

摘要语言视角下医学突破性论文识别研究^{*}

林紫洛 杨雪梅 于诗睿 陈逸菲 唐小利

(中国医学科学院/北京协和医学院医学信息研究所/图书馆 北京 100005)

〔摘要〕 遴选具有代表性的突破性论文集,分析论文摘要语言学特征,提取特征句式并划分摘要语步,利用深度学习算法构建医学突破性论文摘要识别模型。实证结果表明模型在具体医学领域中适用,为突破性论文早期发现提供参考。

〔关键词〕 突破性论文;摘要文本;语言学特征;自动识别

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2023.05.007

Study on Identification of Medical Breakthrough Papers from the Perspective of Abstract Language LIN Ziluo, YANG Xuemei, YU Shirui, CHEN Yifei, TANG Xiaoli, Institute of Medical Information & Library, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100005, China

〔Abstract〕 A representative collection of breakthrough papers is selected, the linguistic features of abstracts are analyzed, and the characteristic sentence patterns are extracted and the abstract moves are divided. Thus, an abstract recognition model for medical breakthrough papers is constructed by using deep learning algorithm. The empirical results show that the model is applicable in specific medical fields and provides references for early discovery of breakthrough papers.

〔Keywords〕 breakthrough papers; abstract text; linguistic features; automatic recognition

1 引言

近年来,我国对突破性研究的重视程度逐步提

高。具体到生物医学领域,在一些情况下生物医学创新受到关键科学发现的指导,当医疗供应方取得科学突破或新进展时,可能促成医学创新^[1]。学术论文作为基础研究工作的重要呈现形式,具有发现新观点、传播新方法、承载新理论、推广新技术等作用^[2],若能在论文发表初期把握其中反映的重大突破,有助于推进突破性研究的遴选与培育,为相关机构加强重要研究项目资助提供参考依据。在论文写作过程中作者有时会采用特定描述声明一项研究的首创性,分析这些描述的语言学特征可以揭示论文新发现的语言模式^[3]。因此,通过论文文本的语言学特征识别潜在突破性论文是目前值得探索的方向。

〔修回日期〕 2023-03-21

〔作者简介〕 林紫洛,硕士研究生;通信作者:唐小利,研究馆员,硕士生导师。

〔基金项目〕 中国医学科学院医学与健康科技创新工程“生物医学文献信息保障与集成服务平台”(项目编号:2021-I2M-1-033);国家科技图书文献中心专项“下一代开放知识服务平台关键技术优化集成与系统研发”(项目编号:2022XM28)。

2 相关研究

突破性研究指根本性突破或重要进展，影响力可辐射其他领域，同时具有原创性^[4]。记录突破性研究、表现突破性意义的学术论文为突破性论文，其识别方法分为基于文献计量学和基于文献内容两种。

2.1 基于文献计量学的识别方法

基于文献计量学的识别方法主要可分为引文频次分析、引文网络分析和多维分析。引文频次分析从论文被引频次、引用时间分布入手，进行统计分析或绘制引文曲线。早期识别潜在突破性论文的方法是找到高被引论文^[5-6]。随着引文分析方法的发展，研究者逐渐意识到仅使用引文频次不足以衡量论文价值，不再通过设置被引阈值筛选论文，如 Bornmann L 等^[7]通过计算学科领域和时间的动态标准化影响评分来识别突破性论文。引文网络分析的基础是文献之间的相互引证。Marx W 等^[8]引入引文出版年光谱，通过引文高峰期找到关键的开创性论文。Huang Y H 等^[9]认为突破性研究会引起现有范式论文引用链遭到破坏，计算“破裂分数”以识别这种变化。引文分析方法表现良好，但在论文发表初期，与引用相关的特征并不明显，需要一定时间窗口的积累。多维分析从论文、期刊、作者等多个维度遴选文献外部特征，采用回归分析或机器学习方法进行突破性识别。如 Wolcott H N 等^[10]考察引用计数、是否开放获取、合著者数量等一系列指标，使用随机森林模型评估其识别突破性论文的能力。梁国强^[11]遴选变革性研究的参考文献及施引文献特征，采用熵权法、因子分析法等构建变革性研究的早期识别模型。这类方法避免了使用单一指标无法全面测度研究的问题，但是部分特征不易获取或计算复杂，在推广应用上有所局限。

