基于词性自动机的关键短语抽取方法

王凌霄 王弋波 朱礼军
（中国科学技术信息研究所，北京 100038）

摘要：关键短语抽取是一种识别目标文本中具有特殊价值的关键词组合的自然语言处理任务场景，对科技文献情报挖掘具有重要的实践价值。由于缺少足够的标注数据、知识库、预训练模型，针对前沿细分学科颠覆性内容的关键短语抽取还存在着许多挑战。将有限状态自动机概念引入关键短语抽取任务中，把关键短语的词性标注组合模式抽象为一系列有限状态自动机文法。这种基于词性自动机的无监督关键短语提取算法，能够在不依赖标注数据和高性能计算设备的条件下，通过高度自定义的词性组合模式，抽取不定长度的细分领域关键短语。这种算法具备运行速度快、环境依赖低、匹配模式多、提取效果好等特点。使用 SemEval−2017 数据集和智能新药发现领域的文献摘要作为测试数据，将研究所提出的算法和几种广泛应用的关键短语抽取算法进行对比。对比结果显示：这种算法在所有关键词中的准确率达到 30.8%，召回率达到 34.1%，*F*1 值达到 32.4%；在关键短语中的准确率达到 30.8%，召回率达到 52.0%，*F*1 值达到 38.7%。召回率指标与 *F*1 指标相比关键词抽取开源算法库有显著提升。

关键词：命名实体识别；关键词抽取；关键短语抽取；有限状态自动机；词性标注

DOI：10.3772/j.issn.1674−1544.2023.05.004 **CSTR：**15994.14.issn.1674.1544.2023.05.004

中图分类号：TP391 **文献标识码：**A

Keyphrase Extraction Algorithm via Tagging Finite Automation

WANG Lingxiao, WANG Yibo, ZHU Lijun
(Institute of Scientific and Technical Information of China, Beijing 100083)

Abstract: Keyphrase extraction is a natural language processing task scenario for identifying keyword combinations with special value in target texts, which has important practical value in mining scientific and technological literature information. Due to the lack of sufficient labeled data, knowledge base, and pre-training models, there are still many practical challenges in the extraction of keyphrases for subversive content in cutting-edge sub-disciplines. In this paper, the concept of finite state automata is introduced into the key phrase extraction task, and the part-of-speech tagging combination patterns of keyphrases are abstracted into a series of finite state automata grammars. This unsupervised key phrase extraction algorithm based on part-of-speech automaton can extract keyphrases of indeterminate length in subdivision fields through a highly customized part-of-speech combination mode without relying on labeled data and high-performance computing equipment. The algorithm has the characteristics of fast running speed, low environment dependence, many matching modes, and good extraction effect. This paper uses the SemEval-2017 dataset and literature abstracts

作者简介：王凌霄（1996—），男，中国科学技术信息研究所研究实习员，硕士，研究方向为机器学习与自然语言处理；**王弋波**（1985—），男，中国科学技术信息研究所副研究员，硕士，研究方向为科技资源管理、生物医学数据治理（通信作者）；**朱礼军**（1974—），男，中国科学技术信息研究所研究员，博士，研究方向为管理信息系统。

基金项目：中国科学技术信息研究所创新研究基金资助项目“基于文本实体挖掘的新药发现领域人工智能技术应用识别方法”（QN2022−06）。

收稿时间：2022 年 8 月 25 日。

in the field of intelligent new drug discovery as test data, and compares the algorithm proposed in this paper with several widely used keyphrase extraction algorithms. The accuracy rate of this algorithm in all keywords reaches 30.8%, the recall rate reaches 34.1%, the *F*1 value reaches 32.4%, the accuracy rate in key phrases reaches 30.8%, the recall rate reaches 52.0%, and the *F*1 value reaches 38.7%. Compared with the open source algorithm library for keyword extraction, the recall score and the *F*1 score are significantly improved.

Keywords: named entity recognition, keyword extraction, keyphrase extraction, finite state machine, part-of-speech tagging

0　引言

关键短语抽取（Keyphrase Extraction）指的是从文本资料中提取特定短语的自然语言处理过程，是知识图谱、推荐系统、搜索系统等复杂工程的基础任务，在信息管理、情报分析、搜索推荐等领域中有着重要意义。关键短语抽取算法的设计过程需要紧密联系具体任务场景和任务需求，如新闻舆情关键短语抽取任务需要提取高频名词短语，而颠覆性技术关键短语抽取任务需要保留低频的潜在技术关键短语。如果把抽取算法应用在场景不匹配的任务中，那么可能会得到难以满足需求的结果。本文将针对前沿细分领域的科技情报文献关键短语提取场景，提出一种运行速度快、环境依赖低、匹配模式多的轻量级关键短语抽取算法。

关键短语抽取领域有着较好的研究基础[1−2]，其解决方案涵盖多种范式，包括传统监督学习方法、无监督学习方法、现代深度学习方法等。与新闻舆情、社交媒体等领域不同，科技文献情报领域缺乏大规模高质量的标注数据，也缺乏对应的大规模预训练神经网络模型，特别是在前沿细分领域文献挖掘任务中。囿于上述限制，科技文献情报方向的关键短语抽取任务难以在标注语料库上进行监督学习训练，也无法直接在特定领域的文献文本中进行大规模预训练模型参数微调。

