁育数据集上测试，并采用召回率来评价两种方法对低频术语抽取能力。从表4中可以看出，当词频阈值为1时，Termolator的低频术语召回率虽达到53.55%，但其抽取结果中包含大量干扰项，这导致其总体术语抽取的F值仅为22.13%。当将阈值提高至3，Termolator的F值提升了12.87%，但过于严格的阈值在筛除垃圾项的同时会误筛低频术语，导致低频术语召回率仅为26.38%。然而，本文方法在抽取术语最低词频为1的情况下，低频术语召回率达到74.88%，高出Termolator最佳结果21.33%；同时，总体术语F值达到77.30%。这表明本方法更适应于领域术语的长尾特征，能够在有效抽取低频术语的同时实现高质量的术语抽取。

方法	总体术语F值	低频术语召回率
本文方法	77.30	74.88
Termolator-1	25.99	53.55
Termolator-3	35.00	26.38

表 4: 在良种繁育数据集的抽取效果*

表5展示了本文提出的方法在两种体裁的语料上抽取到的林业术语示例。从展示样例可以看到，本文方法成功地抽取了诸如“木质化”、“树冠”和“间伐”等常用林业术语。同时，本方法还具有一定的林业新兴术语识别能力，例如“林业碳票”和“北京绿林认证”等，这些术语都是近两年应时代需求出现的林业新兴术语，在一定程度上反映了当前林业领域的热点问题。此外，本问方法还能成功抽取像“浙江红花油茶”这样的品种名称。

技术类书籍识别术语	**林业新闻识别术语**
木质化、韧皮部、腹接、嵌合枝接、常绿树、容器育苗、苗期、树冠、催芽、徒长枝、浙江红花油茶	林业碳票、北京绿林认证、林草碳汇、森林碳库、森林蓄积量、碳泄漏、间伐、中幼龄林、混交林

表 5: 识别术语样例

在实验结果中，我们还发现了一些语义相同但使用混乱的术语，如“无人机飞防”、“无人机

*Termolator-1,Termolator-3分别表示词频阈值设置为1和设置为3的Termolator方法。

防治”和“飞机防治”。这三个术语都表示“使用无人机喷洒农药以防治森林病虫害”，但由于缺乏统一标准，导致在具体使用中出现了用词混乱。这些词语识别将有助于反映了林业术语标准化需要关注和改进的方向。

5 结论

本文针对低频术语抽取所面临的挑战，提出了一种基于词向量的自适应术语抽取方法。这种方法不局限于特定领域，仅利用词典信息即可完成术语的无监督抽取。该方法分为候选术语单元性识别与候选术语领域性筛选两步。与现有研究成果相比，基于假设检验识别候选术语单元识别无需大规模语料库训练，能够根据目标文本中字符的统计特性自适应地设置参数，消除了由于主观设置的阈值而导致的低频词被筛除问题，从而克服了现有统计方法在低频术语识别方面的局限性。在判断候选术语领域性时，本文方法采用基于词向量的密度聚类方法筛选候选术语的领域相关性，该方法能够有效判别候选术语的领域性；针对未在词向量集内的未登录候选术语，本文采用基于掩码语言模型的方法获取其词向量，有效保证了新兴术语、非规范术语等未登录术语的筛选。

本方法在林业领域的三个数据集上开展实验，分别取得了73.71%、77.30%、75.63%的F值，并在低频术语抽取实验中达到74.88%的召回率。实验结果表明，本方法在不同体裁和主题的文本上表现稳定；相对于已有方法，本方法能够更为有效抽取低频术语，达到全面识别各频率术语的抽取效果。

参考文献

Hongliang Dai, Yangqiu Song, and Haixun Wang. 2021. Ultra-Fine Entity Typing with Weak Supervision from a Masked Language Model. pages 1790–1799, August.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. Automatic recognition of multi-word terms:. the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130, August.

Dongchen Jiang, Aoyuan Jiang, and Shuai Tang. 2022. An adaptive method for Chinese new word detection based on hypothesis testing. *Pattern Analysis and Applications*, 25(4):993–999, November.

Rodrique Kafando, Rémy Decoupes, Sarah Valentin, Lucile Sautot, Maguelonne Teisseire, and Mathieu Roche. 2021. ITEXT-BIO: Intelligent Term EXTraction for BIOmedical analysis. *Health Information Science and Systems*, 9(1):29, December.

