应对不确定和不连续

奇幻、玄幻、科幻,你喜欢读哪一种?为什么不少企业家推崇《三体》?

周末读北京大学管理学教授路江涌的新书《科幻中的战略思维》,他写道:科幻小说更注重用科学理论解释虚构事物的合理性。其中以描写新技术、新发明给人类社会带来影响的科幻作品被称为“硬科幻”,以哲学、心理学等人文学科为主的科幻文学被称为“软科幻”。无论软硬,都要求具备“逻辑自洽”“科学元素”“人文思考”三要素。

科幻大师海因莱因曾提出一个区分科幻和奇幻/玄幻小说的方法,他说:科幻小说的作者要能够解释,这件事是怎么发生的,你应该怎么做,而奇幻/玄幻小说的作者不需要解释。

中国的科幻小说早在20世纪初就有几次小高潮。1902年,梁启超写了本书叫《新中国未来记》,后来鲁迅也翻译了凡尔纳的科幻作品《从地球到月球》,等等。

《科幻中的战略思维》书中有一些深入的思考和洞察,比如:科幻小说里最常见的挑战是不确定性和不连续性,反观人类文明史,又何尝不是一部“在不确定性中寻找确定性,在不连续性中创造连续性”的历史呢?

本期读+采访和讨论建设中文数据语料库,中国的人工智能要解决中国的问题,AI大模型要说中国话,同样也体现了应对发展中的不确定性和不连续性的战略思考。

王永芳

要让AI“讲中国话”,必须建好中文语料库

——专访智源人工智能研究院副院长兼总工程师林咏华

□长江日报记者李煦

访谈

中国的人工智能,理所当然、天经地义地要“讲中国话”。从语言学层面说,它应该理解中国话的丰富、微妙和深刻,不应该再闹出“胸有成竹”就是胸前长出几棵竹子的笑话;而从价值观层面说,中国的人工智能应该为中国人思考、为中国人服务,这才叫作“讲中国话”。

怎样才能做到这一点?这就要靠语料库了。语料库是人工智能的粮食和养料,两台一模一样的电脑,使用一样的算法,但是用不同的语料库“喂养”,结果必然是大相径庭。

上周,长江日报《读+》周刊专访了北京智源人工智能研究院副院长兼总工程师林咏华,智源研究院由科技部和北京市支持,依托北大、清华、中国科学院、百度、小米、字节跳动、美团点评、旷视科技等北京人工智能领域优势单位共建,前不久发布了中文互联网语料库CCI3.0(Chinese Corpora Internet,简称CCI)。

林咏华

从“胸有成竹”“麻婆豆腐”的尴尬说起

说起中国生成式人工智能的发展历程,人们不该忘记那尴尬、遗憾的一幕。

在GPT第一波大潮拍岸而来的时候,一批中国GPT也随之起舞。人们振奋欣喜之余,好奇地上手体验,不料出了一些笑话:输入“麻婆豆腐”,生成的图片是一位有麻子的老太太和一碗豆腐;输入“胸有成竹”,生成的图片是胸前长出几棵竹子。

人们又用“夫妻肺片”“车水马龙”去测试,结果还是笑话。

好在,哄笑之余,自嘲之余,该做事的还是要做事,该调整的自然也要调整。中国人工智能是走在正确的道路上,大方向没问题,最初几步虽然走得还踉踉跄跄,但确实是在追赶世界先进水平。

承认技不如人,这不丢人。更重要的是,有识之士看到了差距,开始奋起直追。人工智能三大关键:数据、算法、算力,每一项都不能有短板。而“胸有成竹”“麻婆豆腐”等问题,正是源于数据,也就是语料库问题。

AI大语言模型,是由具有许多参数(通常数十亿个或更多)的人工神经网络组成。其基础是人类已经产生的各类知识产品,包括各类公开发表的或在网上出现的报纸文章、著作,以及各类网文、帖子等。国外大模型能先领风骚,主要依赖于英文语料库;如果中国的大模型AI发展要获得突破,必然要依赖于中文。全球目前最有科学性和经过验证的语料来自学术资料库,包括期刊等出版物,但遗憾的是,在这些载体上发表文章的语言绝大部分是英语。