基于此，本文提出了词性自动机（Tagging Finite Automation，TFA）模型。这个模型能够在不依赖标注数据和高性能计算设备的条件下，通过高度自定义的词性组合模式，抽取不定长度的细分领域关键短语的无监督方法。本文将有限状态自动机（Finite Automation）引入关键短语抽取任务中，把关键短语的词性标注（Part−of−Speech Tagging）[3]组合模式抽象为一系列有限状态自动机文法。

虽然许多关键短语提取算法在工作流程中使用了无监督词性标注，但是其词性标注的利用程度较低，一些算法仅仅约定了认可的词性标签范围，或只是有限枚举了认可的词性标签组合。本文提出的词性自动机使用状态自动机的归纳规则，能够支持多种关键短语词性组合模式的灵活定义，能够高度抽象地定义词性标签的组合模式，从而实现召回程度较高的关键短语抽取。

本文使用SemEval−2017数据集和智能新药发现领域的Web of Science文献摘要作为测试数据，将本文提出的抽取算法和几种广泛应用的关键短语抽取算法进行对比，发现其优势，以在科技文献情报的关键词挖掘中发挥实际应用价值。

1　相关工作概述

关键短语抽取算法的通用流程大致由5个步骤组成，即预处理语料文本、划定候选关键短语范围、选择关键短语特征、根据规则给候选关键短语计分、抽取最终关键短语并进行效果评估。关键词短语抽取的大量研究工作主要集中在短语特征选择与评分步骤。此环节所应用的方法可以分为3种范式，分别是无监督学习式、监督学习式、深度学习式。深度学习式算法实际上是使用监督学习进行的，但是按惯例进行单列讨论。

1.1　无监督学习型关键短语抽取算法

无监督型关键短语抽取算法所使用的关键短语特征通常是启发式的结构信息、频率信息等，并不依赖标注数据训练。无监督算法的计算量较少，程序实现灵活，适用于缺乏标注数据任务场

景。无监督算法可以再继续细分为多个子类，如基于统计特征的抽取算法、基于图网络的抽取算法等。

基于统计特征的关键短语抽取算法代表是TF−IDF抽取法[4]、KPMiner算法[5]、YAKE算法[6]等。这类算法使用自然语言特征、文本位置信息等多种启发式特征生成候选关键词并赋予不同的权重。TF−IDF抽取法[4]是最容易实现、最经常使用的短语特征，它要求关键短语在当前文本中有着较高的频率，但是在整个语料库中又不过于频繁出现。KPMiner算法[5]优化了TF−IDF抽取法的表达式，按比例增大了多单词候选短语的权重。YAKE算法[6]设计了5种特定的规则评分项来计算候选短语的权重，并使用Levenshtein距离来融合相似的候选短语。

基于图网络的关键短语抽取算法的代表是TextRank[7]、SingleRank[8]、TopicRank[9]、Topical PageRank[10]、PositionRank[11]、MPRank[12]等。这类算法把整个文档集构建成一张图网络，图的点是候选短语，图的边是候选短语之间的共现关系。候选短语的评分依照不同的图中心度衡量指标来实现，如TextRank[7]采用了原始的PageRank指标[13]。

1.2 监督学习型关键短语抽取算法

监督学习型关键短语抽取算法把关键短语抽取任务建模为二分类问题，即判断候选短语是否为真正的关键短语。比较具有代表性的算法有KEA[14]、GenEx[15]、CeKE[16]、KeyEx[17]等。其中，KEA[14]使用TF−IDF与朴素贝叶斯进行短语判别，GenEx[15]使用遗传算法进行短语判别。监督学习型算法适用于具有标注数据、通用词表的任务场景中，如新闻舆情、社交文本、电商评论等。这些任务中的标注规则简单，容易招募足够数量的标注人员，积累适当规模的标注数据就能够支撑大量无标注数据的关键短语提取。

1.3 深度学习型关键短语抽取算法

随着神经网络结构的更新、深度学习软硬件生态的发展，许多深度学习方法被运用到关键短语抽取任务中，如循环神经网络被运用于推特关键短语抽取中[18]。随着预训练技术的发展，BERT[19]等大规模预训练模型在绝大多数自然语言处理任务中的表现都超过了朴素的循环神经网络，BERT也被应用到关键短语抽取任务之中[20]。

2 抽取关键短语的主要方法

为了提高科技文献情报领域，特别是通用标注数据少、专家咨询成本高的前沿细分领域的关键短语抽取能力，本文提出了词性自动机（Tagging Finite Automation，TFA）模型。词性自动机借助状态转换图，配合少数初始定义和少数递归规则，高度抽象地定义关键短语的词性标签组合模式，简明地表示了大量关键短语词性组合模式。词性自动机能够在不依赖标注数据和高性能计算设备的条件下，完成前沿细分领域的关键短语抽取。

2.1 词性自动机

2.1.1 词性标签匹配

受限于稀缺的标注数据与高昂的人工成本，监督学习式（Supervised Learning）关键短语抽取算法在前沿科技文献情报挖掘等许多场景中难以开展应用。相比之下，无监督学习式（Unsupervised Learning）关键短语抽取算法仅需要少数启发式特征就足以运行，在缺乏标注数据的场景中往往具备较高的可行性。

