2023 年 1 月 东南大学学报(哲学社会科学版) Jan. 2023
第 25 卷第 1 期 Journal of Southeast University(Philosophy and Social Science) Vol. 25 No. 1

“互联网+多语”公共卫生监测系统:价值定位、实现路径及应用前景

祝晓宏 郭 熙

(暨南大学 华文学院/国家语委海外华语研究中心, 广东 广州 510610)

[摘 要] 新冠疫情暴露了现有传染病监测系统的一些漏洞,研发新型公共卫生监测系统是势之所趋。建设“互联网+多语”公共卫生监测系统,有利于筑牢我国公共安全监测体系。结合 WHO 公共卫生监测指标和我国语言能力现状,将新系统的价值定位为识别及时、覆盖广泛、简单有效,实现路径依赖信息收集、处理、分析、传播四环节和多学科支撑,尤其需要语言学科在谣言识别、翻译和本体知识构建、异常疾病事件识别和检测、信息发布的语种和语用策略方面进行知识供给。新系统可以提高社会治理效能,促进语言产业发展,有着广阔的应用前景。

[关键词] “互联网+多语” 公共卫生监测 语言应用

一、引言

2019 年末,一场突如其来的新冠肺炎疫情严重地打乱了世界进程。面对疫情,我国政府紧急应对,科学施策,始终以人民生命安全为首位,取得了举世瞩目的抗疫战绩。面对疫情,我国学界也纷纷响应,包括医学、情报学、传播学、语言学等在内的众学科都各展其长,发布了很多应急研究成果,有力地支撑了科学抗疫。整体来看,大量研究都是助力“抗疫”而作,针对“防疫”、深刻反思我国公共卫生和情报治理体系问题的研究还不多①。

我国的抗疫成绩表明:如果从源头上国家就拥有一套快速、精准、高效的防疫监测系统,一旦识别险情马上预警,处置有方,事态或许不会失控蔓延,酿成大灾。正是基于这样的考虑,2003 年非典之后,卫生部中国疾控中心即建立了覆盖全国的传染病监测信息系统。这套系统简称“网络直报系统”:医院发现传染性病例,医生在这套网络系统点击报告病例,中国疾控中心在内的各级疾控部门就能实时了解情况。需要注意的是,这套网络直报系统并不是逐级报告,而是只要医院在内部网络系统中点击报告了病例,中国疾控中心第一时间就应该收到。然而遗憾的是,一直运行良好的这套系统却在新冠疫情早期预警中信息失灵②,从而影响了后续的应急反应。

目前,国外基于非正式网络信息的疫情监测工具(如 BlueDot、HealthMap、Flowminder 等)功用日显;国内研发“境外”“全球”疫情监测预警系统呼声日浓③。在认真梳理既存问题的基础上,结合国

[基金项目] 国家社科基金重大项目“境外华语资源数据库建设及应用研究”(19ZDA311)、教育部人文社科项目“东南亚华语传承口述史数据库建设研究”(19YJC740125)成果之一。

[作者简介] 祝晓宏(1981—),安徽巢湖人,暨南大学华文学院副教授,博士,研究方向:社会语言学、海外华语及华文教育。

① 参见丁蕾、蔡伟、丁健青等《新型冠状病毒感染疫情下的思考》,《中国科学:生命科学》2020 年第 3 期,第 11 页;苏新宁、蒋勋《情报体系在应急事件中的作用与价值——以新冠肺炎疫情防控为例》,《图书与情报》2020 第 1 期,第 6-14 页;周晓英《新冠肺炎疫情防控中的应急信息管理问题与对策研究》,《图书与情报》2020 年第 1 期,第 51-57 页。

② 信娜、王小、孙爱民等:《传染病网络直报系统投资了 7.3 亿,为何失灵了 28 天?》(2020-02-25)[2022-10-26],https://new.qq.com/rain/a/20200225A0N9JF00。刘玉海:《中国疾控中心原副主任杨功焕:SARS 之后国家重金建立、传染病网络直报系统,应关注其在这次疫情中如何运行》(2020-01-30)[2022-10-26], https://baijiahao.baidu.com/s?id=1657144472405928296&wfr=spider&for=pc。

③ 参见施雯《新兴疫情监测工具或可成为“吹哨人”》,《竞争情报》2020 年第 2 期,第 2-8 页;边文越、冷伏海《面向突发重大公共卫生事件应急决策的境外公共卫生战略情报体系研究——以应对新冠肺炎疫情为例》,《图书与情报》2020 年第 2 期,第 13-18 页;巴剑波、连凌、栾洁等《全球传染病疫情网络监测预警系统的设计与实施》,《海军医学杂志》2021 年第 1 期,第 54-57 页。

