生物技术与装备知识图谱系统的设计与实现

佟 凡,毛逸清,阳沛湘,李江域,赵东升

[摘要] **目的** 通过构建生物技术与装备知识图谱,为生物技术与装备研究、应用与教学提供指导。**方法** 综合利用多源数据自动跟踪采集、领域本体和知识图谱模型构建、命名实体识别和关系抽取、知识图谱可视化展示和分析技术。**结果** 开发的生物技术与装备知识图谱开发和应用系统,具备数据采集、知识加工标引、知识检索与知识可视化展示功能。**结论** 系统可为生物技术与装备研发与应用提供有效的信息和知识服务。

[关键词] 生物;技术;装备;知识图谱
[中图分类号] TP391.1;TU244.5 [文献标志码] A
[文章编号] 1674-9960(2023)07-0539-06 DOI:10.7644/j.issn.1674-9960.2023.07.010

Design and deployment of biological technology and equipment knowledge graph system

*TONG Fan, MAO Yiqing, YANG Peixiang, LI Jiangyu, ZHAO Dongsheng**
(Academy of Military Medical Sciences, Academy of Military Sciences, Beijing 100850, China)
*Corresponding author, E-mail: dszhao@bmi.ac.cn

[**Abstract**] **Objective** To construct a biological technology and equipment knowledge graph in order to facilitate the research, application and teaching related to biological technology and equipment. **Methods** Such technologies as multi-source data collection, ontology and knowledge graph construction, text mining and analysis, knowledge annotation and curation, knowledge graph visualization and analysis were employed. **Results** The biological technology and equipment knowledge graph system was capable of data collection, knowledge labeling, knowledge retrival and knowledge graph presentation. **Conclusion** The system provides practical information and service for developing and deploying biological technology and equipment.

[**Key words**] biology; technology; equipment; knowledge graph

得益于生命科学的深入研究,生物技术与装备获得长足发展[1,2]。近年来,国际生物威胁呈多样化态势发展[3~6],同时生物恐怖、实验室泄漏等非传统生物安全威胁持续存在[7~10],我国必须充分做好应对复杂困难局面的准备,深度开展自主可控生物技术装备研发。高效的信息和知识服务,是开展相关研究和应用的重要支撑。

知识图谱是一种用图模型来描述知识、建模世界万物之间关系的技术方法,它吸收了本体和语义网在知识组织和表达方面的理念,以符号形式描述现实世界中的概念及其关系,使得知识更易于在计算机之间和计算机与人之间交换、流通和加工[11]。知识图谱以从非结构化信息中抽取结构化知识的技术优势,顺应了时代对知识服务高涨需求,成为大数据时代人工智能的前沿和基础性研究领域[12]。传统的生物信息与数据库资源聚焦于智库、指南、法律、规范等宏观层面,数据库构建也仅止步于相关情报的收集、获取、存储与展示过程,面向生物技术和装备领域的知识图谱(如构建工具、知识库及应用系统)研究较少,使得生物技术研究与装备研发人员难以快速和准确获得国内外最新进展、发展态势等信息,在一定程度上制约了科研进展。为此,本研究综合利用自然语言处理、机器学习与知识图谱技术,构建了面向生物领域研究人员的知识图谱开发和应用系统,具备数据采集、知识加工标引、知识检索与可视化展示功能,现报道如下。

1 系统功能与数据模型设计

1.1 功能设计

生物技术与装备知识图谱系统围绕各类信息资源采集管理、知识加工标引、知识图谱应用等功能开展。系统功能如图1所示。

[作者简介] 佟 凡,男,助理研究员,研究方向:生物医学大数据和知识图谱,Tel:010-66930378,E-mail:tongfan@bmi.ac.cn
[作者单位] 军事科学院军事医学研究院,北京 100850
[通信作者] 赵东升,Tel:010-66931123,E-mail:dszhao@bmi.ac.cn