词性标签（Part−of−Speech Tagging）[3]是无监督学习式算法经常使用的一种自然语言特征。词性分类、词性标注等研究工作在语言学、自然语言处理等领域中已经有了坚实的基础，能够较好地完成关键短语挖掘任务。下面介绍两种借助词性标签匹配实现关键短语挖掘的方法。

一是无监督抽取算法使用常见词性标签的出现频次来挖掘关键短语。如连续出现形容词、动词、名词三者之一的达到n次的单词序列就可以被视为有效关键短语[21]。如单词序列“gaussian random variable”所对应的词性标签序列是“形容词 形容词 名词”，若设置关键短语的最小单词数为2，则这个单词序列满足匹配规则，被识别为有效关键短语。又如单词序列“则该单词序

new deep learning methods”对应的词性标注序列是“形容词（重复4次）名词”，也被识别为有效关键短语。

二是无监督抽取算法采用词性标签组合模式作为抽取条件。如使用“形容词—名词”词性组合抽取短语“deep learning”，使用“名词—名词”词性组合抽取短语“regression coefficients”。这种方法要求关键短语的词性标签必须严格符合某种预先设定好的组合模式，以免过度识别一些无意义的单词序列。举例说明，如果预定义了“形容词—形容词—名词”词性模式，那么单词序列“那new deep learning methods”被识别的部分就仅为“deep learning methods”，而不是全部的5个单词。

上述两个方法展示了使用词性标签匹配进行无监督关键短语挖掘的过程，但是它们还没有充分挖掘词性标签的潜力。其原因：一是有些算法仅仅简单地判断某些单词的词性是否在事先规定的词性标签集合中，忽略这些词性标签的前后顺序、重复次数等模式关系。二是有些算法只是手工枚举了一些词性标记组合，但这种方式往往缺乏扩展性和可维护性。

表1展示了常见的英文关键短语词性组合模式[21]。这些手工枚举的组合模式存在很多的局限性。如重复名词2次和重复名词3次都是常见的词性组合模式，但是在实际任务场景中还有可能出现重复名词4次的关键短语，如tissue image analysis technology等。再如形容词—名词和重复形容词2次—名词都是常见的模式，但是也可能出现重复形容词3次—名词的关键短语，如high−level quantum mechanical energies等。

表1　常见的英文关键短语词性组成模式

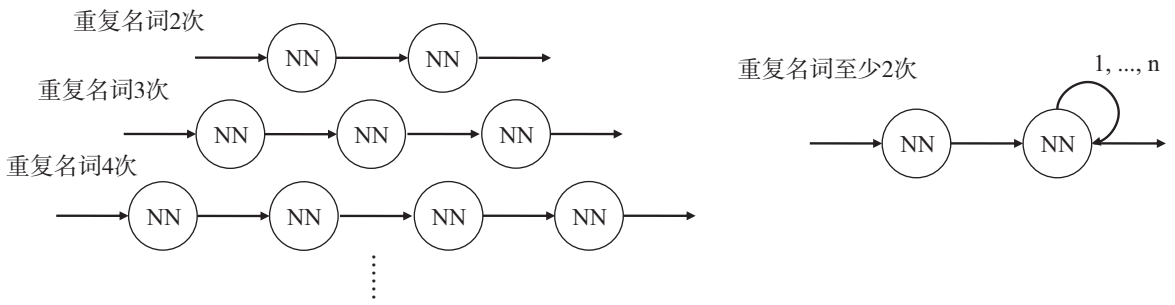
词性组合模式	词性组合模式示例
名词—名词	regression coefficients
名词—名词—名词	class probability function
形容词—名词	linear function
形容词—形容词—名词	gaussian random variable
形容词—名词—名词	rational drug design
名词—形容词—名词	mean squared error
名词—介词—名词	degrees of freedom

本文提出的词性自动机模型，希望能在专家经验总结的词性组合模式的基础上，尽可能地提高词性组合模式的扩展性，以便更好地抽取新出现的低频学术关键短语。

2.1.2　词性自动机的状态转移图

引入有限状态自动机理论（Finite State Automation）能够很好地解决上述问题。如图1所示，前述问题中的关键名词“重复1次”“重复2次”甚至是“重复n次”都可以用自动机的状态转移图抽象表示，并且具备更好的扩展性。

有限状态自动机是现代计算机科学的基石，并且在自然语言处理和情报分析中也有广泛的应用[22]。有限状态自动机维护着一张状态转移表，并根据从外部读取的符号序列不断更新内部状态，根据内部状态做出一系列决定，如接受符号组合、拒绝符号组合、继续读取符号等。有限状态自动机有2种具体的分类，即确定有限自动机（DFA）和不确定有限自动机（NFA）。DFA的优势是方便计算机进行模拟，其缺点是不容易设计



a.手工枚举的关键短语词性组合模式　　b.自动机表示的关键短语词性组合模式

图1　关键短语词性组合的抽象表示

复杂的匹配模式。NFA的优势是容易设计匹配模式，其缺点是计算机模拟不够直观。

本文把有限状态自动机应用在关键短语词性组合的模式识别场景中，故将这种特殊的状态自动机称为词性自动机（Tagging Finite Automaton，TFA）。图2展示了常见的关键短语词性组合的词性自动机转移状态图，由于篇幅限制只展示了状态转移图中主要边（箭头表示）和节点（圆圈表示）。其中，单实线圆圈表示状态，双实线圆圈表示终结状态，圆圈内数字表示状态编号，箭头代表状态转换函数，箭头旁的符号∈代表触发状态转移的符号。