Victoria Kosa, David Chaves-Fraga, Gennadiy Dobrovolskiy, and Vadim Ermolayev. 2020. Optimized Term Extraction Method Based on Computing Merged Partial C-Values. pages 24–49. January.

M. Kucza, J. Niehues, T. Zenkel, A. Waibel, and S. Stüker. 2018. Term extraction via neural sequence labeling a comparative evaluation of strategies using recurrent neural networks. *19th Annual Conference of the International Speech Communication, INTERSPEECH 2018; Hyderabad International Convention Centre (HICC)Hyderabad; India; 2 September 2018 through 6 September 2018*, page 2072.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *undefined*.

Ruixuan Luo, Jingjing Xu, Yi Zhang, Zhiyuan Zhang, Xuancheng Ren, and Xu Sun. 2019. PKUSEG: A Toolkit for Multi-Domain Chinese Word Segmentation, June.

Elise Michon, Josep Crego, and Jean Senellart. 2020. Integrating Domain Terminology into Neural Machine Translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3925–3937, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Tomas Mikolov, Kai Chen, G. Corrado, and J. Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *undefined*.

Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. 2018. Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 175–180, New Orleans, Louisiana, June. Association for Computational Linguistics.

Jianlin Su. 2020. WoBERT: Word-based Chinese BERT model - ZhuiyiAI. Technical report.

Jake Ryland Williams, Paul R Lessard, Suma Desu, Eric M Clark, James P Bagrow, Christopher M Danforth, and Peter Sheridan Dodds. 2015. Zipf's law holds for phrases, not words. *Scientific reports*, 5:12209, August.

Zhiyong Wu, Yun Chen, Ben Kao, and Qun Liu. 2021. Perturbed Masking: Parameter-free Probing for Analyzing and Interpreting BERT, May.

任秋彤, 王昊, 熊欣, and 范涛. 2021. 融合GCN远距离约束的非遗戏剧术语抽取模型构建及其应用研究. 数据分析与知识发现, 5(12):123–136.

俞琰and 赵乃. 2018. 基于通用词与术语部件的专利术语抽取. 情报学报, 37(7):742–752.

俞琰, 陈磊, 姜金德, and 赵乃. 2020. 融合论文关键词知识的专利术语抽取方法. 图书情报工作, 64(14):104–111.

刘书剑and 彭道黎. 2011. 林业信息术语标准化研究. 林业调查规划, 36(01):104–107+116.

刘伟童, 刘培玉, 刘文锋, and 李娜娜. 2019. 基于互信息和邻接熵的新词发现算法. 计算机应用研究, 36(05):1293–1296.

吴俊, 程, 郝瀚, 艾力亚尔·艾则孜, 刘菲雪, and 苏亦坡. 2020. 基于BERT嵌入BiLSTM-CRF模型的中文专业术语抽取研究. 情报学报, 39(4):409–418.

孙水华, 黄德根, and 牛萍. 2016. 中医针灸领域术语自动抽取研究. 中文信息学报, 30(3):118–124.

木合亚提·尼亚孜别克, 古力沙吾利·塔里甫, and 达吾勒·阿布都哈依尔. 2016. 采用CRF模型的哈萨克语信息技术术语自动抽取技术研究. 西北师范大学学报*(*自然科学版*)*, 52(01):53–56.

李思良, 许斌, and 杨玉基. 2018. DRTE:面向基础教育的术语抽取方法. 中文信息学报, 32(3):101–109.

李贞贞, 钟永恒, 王辉, 刘佳, and 孙源. 2022. 基于深度学习与统计信息的领域术语抽取方法研究. 数据与计算发展前沿, 4(2):87–98.

游新冬, 杨海翔, 陈海涛, 孙甜, and 吕学强. 2021. 融合术语信息的新能源专利机器翻译研究. 中文信息学报, 35(12):76–83+93.

王大亮, 蒋宏潮, 涂序彦, 郑雪峰, and 佟子健. 2008. 基于选择倾向性的词汇获取方法. 计算机工程, (12):169–171.

董洋溢, 李伟华, and 于会. 2017. 文本特征和复合统计量的领域术语抽取方法. 西北工业大学学报, 35(4):729–735.

蒋婷. 2021. 学术文献术语抽取方案比较研究. 信息资源管理学报, 11(1):112–122.

黄菡, 王宏宇, and 王晓光. 2019. 结合主动学习的条件随机场模型用于法律术语的自动识别. 数据分析与知识发现, 3(6):66–74.