2023年11月29日,智源研究院联合拓尔思等单位共建的“中文互联网语料库”CCI1.0正式发布,规模为104GB,中文互联网语料库是在中国网络空间安全协会人工智能安全治理专业委员会数据集工作组、北京市委网信办、北京市科委、中关村管委会、海淀区政府的鼎力支持下构建的,旨在为国内大数据及人工智能行业提供一个安全、可靠的语料资源,共同推动大数据和人工智能领域的健康发展。

今年4月,迭代到CCI2.0,规模约500G,涵盖1.25亿个网页;今年9月,发布3.0,规模1000G,收录超过2.68亿个网页,涵盖新闻、社交媒体、博客等多个领域。

“弱智吧”是优质数据吗?

1000G,似乎不算多啊;现在一部蓝光电影,可能就有几十G;普通的办公电脑,硬盘也有1000G。那么,1000G的语料库,可以涵盖多少数据呢?

《四库全书》8亿字,换算成电脑字节,还不到200M;五部《四库全书》加在一起,也才1G。一个文本语料库,不需要保留格式、图片,那么1000G已经是相当惊人了。

在语料库建设热潮中,还出现过一个有趣的说法,被誉为“最大原创段子手孵化地”的百度贴吧“弱智吧”,竟成了“最佳中文AI训练语料库”。

据说,由中国科学院自动化研究所和众多高校组成的联合研究团队,前往知乎、小红书、豆瓣、百度贴吧等社交平台取材,以提升大模型对中文指令的理解和响应能力。其中,“Ruozhiba(弱智吧)”数据集在头脑风暴、分类、生成、总结等八个评测项目中夺得最高分,且总评分稳居第一。

“弱智吧”有各种段子、双关、谐音、反转、幽默,比如:“老头哭了是因为老头乐坏了”“刘备被称赞爱民如子,将围观群众挨个摔倒在地”“每天吃一粒感冒药,还会感冒吗?”“我买了一斤藕,为什么半斤都是空的?”

有专家认为,“弱智吧中的帖子有不少脑筋急转弯和双关语,包含更复杂的逻辑,并且表达简洁干练,数据较为干净,属于一种高质量数据。用这样的数据训练出的大模型,在部分场景的确会让人感觉逻辑推理能力更强。但在解决实际问题时,往往需要更广泛地覆盖不同场景和类型的数据”。

一个大模型,如果听不懂这些俏皮话,人们会觉得它不够聪明;但是,要是一个家伙开口闭口都是“抖机灵”,那未免过于油腻,而且要这么个家伙有何用呢?现阶段人们使用人工智能,主要还是希望从中得到工作、学习上的助益。

业界共识是:更为严谨的并获得反复验证的百科全书式知识信息,才是大语言模型最为可靠的语料库。

林咏华在接受长江日报记者采访时也表示,智源CCI不排除段子,但是没有刻意去收录“弱智吧”的数据。

语料库是人工智能的“压舱石”

读+:在智源的网站上,将语料库称为人工智能的“压舱石”,为什么这样说?

林咏华:人工智能大模型是“三驾马车”,数据、算法和算力,数据是其中很重要的一块。从算法来看,现在大家有各种各样的变化,但是最终模型的性能还是跟这个模型学到了什么样的语料很相关。

就如同小孩子的成长,和他吸收了什么样的知识,这些知识的质量高低以及全面性、认知正确性有很大关系。其实这些都是外界给孩子的输入,包括课本、老师的讲解,以及他听到看到的所有,等等。跟这个类比,实际上数据就扮演了人类成长过程中外部给他的知识,这个道理几乎是一样的。

所以说语料库是人工智能的“压舱石”,这是从第一个维度“知识输入”来讲的。

第二个维度是,咱们在学习一个知识的时候,它正确与否也很重要。老百姓去查一个医疗问题,查到的到底是民科的解释、没有经过核实的解释,还是专业的解释?所以对于同一个知识点,它的来源也是很重要。这就涉及语料库的质量和正确性,它决定了大模型通过训练之后,能不能够生成正确的判断、正确的认知。