2.2 基于文献内容的识别方法

相比起引文特征，论文的文本特征不需时间积

累，更有利于突破性论文的早期识别，基于内容的识别方式应运而生。目前基于文献内容的识别方法所用语料主要是引用语句。Small H 等^[12]将引用时使用“discovery”“discover”和“discovered”等术语的句子称为发现句，并提出用机器学习方法识别描述科学发现的论文。王雪等^[13]获取突破性文献集引用语句中表征突破性评价意义的特征词，提出基于引用语句，结合深度学习算法的突破性论文识别方法。引用语料的获取受限于论文全文的可访问性，相比之下，摘要的可获取性更强。在摘要中作者可能讨论研究目的、总结研究成果、评估研究意义或影响，从而突出论文价值。杜建等^[14]在分析典型变革性研究论文中的表达时发现，其标题或摘要在表述上突显变革性，如“propose a new type of…，we overcome the limit by…”。从这个角度考虑，论文标题或摘要若出现类似术语可能是潜在的突破性论文。Fischer I 等^[15]将随机试验报告按照重要性分类，使用神经网络方法学习摘要语言模式，发现报告的重要程度与摘要语言的情感相关。已有研究证明基于内容识别突破性论文具备可行性，且摘要语料更易获取，能够弥补引文滞后性导致的引用不足，因此本文旨在探索使用论文摘要文本识别生物医学领域潜在突破性论文的方法。

3 研究设计（图 1）

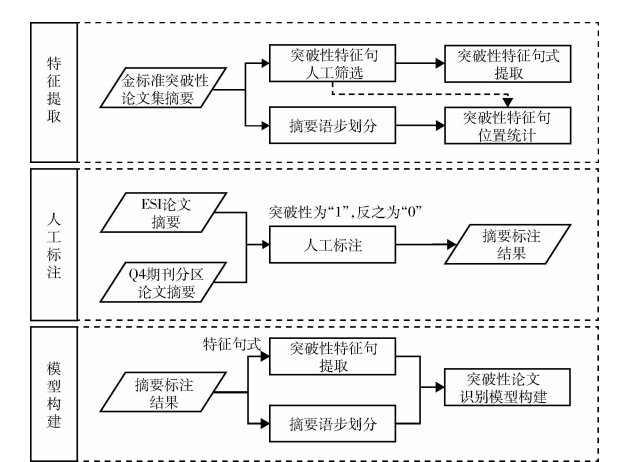


图 1 突破性论文识别方法路径

3.1 突破性论文摘要特征提取

3.1.1 金标准数据集 突破性论文摘要特征提取的首要任务是选择具有代表性的突破性论文作为金标准数据集，这些论文相较于其他论文可能具有突出的语言学特征。本研究从两方面考虑，将以下 4 个来源的论文纳入金标准数据集。一方面是具有滞后性的重大科学贡献出版物评审，如科学界最高奖项诺贝尔奖，其官网自 1981 年开始在介绍当年获奖者时会列举其关键论文；《自然》杂志不定期发布“里程碑”系列专题，总结某领域在过去取得的突破性进展。另一方面是针对近期发表论文的评审，如《科学》杂志自 1996 年起每年评选出当年最具代表性的 10 个科学突破并提供参考文献；Faculty Opinions 数据库遴选生物医学领域重要论文，由领域专家对论文进行评价和推荐，少数论文被推

