doi:10.3772/j.issn.2095-915x.2023.05.008

基于 BLSTM-CRF 的自举式术语识别方法研究

开放科学（资源服务）标识码（OSID）

陈翀[1,2] 高欣妍[1] 黄红[1]

1. 北京师范大学政府管理学院 北京 100875;
2. 富媒体数字出版内容组织与知识服务重点实验室 北京 100038

摘要：[目的 / 意义] 自动识别优质术语一直是多领域普遍关注的问题，其中一个突出困难是缺乏领域标注语料，为此本文提出一种基于 BLSTM-CRF 的自举式领域术语识别方法。[方法 / 过程] 首先选取少量种子术语标注语料，训练 BLSTM-CRF 模型，识别候选术语；再基于术语质量特征构造筛选准则，从候选术语中挑出优质且新增的结果加入到新一轮训练的标注词汇集合，迭代标注训练，直到新增术语量小于某一阈值或迭代达到特定次数。本文还检测了模型迭代训练效率及在其他领域的推广性，将在计算机领域语料训练出的模型用于新兴的融合出版领域的技术术语识别。[局限] 术语质量特征量化方法待综合多指标优化，模型改进学习机制未引入负例且迭代不易收敛等。[结果 / 结论] 本文最终通过标注数量和标注语境丰富度实验表明了采用新增标注数据进行迭代的有效性。以 50 轮迭代训练后结果为例，在计算机测试语料上识别术语及其所有标注序列的 F1 值为 0.43 和 0.59，新术语率为 0.79，均优于基准 BLSTM-CRF 模型、BERT-BLSTM-CRF 模型效果，证实了新方法启动成本低，领域适应性好，能够有效解决术语识别中训练语料缺乏的问题。在模型迁移效能评价中，抽样判断的术语识别平均正确率为 87.7%，说明了迁移学习方法的应用潜力。

关键词：术语识别；自举；BLSTM-CRF 模型；识别性能评价；术语质量筛选准则

中图分类号：G35; TP391

A BLSTM-CRF-based Bootstrapping Terminology Recognition Approach Research

CHEN Chong[1,2] GAO Xinyan[1] HUANG Hong[1]

1. School of Government, Beijing Normal University, Beijing 100087, China;
2. Key Laboratory of rich-Media Knowledge Organization and Service of Digital Publishing Content, Beijing 100038, China

Abstract: [Objective/Significance] Automatic extraction of domain terms has been a research hotspot in the field of information technology. An urgent problem to be solved is the shortage of terms for labeling training corpus, which limits the application of

基金项目 富媒体数字出版内容组织与知识服务重点实验室 2022 年度开放基金项目。

作者简介 陈翀（1975-），博士，教授，研究方向为知识组织，文本挖掘，E-mail：chenchong@bnu.edu.cn；高欣妍（2002-），本科生，研究方向为信息管理与信息系统；黄红（1996-），研究生，研究方向为文本挖掘。

引用格式 陈翀 , 高欣妍 , 黄红 . 基于 BLSTM-CRF 的自举式术语识别方法研究 [J]. 情报工程 , 2023, 9(5): 97-111.

doi:10.3772/j.issn.2095-915x.2023.05.008

neural network models in domain term extraction. To solve this problem, this paper proposes a BLSTM-CRF-based bootstrapping term recognition approach. [Methods/Processes] First, inputting a small number of seed terms for corpora annotation and training BLSTM-CRF model to identify candidate terms; Then, constructing a set of criterions based on the quality of terms in order to select high-quality new terms from candidate terms, and adding these quality terms to the annotation set for next round training. Thus, the corpus is relabeled for iteratively model training until the number of new terms is less than a certain threshold or a specific number of iteration rounds is reached. In addition, the model trained on the corpus of the computer science domain can be transferred to recognizing technical terms on the new-emerging domain of fusion-publishing. [Limitations] There are still issues such as the quantification method of term quality features to be optimized by integrating multiple indicators, the learning mechanism of model improvement does not introduce negative examples and the iteration is not easy to converge, etc. [Results/Conclusions] The decision of iteration approach is supported by the experiments on the amount of annotation and the contextual richness, which show that the performance of term recognition can be improved when new annotation data increases. Taking the model obtained after 50 rounds of iterative training as an example, the F1 of the recognized terms and all the annotation sequences are 0.43 and 0.59 on the test set of the computer science domain, and the new-term rate is 0.79, which are better than the benchmark BLSTM-CRF model and the BERT-BLSTM-CRF model. It is confirmed that the new method has low starting cost and good domain adaptability, which effectively solves the problem of lack of training corpus in term recognition. In the model transfer efficiency evaluation, the average correct rate on the sampled term recognition is 87.7%, which illustrates the application potential of the transfer learning method.

Keywords: Term Extraction; Bootstrapping, BLSTM-CRF Model; Performance Evaluation of Term Recognition; Term Quality Criterions

引言

术语是指各门学科的专门用语。术语识别是在领域语料中找到能概括该领域知识的词汇或词组。优质的术语不仅可以用于概括领域知识体系，还可以揭示资源的内容主题，描述资源或学者等实体的特征，因而是构建领域本体、词表及知识图谱等基础应用的核心。在当今学科交叉，新兴领域和已有领域中新的术语不断出现的情况下，人工编制术语表显然不能满足现实需要，而由于领域表述差异、标注训练数据稀缺等原因，也没有通用性好的术语识别工具[1]，因此自动识别术语一直是自然语言处理、知识组织等研究中普遍关注的问题。

术语识别方法面临的一个瓶颈是各领域优质标注语料匮乏，因为多数方法需要从标注语料中学习术语的上下文和构词特征以便训练模型。本文提出了一种自举式（Bootstrapping）术语识别方法，以解决神经网络模型在标注数据不足情况下的术语识别问题。该方法依据术语的流行性、单元性、信息性、领域性及词法构成等质量特征定义规则，用于在每轮迭代中筛选模型所识别的候选术语，将其扩充到标注词集合，以便训练新的识别模型，从而达到提升模型训练效果的作用。

本文还探索了 BLSTM-CRF 模型对标注词数量及语境丰富度的依赖性，实验结果为本文选择迭代方案提供依据。本文用精度、召回率、F1 值和新词识别率衡量所提出的自举式方法与 BLSTM-CRF、BERT-BLSTM-CRF 相比的