现举两例说明词性自动机的状态转移图。图2顶部的状态转移链路“状态1—状态2—状态3”代表词性自动机首先匹配一个以上形容词（JJ），随后匹配一个以上名词（NN），最后匹配一个名词复数形式（NNS）。图2底部的状态转移链路“状态14—状态15”代表词性自动机首先匹配一个名词，然后再匹配一个以上名词，即匹配由两个以上名词组成的关键短语的全部可能组合情况。

2.1.3 词性自动机的正则语义表示

状态转移图能够表示任意的词性组合重复次数，其表现能力远大于人工枚举的模式匹配。为方便用户根据实际场景调整词性组合模式，状态转移图需要借助正则表达式（Regular Expression）进行文本表示。正则表达式是用于描述某种序列模式的符号序列，由表示符号集合的常量和表示符号集合上运算的算子组成。

人工枚举的词性组合模式通过状态图抽象和正则语义表示之后，不仅内容篇幅得到精简，而且匹配能力得到了很大程度的加强。表2展示了

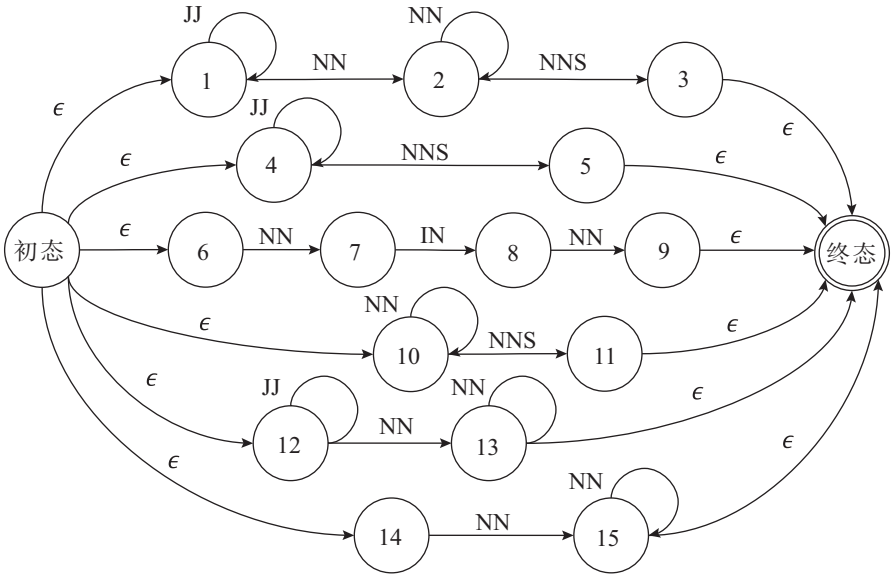


图2 常见的关键短语词性组合模式的状态转移图

表2 常见的英文关键短语词性组成模式（正则简化版）

词性组合模式（枚举表示）	词性组合模式（正则表示）	词性组合模式示例
名词—名词	名词{2，}	regression coefficients
名词—名词—名词		class probability function
形容词—名词	形容词+名词+	linear function
形容词—形容词—名词		gaussian random variable
形容词—名词—名词		rational drug design
名词—形容词—名词	名词—形容词—名词	mean squared error
名词—介词—名词	名词—介词—名词	degrees of freedom

表1中英文关键短语词性组成模式的正则简化版，7条规则被归纳为4条规则，并且强化了每条规则的匹配能力。在词性自动机的正则表达式中，加号+表示被标注为某个词性的单词重复1次及以上，花括号{n,}表示被标注为某个词性的单词至少重复n次。

词性自动机状态转换图（图2）展示的词性组合模式在正则语义下的表现形式可见表3。正则标识能够灵活地与词性自动机数据结构进行互相转换。程序在模拟执行词性自动机规则时，先把正则表达式转化为不确定有限状态机（NFA），最后转换为确定有限状态机（DFA）交由计算机处理。

表3　本文涉及的词性组合模式的正则语义表示

语法状态	简写
形容词+名词单+名词复	(JJ)+(NN)+NNS
形容词+名词复	(JJ)+NNS
名词单+名词复	(NN)+NNS
名词单 介词 名词单	NN IN NN
形容词+名词单+	(JJ)+(NN)+
专有名词单 名词单+	NNP (NN)+
动词过去式 名词单+	VBD (NN)+
名词单{2,}	NN{2,}

2.1.4　配置词性自动机

本文提出的词性自动机还具备灵活配置的优点，研究人员可以根据文本场景的变化，选用不同的词性标签集，并根据词性标签编辑对应的词性组合正则表达式。

不同的词性标注程序往往有不同的词性标签集。常见的英文词性标签集有Penn Treebank Project等，常见的中文词性标签规范有《北京大学现代汉语语料库基本加工规范》[23]。词性自动机只是要求用户编写的词性组合正则表达式与标注软件输出的词性标签满足相同的标准，而不严格限制具体的词性标注集。

2.2　关键短语抽取流程

基于词性自动机的关键短语抽取算法由以下步骤组成：第一步，切分待处理文本并标注文本词性；第二步，通过词性自动机抽取满足词性组合模式定义的关键短语；第三步，集成下游抽取算法。