荐为“里程碑”。资料获取时间为 2022 年 10 月。删除其中综述、评论、新闻等类型的文献并去重，最终得到 968 篇（89.5%）可获取摘要的突破性论文。

3.1.2 突破性特征句式 为说服读者接受文章观点，作者会采用显性或隐性描述阐明研究的重要意义或新颖成果^[16]。对论文摘要进行语言学特征分析，分析角度包括常用于表述突破性研究的特征词及其搭配、特征句在摘要中的位置等。采用人工判别方式阅读金标准突破性论文摘要，找出表明研究具有新发现、新观点、新方法以及存在重要意义或贡献的句子。单词具有广泛搭配，某个词在词组中的含义可能与原来不同，因此从前述筛选的句子中提取出特征词及其搭配，根据句中单词的依存关系组成特征句式。提取时，将名词分别以单数和复数形式表示，将动词以不同时态表示，见表 1。

表 1 突破性特征句式提取过程示例

突破性特征句	翻译	突破性特征句式
We found that a dendritic cell vaccine led to an increase in naturally occurring neoantigen – specific immunity and revealed previously undetected human leukocyte antigen (HLA) class I – restricted neoantigens in patients with advanced melanoma.	我们发现树突状细胞疫苗导致自然产生的新抗原特异性免疫力增加，并揭示了晚期黑色素瘤患者先前未发现的人类白细胞抗原（HLA）I 类限制性新抗原。	①we find ②we found ③reveal…previously undetected ④reveals…previously undetected ⑤revealed…previously undetected ⑥revealing…previously undetected

3.1.3 突破性特征句位置特点 摘要是对文献内容的高度概括，为尽可能多提供信息，作者可能会将背景、目的、方法、结果、结论等内容组织进摘要中，描述这些内容的句子称为语步。科技文献知识人工智能引擎 SciAIEngine^[17]集成自动识别摘要语步类型功能。使用其对摘要进行语步划分，并统计突破性特征句在摘要中的位置，进一步考察其位置特点。

3.2 训练数据获取及人工标注

基本科学指标数据库（essential science indicators，ESI）是评估学者、机构等学术水平及影响力的重要指标，入选 ESI 一定程度上说明论文具有较高学术价值，更有可能是突破性论文。与之相对，

影响因子较低的期刊论文中存在突破性论文的可能性较低。因此从近 10 年生物医学领域 ESI 论文中随机抽取一部分含摘要的非综述类论文，同时获取与 ESI 论文发表年份和学科领域相同，被《期刊引证报告》（journal citation report，JCR）收录为 Q4 期刊分区的论文。将突破性论文的识别任务转换为“是否为突破性论文”的二分类任务，对摘要进行人工标注，若摘要表征文章具有重大发现或首创性，则标注为“1”，反之标注为“0”。

3.3 突破性论文识别模型构建

3.3.1 摘要语步划分 由于学术论文摘要之间存在体量差异，一些较长摘要中存在诸多非关键信息句，可能影响利用摘要识别突破性研究的效果，因

此引入语步识别方法拆分摘要结构，得到背景、目的、方法、结果、结论，对文本进行降维处理。此外，通过前期获取的突破性特征句式匹配摘要中的特征句，特征句和摘要语步共同构成输入语料。

3.3.2 识别模型构建 采用深度学习算法进行识别模型训练，自动学习语言学特征并实现分类。前期使用小样本测试支持向量机（support vector machine, SVM）、卷积神经网络（convolutional neural networks, CNN）、深度神经网络（deep neural networks, DNN）、双向编码器表征（bidirectional encoder representation from transformers, BERT）等多种深度学习算法，发现 BERT 和 DNN 表现较优，并且 DNN 能以更速度与更优性能得到和 BERT 相近的结果，因此选择 DNN 算法进行识别模型构建。DNN 是一种多层深度神经网络算法，包括输入层、隐藏层和输出层，层与层之间存在相互连接、具有某种线性关系的神经元，再经由激活函数得到分类结果。在模型训练阶段，首先由隐藏层将语料中的词转化为固定维度的词向量，以表示不同词之间的语义关系，然后模型自动学习输入语料的语言特征与人工标注结果的关系；在测试阶段，模型将对语料进行同样处理，输出标签为“1”和“0”的分类结果及对应概率。

4 实证分析

4.1 突破性论文摘要语言学特征

通过人工判别方式从 395 篇金标准数据集摘要（占总数的 40.8%）中筛选出 477 个显性描述文章突破性意义的句子，并提取出 1 192 个突破性特征句式。在含有显性突破性特征句的摘要中，特征句在各