优势，并用优质术语率和产出投入率对该方法每轮的迭代效率进行评估。最后，将训练好的模型用于新兴的融合出版领域的术语识别，抽样 1000 词，人工评价识别效果，平均正确率是 87.7%，说明该方法具有一定的领域推广性。本文的创新点一是筛选优质术语为每轮迭代训练模型提供了更直接的优化目标，二是选择了迭代训练的策略，使得术语识别方法所需的人工标注词汇少，启动成本低，领域适应性好。

1 相关研究

1.1 术语识别的方法

术语识别问题提出已久，开展的相关研究也很多[1-3]，但它至今并未被完美解决。与关键词或短语的识别相比，术语识别需要考虑术语所具有的领域性，而不仅仅是在特定文档中的重要性。与一般的命名实体识别相比，术语的构成字词比人名、地名、机构名等实体名中的取词范围更广，边界特征不明显，而且标注语料稀缺。2016 年，Astrakhantsev 在 7 个开放数据集上比较了 13 种当时最优的方法，包括基于统计特征、基于主题模型、基于外部知识库等，最终发现并没有某种单一方法能在所有数据集上获得最佳的平均精度[1]。

现有的术语识别方法主要分为三类，即基于语言学规则的方法，基于统计的方法和基于机器学习的方法[3,10]。基于语言学规则的方法主要是依据语言学规律识别术语[2]，例如术语的词性、词长、构词特点、句法依存关系等。该类方法适用于小数据集，随着语料库的增大，术语出现情况增多，对应的语言规则会越来越复杂，召回率也低。基于统计的方法主要是利用一些统计指标对词汇打分，从语料库中挑选出重要的词汇作为领域术语，统计指标包括词频、文档频率、互信息、信息熵、TF-IDF、C-Value[4] 等。该类方法不依赖于特定领域，具有通用性，但需要大规模语料的支撑，且对语料库的质量要求也比较高。机器学习兴起后，基于规则和统计的方法可以作为机器学习方法中的组成环节，用以提升术语识别性能。

机器学习方法通常将术语识别转化成两种类型的问题。一种是二分类问题，给定训练数据，采用朴素贝叶斯、决策树、支持向量机等方法根据训练实例的特征学习分类模型，进而识别术语片段。词汇特征可以是基于语言学特征、统计特征及两者的组合特征，还可以是来自外部知识库的特征。另一种思路是转换为序列标注问题。用于序列标注的机器学习方法主要有隐马尔可夫模型、最大熵模型、条件随机场（Conditional Random Field，CRF）等。CRF 通过学习被观察的输入数据序列 $X=(x_1, x_2, \ldots, x_n)$ 及其对应的隐含状态序列 $Y=(y_1, y_2, \ldots, y_n)$ 来构建条件概率模型 $P(Y|X)$，当给定了某个观测序列 x，求解 $P(Y|X)$ 最大时的状态序列 y 作为对 x 的标注结果。在术语标注中观测序列是连续的文本，隐含状态序列是文本中词汇的类型标签，例如是否术语的开头、结尾等。

随着词嵌入技术和深度学习方法的发展，大量神经网络模型如循环神经网络（Recurrent Neural Networks，RNN）、双向长短时记忆（Bidirectional Long Short-Term Memory，BLSTM）网络等被应用于领域术语识别。Col-

doi:10.3772/j.issn.2095-915x.2023.05.008

lobert 等 [5] 是将神经网络模型用于词性标注、分词、命名实体识别等自然语言处理任务的早期代表之一。而 Meng 等 [6] 为了能更好地捕获文本中的语义信息和句法特征，提出了 Copy-RNN 关键短语预测生成模型，生成文本中不一定原样出现过、但是概括了文本语义的词汇。

深度学习方法的优势在于不需要特定的人工构建规则和繁琐的特征工程，易于提出隐含的语义信息，方法的领域独立性强。这些提取方法大多基于“编码器—译码器”框架展开，“编码器—译码器”框架包括输入的分布式表示、上下文编码器和标签解码器三个部分 [6-8]。第一部分，输入的分布式表示可以用 Word2vec、BERT 等语言模型生成表示向量，或融入词性、依存句法和大小写等语言学特征 [9] 来人工构建特征向量。第二部分，上下文编码器主要是用 CNN、RNN、LSTM 等神经网络模型去挖掘输入数据中的隐藏信息。其中 CNN 能有效提取输入数据中的局部特征，而 RNN 能够记忆信息，LSTM 模型则通过内部状态向量引入了内存机制，以及输入门、遗忘门和输出门三个门控机制，这使得 LSTM 能够有选择性地记忆长距离信息，具有更强的信息捕捉能力，改善了 RNN 训练困难及短时记忆的弊病，一般适于处理长序列数据。BLSTM 模型是特殊的 RNN，不但具有信息记忆能力还能挖掘隐含信息，因此被广泛应用于自然语言处理研究中。Chiu[7] 提出了一种融合 BLSTM 和 RNN 的方法来提取词汇级别和字符级别的特征以完成命名实体识别任务，最终在 CoNLL-2003 数据集和 OntoNotes 数据集上都取得了不错的效果。第三部分，标签解码器用来解码神经网络模型最终的输出。常用的标签解码器有全连接层加上 softmax、CRF、RNN 等。其中全连接层加上 softmax 会独立地预测每个单词的标签，导致最终的预测结果出现标签偏置的问题，而 CRF 由于其考虑全局最优化，能够有效解决这个问题，纠正了 BLSTM 模型可能产生不合理的标签组合的问题，因而目前 BLSTM-CRF 模型因其序列预测效果好而成为自然语言处理领域的经典模型 [7-10]。为了进一步优化，研究者还引入了注意力机制，该机制赋予 BLSTM-CRF 模型输出向量不同的权值，突出重要词汇的重要性，使得模型的术语识别性能得到提升 [9]。

1.2 标注语料的扩展

识别术语面临训练机器学习模型所需的标记语料稀缺的问题，大多数应用场景都缺少可用的术语表或外部知识库。一种折中的办法是采用自举（Bootstrapping）方式增量扩展训练集，即初期用少量的标注数据，在迭代中不断获取置信度较高的标注结果，优化训练模型，提升模型最终效果 [11]。

在基于 CRF 序列标注的术语识别迭代中，扩展标注数据的方式与采用的模型有关。第一种是直接利用输出的观测序列，即标出的术语，制定规则来获得高置信度的标注结果。例如规定将相邻术语片段或满足前后缀特征的片段进行组合得到新的术语 [12]；第二种是利用输出的隐含状态序列，即词的标签，将标签被选择的概率作为置信度来筛选具有特定标签特征的词汇 [13]。前一种方式直接将人类对术语特征的观察经验融入筛选规则，适用于对术语特征有一定知识的情况，其优点是提供了更为直接的学