内外公共安全监测系统研发现状,我们认为,有必要强调一种新的监测思路,建设一套基于“互联网+多语”的公共卫生监测系统,以弥补现有监测系统的疏漏,提升我国公共安全预警能力。

二、建设公共疫情多语监测系统是势之所趋

(一)我国缺少“互联网+基于事件”的主动监测系统

信息失灵指供需双方在产品信息的掌握程度上的不对称。如果将事关人民生命安全的传染病信息当作某种特殊的公共产品,那么疾控中心是信息需方,医院则是信息供方。在疫情初始阶段,倘若医院因为某些缘故(特别是对于不明的新发疫病)而漏报、未报或误报,疾控中心就无法第一时间掌握信息,预警响应也就无从谈起①。

完全依赖医院直报和基于医生专业判断,自下而上信息传达,本质上是一种被动监测。被动监测极有可能产生信息盲点和信息延误②。研究证实,私人诊所报告的时效性最差;个体门诊几乎是传染病疫情报告的空白③;“疫情上报第一人”张继先的情报也没能即时上达中国疾控中心系统④。被动监测的消息源比较单一,容易丧失先机。可以说,不能主动出击利用多方面信息进行监测,是这套系统的短板。

人类正在全面深度进入互联网时代。互联网时代需要强调一种新的监测思路,即自上而下地获取信息,利用互联网世界里有关公共卫生的实时信息资源,采取文本挖掘、深度学习等多种技术手段,计算机自动识别,辅以人工干预,政府顶层主动监测。

公共卫生监测有不同的类型:常规报告和哨点监测,主动监测和被动监测,基于指标的监测(indicator-based surveillance)和基于事件的监测(event-based surveillance)。后两类监测分别对应被动监测和主动监测,它们在工作原理和数据来源方面存在诸多差异:前者的监测过程是系统、常规的,采用有组织、正式、有限、预先界定的疾病数据;后者监测过程通常是即时、灵活的,采用多元(正式和非正式)、非限定的、所有事关危险的数据⑤。一般来说,前者的优势在于监测信息准确可靠,但时效性有所滞缓,而后者的优势在于及时敏感,且耗费成本较低。

当前,国际社会愈益重视监测信息来源的多元化。全球基于互联网的公共卫生监测项目发展迅速,监测准度大大提升,很多国家包括世界卫生组织都开始重视监测源于互联网的非官方渠道的传染病信息⑥。特别是人工智能为信息技术赋能,“互联网+大数据+基于事件”的监测优势越发明显,成为传统监测体系的重要补充。

(二)全球公共卫生监测系统研发简况

在公共卫生监测领域,上海市率先开始探索建立“互联网+基于事件”的监测模式⑦,一些公共卫生和信息安全专家也提出建立“互联网、大数据、人工智能”的监测系统设想⑧。但是,这些设想都是限于局部地区或境内,是单语思维监测,监测范围和监测人群还非常有限。

---

① 许雯、陈思:《中疾控独家回应:“人传人”早有推论 保守下结论有原因》(2020-01-03)[2022-10-26],https://www.sohu.com/a/369779073_114988。

② 杨海:《武汉早期疫情上报为何一度中断》(2020-03-05)[2022-20-26],https://baijiahao.baidu.com/s?id=1660338140527663778&wfr=spider&for=pc。

③ 参见俞新莲、刘红莲、颜玉炳等《网络直报6年后厦门市传染病监测时效性评价》,《中华疾病控制杂志》2010年第8期,第742-744页;中国疾病预防控制中心《中国疾病预防控制传染病监测信息系统介绍》(2007-2-15)[2022-10-26],https://www.chinacdc.cn/ztxm/ggwsjc/jcxt/200702/t20070215_41341.html。

④ 隋唐:《“疫情上报第一人”张继先:我这次把一生的眼泪流光了!》(2020-02-09)[2022-10-26],http://society.people.com.cn/n1/2020/0209/c1008-31578284.html。

⑤ WHO., *Early detection, assessment and response to acute public health events: Implementation of Early Warning and Response with a focus on Event-Based Surveillance*, WHO,2014, p.13.