语步中出现的概率分别为背景 2.0%、目的 16.6%、方法 1.1%、结果 13.7%、结论 66.7%，说明作者撰写摘要时更可能在目的、结果或结论中使用显性表述阐明研究的突破性价值。在目的中出现次数较多的特征句式有“第 1 次（for the first time）”等，表明研究的重要意义；结果中出现较多的特征句式有“我们找到（we found）”“我们发现（we discovered）”等，说明研究的关键发现；结论中出现较多的特征句式有“据我们所知（to our knowledge）”“提供……证据（provide...evidence）”等，突出研究的特殊价值或创新性贡献。

4.2 突破性论文摘要识别模型训练结果

获取 5 000 条 ESI 论文摘要（占近 10 年生物医学 ESI 论文总数的 6%）和 5 000 篇被 JCR 收录的 Q4 期刊分区论文摘要，923 篇摘要经人工标注后标为“1”，9 077 篇摘要标为“0”，正负样本比例约为 1:9。样本不平衡可能影响预测结果，可以采用减少负样本或增强正样本的方法解决。考虑到减少负样本可能忽略数据中的有用信息，本研究使用美国 Protago 实验室提出的简单数据增强技术（easy data augmentation, EDA）增加正样本数量^[18]，使正负样本比例达到 1:1，模型性能得到一定提升。

对摘要进行语步划分，其中含目的语步的摘要有 7 975 篇、结果 8 658 篇、结论 9 738 篇，使用不同语步作为输入语料进行自动识别模型训练，见表 2。可以看出，使用语步训练的模型识别效果优于使用摘要，而加入特征句能够有效提升识别效果。整体而言，使用目的、结论语步和特征句的语料训练模型识别效果最好，F1 值为 0.835 1。

表 2 突破性论文摘要识别模型训练结果

训练语料	对正样本的识别效果			整体识别效果		
	精确率	召回率	F1	精确率	召回率	F1
摘要	0.892 7	0.687 9	0.777 0	0.810 9	0.799 1	0.793 3
摘要 + 特征句	0.881 1	0.712 1	0.787 7	0.811 4	0.804 0	0.799 6
目的 + 结论	0.902 0	0.732 3	0.808 3	0.830 0	0.822 8	0.818 4
目的 + 结论 + 特征句	0.881 7	0.789 7	0.833 2	0.838 0	0.837 1	0.835 1
结果 + 结论	0.899 8	0.714 7	0.796 7	0.822 9	0.814 0	0.809 0
结果 + 结论 + 特征句	0.903 7	0.727 1	0.805 8	0.829 3	0.821 3	0.816 7
结论	0.867 1	0.755 6	0.807 5	0.815 1	0.813 7	0.810 6
结论 + 特征句	0.872 4	0.773 6	0.820 0	0.824 8	0.824 1	0.821 6

4.3 T 细胞领域实证测试集

T 细胞作为特殊淋巴细胞,在免疫系统中发挥着重要作用,识别 T 细胞生物学发展进程中的重大突破,有助于未来研究人员对高效免疫治疗方法的探索。实证测试集的正样本包括两个部分。一是 Faculty Opinions 数据库中每篇受推荐的论文都有相应评级与分类,评级说明研究的重要程度,exceptional 等级重要程度最高,分类包括 controversial、good for teaching、new finding 等 10 类,其中 new finding 类表明该研究存在新发现,被推荐为 new finding 类的次数越多可能表明研究首创性越强。评级为 exceptional 且被推荐为 new finding 类次数大于 6 的 T 细胞领域论文共 125 篇,约占该数据库中 T 细胞领域受推荐文章的 2%,是较顶尖的医学论文,可视为潜在突破性论文。二是《自然》杂志于 2022 年 12 月推出 T 细胞“里程碑”系列专题,回顾过去 60 年内 T 细胞研究关键进展,并总结 34 篇相关里程碑文章。两部分内容去重后得到 154 篇正样本论文。负样本获取方式与训练数据相同,以相同数量 T 细胞领域的 Q4 期刊分区论文作为负样本,将测试集的摘要文本语料输入识别模型后得到预测结果。