习依据，但不足是所加入的规则覆盖能力有限。后一种方式适用于对术语特征缺少人类经验知识的情况，完全从标注语料学习，并依赖 CRF 模型算出的标签概率来构造术语筛选的判决策略，其不足是在隐含状态层面的判决策略传递到观测序列上并不一定都能构成有效筛选。

用迁移学习也是一种解决标记语料稀缺的方法，在标注语料丰富的领域数据集中训练模型，将训练好的模型参数迁移到标注语料少的领域上，从而实现在少量标注数据条件下训练出高质量的术语识别模型的目标。如果一个系统具有迁移学习能力，则该系统能够将源领域或任务上学习到的知识或模式应用到相关的目标领域或问题中 [14]。迁移学习的可行性基础是源领域和目标领域要有相似性。按照源任务和目标任务、源域与目标域的相同与否分为归纳式、直推式和无监督迁移学习几种，其中，直推式迁移学习的源任务和目标任务是相同的，源域和目标域不同但相关，因而目标域和源域的任务之间共享相同的模型参数或者是服从相同的先验分布 [14]。刘宇飞等 [15] 基于 BLSTM-CRF 模型在 CoNLL-2003 使用的新闻领域 NER 英文公共数据集上训练模型，在数控系统领域的专利文献数据集上获得了术语标注结果。

2 研究设计

自举式方法通过逐步增加优质标注词不断训练术语识别模型 F，最终得到的模型记为 B_F。本文的基础模型 F 选择公认性能优越、使用广泛的序列标注模型 BLSTM-CRF [10]。在迭代中，利用术语的质量来构造筛选条件，选择新增的标注词。本文关注的问题主要有：（1）能反映术语质量的筛选条件如何定义；（2）迭代方法的效果及效率评估，包括对识别新词的能力和标注投入与术语产出率的评估；（3）在领域 S 中训练的模型 B_F，迁移到相似领域 D 的同类任务中具有何种效果。为此首先在本节阐明迭代训练过程和术语筛选规则，在第 4 节实验中评价迭代方法在源领域 S 以及迁移到相似领域 D 的效果。为了便于说明，定义符号及含义如表 1 所示。

表 1 符号表

S	源领域语料，用于训练模型的子集为 S_1，用于测试的子集为 S_2' 和 S_2, $S_2 \subset S_2'$
T	源领域测试语料 S_2 中的人工标注术语，用于模型识别术语效果评价
T'	源领域测试语料 S_2' 中以启发式方法构建的优质词汇表，用于评价模型识别新词的能力
D	目的领域语料，与 S 具有相似性但缺乏标注数据，用于检验模型直推式迁移效果
i	迭代轮次，$i \in [1,m]$
m	停止迭代时的轮次
T_0	初始标注训练集的种子术语
T_i'	BLSTM-CRF 模型在第 i 轮训练时，在训练集标出的词汇
T_i	T_i' 中经过术语质量筛选条件过滤后保留的优质术语，$T_i \subseteq T_i'$
$\tilde{T}_i$	T_i 中的已知术语，$\tilde{T}_i = T_i \cap T$
ΔT_i	第 i 轮识别出的优质新术语 $\Delta T_i = T_i - T_{i-1}$

2.1 基于BLSTM-CRF的自举式术语识别过程

选取特定领域，以论文摘要构成的句子集合代表领域语料 S，同时获取论文的关键词，从中挑选种子术语集合 T_0。本文提出术语识别的方法是（1）在第 i 轮模型训练中，利用集合 T_{i-1} 标注训练语料 S_1；（2）用标好的语料训练 BLSTM-CRF 模型，记为 F；（3）用模型 F 在 S_1 上产生第 i 轮的词汇集合 T_i'，它可视为本轮

doi:10.3772/j.issn.2095-915x.2023.05.008

找出的术语候选词；（4）按照质量规则筛选 T_i' 得到优质词集合 T_i，将 T_i 中相对于 T_{i-1} 而言的新术语 ΔT_i 添加到 T_{i-1} 作为下一轮训练的标注词集合；（5）判断抽出的新术语数 $|\Delta T_i|$ 如果少于给定的阈值或模型迭代达到指定的次数，则停止迭代，否则重复步骤（2）—（5），图 1 为本文的术语识别流程。假设在第 m 轮停止迭代，得到的最终模型即是本文所提的自举式方法的训练结果，记为 B_F。训练结束后，在测试语料 S_2 上评价模型训练效果。

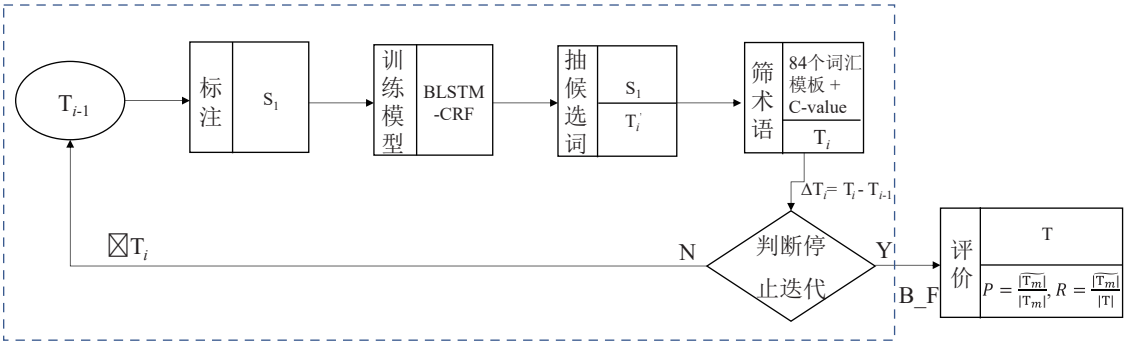

图 1 基于 BLSTM-CRF 的自举式术语识别方法 B_F 的训练过程

为模型提供标注数据训练并不是要其记住标注术语本身，而是要学习术语及上下文的语言特征，建立词汇及其标注状态的条件概率模型，因此在一轮迭代中，作为输入的标注术语并不一定都会被标出。在这种情况下，如果用本轮输出的术语词直接为下一轮训练做标注，将有可能丢失原始优质种子术语贡献的上下文语言特征。为了保证标注质量不下降，每轮训练的标注数据都保留了那部分经人工挑选的种子术语 T_0，且吸收本轮输出的优质新术语 ΔT_i，用 $T_{i-1}+\Delta T_i$ 为下一轮训练做标注，这些新术语能为模型提供更多学习实例。