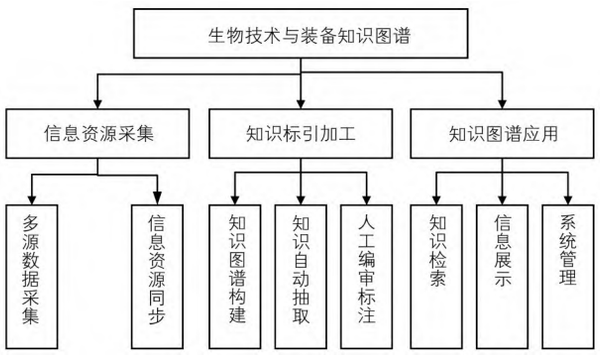

图1 生物技术与装备知识图谱系统功能图

1.1.1 *信息资源采集* 多源数据采集,实现包括研究文献类、期刊类、机构类、技术类、装备类、病原微生物类、物种类、传染病类和法律法规等在内的多类数据源、多种结构的数据采集功能;信息资源同步,实现定期自动采集数据,支持不同数据源采集模型的注册管理和调度管理,支持采集周期等信息配置。

1.1.2 *知识标引加工* 知识图谱构建,具备实体定义、关系定义等功能;知识自动抽取,支持生物技术与装备专业领域的实体识别、关系抽取等人工智能算法,从文本信息中自动提取实体和关系;人工编审标注,实现对提取知识的人工修正,并提供意见冲突时的二次审核机制。

1.1.3 *知识图谱应用展示* 知识检索,支持对采集信息的全文检索、关联检索、知识卡片和知识推荐;信息展示,提供友好的信息浏览方式,实现图形化的知识图谱展示和分析交互界面,支持研究机构、研究人员的合作网络分析,研究热点趋势分析;系统管理,对用户、角色、权限与资源的分配、回收等行为操作。

1.2 *数据模型设计*

生物技术与装备知识图谱系统的数据模型设计包括知识范围(数据来源)界定、实体类型(概念)和实体关系定义、数据库(数据存储模型)设计三方面,如图2所示。

1.2.1 *知识范围界定* 以生物技术与装备领域权威数据源、现有生物技术与装备专业数据库和其他来源的相关数据为基础,通过数据采集模型生成采集数据集。选取:①重点国家或地区CDC、梅斯、健康界等中外重要官方网站共计23个信息来源;②*Lancet*、*Nature*、*Science*等高影响力期刊及其子刊共计43个期刊来源;③ThermoFisher、QIAGEN、Roch等高市场占有率品牌共计27个技术与设备来源;④GO[13]、LOINC[14]、MeSH[15]等共计7个高公信力本体与术语词表来源。

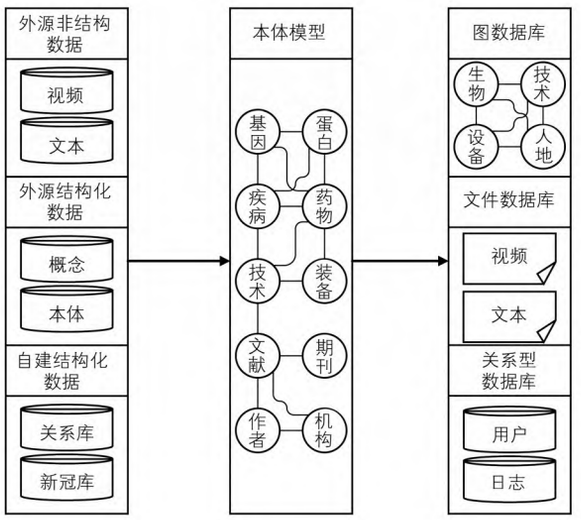

图2 系统数据架构图

1.2.2 *实体类型和实体关系定义* 由领域专家结合领域知识、现有国际通用本体库和已采集数据集,利用protégé本体构建工具[16]完成本体数据建模,并形成生物技术与装备本体库。本研究聚焦“防控重大新发突发传染病、动植物疫情”“生物技术研究、开发与应用安全”等6大领域内容,构建疾病、化合物、药物、技术、装备、物种等共计30类主要实体,以及contained in、depend on、derive from等共计25种实体间关系,name、code、definition等共计294个实体属性的知识图谱模型,如图3所示。