第三个维度是价值观的输入。在这么多的数据语料里头,它会有一些不同的价值观,这些不同的价值观如果输入到大模型里头,大模型就会直接去学,没有人会告诉它今天学到的价值观是正确还是错误。因此我们一开始输入的语料,最好就是一个正确的、合适的世界观、价值观,从这个意义来说,语料库就是人工智能的“压舱石”。

中文语料库会收录《四库全书》吗

读+:2023年初发生的“麻婆豆腐”“胸有成竹”等问题,是因为当时用的是英文语料库吗?

林咏华:这是一个“文生图”问题,属于图文跨模态。在训练这种“文生图”模型的时候,就需要海量的图片,再加上对这个图的一个描述。当时很多的模型,它的数据是来源于一个叫LAION的数据集,这是全球最大的一个数据集,人家把五十亿个图文对的数据开源出来,由国际上一个非营利组织在2021年开始构建,对全球的图文模型起到了很重要的作用。如果当初没有50亿条训练数据作为启动的基础,大模型不可能有今天这么强的“文生图”能力,但是我们发现,这个数据集里超过95%的图文对是英文的,中文的图文对在当时很少,并且它的标注质量很低。

以咱们的中国画为例,我看到有各种各样的牡丹,但它的标注就只有短短的几个字,中国画。至于这里画的是一棵中国的牡丹,包括它是什么颜色、多少头,这种很细的细节,它都缺失了,所以这就导致了在2023年之前,高质量的中文描述图片语料就特别少。

但是大模型有很强的跨语言能力,训练语料里头没有中文,但它会自动把英文的意思给转成中文的描述。面对“麻婆豆腐”,可能在它的训练数据里头没有四川那盘麻婆豆腐的图片,但是它会用跨语言能力去把麻婆豆腐进行解读,然后在英文的图片里面去寻找英文麻婆豆腐匹配的图片,于是它就会出现长了麻子的老太太和一盘豆腐。

所以在2023年年初,国内第一批中文大模型上线,大家进行各种各样的测试,发现它不能够很好去理解“中国话”,智源研究院当时比较紧急地启动了一个公益项目,连同国内几十家厂商一起去发动构建中文互联网语料库,我们都把我们手上认为有价值的中文的数据放在一起。

读+:现在这个语料库已经迭代到3.0,在世界上是什么水平?

林咏华:CCI3.0已经成为最大的全球高质量中文数据集,有1TB的数据,应该说是初步解决了高质量中文语料缺失的问题。但是我们觉得还不是足够。今年上半年国外发布了Llama3语言模型,它用的数据量已经是十几TB了。我们觉得缺失的,是一些高质量的、大量的文献数据,因为这些数据很多都有版权的制约。

读+:咱们的语料库会收录《四库全书》和《三体》吗?

林咏华:《四库全书》这样的公版数据应该都在里头,但是非公版的书籍必须考虑版权问题,我们自然不会故意去把非公版的数据放到我们的语料里头,但是我们的语料的确很庞大,在互联网上抓取的网页里可能有某一篇文章内嵌了这些内容,我们难以完全避免。

破解中文“数据孤岛”还需各方努力

读+:有一种说法,认为中文语料库的数量质量不如英文,是不是这么一回事?

林咏华:全球使用中文来写的数量大概是在5%,很多的统计都有类似的结果,因为直接用中文来进行创作的数量本身就比英文要少很多,所以中文的数据集天然会比英文的要小。

质量上,长期以来,语料数据集的积累和开源基本上都是在国外做的,他们在做的时候,可能更多的只是关心英文语料的质量,比较少去关心其他的语言的语料质量,所以我们联合其他团队一起奋起直追,要去解决这个问题。

不单是智源发布了中文语料库,其他的单位也有在做。大家整体都是向更大的、更高质量、更多样性的语料库方向去往前走。

一方面我挺开心,我们3.0的语料库跟其他已有的中文开源语料库相比,训练之后得到的模型在中文质量上能力上更优;另一方面,我们也觉得很难一家去把整个问题解决,最好是出现更多的高质量中文语料库,让不同的企业可以去做更多的选择。