2.2 术语筛选的规则

为了提高模型识别术语的准确率，本文从术语质量上定义筛选规则。Kageura 等 [2] 将优质术语的特征归纳为单元性和领域性，单元性指如果一个单词序列频繁地出现在一起，则它可能表达了一个独立完整的语言含义；领域性衡量术语候选词与特定领域的相关程度。因为术语在无关领域的出现机会比较少，人们提出了 Domain Pertinence[16] 用词汇在不同领域出现频次来度量，同类指标 Weirdness[17]、Relevance[18] 考虑了频次值在不同领域的规范化。C-value/NC-value 指标混合了词长、词频、上下文等多种因素的考虑 [4]，其特点是偏向长词的选择，这比较符合领域术语的表述特点。

本文筛选术语质量首先基于统计规则，认为优质术语应具有流行性、单元性、信息性和领域性特征，选择 C-value 和 Weirdness 来量化上述特征，理由及对照关系如表 2 所示，定义词汇的筛选规则如式（1）所示。其次基于构词规则，由于术语是特定领域对象、概念的专指名称，根据构词特点，一般是由多个名词结构或动名词结构等形式组成，所以人工挑选种子术语后，以其词法特征作为术语的构词模板，

即分析领域 S 中种子术语的词法构成，从中挑选出现频次较高的词法模板，作为术语的构词特征筛选标出结果，规则详见 3.2 节表 4。通过上述两类规则过滤的候选词被认为是优质术语，可作为训练模型所用的标注数据。

统计规则 Score(w) =a*C-Value (w)+(1-a)* Weirdness (w)　（1）

Score(w) 代表候选术语 w 的分值，一部分来自 w 在领域内的重要性，一部分来自 w 相对于其他领域而言所具有的特殊性。这两部分借助 C-value 和 Weirdness 指标及其加权来表示，a 是权重系数。

C-value 的计算如式（2），$|w|$ 表示 w 的字符长度，$f_S(w)$ 表示 w 在语料库 S 中出现的频次；T_w 表示包含 w 的候选术语集合。如 w 为“神经网络”，T_w={“BP 神经网络”，“循环神经网络”，“卷积神经网络”} 是都包含 w 的词汇集合，即父串，$|T_w|$ 为该集合的元素数。其中式（3）反映了既有独立性又同时作为其他术语成分出现的术语 w，在计算时需要排除包含 w 的父串术语造成的影响。

$$\mathrm{C-value(w)}=\begin{cases} log_2|w|\cdot f_S(w), & w\text{无父串} \\ log_2|w|\left(f_S(w)-\frac{1}{|T_w|}\sum_{b\in T_w}f_S(b)\right), & w\text{有父串} \end{cases} \quad (2)$$

Weirdness 计算如式（3），用 w 在 S 及其参照领域 reference 中出现的频次的归一化值之比来表达 w 是否具有领域特性，其中 N_S、$N_{reference}$ 分别代表两个领域的语料中的词汇数。

$$\mathrm{Weirdness}(w)=\left(\frac{f_S(w)}{N_S}\right)/\left(\frac{f_{reference}(w)}{N_{reference}}\right) \quad (3)$$

a 的选择取决于对领域 S 和参照领域 reference 的比较。考虑到目前领域交叉广泛存在，Weirdness 对两个领域中的共有术语 w 并不灵敏，参照领域的选取标准不能简单确定，领域性的度量结果赋权有待深入探讨。本文实验中为简化起见，限于对特定领域的术语识别，将 a 赋为 1。

表 2 术语质量的特征、含义及与量化指标对照关系

质量特征	含义	对应指标
流行性	术语 w 因具有领域共识因而在语料中具有一定出现频度。	$f_S(w)$，f_S 表示术语 w 在领域 S 中的频度
单元性	术语词汇的完整性以及词间搭配的稳定性，可以近似用语料中术语 w 的“独立出现频度”反映，这里“独立出现”是指刨除了“包含 w”的词汇出现带来的干扰，体现 w 作为完整的词汇单元表达独立的语义。	$f_S(w)-\frac{1}{\lvert T_w\rvert}\sum_{b\in T_w}f_S(b)$
信息性	术语在特定领域有所指，能代表领域内特定知识实体、概念、过程或现象等。衡量术语的信息性本应结合内容上下文进行语义分析，但为计算方便起见，可以用术语的长度近似替代。	$\lvert w\rvert$
领域性	术语的领域特有性，表现为在特定领域中高频出现，而在与该领域相关度较小的领域中低频出现或不出现。如果是在单个领域的语料中识别术语，也可以暂不考虑这一特征。	$\left(\frac{f_S(w)}{N_S}\right)/\left(\frac{f_{reference}(w)}{N_{reference}}\right)$

2.3 效果评价

2.3.1 每轮迭代的评价

理想的方法是标注成本低、标出的术语多且质量高。迭代方法使用少量的初始种子可以获得大量优质术语。为了弄清训练的成本，即投入标注词与获得新术语的数量关系，可以计算每轮迭代的优质术语率 λ 和每轮训练的产出

doi:10.3772/j.issn.2095-915x.2023.05.008

投入率 δ，定义如式（4）和（5）。式（4）中，ΔT_i 是第 i 轮迭代训练所得的优质新术语，T_i'-T_{i-1} 代表第 i 轮中识别出的新词，λ_i 代表在第 i 轮识别出的新词中优质术语的占比。式（5）代表第 i 轮产出的优质新术语与输入的标注词之比。迭代模型的优质术语率 λ 越高、产出投入率 δ 越大，则说明迭代越有效、实用性越好。

优质术语率：$$\lambda_i = \frac{|\Delta T_i|}{|T_i' - T_{i-1}|} = \frac{|T_i - T_{i-1}|}{|T_i' - T_{i-1}|} \quad (4)$$