⑥ 参见熊玮仪、冯子健《中国传染病监测的发展历程、现状与问题》,《中华流行病学杂志》2011年第10期,第957-960页。

⑦ 参见何懿、陆殷昊、何永超等《上海市公共卫生安全保障基于事件的监测体系的构建》,《上海预防医学》2019年第11期,第874-880页。

⑧ 参见尉景、王松俊、赵东升《面向传染病监测的互联网情报动态追踪系统的设计与实现》,《军事医学》2014年第12期,第976-980页;屈晓晖、袁武、袁文《时空大数据分析技术在传染病预测预警中的应用》,《中国数字医学》2015年第8期,第36-39页;祝丙华、王立贵、孙岩松《基于大数据传染病监测预警研究进展》,《中国公共卫生》2016年第9期,第1276-1279页。

世界范围来看,随着人工智能和语言资源处理技术的进步,传统监测系统和基于“互联网+”的新型监测系统互相配合,协同作用,正在成为生物安全监测(Bio- Surveillance)领域的趋势,一些大国和地区都在建设和布控基于互联网信息资源的公共安全监测系统,以服务国家利益和国家战略①。表1是全球六大颇有代表性的生物监测系统情况表。

表1 全球“基于事件的生物监控系统”情况表

系统名称	ProMED	GPHIN	MedISys	Healthmap	IBIS	PADI-web
所有者	国际传染病协会(美国)	公共卫生部(加拿大)	联合研究中心(欧盟)	波士顿儿童医院(美国)	墨尔本大学(澳大利亚)	ESA监测平台(法国)
推出年份	1994	1997	2004	2006	2013	2016
访问政策	公开	限制	公开	公开	公开	公开
威胁类型	动物/人类/植物	动物/人类/植物/环境	动物/人类/植物/环境	动物/人类/植物/环境	动物/植物	动物
语言数量	7	9	50	7	1	6
数据类型	官方/非官方	官方/非官方	官方/非官方	官方/非官方	非官方	非官方
界面数据源语言	母语	英语	母语	英语/母语	母语	英语/母语
界面数据输出	无	无	时间序列/地图	时间序列/地图	地图	时间序列/地图

这些系统在全球公共安全领域内发挥了积极的作用。其共同点是,都非常重视互联网等非官方数据,监测的语言多达50种,用户登录系统界面基本可以使用母语查看数据源。这些系统反映了这些国家在全球公共安全监测方面的战略部署,而非国家语言实力。实际上,这些监测系统涵盖的语种也主要是联合国常用工作语言或欧盟语言,在语种代表性方面都还很不够。例如,马来语是东南亚的通用语,并不在上述系统之内。一些学者发现,ProMED和HealthMap两大知名监测系统漏掉了马来语地区2016年8月—2017年8月内多个传染病暴发新闻②。

除了上述多语监测系统,2006年日本国立情报学研究所也推出了能够辐射东南亚的BioCaster疫情监测系统(8种语言);2015年,英国提出要提高国家语言能力,扩大在公共卫生领域语言监测方面的影响力。一些民间机构也在应势而动。加拿大人工智能创业公司BlueDot宣称,早在2019年12月31日,他们就利用非官方和官方源消息,发出了新冠肺炎疫情暴发预警③。一些有识之士指出,我国应从战略高度加强应对全球公共卫生危机的情报能力建设④。钟南山院士也呼吁,我国需要建立一个预警潜在传染病的全球“哨兵”系统⑤。语言是信息的载体,全球网络空间的信息载体是多语的,多语监测既已成为互联网公共安全监测的底层思维,从语言介质而非地理空间入手,可以实现监测效益的最大化⑥。就我国来说,未来的监测系统应该跨越语言藩篱,能够处理多语多方言信息。病毒不分国界,未来的世界安全与全球治理需要更多的中国担当和中国方案,鉴于国内和国外两个大局,建设“互联网+多语”的公共卫生监测系统,应是势之所趋,有助于筑牢我国立体辐射的公共安全监测体系。

---

① 参见帕特里克·沃尔什《生物安全情报》,北京:金城出版社,2020年,第135-140页.

② Sulaiman, Feroza Binti, NK Semara Yanti, et al., “Language Specific Gaps in Identifying Early Epidemic Signals: A Case Study of the Malay Language”, *Global Biosecurity*, 2019(3), p. 235.

③ Collier, N., “Towards cross-lingual alerting for bursty epidemic events”, *Journal of Biomedical Semantics*, 2011(2). https://doi.org/10.1186/2041-1480-2-S5-S10。The Cambridge Public Policy Strategic Research Initiative., *The value of Languages: Ideas for a UK Strategy for Languages*, Cambridge University, 2015, p. 18. McCormick, J., How AI Spotted and Tracked the Coronal-virus Outbreak (2020-02-06) [2022-10-26], https://www.wsj.com/articles/how-ai-spotted-and-tracked-the-coronavirus-outbreak-11580985001.

