大模型发展提速 中文语料够“吃”吗

深瞳工作室出品
采写：本报记者 龚 茜
策划：何 屹 房琳琳

继去年“百模大战”之后，今年国内大模型产业应用进入爆发元年。

然而，大模型产业发展如火如荼的同时，其训练数据规模的增长速度跟不上、语料质量参差不齐，尤其是高质量中文语料短缺的问题日益凸显，成为各方关注焦点。

阿里研究院5月发布的《大模型训练数据白皮书》（以下简称《白皮书》）显示，互联网上中文语料和英文语料占比存在显著差异：在全球网站中，英文占比高达59.8%，而中文仅占1.3%。

同样，语料的质量会显著影响大模型的性能。在大模型领域，输入低质量数据，必然会输出低质量结果。

在近日举办的第六届北京智源大会上，中国互联网协会理事长尚冰指出，高质量数据的生成速度远低于AI大模型训练数据需求量的增长速度，数据短缺问题已初现端倪。

如何获取规模化高质量中文数据？建设高质量中文数据集的难点和堵点是什么？加速数据流通，推动中国特色大模型创新发展与应用的意义何在？对此，科技日报记者进行了采访。

高质量中文语料供给严重匮乏

语料即大模型训练所需数据，是大模型训练的基础，也是决定大模型性能和专业性的关键因素。商汤科技大装置事业群高级总监张行程告诉记者，中文高质量语料相对缺乏是国内外大模型面临的共同问题。中文语料库不仅规模较小，且其电子化和网络化程度明显不足。此外，受版权、隐私等限制，许多优质中文语料库也无法公开获取。

其中，有一类型的中文语料极为重要，但又非常短缺——中式价值观类语料。《白皮书》主要编写成员、阿里研究院数据经济研究中心副主任王峥解释说，为了更好理解客观世界和掌握客观规律，大模型需要学习大量知识和价值观层面的数据，这些数据深受人类主观意志的影响。

在王峥看来，文言文、古汉语、电子书等反映优秀传统文化的内容，以及主流媒体发布的反映本土价值观的内容，都可视为具有中式价值观的高质量语料。

“训练中融入更多这类中式价值观语料，有助于大模型深入理解和反映中文使用者的文化背景和价值取向，从而在全球化背景下保持中国文化的独特性。”王峥说，“更重要的是，能更好地服务中国本土用户，满足行业发展的需要。”

但目前面临的实际困难是，这类语料开放共享与开发利用的程度远远不够，且无法通过机器翻译弥补其短缺问题。《白皮书》指出，中文语料量的短缺尚有可解决方案，但中式价值观类语料的短缺，则会成为制约我国大模型发展的短板。

高质量中文语料的供给是中国大模型本土化的关键。“我们希望行业能加强企业间合作以及产业上下游协同，共同推动高质量中文数据集的共享、开放，鼓励数据提供方将高质量中文语料库在一定范围内公开，为各行各业大模型技术创新和应用奠定坚实的基础，形成中国特色的AI大模型创新路径，不断提高国际竞争力。”张行程说。

供需双方合作机制尚待完善

一方面，大模型厂商需要高质量数据支撑，以解“巧妇难为无米之炊”的困境；另一方面，高质量中文语料库的数据拥有者，如拥有各类图书、文献的出版商等，也期望在智能化时代实现数据增值。因此，探索数据供需双方合作模式是关键。

然而，要推动数据供需双方建立合作并非易事。“拦路虎”到底是什么？

当前，大模型数据获取主要有合理爬取、版权采购等途径。

张行程透露，商汤目前的解决方案是联合各机构尽量挖取、寻找现存的中文高质量语料，比如精心编校过的书本、论文等，以及向供应商购买版权语料。“虽然购买数量有限，但质量很高。”张行程说。这是以前置协商付费方式来获取版权类语料的传统商业模式。阿里巴巴“通义千问”大模型也采取了类似做法。

王铮还提到第二种潜在的方式，即与版权方协商，以训练后的模型为版权方提供服务的方式进行对价。

然而，关于版权类语料使用，数据提供者和大模型厂商持有不同见解。王峥认为，大模型对版权类训练语料的使用属于转换性使用，而非复制式拷贝，应构成“合理使用”或“法定许可”。