2.2.1　标注待处理文本的词性

词性标注是本算法的运行基础。这一步有两个子步骤：首先是进行切词，然后是进行标注。

切词是把待处理文本切成长度为N的词符序列，词符既可以是单词，也可以是某个标点符号。对于中文文本而言，切词步骤还应当额外增加中文分词流程，使得词符是完整的中文词语而不是单个字符。

标注是把前面形成的词符序列标记为对应的词法标签序列，如形容词、名词、介词等。词法标签序列和词符序列应当一一对应，有着相同的长度N。常用的词性标注方法包括最大熵模型、隐马尔科夫模型、条件随机场等。本文没有对词性标注算法提出特殊的要求。

2.2.2　生成候选关键短语

词法标签序列里的词法标签被依次读入词性自动机中。词性自动机从初始状态开始，根据读入的词性标签选择对应的状态转移方向。如果某个词性标签不在当前状态的转移表中，那么意味着当前的词性标签不属于任何关键短语模式。此时状态机将会重置输入指针，并且把状态重置为初始状态，等待匹配新的关键短语模式。图3展示了基于词性自动机的无监督关键短语抽取算法的全部流程。

2.2.3　词性自动机与其他提取算法的结合

词性自动机不仅能够独立提取关键短语，而且能够便捷地结合其他关键短语提取算法。许多基于图网络的关键短语抽取开源算法[24]都把词性标注作为前置预处理步骤，词性自动机能够与这些词性标注相关的前置步骤很好地结合起来。

以开源算法库PKE[24]为例。此算法库首先把形容词、名词和专有名词定义为默认词性标签集合，然后把具有以上词性的最长单词序列作为候选关键短语。经图网络运算，这些候选关键短语就会被进一步筛选成为最终关键短语。此算法能够提升候选关键短语挑选阶段的工作效率，提高后续流程的效果。

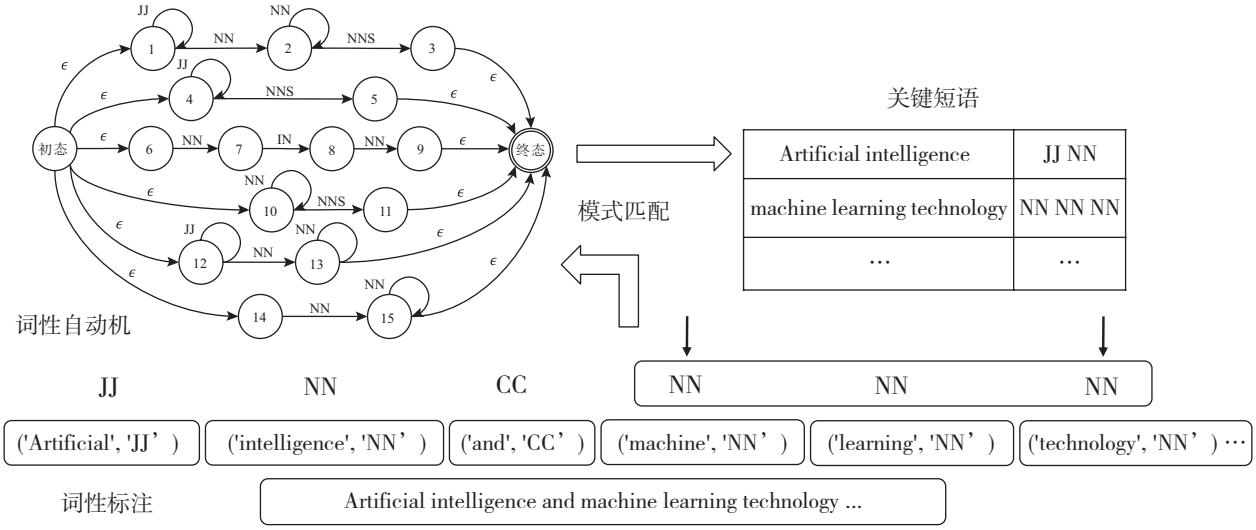


图 3　词性自动机关键短语抽取流程

这些开源算法在使用词性标签的信息时，仅仅把它当成简单集合来使用，而并没有使用更精确、更有价值的词性组合模式。PKE算法库所实现的候选关键短语抽取规则，其实就是前文表 3 的第 5 行、第 6 行、第 8 行所定义的规则，而词性自动机所支持的其他更精确的词性组合模式并没有被使用。

3　对比实验与场景分析

本章节首先介绍这种算法在关键短语数据集SemEval−2017 上的测试效果，并且与 10 种算法进行比较。然后使用智能新药发现领域的Web of Science文献摘要作为测试数据，展示此算法提取关键短语高频子模型的具体效果，并且对该算法的优缺点展开讨论分析。

3.1　SemEval-2017 关键短语数据集对比试验

本章节在SemEval−2017 数据集上测试词性自动机算法（TFA）和 10 种开源无监督关键词提取算法，并且对比它们的准确率、召回率和 *F*1 值。

测试结果显示，此算法的关键短语提取准确率达到 30.85%，召回率达到 34.17%，*F*1 值达到了 32.43%，而非单词型关键短语提取准确率达到 30.85%，召回率达到 52.00%，*F*1 值达到了 38.72%（表 4）。

SemEval[25]是自然语言处理领域重要的基准数据集系列。SemEval−2017 数据集记录了大量学术文献关键短语及其相互关系的人工标注，为此算法提供了很好的测试场景。SemEval−2017 数据集的内容字段以txt文本文件表示，人工基准以ann标注文件表示。