1.2.3 *数据库设计* 由生物技术与装备基础数据库、生物技术与装备知识库组成。生物技术与装备基础数据库通过对本体数据和采集数据的多级数据ETL处理,采用关系型数据库与文件数据库存储技术,初步实现数据标准化、格式化、归一化;生物技术与装备知识库是把基础数据库的数据通过机器学习算法、人工智能算法和人工审核等方法,利用图存储技术完成数据关联,形成生物技术与装备知识网络,为后续的数据应用提供有力支持。

2 系统实现

2.1 *系统技术架构*

系统技术架构如图4所示,由底而上分为数据存储层、数据处理层与数据展示层3个层次。

2.1.1 *数据存储层* 基于国产自主可控服务器与开源Linux操作系统,采用MySQL关系型数据库存储用户相关信息,MongoDB文档数据库[17]存储文本类、视频类非结构化信息,dGraph图数据库[18]存储知识图谱实体、实体属性及实体间关系信息,实现聚焦数据特点的不同数据类型的数据存储。

2.1.2 *数据处理层* 对于文本类来源数据(如

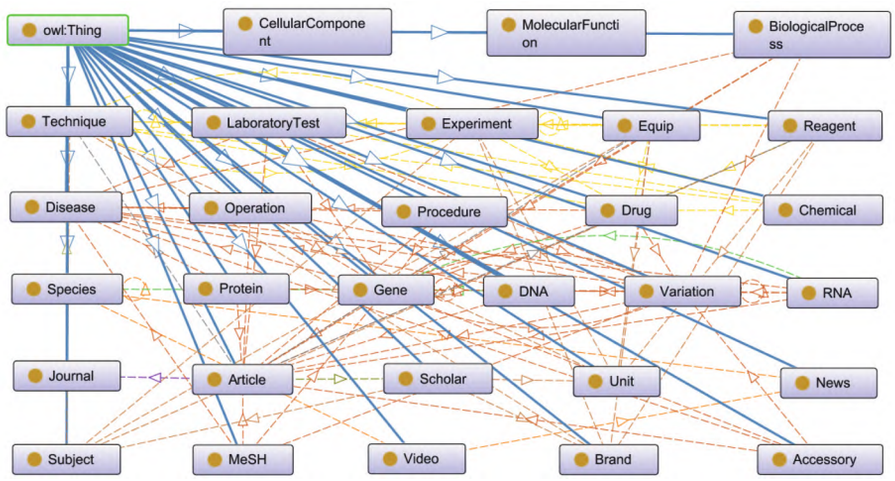

图3　生物技术与装备知识图谱模型

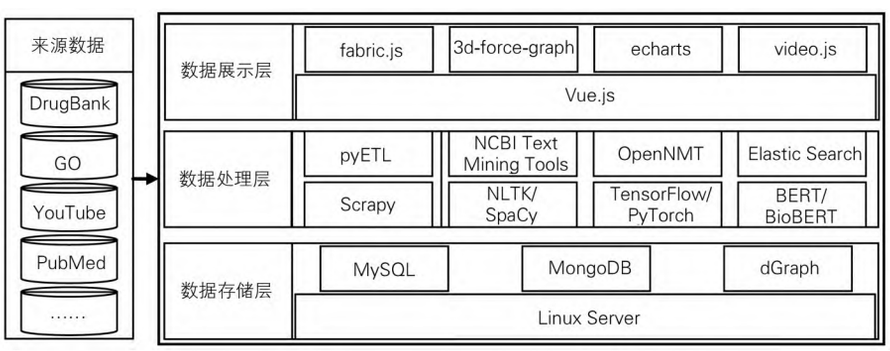

图4　系统技术架构图

PubMed)、视频类来源数据(如YouTube)等非结构化数据,利用Scrapy[19]数据获取加工处理工具,对于本体类来源数据(如Gene Ontology)、概念类来源数据(如DrugBank[20])等结构化数据利用pyETL数据抽取转化清洗工具进行存储入库;利用NLTK[21]、SpaCy[22]等自然语言处理工具与BioBert[23]、OpenNMT[24]等深度机器学习模型,实现命名实体识别、关系抽取、中英翻译与全文检索等功能。

2.1.3　*数据展示层*　基于Vue.js框架建立可视化用户交互页面,知识图谱展示与交互主要通过fabric与3d-force-graph库实现,热搜主题词热区图、用户访问浏览曲线等数据统计图表,展示由echarts库完成,视频播放与字幕嵌入功能由video库实现。

2.2　关键处理流程和技术实现