上海世纪出版集团数字出版部副主任刘寅春对此持有保留意见。她指出，大模型的深度学习机制与人类学习有相似之处，使用版权类数据进行训练，类似于人类阅读文献后撰写论文而不标注参考文献。“从学术规范上来说，这种做法很难说没有瑕疵。”她说。

此外，大模型厂商训练大模型的最终目的是商用，这与“合理使用”的初衷和前提并不相符。“法定许可”需要满足一定条件，包括说明作品的出处、作者姓名，并支付报酬。如果这些条件无法满足，那么在显性法律释义下，这种行为很难构成“法定许可”。

在人工智能时代，高质量数据集是出版行业的核心资产。刘寅春认为，在有利于行业健康、可持续发展的前提下，切实保障知识产权，对高质量数据集进行有效开发和高质量转化，是出版行业的核心。

“出版行业为大模型提供语料，相应地，大模型的技术进步、功能提升，也应惠及包括出版行业在内的更广泛群体。”刘寅春提倡以合作共赢的方式与大模型厂商开展数据交易，通过订立授权协议，明确授权范围和条件，实现共同发展。

“如何将出版物进一步加工为数据要素并有效、有序流通，是摆在出版人面前的新问题。”中国出版传媒股份有限公司副总经理张纪臣说，“但我认为这同样是新机遇，因为我国出版行业一直强调知识服务这一理念。将出版物作为语料使用，从而提供产品和服务能力，正是出版知识服务的产品化体现。”

数据开源分享动力不足

目前，我国可供大模型训练的优质数据资源呈碎片化、分散状态。

“特别是语料和科研成果等中文高质量数据集开放程度低，企业在训练大模型时使用的语料来源不透明、权属不明确，开源后存在合规风险，这导致企业更倾向于自行采集和使用数据，大模型数据流通机制尚未形成。”王峥说。

北京理工大学管理学院副研究员尹西明认为，需要构建一个市场化、互利共赢的数据共享机制，以促进高质量中文数据的积累和有效利用。

“确立清晰的数据要素市场制度对于激发高质量数据集构建至关重要。”在复旦大学教授、上海市数据科学重点实验室主任肖仰华看来，只有当市场机制能够确保数据贡献者获得合理回报时，才能吸引更多的数据流入市场，充分挖掘并实现数据共享的巨大潜力与价值。

2023年12月31日，国家数据局等部门印发《“数据要素×”三年行动计划（2024—2026年）》，强调坚持需求牵引、注重实效，试点先行、重点突破，有效市场、有为政府，开放融合、安全有序4方面基本原则。

该行动计划进一步明确，要提升数据供给水平，在科研、文化、交通运输等领域，推动科研机构、龙头企业等开展行业共性数据资源库建设，打造高质量人工智能大模型训练数据集。

事实上，作为数据流通领域中最大的“富矿”，公共数据开放的步伐正不断加快。《全国数据资源调查报告》显示，2023年，我国公共数据开放量同比增长16%；省一级政府的开放数据量同比增长了18.5%，北京、浙江等15地数据管理部门开始探索公共数据授权运营机制。

今年初开始实施“数据入表”政策。张纪臣认为，随着“数据入表”政策的实施，出版企业的数字资源经过确权、评估、标准化后入表，成为出版企业的数据资产。在此基础上构建大模型训练使用方与出版企业共赢的商业模式，能发挥中国价值核心数据在人工智能时代的智能服务话语权。“这样一来，‘数据入表’可能成为加速数据有效流动、共享并实现共赢的关键一步。”他说。

数据流通环节问题突出

算法、算力、数据和场景是大模型发展的4个核心要素。当前，我国大模型算力算法能力显著提升，高质量发展取决于数据和场景，应构建“供得出、流得通、用得好”的高质量数据集。

尹西明表示，大模型变强用好，前提是建立以场景驱动创新的思维，引领高质量数据持续在各种应用场景中发挥价值。那么，解决数据“供得出”难题后，应重点确保高质量数据“流得通”，真正面向场景释放数据乘数效应和大模型对新质生产力的引擎价值。

数据要素在生产中的地位愈发重要，数据要素流动带来的开放性与动态性问题，为传统数据理论与相应技术带来新挑战和新要求。

“其中之一便是数据确权。”肖仰华表示，相比其他生产要素，数据要素在流通过程中主体更加多样，涉及数据生产者、采集者、加工者、使用者、运营者和其他产权人，权属界定复杂。