④ 陈超:《疫情防控阻击战中的信息情报》,《竞争情报》2020年第1期,第1页。

⑤ 贾璇:《钟南山:新冠疫情有望在四月份结束》(2020-02-11)[2022-10-26], https://news.sina.cn/2020-02-12/detail-iimxyqvz2110004.d.html。

⑥ Gaël Lejeune, Hatmi M, Doucet A, et al., *A proposal for a multilingual epidemic surveillance system*, User Centric Media, Springer Berlin Heidelberg, 2009. https://doi.org/10.1007/978-3-642-12630-7_43.

三、公共疫情多语监测系统的价值定位

经过多年的实践总结,WHO 对于公共卫生安全监测形成了一套复杂的评估指标,包括及时性、敏感性、完整性、代表性、有效性、特异性、简单性、可接受性、灵活性等①。我国作为 WHO 的会员国,需在国际条约框架下设计监测系统,但是还须考虑具体国情和国家战略。就语言国情而言,我国还称不上是语言强国,国家语言能力还比较薄弱②。疫情以来,我国的大国能力、大国风范感召世人,我国正在加速实现从负责任的大国向负责任的强国转变。强国需要强语,强语辅助强国。着眼于国情和国家战略,“互联网+多语”公共卫生监测系统可以发挥三个方面的优势和特色,分别对标及时性、完整性、有效性。

(一)识别及时

疫情来临,时间就是生命。监测系统必须和时间赛跑,在早期预警阶段,抢在疫情流行之前就能及时识别传染病,将事态遏制在萌芽之中,在疫情成势阶段,也能及时跟踪传染病发展情况。

互联网是健康信息的集散地。全世界每天有数百万人在网上搜索、报告、交流与健康相关的信息,这使网络搜索查询和互动平台成为观测公共卫生趋势极有价值的信息来源。互联网往往也是有关卫生信息最早披露的地方。据估计,有关疫病暴发的初始报告中约有 65%来自网络新闻报道和非正式信息源③。随着自媒体的崛起,每个用户都会成为信息网络的节点。无可争议的是,在时效性、弥散性和动态性方面,网络媒体相比于传统媒体具有绝对优势。正是具有这样的优势,建立在互联网数据源基础上的多语监测就能在识别时间和监测过程上占得先机,尽早响应。

事实是最好的证明。2002 年 11 月,加拿大联合 WHO 研发的“全球公共卫生情报网”(GPHIN),最早发出我国南方发现“不明肺病”的预警报告,比我国官方通报世卫组织早了数月。2010 年,哈佛医学院研究人员利用 HealthMap 系统的数据挖掘平台成功估计海地霍乱疫情的发生情况,比当地卫生工作者的报告早两周④。同样,2014 年,该系统在世卫组织宣布埃博拉疫情 9 天前,就在地图上标记了几内亚的“神秘出血热”并发布警报。如前所述,这些系统正是基于“互联网+多语”监测思路实现的。相反,也有一些研究表明,缺少多语能力的网络监测系统,会延误报告事件⑤。

(二)覆盖广泛

疾病不问地域、不问人群。对于流行病监测来说,足够的地理覆盖范围同样至关重要,因为谁也不知道下次疫情是在何地、将以哪种语言首先披露。法国公共卫生研究所的研究显示:世界上将近一半的重大健康威胁事件,它的初始消息源是以英语之外的语种报道的⑥。所以,能够处理多语言文本的监测系统就非常必要。

中国现有 6 万多各级医疗基站,东部和西部、发达地区和贫困地区医疗资源很不平衡,直报比例差距很大⑦,这就给信息收集留下了盲区。人为期待西部不发达地区的医护人员会“主动”越过信息鸿沟是不现实的,相反,监测系统的制度顶层自身应该牢牢抓住主动权。

目前,我国网络使用的民族语言共有 14 种,国际互联网网页约有 350 种语言。理论上,为应对全球公共安全危机,多语监测系统应该覆盖所有网络语种,国际上也已提出“无国界监测”和“无数

---

① WHO, *Early detection, assessment and response to acute public health events: Implementation of Early Warning and Response with a focus on Event-Based Surveillance*, WHO, 2014, p. 32.

② 李宇明:《提升国家语言能力的若干思考》,《南开语言学刊》2011 年第 1 期,第 1 页。

③ Kawazoe A, Chanlekha H, Shigematsu M. et al., “Structuring an event ontology for disease outbreak detection”, *BMC Bioinformatics*, 2008, (9). https://doi.org/10.1186/1471-2105-9-S3-S8。

④ Aaron, Pressman., How A. I. is aiding the coronavirus fight(2020-3-16)[2022-10-26], https://fortune.com/2020/03/16/ai-coronavirus-health-technology-pandemic-prediction/。

⑤ Collier, N., “What's unusual in online disease outbreak news?”, *Journal of Biomedical Semantics*, 2010(2), p. 16.