产出投入率：$$\delta_i = \frac{|\Delta T_i|}{|T_{i-1}|} \quad (5)$$

2.3.2 方法性能的评价

基本评价指标为精度、召回率和 F1 值。由图 1 所示，B_F 方法在 BLSTM-CRF 识别结果基础上还做了词汇筛选，因此该方法的精度是 $P = |\widetilde{T_m}| / |T_m|$，召回率 $R = |\widetilde{T_m}| / |T|$。此外，本文还关注识别新术语的能力，衡量特定方法识别出的优质术语中，非种子标注词的占比 τ，见公式（6）。在 B_F 中，τ 相当于对迭代过程识别的所有新术语在最终结果中的占比，在 BLSTM_CRF 中，由于没有术语筛选的环节，所以分子是 T_1'-T_0。

新术语率：$$\tau = \begin{cases} \dfrac{|T_1' - T_0|}{|T_1'|}, & \text{BLSTM-CRF} \\ \dfrac{|T_m - T_0|}{|T_m|}, & \text{自举式方法B\_F} \end{cases} \quad (6)$$

2.3.3 迁移性能的评价

进行直推式迁移学习，将在领域 S 中训练好的模型用于融合出版技术领域 D。由于领域 D 缺乏标注数据做评价集，本文将结果采样后由多人打分，在评分一致的情况下，用识别出的术语的平均正确率来衡量该模型在领域 D 识别术语的效果。

3 实验与分析

3.1 数据准备

3.1.1 语料集

训练模型使用了计算机领域的期刊论文数据。取中国计算机学会网站 2020 年列出的计算机领域中文核心期刊，有《计算机学报》等 12 种。从 CNKI 导出上述期刊 1998—2019 年间发表论文的题目、摘要、关键词等元数据，经过筛选清理后一共有 35695 篇论文。将所有论文的摘要切分成句子，过滤掉字符长度小于 30 的句子，作为领域语料 S。随机取 80% 的论文，以其摘要句子作为训练集合，记为 S_1。剩余 20% 论文的摘要句构成集合 S_2'，用于迭代效率和新词识别能力评测。在 S_2' 中随机抽取 5000 句构成集合 S_2，$S_2 \subset S_2'$，用于术语识别性能评测。

此外，为了测试本文方法在相似领域的迁移能力，本文还选取了融合出版技术为目标领域。融合出版技术是出版业与人工智能、大数据、融媒体、数字化等技术结合而形成的新兴领域。2022 年 1 月，全国科学技术名词委发布了该术语的定义，专家表述融合出版典型技术和关联的术语时，提到“移动互联网”“AR 技术”“数字出版”等[①]。因此本文首先以“融合出版”和上述词汇在CNKI与万方数据库进行论文检索，由专业人员从论文关键词中挑选能表示融合出

① 全国科技名词委组织召开融合出版概念及定义审定会议网址：http://www.cnterm.cn/xwdt/tpxw/202201/t20220114_678525.html

版技术的术语 73 个；而后以“融合出版”分别与 73 个术语组合进行扩展查询。经过合并清洗后得到论文 6516 篇，取摘要中长度不小于 30 的句子构造语料 D 来代表“融合出版技术”领域。语料集基本信息见表 3。

表 3 领域语料基本信息

语料	论文数量（篇）	句子数量（句）
S_1	28556	86689
S_2’	7139	21819
S_2	-	5000
D	6516	28770

3.1.2 种子术语T_0

用于训练语料的初始标注词集。本文从训练集 S_1 对应的论文高频关键词中选择种子术语。为保证种子术语的质量，综合词长、tf-idf 词权及词汇规范程度的因素进行选择。先取长度为 4~10 个字符且在摘要文本中 TF-IDF 权重降序排前 10% 的词汇，再借助百度百科作为词典进行筛选。这样可筛除论文关键词中不是领域术语的词汇，如“评价”等，共得到 1410 个种子术语。

3.1.3 术语评测集T

构造 T 是为了评估模型识别术语的效果，由一名计算机相关专业的标注者依据领域知识和优质外部词表，在 5000 个句子构成的测试集 S_2 上标出的全部术语，共 1069 个术语，出现 5291 次。外部词表包括种子术语集 T_0，原始论文关键词集合，百度百科信息科学词汇集合，术语在线，知网计算机科学技术领域术语集合，国家自然科学基金中的计算机科学、人工智能、电子学与信息系统、交叉学科中的信息科学分支等领域受控词。

3.1.4 新词评测集T

为了衡量迭代效率，评估模型在每轮迭代中找出优质新词的能力，希望获得语料上尽可能全的词汇，因而用半自动的方法识别其中的优质词汇。在 S_2' 上首先用 jieba 分词功能将每个句子切成词汇序列，在此基础上做 unigram 和 bi-gram 的词汇组合得到候选词汇集合。其次统计词频和 C-value 值，保留词频不小于 3 且 C-value 值不小于 3.5 的词汇，共有 17899 个。最后由四名计算机相关专业的学生进行筛查，对无法判别的词汇用百度百科、维基百科等网络词典检验，并剔除如“支持向量”“向量机”等不完整词汇，以及如“算法 KNN”等组成顺序不当的词汇，最终选出 5336 个计算机领域优质词汇作为此评测集。

T' 与 T 的区别是 T' 以较大的测试语料 S_2' 中的分词结果组合为基础，目的是产生词典中可能未收录的优质新词，组合过程中产生的不合理词汇需要人工去除。上述半自动处理过程得到的词汇可被视为优质词或候选术语。而 T 强调全面准确，是由人工在较小的测试语料 S_2 标出的术语集，用于评测模型术语识别能力。

3.2 条件设置

3.2.1 语料标注方式

为避免切词造成的错误，用字符级别的标注方式标注每一个输入句子。标签 BME 分别表示术语的开头、中间部分和结尾，O 表示非术语部分。标注过程用 python 程序自动完成。

3.2.2 术语筛选所用的词法模板

对应 3.2 节术语筛选规则中的构词规则。

doi:10.3772/j.issn.2095-915x.2023.05.008

为了在每轮迭代训练中筛选出优质词汇，分析种子术语的构词特征，用 jieba 工具包统计分析了 T_0 中词汇的词法结构，共获得 214 个词法模板。从中挑选出现频次不小于 2 的模板共 84 个，可覆盖 90.8% 的种子词汇。出现频次排名前 10 的词法模板见表 4。