4.4 T 细胞领域实证识别结果分析

4.4.1 识别效果评估 利用训练中识别效果最优的“目的+结论+特征句”模型对 T 细胞领域测试集进行识别,共识别出 90 篇突破性论文,其中 65 篇被收录在 Faculty Opinions 数据库或《自然》杂志“里程碑”系列,精确率为 72.2%。在全部 154 篇正样本中,有 89 篇未被识别为突破性论文,其中 37 篇摘要没有表征论文突破性意义的句子,调整后对突破性论文识别的召回率为 52.3%。总体来看,本研究构建的摘要识别模型一定程度上能够识别出潜在突破性论文,精确率略高于召回率。

4.4.2 正样本识别结果分析 进一步分析正样本中的两类论文, Faculty Opinions 数据库的潜在突破性论文精确率为 76.1%,召回率为 43.2%,而《自然》杂志“里程碑”系列论文精确率为

61.9%,召回率为 39.4%,识别效果弱于 Faculty Opinions 数据库论文。可能的原因是“里程碑”往往是某一研究方向的开创者,颠覆了过去的研究视角或模式,研究成果有待时间检验,作者在撰写摘要时倾向于避免使用过于积极的话语表述,但具有开创性的里程碑论文毕竟是少数,大多数突破性论文属于领域内的渐进式重大进展,相对而言作者在摘要中使用显性特征句说明研究贡献的概率更高。

5 结语

本研究在分析金标准突破性论文摘要语言学特征的基础上,提取摘要突破性特征句式,并利用深度神经网络 DNN 训练突破性论文自动识别模型。实践说明通过摘要的语步及语言学特征识别突破性论文具有可行性,能够避免利用引文相关指标进行识别的时滞性问题,为突破性论文的早期识别提供参考。但本研究也存在局限性,在具体医学领域中进行实证时召回率较低,后续研究中可以将语言特征和计量特征相结合,验证二者结合的方法是否能兼顾精确率与召回率,并尽可能实现早期识别。此外,模型语料中使用语步结构,语步识别工具精度会对模型实证研究结果造成一定影响。

参考文献

- 1 COMINS J A, LEYDESDORFF L. Citation algorithms for identifying research milestones driving biomedical innovation [J]. Scientometrics, 2017, 110 (3): 1495 - 1504.
- 2 罗卓然,王玉琦,钱佳佳,等. 学术论文创新性评价研究综述 [J]. 情报学报, 2021, 40 (7): 780 - 790.
- 3 毛琛瑜,乐小虬. 领域内中文科技文献中新发现语言描述特征分析 [J]. 现代图书情报技术, 2016 (5): 47 - 55.
- 4 SCHNEIDER J W, COSTAS R. Identifying potential “breakthrough” publications using refined citation analyses: three related explorative approaches [J]. Journal of the association for information science and technology, 2017, 68 (3): 709 - 723.
- 5 PONOMAREV I V, WILLIAMS D E, HACKETT C J, et al. Predicting highly cited papers: a method for early detection of candidate breakthroughs [J]. Technological forecasting and

social change, 2014, 81 (1): 49–55.

6 GUO G M, CHEN H S. Methods for recommending and predicting Nobel Prize candidates [J]. Program: automated library and information systems, 2014, 48 (2): 185–205.

7 BORNMANN L, YE A, YE F. Identifying landmark publications in the long run using field – normalized citation data [J]. Journal of documentation, 2018, 74 (2): 278–288.

8 MARX W, BORNMANN L, BARTH A, et al. Detecting the historical roots of research fields by reference publication year spectroscopy (RPYS) [J]. Journal of the association for information science & technology, 2014, 65 (4): 751–764.

9 HUANG Y H, HSU C N, LERMAN K. Identifying transformative scientific research [C]. Dallas: IEEE 13th International Conference on Data Mining (ICDM), 2013.