⑥ The Cambridge Public Policy Strategic Research Initiative., *The Value of Languages: Ideas for a UK Strategy for Languages*, Cambridge University, 2015, p. 18.

⑦ 王婧、赵琦、赵根明:《传染病监测和预警系统研究进展》,《中国预防医学杂志》2010 年第 7 期,第 755 页。

字边界监测”的倡议[①]。但是,考虑到我国实情,借助较为国际通用的语言作为试点,“能监全监”将是可行的方案。我国正在努力提升国家语言能力,多语监测系统的研发也将带动我国国家语言能力的提升。

(三)简单有效

监测系统需要考虑用户的接受程度。此次疫情暴发后,不少医生直言根本不知道CDC的传染病监测系统,也有很多医生表示无暇或无力填报。“互联网+多语”监测系统,直接从网络大数据获取信息,结合自然语言处理和专业分析提取事件,无须依赖医护人员直报。该系统自动搜集、处理数据,可接受性高。

互联网监测,最大的争议在于信息来源驳杂,真假难辨。随着技术进步,数据降噪处理已经取得很大进步。相对于网络信息不纯问题来说,更加关键的是信息壁垒和数据孤岛问题。《国际卫生条例》对于疫情监测信息传播规范设置了公开、限制和保密三个层级,这就给关键信息无法及时公开共享预留了法律解释空间。发达国家和不发达国家、政府和民间在监测方面存在着信息鸿沟,如果缺乏有效的信息交流机制和对接平台,就会造成“信息孤岛”。为此,Bill Gates[②]呼吁,世界各国还需要在疾病监测方面建立分享信息的规则。

在攸关人民生命和国家利益的重大问题面前,我们对于疫病信息的掌握应该做到主动、有效。瘟疫是世界人民公敌,应该在“人类命运共同体”的格局下思考监控体系的研发。“互联网+多语”的监测系统正是着眼于人类命运相连,为突破信息壁垒打造公共产品,分步推进,有效地辐射更多的国家和人群,统筹国内外大局。

监测系统是保障国家和人民生命安全的屏障。就一套监测系统来说,识别及时是其“生命”,覆盖广泛是其“活力”,而简单有效可谓其“灵魂”。新型公共卫生多语监测系统当定位如此,并随着国家整体战略而不断优化。

四、公共疫情多语监测系统的实现路径

(一)基本路径

基于互联网大数据的公共卫生监测,各家系统研发手段和过程各不相同[③],但基本路径相似:(1)从互联网收集和存储数据;(2)处理这些数据以产生信息;(3)将这些信息汇总分析;(4)将分析结果传播给最终用户。我国建设“互联网+多语”卫生监测系统可以参照这个思路,如图1所示:

步骤1:确定数据源,从互联网中锁定与疫情相关的数据来源,包括社交媒体、网站新闻、疾病论坛、引擎搜索记录、众包新闻平台、RSS新闻提要列表等非官方媒介。为了过滤不相干信息和假新闻,需采集官方信息进行验证和补充。数据采集涵盖时空两个向量:时间上实现“7×24小时”全天候监测,每周形成一类数据集;空间上尽可能覆盖全球范围。这就需要确定“多语”监测的品种。目前,全球互联网语言使用人数在0.1%以上的为40种[④]。另一方面,世界上仍有许多重要的区域通用语和跨境语不在这40种之内,如马来语、菲律宾语、斯瓦希里语、豪萨语、蒙古语等。这些语言有的是东盟和“一带一路”沿线国家工作语言,有的则处在传染病多发且医疗资源贫乏地区。根据公共卫生监测的需要,它们不应该被忽视,另外还要考虑我国语言能力现状。这次防疫中发挥积极作用的“疫情防控外语通”就是我国语言能力的一次成功展示,他们根据急用、广用、宜用三原则选择了41种外

① Brownstein S, Freifeld C, Reis Y, et al., “Surveillance Sans Frontières: Internet-based emerging infectious disease intelligence and the HealthMap project”, *PLoS Med*, 2008, (7), pp. 1019-1024. Davies S., Reporting Disease Outbreaks in a World with No Digital Borders, Colin McInnes, Kelley Lee, and Jeremy Youde (eds.) *The Oxford Handbook of Global Health Politics*, Oxford University Press, 2020, pp. 513-530.

② Bill Gates. “Responding to Covid-19: A Once-in-a-Century Pandemic?”, *The New England Journal of Medicine*, 2020(382), p. 1678.