PKE[24]是开源的关键短语抽取算法软件库，实现了包括TextRank[7]、SingleRank[8]、TopicRank[9]、TopicalPageRank[10]、PositionRank[11]、MPRank[12]等著名的关键短语算法，为此算法提供了很好的横向对比条件。

表 5 展示了PKE算法库所提供的 10 种无监督关键短语抽取算法在SemEval−2017 数据集上的准确率、召回率和*F*1 值。因为PKE中的几种算法需要显示指定输出的关键词数量，所以表 5 分别展示了前 5 位（@5）、前 10 位（@10）和前 15 位（@15）关键短语对应的准确率、召回率和 *F*1 值。

对比表 4 与表 5，可以发现TFA算法的召回率和*F*1 值明显高于PKE对比算法的前 5 位关键短语（@5）和前 10 位关键短语（@10）结果。TFA算法在长度大于等于 2 的词组型关键短语上

表 4　TFA 算法在SemEval-2017 数据集中性能表现

单位：%

算法名称	准确率*P*	召回率*R*	*F*1 值
TFA	30.85	34.18	32.43
TFA−2	30.85	52.00	38.73

表 5　不同关键短语提取算法在SemEval-2017 数据集中性能表现

单位：%

算法名称	*P*@5	*R*@5	*F*1@5	*P*@10	*R*@10	*F*1@10	*P*@15	*R*@15	*F*1@15
FirstPhrases	27.26	11.75	16.42	25.94	22.14	23.89	23.93	29.89	26.58
Tfldf	22.83	18.63	20.52	21.02	27.01	23.64	19.87	32.42	24.64
KPMiner	18.34	15.11	16.57	14.81	17.21	15.92	14.41	17.94	15.99
YAKE	24.70	10.92	15.15	21.19	20.62	20.90	19.25	28.33	22.92
TextRank	37.66	10.34	16.22	35.17	20.82	26.16	32.88	30.52	31.66
SingleRank	39.28	13.31	19.88	34.99	25.79	29.69	31.88	37.45	34.44
TopicRank	31.33	18.04	22.90	28.70	26.38	27.49	26.33	31.64	28.74
TPRank	35.62	12.14	18.11	33.18	25.26	28.68	29.67	35.06	32.14
PositionRank	39.25	15.31	22.03	33.68	27.16	30.07	28.22	36.71	31.91
MPRank	32.68	18.72	23.81	29.14	28.13	28.63	25.83	33.54	29.18

的召回率和*F*1 值高于全部对比算法，包括前 15 位关键短语结果（@15），并且非单词型关键短语提取*F*1 值在对比算法中最高，关键短语提取*F*1 值达到对比算法第 4 位。TFA算法是召回率超过 50% 的唯一算法。TFA算法的不足之处体现在准确率指标上，TFA的准确率在 30 个实验对照组中位列第 9 名，处于中上水平。关于TFA算法在准确率指标和召回率指标上的权衡问题将在后续讨论。

由此可见，TFA该算法的准确率、召回率和*F*1 值都达到了关键词抽取开源算法库的平均水平，并且在召回率指标中表现出显著优势，比召回率排名第 2 的算法高出 14.55%。此算法能够在科技文献情报的关键词挖掘中发挥实际应用价值。

3.2　智能制药关键短语提取场景分析

为展示TFA算法在具体应用场景提取关键短语高频子模型的实际效果，本章节使用了Web of Science智能新药发现领域的文献摘要作为案例分析数据。新药发现（Drug Discovery，DD）是生物制药领域中的一个细分方向，基于人工智能技术的新药发现（Artificial Intelligence Drug Discovery，AIDD）是此方向与人工智能相结合的交叉方向。因此，智能新药发现领域没有足够的基准标注数据，没有权威的叙词表，甚至很难找到足够数量的复合型专家。TFA算法适合在这类领域场景中挖掘关键短语，特别是能引起行业颠覆性创新的关键技术术语。

图 4 展示了TFA算法在智能新药发现领域的关键词提取片段。此实验使用NLTK工具包进行文本预处理与词性标注，再将单词序列与对应词性序列输入词性自动机。词性自动机依次读取词性标签，并刷新内部状态转换图，在完成词性组合模式匹配时输出对应的单词。TFA算法抽取所有符合预定义词性组合模式的关键短语（图 4 中的加粗部分所示），如off−target delivery、microarray data、clinical trials等。TFA算法也抽取到少数噪声数据，如 crucial role 和 other words 等。为解决此问题，TFA支持统计各个词性组合模式中真实出现的高频子类型，为优化词性组合模型和筛选噪声特征提供帮助。

表 6 统计了TFA算法在不同词性组合模式下真实出现的高频子类型。表 6 中的数据仍来自

Drug designing and development is an **important area** of research for **pharmaceutical companies** and **chemical scientists** . However , **low efficacy** , **off-target delivery** , **time consumption** , and **high cost** impose a hurdle and challenges that **impact drug design** and discovery . Further , complex and **big data** from genomics , proteomics , **microarray data** , and **clinical trials** also impose an obstacle in the **drug discovery pipeline** . **Artificial intelligence** and **machine learning technology** play a **crucial role** in **drug discovery** and development . In **other words** , **artificial neural networks** and **deep learning algorithms** have modernized the area .