2.2.1　*多源数据自动跟踪采集*　编写python爬虫软件实现跟踪采集功能,先划定数据范围和数据源,并筛选网站数据内容和重要数据项,考察数据质量,确定数据采集范围、数据标签和离线/实时采集需求;再根据网站数据记录,估计网站更新情况,确定数据采集/更新策略;然后开始数据采集工作,由采集作业调度器负责多数据源采集任务的分配、触发、调度、监控和管理,同时进行负载均衡和任务异常处理;每个数据源的离线采集和实时采集采用不同的爬取方式,各自建立独立爬虫进行数据采集,爬虫包括引擎、调度、下载、缓存和分布式处理模块。

2.2.2　*知识自动抽取*　实体抽取工具主要采用tmChem(化学实体标注[25])、DNorm(疾病实体标注[26])、GNormPlus(基因标注[27])、SR4GN(物种标注[28])、tmVar(变异标注[29])、BioByGANS(多实体标注[30]),上述工具已经过多次版本更新,在NCBI在线注释等系统中使用;系统关系抽取工具主要采用BioBERT[30],较传统方法,BioBERT作为预训练模型的关系抽取算法,各性能指标都有显著提升,它在段落级别的文章关系抽取中能抽取出更多实体内容。

2.2.3　*知识自动注释和人工编审*　对生物技术与装备的新闻资讯和文献等文本资源进行编审,支持专家单篇或批量确认。系统根据当前激活算法和全局配置情况,启动对信息的知识抽取,抽取后专家需对抽取结果进行确认,并进行人工编辑。

2.2.4　*知识检索与可视化分析*　采用Vue前端框架,调用网络图绘制库、统计图表绘制库、PDF文件解析库、视频文件解析库等一系列开源库实现知识

图谱的可视化与交互性，辅助用户对于感兴趣的概念进行全文检索和知识图谱检索，并在此基础上进行研究机构、研究人员合作网络分析，研究热点趋势等后续深层数据分析。

2.3 结果

系统部署于飞腾S2500 64核CPU、紫光国芯32G内存、1T机械硬盘服务器，操作系统选用CentOS 7.6，其他工具均选用开源、稳定、长期维护版本，如Apache 2.4.46、MySQL 8.0.21、PHP：7.2.33、Python：3.9.5，3.7.4与Java openjdk 11.0.14.1。

系统现包含生物相关技术619项、装备2698项、文献6.7万篇、资讯1.4万条、医学主题词2.9万条、生物相关实体80余万条，实体间关系56余万条，实体属性190余万条。系统交互界面如图5所示。