表 4　术语筛选所用的词法模板 top10

模板	出现频次	词语样例
n+n	206	路由协议、分布式数据库
n+v	159	机器学习、图像分割
n	130	神经网络、虚拟现实
eng	116	LINUX、ADABOOST
v+n	83	拓扑结构、组播路由
eng+n	73	BP 神经网络、IP 协议
n+vn	61	数据广播、主元分析
l	46	操作系统、软件工程
v+v	32	迁移学习、并发控制
vn+n	21	循环神经网络、传输协议

jieba 词性说明：n 名词、v 动词、eng 英语、vn 动名词、l 习用语

3.2.3　BLSTM-CRF模型参数设置

使用 Tensorflow 2.2.0 搭建 BLSTM-CRF 模型，隐藏层设为 100 维，学习率设为 0.001，*dropout* 值设为 0.5，*batchsize* 设为 50，*timesize* 设为 100。词嵌入层使用 Tensorflow 封装的词嵌入层生成词向量，向量维度设为 100 维度，词向量跟随着模型一起进行训练。模型训练的停机条件设置为 5 轮验证集上的准确率未有改变。

3.2.4　自举式方法的迭代控制

每轮迭代中，模型抽出的词汇为候选词要按照术语质量规则筛选，满足如下条件的词汇构成当前轮次获得的术语集合 T_i：（1）符合从种子词汇集合中挑选出的 84 个词法模板之一，且（2）词汇的 C-value 值大于阈值，本实验中设置为经验值 3.5。每轮得到的术语与当前轮次的标识词相比，新增的词汇小于 10，或迭代达到指定次数时，自举式方法停止迭代，并将最后一轮标出的术语作为方法的输出结果。

3.3　标注量和词汇丰富度BLSTM-CRF模型的影响

本文选用了 BLSTM-CRF 作为基准模型，为了掌握它对标注语料的需求，设计两组实验，在 S_2 上分别观察标注量和词汇丰富度对训练效果的影响，作为本文设计迭代模型不断引入新标注数据的实证依据。两组实验均以 4.1 节所述的词表 T 为评价基准。实验一是比较在不同标注量的训练下，BLSTM-CRF 模型的术语识别性能差异。随机采样种子术语总量的 25%、50%、75% 以及 100% 的情况下模型识别术语的性能，见表 5。随着训练语料标注量从 25% 增加到 100%，标出术语数、P、R 和 F1 值都有了大幅度的提升，P 从 0.55 增至 0.81，R 从 0.02 增至 0.17，F1 值从 0.04 上升到 0.28。即模型的术语识别能力随标注术语量增加而提升，不但识别了更多的术语，而且对词典中的术语覆盖面更大。此外，从所识别术语中的非标注词占比来看，新术语识别能力也随着标注词用量的增加而增加，$\tau(S_2)$ 从 0.16 变为 0.27。以上说明训练语料的标记量对 BLSTM-CRF 模型的术语识别性能影响很大。因此本文拟采用迭代的方法，通过每轮增加标注数据来提升模型识别术语的能力。

表 5　标注数据的数量对 BLSTM-CRF 的性能影响

种子标注集	投入术语数	标出术语数	P	R	F1	τ（S_2）
25%T_0	353	47	0.55	0.02	0.04	0.16
50%T_0	705	76	0.76	0.05	0.10	0.17
75%T_0	1059	142	0.67	0.09	0.16	0.17
100%T_0	1410	218	0.81	0.17	**0.28**	**0.27**

实验二是分别以在训练集 S_1 中出现频次降序为前 15% 和后 85% 的种子术语来标注，它们在语料 S_1 中的出现次数接近，前者是 51499 次，后者是 51416 次。由于后者不相同词汇多，对应的语境可能更丰富，因此实验二是比较在标注量大致相同但标注上下文丰富度不同的情况下，训练 BLSTM-CRF 模型的术语识别性能差异，如表 6 所示。结果显示，仅用频次前 15% 的术语标注和仅用后 85% 的术语标注，对应的 F1 值从 0.05 升至 0.2，说明标注词的语境丰富对 BLSTM-CRF 模型的性能有影响，不断引入新的标注词训练模型比用单批标注数据的训练更有助于提升性能。

表 6　标注数据的丰富度对 BLSTM-CRF 的性能影响

训练术语	投入术语数	标出术语数	P	R	F1
前 15%	207	32	0.81	0.02	0.05
后 85%	1203	179	0.71	0.12	0.20

3.4　自举式术语识别方法性能

在自然语言处理领域中，术语识别一直是一项具有挑战性的任务。在近年来的研究中，基于深度学习的模型已经成为该领域研究的热点之一。其中，BiLSTM-CRF 模型 [10] 和 BERT-BiLSTM-CRF 模型 [19-20] 被广泛应用于计算机领域等技术术语识别任务中且表现良好，BERT-BiLSTM-CRF 模型是目前术语识别领域的主流方法之一。为了探究本文所提的基于 BLSTM-CRF 的自举式术语识别方法上的性能，在 S_2 上以 BLSTM-CRF 模型和 BERT-BLSTM-CRF 模型为比较基准进行术语识别效果评价，自举式术语识别模型迭代轮次 i=5、10 和 50。标注集为 T_0，选择 P、R、F1 和 τ 指标，评价对象为模型所识别出的唯一术语和模型输出的标注序列，前者考查识别出 T 中的唯一术语的能力，后者考查识别出术语在测试集中所有出现情况的能力。一个术语在测试语料中可能多次出现，但是能将出现在不同语境中的同一术语都尽可能识别出来的模型是训练的目标之一。本文使用测试集 S_2 中的基准词表 T，|T|=1069，T 在 5000 句测试集上对应的标注序列共 5291 个。

表 7 的“按唯一术语”和“按标注序列”分别对应以术语为对象和以模型输出的标注序列为对象的评价结果。从 i=5、10、50 的结果对比可以看出迭代训练出的自举式模型 F1 值超过 BLSTM-CRF 和 BERT-BLSTM-CRF 模型的效果。BERT-BLSTM-CRF 模型虽然提取出的术语量多于自举式方法初期迭代识别术语量，但精度和召回率较低，说明 BERT-BLSTM-CRF 模型提取术语能力较强，但识别术语的质量低于自举式方法。随着迭代次数增加，由于自举式方法识别术语量或标注序列增加，呈现出精度 P 下降，召回率 R 上升，F1 值持续上升，这说明了模型的效果随着迭代次数增加而提升。

doi:10.3772/j.issn.2095-915x.2023.05.008

表 7 BLSTM-CRF 与自举式方法的性能比较

方法		轮次	按唯一术语				按标注序列			
		i	术语数（个）	P	R	F1	术语出现（次）	P'	R'	F1'
基准模型	BLSTM-CRF	1	218	0.81	0.17	0.28	1846	0.87	0.30	0.45
	BERT-BLSTM-CRF	1	921	0.32	0.28	0.30	4028	0.53	0.40	0.45
自举式训练的模型 B_Fi		5	384	0.73	0.26	0.38	2842	0.77	0.41	0.54
		10	475	0.64	0.29	0.40	3462	0.68	0.45	0.54
		50	966	0.45	0.41	**0.43**	6278	0.54	0.64	**0.59**