图 4　智能新药发现领域关键词提取片段

表 6　词性自动机在测试数据集上的关键短语抽取结果统计

词性组合模式	高频子类型	词频	占比/%	关键短语示例
（JJ）+(NN)+ NNS	JJ NN NNS	4 943	0.05	deep learning algorithms
	JJ{2} NN NNS	788	0.01	explainable artificial intelligence techniques
	JJ NN{2} NNS	726	0.01	silico drug discovery efforts
（JJ）+ NNS	JJ NNS	24 796	0.24	recurrent networks
	JJ{2} NNS	4 442	0.04	quantitative structure−property relationships
	JJ{3} NNS	284	0	high−level quantum mechanical energies
（NN）+ NNS	NN NNS	6 490	0.06	combination therapies
	NN{2} NNS	1 401	0.01	drug discovery efforts
	NN{3} NNS	94	0	amino acid contact energies
NN IN NN	NN IN NN	3 519	0.03	popularity of machine
（JJ）+ (NN)+	JJ NN	26 540	0.25	molecular docking
	JJ NN{2}	5 551	0.05	sophisticated data quality
	JJ{2} NN	3216	0.03	computer−aided molecular design
NNP (NN)+	NNP NN	5 223	0.05	PLI prediction
	NNP NN{2}	885	0.01	k−modes clustering algorithm
	NNP NN{3}	77	0	DNA polymerase activity inhibition
VBD (NN)+	VBD NN	364	0	supervised machine
	VBD NN{2}	115	0	visualized knowledge graph
	VBD NN{3}	16	0	predicted substance abuse treatment
NN{2，}	NN{2}	11 119	0.11	drug discovery
	NN{3}	1 935	0.02	drug dose selection
	NN{4}	139	0	tissue image analysis technology

智能新药研发测试数据集，枚举了各个词性组合模式在数据集中对应的高频子类型，并统计了这些高频子类型的词频和示例。借助有限状态自动机的强大表示能力，TFA算法能够匹配手工词性组合特征无法挖掘的许多关键短语，如 284 个JJ{3} NNS关键短语（重复 3 次形容词再后接名词复数）、885 个NNP NN{2}关键短语（专有名词后接 2 个名词）等。由此可以看出，TFA算法实际能够抽取的词性组合类型要远远多于手工定义的传统词性标注抽取方法。在面对一些可能会抽取到噪声的形容词时，如other、important等，TFA算法的统计功能支持用户针对性地过滤相关词性组合模式，或者通过更改上游词性标注算法进行排错。

TFA算法的设计思路和面向的应用场景主要围绕文献情报挖掘的颠覆性技术识别方向展开。当科技文献情报挖掘导向的关键短语抽取任务在文本中出现频率较低的名词短语时，要特别谨慎。这些名词短语既可能是低价值噪声，又可能是隐而未发的颠覆性技术信号。为避免简单地过滤掉低频的信号词汇，导致许多具有潜在价值的科技趋势情报的遗漏，TFA算法对准确率指标实行了比较宽松的标准。TFA算法可以通过补充词性筛查，或者与其他下游算法集成来进一步提高最终的综合效果。

4　结语

本文提出了TFA的无监督关键短语抽取算法。该算法的主要贡献是引入了有限状态自动机来构造关键短语抽取所需的词性组合模式。与监督学习型算法与深度学习型算法相对比。这种算法能够在缺乏标注数据、缺乏通用词表、缺乏预训练模型的细分交叉领域中取得良好的效果。与无监督学习算法相比，此算法能够具有很高的设计灵活性，实际使用者能够通过定义词性组合的正则表达式来达到等同于启发式特征设计的效

果。因为TFA算法的状态转移图能够表示任意的词性组合重复次数，所以TFA算法的表现能力远远大于人工枚举词性标签的传统关键短语抽取算法。TFA算法使用词性自动机抽取了大范围的关键短语，既保证了关键短语的纳入范围，又通过词性规则把控了候选关键短语的质量。

TFA算法在SemEval−2017数据集上表现出了很好的效果。在关键词评测中的召回率达到34.17%，在关键短语评测中的召回率达到52.00%。此算法的准确率、召回率和*F*1值都达到了关键词抽取开源算法库的平均水平，并且在召回率指标中表现出显著优势，比召回率排名第二的算法高出14.55%。因此，算法能够在科技文献情报的关键词挖掘中发挥实际应用价值。

本文的不足之处是暂时没有形成体系化的TFA后续噪声过滤体系。在未来研究中可以继续围绕词性统计数据或者其他集成算法来进一步提高词性自动机效果的最终呈现效果。

参考文献

[1] PAPAGIANNOPOULOU E, TSOUMAKAS G. A review of keyphrase extraction[J]. Wiley interdisciplinary reviews: data mining and knowledge discovery, 2020, 10(2): e1339.

[2] ALAMI M Z, FRIKH B, OUHBI B. Automatic keyphrase extraction: a survey and trends[J]. Journal of intelligent information systems, 2020, 54(2): 391−424.

[3] HULTH A. Improved automatic keyword extraction given more linguistic knowledge[C]//Proceedings of the 2003 conference on Empirical methods in natural language processing. Somerset, NJ: Assoc Computational Linguistics, 2003: 216−223.

[4] 徐文海, 温有奎. 一种基于TFIDF方法的中文关键词抽取算法[J]. 情报理论与实践, 2008, 31(2): 298−302.