用户在系统首页（图5a）可通过全文检索与图谱检索2种方式检索感兴趣的实体或概念，如COVID，其中全文检索主要面向技术、装备、文献、资讯和视频类资源，可按照相关度、阅读量和发布时间3种排序方式提供建检索结果（图5b）；而图谱检索则主要针对知识图谱模型中的实体展开，包括DNA、RNA、蛋白质、疾病、药物等（图5c）。进入到具体文献、资讯详情页后，也可进一步查看知识标引与知识图谱结果（图6a）。对于知识标引与知识图谱结果，可放大查看其详情（图6b）。

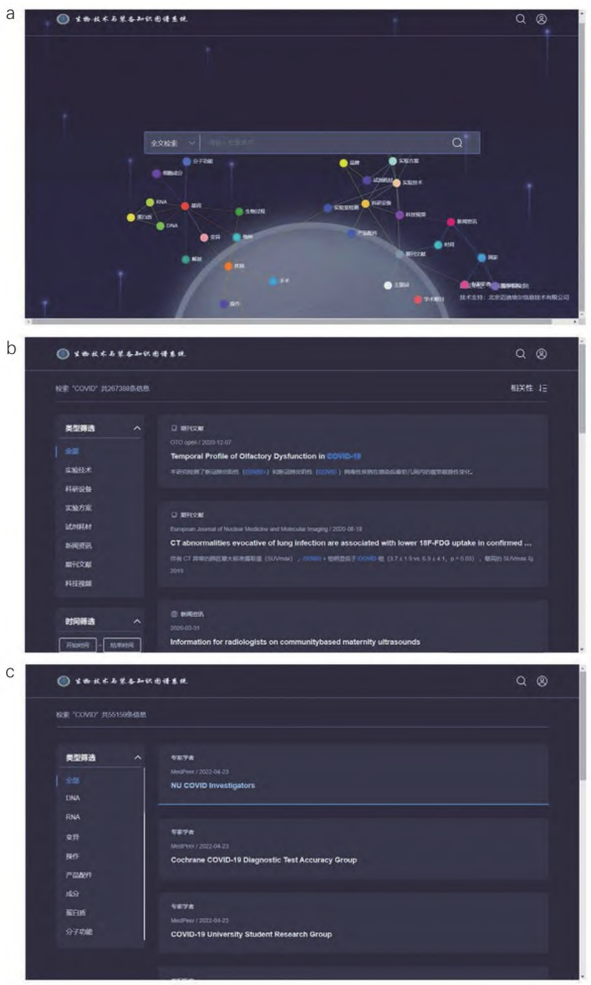

图5 系统代表页面

a.系统首页；b.全文检索结果页；c.图谱检索结果页

3 讨论

本研究面向生物技术装备研发和应用需求，设计开发了生物技术与装备知识图谱系统，具备数据采集、知识加工标引和知识可视化展示功能，且配备数据定期更新机制与算法自动抽取流程，可为生物技术与装备研发提供有效而持续的信息和知识服务。

在生物研究领域中，本系统一是提供了覆盖广泛的技术装备知识图谱模型：在前期需求要素与数据资源充分调研的基础上设计了涵盖操作、物种、疾病、化合物、装备、技术、基因、蛋白质、文献、机构等重点关注的实体、实体属性与实体间关系的知识图谱概念模型，填补了生物技术与装备领域技术与装备本体研究方面的空白，为生物技术与装备知识图谱的研发打下坚实基础。二是提供了性能良好的技术装备知识图谱知识抽取算法库：基于前期理论与技术积累，构建了生物技术与装备领域专用多命名实体识别算法库与生物技术与装备领域专用