表 8 显示了模型识别新术语的能力。在 S_2 上对新术语占比的评价结果可以看出，$\tau(S_2)$ 从 0.27 升至 0.79，远高于 BLSTM-CRF 和 BERT-BLSTM-CRF 的新术语占比。而在更大的测试集合 S_2' 上再次评价，$\tau(S_2')$ 从 0.86 升至 0.94，均高于 BLSTM-CRF 和 BERT-BLSTM-CRF 的新术语占比，说明迭代模型识别出新术语的能力更高，这与每轮迭代中引入新的标注术语来训练模型有关。BERT-BLSTM-CRF 的新术语占比略低于 BLSTM-CRF，可能是由于 BERT 的嵌入使模型在使用大规模的语料进行预训练过程中，能够更深入地学习种子词汇的特征，从而导致其在识别新术语时表现不佳。

表 8 BLSTM-CRF 与自举式方法的新术语率

方法		i	（S_2）	（S_2'）
基准模型	BLSTM-CRF	1	0.27	0.73
	BERT-BLSTM-CRF	1	0.22	0.69
自举式训练的模型 B_F_i		5	0.46	0.86
		10	0.56	0.93
		50	**0.79**	**0.94**

3.5 自举式方法迭代训练过程的效率

自举式方法训练过程中逐步筛选优质术语充实标注词，为了衡量这些加入的词汇所发挥的效用，取在 S_1 上迭代训练出的模型 B_F_i，$i \in [1,10]$，计算它在第 i 轮识别出的新术语中的优质术语率 λ，以及训练的术语产出投入率 δ，见表 9。在迭代训练模型时，第 i 轮次标注集为 T_{i-1}，模型直接标出的词汇集合是 T_i'，经质量筛选后作为方法 B_F_i 的输出结果是 T_i。需要说明的是，第 i 轮识别出的新术语数 $|\Delta T_i|$ 是筛出的术语 T_i 与第 i 轮用于标注的词汇集合 T_{i-1} 的集合差，而并非数值差。

表 9 训练中标出的优质术语率及产出投入率

迭代轮次	$\lvert T_{i-1}\rvert$	$\lvert T_i\rvert$	$\lvert T_i'\rvert$	$\lvert\Delta T_i\rvert$ $(=\lvert T_i-T_{i-1}\rvert)$	λ_i (%)	δ_i (%)
1	1410	2766	7982	1687	25.67	119.65
2	3097	4806	14060	2022	18.44	65.29
3	5119	5206	13261	765	9.40	14.94
4	5884	6313	16934	953	8.62	16.20
5	6837	7525	20797	1214	8.70	17.76
6	8051	8057	20450	848	6.84	10.53
7	8899	7883	17989	363	3.99	4.08
8	9262	9293	24154	889	5.97	9.60
9	10151	10205	25770	1073	6.87	10.57
10	11224	11382	28788	1238	7.05	11.03

从表 9 中看出，在前 10 轮中输入标注术语均比输出标注术语少，即迭代方法能基于较少的标注词产生较多优质术语。λ 和 δ 的增长速度随迭代次数增加而趋于降低，说明到了一定

标注规模后，增加标注词虽然带来优质数据绝对数量的增长，但训练效率降低。特别是到了第 6 轮，输入的标注词数量 $|T_{i-1}|$ 和最终获得的优质术语数量 $|T_i|$ 大致相当，虽然该轮结果中仍产出 848 个新的优质术语，使得投入产出率为 10.53%，但方法在总体上迭代优势下降。如果目标是希望尽可能多地获得语料中的优质术语，迭代仍可以继续；如果是兼顾计算开销和时间成本，可以根据 λ 或 δ 制定停止条件。

3.6 在融合出版技术领域的术语识别结果

选取训练好的模型 B_F_i（本文选 i=50）在语料 D 上识别融合出版领域的术语，输入的标注词集合为 T_0，识别出术语 8561 个，对 D 上的论文语料而言，篇均 1.31 个标出术语。评测时，从输出结果中随机抽样 1000 个术语，由两位熟悉领域 S 和 D 的专业人员分别独立判断，术语正确率分别为 88.8% 和 86.6%，平均正确率 87.7%。两位判断者的结果一致性通过检验，kappa 系数为 0.287，显著性 p 值 0.000。表 10 是两个判断者的交集中部分术语，可以看出在计算机领域语料训练的模型在融合出版技术的相关文献上，也有较好的术语识别效果。

表 10　模型迁移学习的术语识别结果（按模型的结果输出顺序取 35 词）

文本结构	协同编纂系统	MPR 技术	发布引擎	协同编纂平台	协同交互	语音识别
识别技术	内容资源	内容标准	动态发布	加工系统	增强现实技术	微信平台
知识图谱	内容编辑	数字内容	资源管理	数据应用	虚拟现实技术	微信传播
知识服务	XML	用户终端	数据库	信息管理	计算机科学	OSID
ERP	ISLI	计算机平台	内容对象	信息化	VR 技术	互动性

本文还从语料 D 的论文关键词中人工筛选出反映融合出版技术的术语共 107 个，有 46 个被上述实验结果涵盖。分析发现未被模型识别出的术语中有些词概念粒度较大，例如“新媒体技术”“三网融合”，一种可能的原因是这些术语来自论文关键词，它们不一定在摘要中出现，因此论文关键词不能作为理想的评价基准。但是模型识别出的术语是来自内容中的有意义的词汇，大部分质量较高，可以为构建领域词表提供论文关键词无法覆盖的补充词汇，也能为领域主题分析等研究提供更丰富的信息和深入理解。

模型识别结果中被认为不符合术语的词汇包括少量残缺词如“模糊现”，以及满足构词特征但不具有信息性的词汇，如“增强管理”“信息数据”等，说明在模型训练中术语质量特征的量化表达仍有值得改进之处。

该实验展示了迁移学习方法在自然语言处理任务中的应用潜力。通过将计算机领域语料上训练的模型应用到融合出版领域，能够有效避免重新训练新的模型，节省时间和资源，并且可以利用源领域的丰富信息来提高目标领域任务的性能。同时，该实验还为其他领域的迁移学习研究打开了新的思路，如使用跨语言语料库进行迁移等。