[5] EL−BELTAGY S R, RAFEA A. Kp−miner: a keyphrase extraction system for English and Arabic documents[J]. Information systems, 2009, 34(1): 132−144.

[6] CAMPOS R, MANGARAVITE V, PASQUALI A, et al. Yake! collection−independent automatic keyword extractor[C]//European Conference on Information Retrieval. Berlin, German: Springer, 2018: 806−810.

[7] MIHALCEA R, TARAU P. Text rank: bringing order into text[C]//Proceedings of the 2004 conference on empirical methods in natural language processing. Barcelona, Spain:Association for Computational Linguistics (ACL), 2004: 404−411.

[8] WAN X, XIAO J. Collabrank: towards a collaborative approach to single−document keyphrase extraction[C]//Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). Manchester, UK: Association for Computational Linguistics, 2008: 969−976.

[9] BOUGOUIN A, BOUDIN F, DAILLE B. Topicrank: Graph−based topic ranking for keyphrase extraction[C]//International joint conference on natural language processing (IJCNLP). Nagoya, Japan: Asian Federation of Natural Language Processing, 2013: 543−551.

[10] JARDINE J, TEUFEL S. Topical page rank: a model of scientific expertise for bibliographic search[C]//Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. Gothenburg, Sweden: Association for Computational Linguistics, 2014: 501−510.

[11] FLORESCU C, CARAGEA C. Positionrank: an unsupervised approach to keyphrase extraction from scholarly documents[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA: Assoc Computational Linguistics−ACL, 2017: 1105−1115.

[12] BOUDIN F. Unsupervised keyphrase extraction with multipartite [C]//Conference on the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.Stroudsburg, PA: Association for Computational Linguistics, 2018: 667−672.

[13] PAGE L, BRIN S, MOTWANI R, et al. The pagerank citation ranking: bringing order to the web[R]. Stanford: Stanford InfoLab, 1999.

[14] WITTEN I H, PAYNTER G W, FRANK E, et al. Kea: practical automated keyphrase extraction[M]//Design and Usability of Digital Libraries: Case Studies in the Asia Pacific. United States: IGI global, 2005: 129−152.

[15] TURNEY P D. Learning algorithms for keyphrase extraction[J]. Information retrieval, 2000, 2(4): 303−336.

（下转第64页）

证分析[J].软科学, 2022, 36(6): 106−114.
[22] 姚建建, 门金来.高校科技人才培养对区域发展的贡献: 基于上海市人力资本和经济发展的分析[J].科技管理研究, 2020, 40(24): 118−126.
[23] 刘兵, 曾建丽, 梁林, 等.京津冀经济发展的动力源泉: 科技人才集聚的关键影响[J].科技管理研究, 2018, 38(3): 120−126.
[24] 赵永平, 李倩倩.人才集聚、区域创新与经济高质量发展[J].哈尔滨商业大学学报(社会科学版), 2023(1): 117−128.
[25] 罗智.我国优秀游泳运动员形态模型研究[J].成都体育学院学报, 2005(5): 93−97.
[26] 李琛, 吴映梅, 高彬嫔.滇中城市群城镇化与资源环境承载力耦合协调研究[J].水土保持研究, 2022, 29(2): 1−9.

（上接第40页）

[16] GOLLAPALLI S D, CARAGEA C. Extracting key−phrases from research papers using citation networks [C]//Proceedings of the AAAI conference on artificial intelligence: volume 28. Palo Alto, CA:Assoc Ad−vancement Artificial Intelligence, 2014.
[17] XIE F, WU X, ZHU X. Efficient sequential pattern mining with wildcards for keyphrase extraction[J]. Knowledge−based systems, 2017, 115(10): 27−39.
[18] ZHANG Q, WANG Y, GONG Y, et al. Keyphrase extraction using deep recurrent neural networks on twitter[C]//Proceedings of the 2016 conference on em−pirical methods in natural language processing. Aus−tin, Texas: Association for Computational Linguistics (ACL), 2016: 836−845.
[19] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre−training of deep bidirectional transformers for lan−guage understanding[C]//2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2019), VOL. 1.Stroudsburg, PA :Assoc Computa−tional Linguistics−ACL,2019:4171−4086.
[20] GROOTENDORST M. Keybert: minimal keyword ex−traction with bert[DB/OL]. [2023−05−22]. https://doi.org/10.5281/zenodo.4461265.
[21] SIDDIQI S, SHARAN A. Keyword and keyphrase ex−traction techniques: a literature review[J]. International journal of computer applications, 2015, 109(2): 18−23.
[22] 宗成庆. 中文信息处理丛书: 统计自然语言处理[M].北京: 清华大学出版社, 2008: 20−23.
[23] 俞士汶, 段慧明, 朱学锋, 等. 北京大学现代汉语语料库基本加工规范[J]. 中文信息学报, 2002, 16(5): 51−66.
[24] BOUDIN F. PKE: an open source python−based key−phrase extraction toolkit [C/OL]. [2023−05−22]. http://aclweb.org/anthology/C16−2015.
[25] AUGENSTEIN I, DAS M, RIEDEL S, et al. SemEval 2017 task 10: ScienceIE−extracting keyphrases and re−lations from scientific publications[C/OL]. [2023−05−22]. https://aclanthology.org/S17−2091.