③ Hartley M, Nelson P, Arthur R, et al. “An overview of Internet bio-surveillance”, *Clinical Microbiology & Infection*, 2013(11), pp. 1006-1013.

④ W3Techs., Usage statistics of content languages for websites (2019-12-25) [2022-10-26], https://w3techs.com/technologies/overview/content_language.

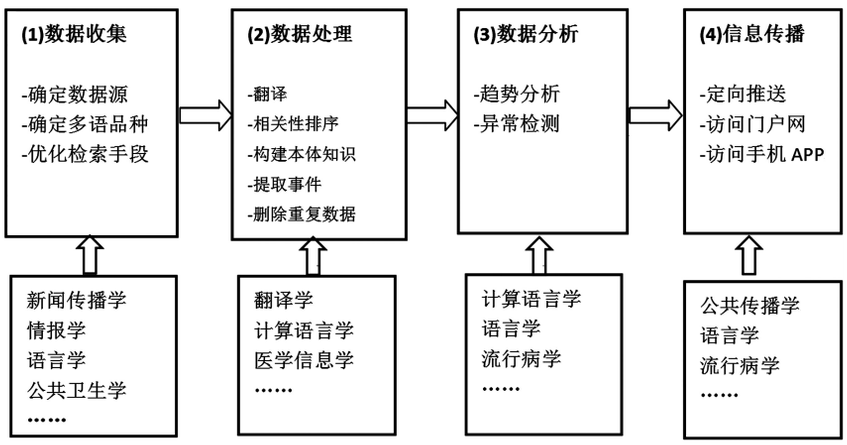

图 1　“互联网+多语”公共卫生监测系统建设路线图

语,基本做到了全球覆盖①。在他们的基础上,我们可先将“多语”范围定在 45 种以内②,将来再根据条件逐步拓展。

步骤 2:从互联网抓取数据后,须对其进行处理使其适合分析。翻译是把当地语言报道的疫情信息翻译成各种语言版本,或者先翻译成区域通用语言,再转换成各种语言版本。相关性排序,是根据用户兴趣评估、排列报道跟疫病事件的相关程度。构建本体知识,是形成机器可读的公共卫生领域概念知识体系③,例如“病源、病原体、分型、人传人、宿主”等,以对出现在新闻报道中的术语做出正确的判断。抽取事件,是根据预定义的关键词从分类文本中抽取出结构化的卫生事件。抽取的过程主要是采用主题或话题分析,如话题自动识别、话题跟踪、事件识别处理,如识别事件发生的几要素——时间、地点、人物、结果等。

步骤 3:数据分析环节,利用自然语言处理技术进行文本的情感倾向性和评价分析,目的是及时推断出哪些事件更为紧急或异常,以便系统用户可以进一步调查,并有可能启动疫情风险评估,评估分析后形成推荐性或结论性的意见、报告。

步骤 4:根据用户需求和报告信息的性质,使用不同的信息传播手段。对于预警、防范信息可用电子邮件、短信推送,对于特定信息用户也可以登录网站界面或 APP 查询(例如疾病类别、感染人数、传播方式、发展趋势、疾病地图等)。信息的查询应该支持多语模式。

建设监测系统的路径流程可简化为信息的收集、处理、分析和传播。

(二)学科支撑

如图 1 所示,建设“互联网+多语”卫生监测系统是个跨学科工程,需要多个学科及其人力资源的互相配合,尤其是需要语言学科进行知识供给。对于“互联网+多语”监测来说,面对海量的大数据,核心问题是要提高监测的准确度和有效性。自然语言处理在基于规则和经验相结合的思路下,至少需要解决的问题有:

第 1 环节,谣言识别问题。当下最流行的阿里 AI 谣言粉碎机,识别成功率达到 81%。其识别原理之一是,将正文里关键的论证提炼为知识点,与知识图谱里的权威知识库做匹配验证,毫无联系、

① 参见刘晓海、田列鹏《应急语言服务领域的语言资源建设与应用——以〈疫情防控外语通〉研发为例》,《云南师范大学学报(对外汉语教学与研究版)》2020 年第 4 期,第 17-25 页。

② 《疫情防控外语通》41 种外语是:阿尔巴尼亚语、阿拉伯语、阿姆哈拉语、阿塞拜疆语、白俄罗斯语、保加利亚语、波斯语、德语、俄语、法语、菲律宾语、芬兰语、哈萨克语、韩语/朝鲜语、豪萨语、吉尔吉斯语、加泰罗尼亚语、柬埔寨语、捷克语、老挝语、罗马尼亚语、马来语、蒙古语、缅甸语、葡萄牙语、日语、塞尔维亚语、斯瓦希里语、泰语、土耳其语、乌尔都语、乌克兰语、乌兹别克语、西班牙语、希腊语、匈牙利语、意大利语、印地语、印尼语、英语、越南语。我们再增加 4 种:汉语、维吾尔语、藏语、壮语。这 4 种语言是我国使用人口最多的语种,而且跨境使用,覆盖地域和人群最广。

③ Collier N, Kawazoe A, Jin L, et al. , “A multilingual ontology for infectious disease surveillance: rationale, design and challenges”, *Language Resources and Evaluation*, 2006(3), pp. 405-413.