北京智源人工智能研究院理事长、中国互联网协会人工智能工作委员会主任委员黄铁军指出：“当前普遍存在一种误解，即将数据视为传统意义上的物理资产。其实，数据并非物理资产，作为数字形态产品，它可以被无限次使用，且不会导致数据损耗。”

他提倡在确保使用合规的前提下，大模型训练阶段可以免费获取数据资源。如果使用数据的过程中并未产生商业利益，则无需支付任何费用；反之，一旦通过数据使用获得了商业收益，便应按照既定比例支付相应的数据使用费用。

“虽然这一模式背后还涉及到数据确权、费率设定、监管机制等复杂问题，这些还有待深入探讨和解决，但‘先使用后收益’更有利于大模型的健康发展。”黄铁军说。

王峥则认为，确保数据流通需政府与企业、开源或非盈利组织、学界、多类型机构等社会力量协同推进。

他建议，在政府侧，对可用于模型训练的公共数据鼓励“应开尽开”，避免在数据开放过程中因为预设特定场景限制了应用范围；在社会力量侧，应坚持“应试尽试”原则，通过不断迭代，探索数据的有效搭配，寻找发挥最大价值的“配方”。

标注专业化、规模化提上日程

从2022年《关于构建数据基础制度更好发挥数据要素作用的意见》出台以来，数据要素建设和市场改革正稳步推进。今年5月，国家数据局提出建设国家级数据标注基地，这一举措对人工智能发展至关重要。

中国信息通信研究院人工智能研究所高级工程师、中国人工智能产业发展联盟数据委员会主任李荪表示，数据标注是推动人工智能进步的核心环节，它能够提升数据质量，挖掘数据核心价值，形成高质量数据集，持续为AI提供数据支持。

也就是说，在一定程度解决数据供给、促进数据共享和打通流通机制后，如何让大模型学习到高质量数据，是接下来各界面临的另一个新挑战。

数据标注的专业性和规模化也被提上日程。

李荪指出，当前国内数据标注产业还比较初级，大部分标注工作以人工为主，劳动密集型特点比较突出。但是，在通用人工智能时代，传统手工标注或简单自动化标注方法无法满足大模型对大规模、高质量、多样化数据的需求，特别是具备模型训练知识、行业领域知识的专业化数据标注人才也相对匮乏。

“大模型训练数据标注人员的学历要求比以前更高，很多是本科毕业。”王峥表示，行业大模型数据标注凸显了专业知识的重要性。

机器在对语言水平这一抽象概念进行评估时，必须依赖预先设定的人类价值判断和标准。电子科技大学智能语言学习与测评实验室与字节跳动合作开发了一款语言水平考试产品。实验室负责人陈大建说，在研发阶段，实验室负责对自行收集的用户音频数据进行标注，其标注内容主要是基于音频所体现的英语能力水平进行分类和标记。标注人员由学校四五十名大学英语教师组成，且均为应用语言学专业的硕博士。

“吃”得好、“吃”得香，还要“吃”得够。只有最终实现了规模化高质量标注，才能切实提升大模型理解中文、传递中国传统文化价值的能力。中国大模型的蓬勃发展也将助力中华优秀传统文化海外传播，架起一座连接古今、沟通中外的桥梁。

中国出版集团中国图书进出口总公司下属中图科信数智技术（北京）有限公司总经理李沄汎认为，从正式出版物如文献、学术专著等入手，依托先进的提取工具和解析技术，将出版物语料化、碎片化、标准化，加工成高质量的语料数据，有别于一般的数据加工。“我们已经实现了大规模和批量开展数据语料化的技术和工具软件，能够更深层次地解析数据，并形成独立的图片、表格、公式数据集，为大模型人工智能服务提供价值更高、标准程度更好的语料供给，使出版数据在人工智能时代焕发出新活力。”他说。

在数字经济大潮中，数据要素的放大、叠加、倍增作用日益显著，成为推动相关产业高质量发展的必然要求。张纪臣认为：“我们正站在新一轮产业科技革命的门口。这是一个不进则退的时代。”

第89届中国国际医疗器械博览会上，观众体验联影智能MR脑功能智能分析系统。

图① AI大模型数字客户系统——AI虚拟机器人。

图② 第二届全球数字贸易博览会前沿趋势馆内，人工智能大模型同场竞技。

图③ 基于Stable Diffusion（AI绘画生成工具）的框架模型开发的软件。

本版图片由视觉中国供图