4 总结与展望

当今各学科迅速发展而导致新词不断出现，

doi:10.3772/j.issn.2095-915x.2023.05.008

各个领域都对高质量术语有迫切需求。术语识别任务是自然语言处理研究中的基础工作，对于领域知识服务、情报分析等应用具有关键作用。现实中，以深度学习模型为基础的自动方法又离不开大量优质术语做标注数据。针对这一矛盾，本文提出了一种基于 BLSTM-CRF 的自举式术语识别方法，能在使用少量术语标注语料的情况下通过迭代的方式强化模型训练，从而达到获取大量优质术语的目的，在一定程度上能够解决神经网络训练语料短缺的问题，对各学科获取新术语生成自身的知识组织体系、提升检索质量有很大帮助。此外，本文关注术语质量的筛选条件定义问题，提出利用术语的质量来构造筛选条件，以便在迭代中选择新增的标注词。本文还提出了用于衡量迭代效率的指标，包括对识别新术语的能力和标注投入与术语产出率的评估。最后探讨了在领域 S 中训练的模型 B_F，迁移到相似领域 D 的同类任务中具有的效果。

本文实验主要探讨了四个相互关联的问题。一是论证迭代方案有效，在 4.3 节通过增加标注词和丰富标注语境的实验证明了少量样本结合自举式迭代方法在识别术语上的有效性。二是评估迭代模型的效果，在 4.4 节对自举式迭代方法和基准 BLSTM-CRF、BERT-BLSTM-CRF 方法识别出唯一术语及其所有标注序列的情况进行评估，用精度 P、召回率 R、F1 值表明迭代模型效果更好，而且迭代模型在识别标注词之外的新术语的能力更强。三是衡量迭代的效率，通过每轮迭代发现的新词中的优质术语率，以及训练数据的产出投入率来判断迭代效率，为更合理地决定迭代终止条件提供量化依据。最后，本文还选取了与源领域 S 相似的领域 D，考查所训练模型的迁移能力，评价显示本文模型 B_F 在同类任务中具有良好效果，具有领域推广能力。

由于时间和条件的限制，本研究还有很多不足之处以及未考虑全面的地方需要在以后的研究工作中进行改进。例如，术语质量特征的量化表示方法应当对多个指标进行综合对比，以完善对术语质量判断规则的经验研究。再如，应当优化模型改进学习机制，在本文训练中只考虑了加入优质术语供模型学习，没有引入负例；而且迭代中只采用了增加新术语的方式来扩充模型识别能力，这会导致训练开销加大，迭代不易收敛，未来还可以增加对标注序列模式的学习。

参 考 文 献

[1] ASTRAKHANTSEV N. ATR4S: Toolkit with state-of-the-art automatic terms recognition methods in scala. Language Resources and Evaluation, 2018, 52(3): 853-872

[2] KAGEURA K, UMINO B. Methods of automatic term recognition: A review[J]. Terminology, 1996, 3(2): 259-289.

[3] 张雪 , 孙宏宇 , 辛东兴 , 等 . 自动术语抽取研究综述 [J]. 软件学报 , 2020, 31(7): 2062-2094.

[4] FRANTZI K T, ANANIADOUS S, TSUJII J.The c-value/nc-value method of automatic recognition for multi-word terms[C]//Proc of the 2nd Eruopean Conference, ECDL'98, 1998: 585-604.

[5] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. Journal of machine learning research, 2011(12): 2493-2537.

[6] MENG R, ZHAO S Q, HAN S G, et al. Deep Keyphrase Generation[C]//The 55th Annual Meeting of the Association for Computational Linguistics,

Vancouver, Canada, 2017: 582-592.

[7] CHIU P J, NICHOLS E. Named entity recognition with bidirectional LSTM-CNNs[J]. Transactions of the Association for Computational Linguistics, 2016, 4: 357-370.

[8] 冯鸾鸾 , 李军辉 , 李培峰 , 等 . 面向国防科技领域的技术和术语识别方法研究 [J]. 计算机科学 , 2019, 46(12): 231-236.

[9] 马建红 , 张亚梅 , 姚爽 , 等 . 基于 BLSTM-attention-CRF 模型的新能源汽车领域术语识别 [J]. 计算机应用研究 , 2019, 36(5): 1385-1389, 1395.

[10] 胡雅敏 , 吴晓燕 , 陈方 . 基于机器学习的技术术语识别研究综述 [J]. 数据分析与知识发现 , 2022, 6(Z1): 7-17.

[11] BRIN S. Extracting Patterns and Relations from the World Wide Web[J]. Lecture Notes in Computer Science, 1998, 1590: 172-183.

[12] 鲍宸洋 , 任明 . 基于 Bootstrapping 的家谱文本信息抽取方法研究 [J]. 图书馆杂志 , 2022, 41(2): 93-102.

[13] 赵洪 , 王芳 . 理论术语抽取的深度学习模型及自训练算法研究 [J]. 情报学报 , 2018, 37(9): 923-938.

[14] PAN S J, YANG Q. A Survey on Transfer Learning[J]. IEEE Transactions on Knowledge & Data Engineering, 2010, 22(10): 1345-1359.

[15] 刘宇飞 , 尹力 , 张凯 , 等 . 基于深度迁移学习的技术术语识别——以数控系统领域为例 [J]. 情报杂志 , 2019, 38(10): 168-175.

[16] MEIJER K, FRASINCAR F, HOGENBOOM F. A semantic approach for extracting domain taxonomies from text[J]. Decision Support Systems, 2014(62): 78-93.

[17] AHMAD K, GILLAM L, TOSTEVIN L. University of Surrey Participation in TREC8: Weirdness Indexing for Logical Document Extrapolation and Retrieval (WILDER)[C]//The Eighth Text REtrieval Conference, 1999.

[18] PEAS A, VERDEJO F, GONZALO J. Corpus-Based Terminology Extraction Applied to Information Access[C]//The corpus linguistics, 2004.

[19] 毛立琦 , 石拓 , 吴林 , 等 . 基于领域自适应的无监督文本关键词提取模型——以“人工智能风险”领域文本为例 [J]. 情报理论与实践 , 2022, 45(3): 182-187.

[20] 翟羽佳 , 田静文 , 赵玥 . 基于 BERT-BiLSTM-CRF 模型的算法术语抽取与创新演化路径构建研究 [J]. 情报科学 , 2022, 40(4): 71-78.