自相矛盾,则减分。如果谣言识别的建模能够考虑更多的语言参量,如谣言的标题特征、用词规律、话语结构、写作风格、生成机制和传播路径等,识别成功率应该还能提高。在这方面,语言学研究应该有所作为。例如有研究基于系统功能语言学理论框架,对新冠疫情期间虚假谣言下的用户评论进行立场分类,构建出用户立场检测分析模型,从而提高谣言传播中的用户评论立场识别准确性,为公共卫生事件的谣言检测补充新思路①。

第2环节,主要是翻译和构建本体知识。翻译的语种类型情况需要翻译学和人类语言学知识。构建本体知识,就是形成人类知识总体中一个有关流行病的子类语言资源。这一环节需要语言学的深度参与。其中,构建本体知识需要多语种的流行病术语表。

不同语言的疾病术语存在差异,需要建成对照表,以便执行关键词搜索提取事件。例如在马来地区,某些疾病(如腮腺炎、登革热和禽流感)的当地术语就和英文术语不同②。各华语区疾病的科学术语、日常名称也有差异。例如,SARS在不同华语区就有非典、沙司、沙斯、萨斯等名称,AIDS就有艾滋病、爱之病、爱滋病等名称。据统计,海峡两岸约有2000个医学术语不一致;新加坡华语对于疾病的日常称呼和普通话有31.5%的差别③。近年来,几种大型辞书《全球华语大辞典》《两岸科技名词词典》相继推出,但还缺乏医学术语对照词典,需要继续进行专门的华语词汇整理研究和协调④。

第3环节,主要是异常疾病事件的识别和检测。异常检测需要将事件的特征置于时空当中,以便确定其重要程度。其工作原理类似于舆情监测。疾病事件新闻通常含有几个要素:时间、地点、人物、症状、传染源、原因以及事件评价等。文本挖掘语义在事件实体上已经取得不俗的成就,但是情感语义分析还是一大瓶颈。

现阶段情感分析主要有基于情感词典和基于机器学习两种方法,还缺乏专用的疾病领域的情感词表。研究表明,不同语言之间在情感词方面存在差异,其总体数量、语义概念等并不一致⑤。华语变体的情感词也有差异,例如普通话的"窝心"在其他华语区反而是"温馨"的语义。因此,在这一环节,还要考虑多语、多变体的情感表达差异,多语言文本语码转换的功能识别,语言歧视的跨语言识解,增加语境和社区参数来解决语言歧义⑥、在线语言交际过程和结构研究等问题。

第4环节,主要是信息发布的语种和语用策略。语言学家可以建议,就形成的重要信息,根据受众需要和安全等级,选用合适的语种、语体风格、交际策略进行发布,以达到最优的传播效果。多语言信息发布不仅包括普通话、方言、民族语和多种外语,还应包括面向一般大众的简易语言和残障人士的特殊语言,以实现信息沟通无障碍。实践表明,疫情来临之时,采用多模态话语形式和多样化话语主体的交际策略,有助于建构积极的政府形象⑦。另外,还应建设历次疾病暴发事件的多语平行语料库,以备训练机器深度学习,为未来进行更准确的公共卫生监测服务。

防疫是场战争。在这场突发战疫中,从最初的"新冠肺炎"术语定名到成立"战役语言服务团",实施医疗语言救援、制作疫情防控语言产品等应急语言实践和研究方面,语言学者没有掉队⑧。然而,战必"动戈",我国自古就有"上战伐谋,其次伐交"的思想,为了谋求国家安全和人民福祉,最好是永远没有防疫战。公共卫生监测系统正是为了"永远没有防疫战",建设这样一个监测系统,需要多学科携起手来,就提供知识经验而言,语言学应该且也当然可以担负使命。

---

① 王丹丹、杨艳妮、张瑞:《系统功能语言学理论视角下突发公共卫生事件谣言用户立场识别研究——以COVID-19疫情为例》,《现代情报》2021年第2期,第19页。

② Sulaiman, Feroza Binti, NK Semara Yanti, et al., "Language Specific Gaps in Identifying Early Epidemic Signals: A Case Study of the Malay Language", *Global Biosecurity*, 2019, (3), pp. 235-247.