10 WOLCOTT H N, FOUCH M J, HSU E R, et al. Modeling time – dependent and – independent indicators to facilitate identification of breakthrough research papers [J]. Scientometrics, 2016, 107 (2): 807–817.

11 梁国强. 变革性研究的引文特征分析与早期识别模型构建 [D]. 大连: 大连理工大学, 2021.

12 SMALL H, TSENG H, PATEK M. Discovering discoveries;

identifying biomedical discoveries using citation contexts [J]. Journal of informetrics, 2017, 11 (1): 46–62.

13 王雪, 杨雪梅, 林紫洛, 等. 基于引文全文本的医学领域突破性文献识别研究 [J]. 情报杂志, 2021, 40 (3): 132–138.

14 杜建, 孙轶楠, 张阳, 等. 变革性研究的科学计量学特征与早期识别方法 [J]. 中国科学基金, 2019, 33 (1): 88–98.

15 FISCHER I, STEIGER H. Toward automatic evaluation of medical abstracts: the current value of sentiment analysis and machine learning for classification of the importance of PubMed abstracts of randomized trials for stroke [J]. Journal of stroke and cerebrovascular diseases, 2020, 29 (9): 105042.

16 张大群. 学术论文中的隐性评价及其识别 [J]. 外语教学理论与实践, 2010 (3): 36–43.

17 中国科学院文献情报中心. 科技文献知识人工智能引擎 [EB/OL]. [2023–02–19]. <http://sciengine.las.ac.cn/>.

18 WEI J, ZOU K. EDA: easy data augmentation techniques for boosting performance on text classification tasks [C]. Hong Kong: Conference on Empirical Methods in Natural Language Processing, 2019.

(上接第 38 页)

7 结语

随着公众对生物医药的关注度持续提升，查新员应首先贯彻落实精准查新理念，在不断提升个人生物医药知识储备的同时正确选择平台或数据库进行药品检索，进一步提高检索质量。此外，查新员应深入研究药品查新新颖性判断标准，完善判断依据及方法；各查新机构应合理制定药品查新服务费用，反映药品查新更高的知识价值。

参考文献

1 任珩, 王晓媛, 王君兰, 等. 我国科技查新机构的发展态势及转型思考 [J]. 情报理论与实践, 2019, 42 (6): 65–70.

2 夏冬, 谢强, 徐英祺. 产品查新与一般项目查新的比较研究——基于查新项目的实证 [J]. 图书情报导刊, 2021, 6 (1): 25–30.

3 陈蔚杰, 丁恩俊, 阮建海. 高校人文社会科学博士学位论文开题查新探讨 [J]. 高校图书馆工作, 2017, 37

(3): 68–70.

4 杨爱英, 徐军锋. 工法查新与一般科技查新的比较研究 [J]. 图书情报工作, 2012 (S2): 297–299.

5 夏冬, 王超, 任波. 标准查新项目与一般查新项目的比较研究——基于查新项目的实证研究 [J]. 标准科学, 2020 (2): 34–38.

6 国家发展和改革委员会. “十四五”生物经济发展规划 [EB/OL]. [2022–08–17]. https://www.ndrc.gov.cn/xwdt/tzgg/202205/t20220510_1324439.html?code=&state=123.

7 安琪. 高校科技查新服务定价调查及收费标准问题 [J]. 图书馆研究, 2021, 51 (3): 85–93.

8 李贞铮, 蔡春梦. 省级新产品查新和传统查新比较——以湖南省为例 [J]. 中国科技信息, 2022 (11): 133–135.

9 国家质量监督检验检疫总局, 中国国家标准化管理委员会. GB/T 32003—2015 科技查新技术规范 [S]. 北京: 中国标准出版社, 2015.

10 夏冬, 史继强, 任波. 尽调型查新与鉴定型查新的比较研究——基于查新项目的实证 [J]. 图书情报导刊, 2022, 7 (5): 28–34.

11 陈振标, 肖春, 范生清. 科技查新“形式化”现象成因探析与对策研究 [J]. 情报探索, 2022 (7): 45–50.