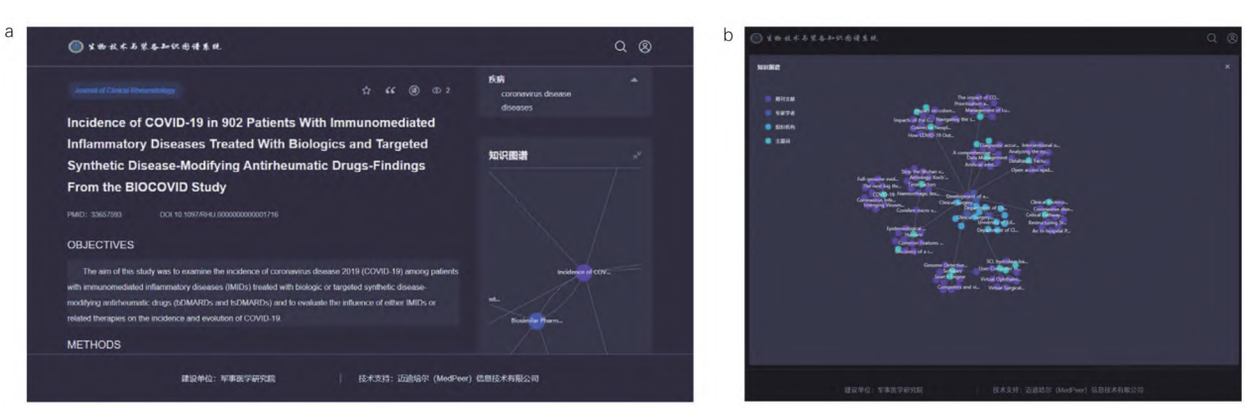

图6 系统知识标引（a）与知识图谱（b）显示页面

多实体关系抽取算法库两套创新算法库，有效提升了生物范畴目标实体及关系的信息提取能力，并为类似生物医学或生物信息学领域相关命名实体识别与关系抽取任务提供有益借鉴。三是提供了内容全面的生物技术装备知识图谱：基于自研技术装备知识图谱构建系统，邀请相关领域与有关学科专家投入时间精力编审机器抽取数据，最终保留一定规模技术、装备、国家、学者、机构、基因（蛋白/变异）、疾病、药物（化合物）、微生物等实体与技术-装备、国家-学者-机构、基因（蛋白/变异）-疾病-药物（化合物）-微生物等关系知识类数据，为生物技术与装备领域科研人员开展教育或研究提供重要可信数据来源。

目前，系统处于原型验证与试运行阶段，涵盖的数据规模与时间跨度仍具有很大提升空间；同时，虽预留并开放数据交换标准与接口，但因缺少实例，与现有业务系统间的兼容性与互补性也有待进一步验证。随着数据规模的增长与拓展性能的提升，以及更多自然语言处理与人工智能技术的引入，系统有望为领域专业人员提供更为全面、精准与智能的知识服务。

【参考文献】

[1] 刘家伟，冯佳佳，孔维华，等.我国生物医药领域中生物医学新技术发展及管理现状的思考[J].医学新知，2023，33(2)：136-142.

[2] 熊　燕.生物技术与信息技术融合发展带来的变革[J].人民论坛，2022(17)：17-21.

[3] 王小理.生物安全时代：新生物科技变革与国家安全治理[J].中国生物工程杂志，2020，40(9)：95-109.

[4] 关武祥，陈新文.新发和烈性传染病的防控与生物安全[J].中国科学院院刊，2016，31(4)：423-431.

[5] 王盼盼.美国生物防御科研项目梳理与分析[D].北京：军事科学院，2021.

[6] Kotwal A, Yadav A. Biothreat & one health: Current scenario & way forward[J]. Indian J Med Res, 2021, 153(3): 257-263.

[7] 孙　琳，杨春华.美国近年生物恐怖袭击和生物实验室事故及其政策影响[J].军事医学，2017，41(11)：923-928.

[8] Peng H, Bilal M, Iqbal H. Improved biosafety and biosecurity measures and/or strategies to tackle laboratory-acquired infections and related risks [J]. Int J Environ Res Public Health, 2018, 15(12): 2697.

[9] Aspland AM, Douagi I, Filby A, *et al*. Biosafety during a pandemic: shared resource laboratories rise to the challenge[J]. Cytometry A, 2021, 99(1): 68-80.