③ 马学博:《海峡两岸医学术语的差异》,《医学情报工作》1995年第5期,第38页;骆珍凤:《新加坡华人医患交际用语初探》,罗福腾主编:《新加坡华语应用研究新进展》,新加坡:新跃大学新跃中华学术中心、八方文化创作室,2012年,第14页。

④ 参见郭熙《域内外汉语协调问题刍议》,《语言文字应用》2002年第3期,第33-39页。

⑤ Pavlenko A., *Emotions and Multilingualism*, Cambridge University Press, 2005, pp. 86-92.

⑥ 徐大明:《语言学理论对自然语言处理的影响和作用》,《云南师范大学学报(哲学社会科学版)》2017年第3期,第1-9页。

⑦ 张卉、陈新仁、张秀芹:《突发公共卫生事件中的政府形象建构——以政务微信"南京发布"为例》,《浙江外国语学院学报》2021年第3期,第55页。

⑧ 参见李宇明、赵世举、赫琳《战役语言服务团的实践与思考》,《语言战略研究》2020年第3期,第23-30页;汲传波、李宇明:《〈疫情防控"简明汉语"〉的研制及其若干思考》,《世界汉语教学》2020年第3期,第311-322页。

五、公共疫情多语监测系统的应用前景

“互联网+多语”公共卫生监测系统有着广阔的应用前景。未来,我国在食品安全、药物警戒、社会突发事件等非传统安全领域面临的风险依然很大①,需要建成更加全面的公共安全监测体系。在这个体系中,接入“互联网+多语”的监测系统,可以提高现代社会治理的效能。例如,一些监测系统通过监视社交网络上的帖子,利用已有的自杀文本数据库和自杀风险预测器来识别、跟踪具有自杀意图的语言标记(有自杀倾向的人更多使用第一人称代词、愤怒的情感词和现在时态),从而尽早干预异常行为。也有研究利用互联网数据挖掘技术,分析香港占中运动暴乱分子在社交网上的语言情感标签,从而预测有关人员的行为倾向②。语言学和互联网大数据、行为科学相结合,正在形成一门新的学问——预测语言学(Predictive Linguistics)。

“互联网+多语”公共卫生监测,不但是国家语言战略,也可以成为一项可持续性的语言产业。新冠疫情发生以来,众多互联网科技公司闻风而动,纷纷研发各类相关的大数据监测系统,及时监测人员流动、防疫资源需求等信息,不仅为防疫提供决策,也为社会有序配置资源,恢复有序生产生活发挥作用。开发、维护这些大数据系统少不了语言学的作用,美国几大公共安全监测系统催生了许多语言分析师③。本次疫情应急语言服务中采用了语料库、音频/文本检索、机器翻译和机器辅助翻译、文本分析等多种技术④,着眼于未来,语言技术和语言资源驱动的公共安全监测,当会进一步带动语言产业的发展。

疫情就是敌人,信息监测系统就是防御敌人的“雷达”。未来疫情仍有可能反扑,面对尚未充分认知的病毒,我们不能完全依赖现有的垂直直报系统,而是需要在公共卫生安全预防体系上“补短板、堵漏洞、强弱项”。习近平总书记指出:“必须加快形成从下到上早发现、早预警、早应对的体系,努力把疫情控制在萌芽状态。要把增强早期监测预警能力作为健全公共卫生体系的重中之重,完善公共卫生应急管理体系。”⑤建设“互联网+多语”公共卫生监测系统,可以作为完善公共卫生应急管理体系的一项基础工程。

(责任编辑　刘　英)

① 参见陈坤《公共卫生安全》,杭州:浙江大学出版社, 2007 年,第 57-73 页。

② O'Dea B, Larsen E, Batterham P, et al. , “A linguistic analysis of suicide-related Twitter posts”, Crisis: *The Journal of Crisis.* Intervention and Suicide Prevention, 2017, (5), pp. 319-329. Carmen Leea, Dennis Chau. , “Language as pride, love, and hate: Archiving emotions through multilingual Instagram hashtags” , *Discourse, Context & Media*, 2018, (22) , pp. 21-29.

③ Nkuchia M, Ruth Lynfield, Chris Van Beneden. , Infectious Disease Surveillance, Blackwell Publishing, 2008, p. 308.

④ 饶高琦:《战疫语言服务中的语言技术》,《云南师范大学学报(对外汉语教学版)》2020 年第 4 期,第 26 页。

⑤ 习近平:《国家中长期经济社会发展战略若干重大问题》,《求是》2020 年第 21 期,第 10 页。