[10] Ta L, Gosa L, Nathanson DA. Biosafety and biohazards: Understanding biosafety levels and meeting safety requirements of a Biobank[J]. Methods Mol Biol, 2019, 1897: 213-225.

[11] 范媛媛，李忠民.中文医学知识图谱研究及应用进展[J].计算机科学与探索，2022，16(10)：2219-2233.

[12] 赵悦淑，王　军，王　蕊，等.中文医学知识图谱研究进展[J].中国数字医学，2021，16(6)：86-91.

[13] Gene Ontology Consortium. Gene ontology consortium: Going forward[J]. Nucleic Acids Res, 2015, 43: D1049-D1056.

[14] Stram M, Gigliotti T, Hartman D, *et al*. Logical observation identifiers names and codes for laboratorians[J]. Arch Pathol Lab Med, 2020, 144(2): 229-239.

[15] Fernandez-Llimos F, Salgado TM. Standardization of pharmacy practice terminology and the Medical Subject Headings (MeSH)[J]. Res Social Adm Pharm, 2021, 17(4): 819-820.

[16] Győrödi C, Győrödi R, Pecherle G, *et al*. A comparative study: MongoDB *vs*. MySQL[C]//The 13th International Conference on Engineering of Modern Electric Systems (EMES). Romania: IEEE, 2015: 1-6.

[17] Musen MA. The protégé project: A look back and a look forward[J]. AI Matters, 2015, 1(4): 4-12.

[18] Ortega V, Ruiz L, Gutierrez L, *et al*. A selection process of graph databases based on business requirements [C]//International Conference on Software Process Improvement. Mexico: Springer, 2019: 80-90.

[19] Myers D, McGuffee JW. Choosing scrapy[J]. J Comput Sys Sci, 2015, 31(1): 83-89.

[20] Wishart DS, Feunang YD, Guo AC, *et al*. DrugBank 5.0: A major update to the DrugBank database for 2018[J]. Nucleic Acids Res, 2018, 46(D1): D1074-D1082.

[21] Bird S. NLTK: The Natural Language Toolkit[C]//Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions. Australia: IEEE, 2006: 69-72.

[22] Vasiliev Y. Natural Language Processing with Python and SpaCy: A Practical Introduction[M]. USA: No Starch Press, 2020.

[23] Lee J, Yoon W, Kim S, *et al*. BioBERT: A pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics, 2020, 36(4): 1234-1240.

[24] Klein G, Hernandez F, Nguyen V, *et al*. The OpenNMT neural machine translation toolkit[C]//Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track). Mexico: IEEE, 2020: 102-109.

[25] Leaman R, Wei CH, Lu Z. tmChem: A high performance approach for chemical named entity recognition and normalization[J]. J Cheminform, 2015, 7(Suppl 1): S3.

[26] Leaman R, IslamajDogan R, Lu Z. DNorm: Disease name normalization with pairwise learning to rank[J]. Bioinformatics, 2013, 29(22): 2909-2917.

[27] Wei CH, Kao HY, Lu Z. GNormPlus: An integrative approach for tagging genes, gene families, and protein domains[J]. Biomed Res Int, 2015, 2015: 918710.

[28] Wei CH, Kao HY, Lu Z. SR4GN: A species recognition software tool for gene normalization[J]. PLoS One, 2012, 7(6): e38460.

[29] Wei CH, Allot A, Riehle K, *et al*. tmVar 3.0: An improved variant concept recognition and normalization tool[J]. Bioinformatics, 2022, 38(18): 4449-4451.

[30] Zheng X, Du H, Luo X, *et al*. BioByGANS: biomedical named entity recognition by fusing contextual and syntactic features through graph attention network in node classification framework. BMC Bioinformatics, 2022 Nov 22, 23(1): 501.

（孙承媛　编辑　2022-10-31　收稿）