读+:还有一种说法,语料库爬取的都是网页,但实际上近10年来,中国互联网很多数据是在移动端,说白了就是在手机上。

林咏华:这个是“数据孤岛”问题,确实存在。自2013年到2024年,11年间的全球主要网站网页内容语言使用历史趋势,中文网页的数量从4.3%下降至1.3%,下降高达70%;中文资讯已越来越“去网页化”,向移动互联网平台高度迁移。在这十余年中,我国9亿多网民迁移至各类移动互联网平台,而迁移背后,各家移动平台为了建立“数据护城河”,先后主动切断与传统网站网页的数据连通。高质量的信息更少出现在网页,更多的可能出现在App里头,例如微信等等。

所以这也是我们觉得需要继续推动的事情,怎么样可以让这些互联网企业愿意把这种高价值的数据“取之于民,用之于民”,我们也会持续去推动,希望这些声音能够被立法所考虑。

前沿观点集萃

生成式人工智能突飞猛进,使得中国学者开始关注、思考语料库问题,《读+》在此摘编介绍其中若干有价值的观点。

国家语料库是重大文化资源

许多国家都将语料库建设视为重要的基础工程,并建设了国家语料库。比如,英国国家语料库于1991年开始建设,1994年完成第一版,此后在2001年和2017年相继发布了第二版、第三版,规模为1亿词。美国国家语料库2003年发布第一版(规模为1110万词),2005年发布了第二版(规模为2200万词),2006年以后开始专注于开放的语料库建设。韩国政府于1998年启动了“21世纪世宗计划”项目,建设2亿词单元的韩国国家语料库,如今已经完成。此外,俄罗斯、匈牙利、泰国、爱沙尼亚等国家也都建设并发布了各自的国家语料库。

我国语料库的建设在20世纪70年代开始起步,目前已建成了多个汉语语料库,如几所高校和科研院所独立建设的规模不等的语料库,但是由于这些语料库在建设之初或多或少存在临时性、局部性、短期性、功能单一性等限制,缺乏长远考虑和全局性设计,未能全面反映国家通用语使用的现状。需要指出的问题主要有以下几点。

其一,语料库采样不平衡,多为书面语语料,缺少口语语料。比如,目前在汉语学界使用率很高的某语料库,其当代语料中报刊语料的占比高达70%以上,而口语语料占比不足0.3%。有的语料库仅将某一年度的微博文本作为口语语料,有的很大规模的语料库甚至没有收录口语语料。

其二,样本大小未加以控制,同等规模的语料库文本覆盖面较窄,影响了语料库的平衡性与代表性。比如,有的语料库对现当代作家作品全文收录,而标准的平衡语料库应避免同一作者的文学作品过多、过长、比重过大,否则会影响语料库的平衡性。如英国国家语料库最多只从单一作者作品的不同部分中抽取45000词的样本。

其三,语料库没有或缺少定期更新计划,难以进行基于历时平衡语料的研究。

其四,语料媒体形式较为单一。国内现有大型语料库中的语料多为文本形式,多媒体形式的语料库规模均比较小,取材受限,无法反映汉语使用的全貌,尤其是鲜活口语使用的面貌。处于国际研究前沿的基于“多媒体、多模态”语料库的研究越来越多,但国家通用语的多媒体、多模态语料库的建设仍相对滞后。

其五,语料库应用系统功能不足。

(中国社会科学院语言研究所所长张伯江、语言研究所副研究员张永伟)

语料库建设必须破解瓶颈

ChatGPT等国外模型训练语料主要取自英文语料库,来源主要包括维基百科、新闻报道、社交媒体、电子书籍、论坛帖子等,中文比重不足千分之一,英文语料占比超过92.6%。由于中西文化的源头、演进路径不同,导致国外生成式人工智能产品可能对中华文化的包容性和多样性的认识不充分,生成交互信息的文化内涵和精神往往带有局限性和片面性;同时,大模型对不同国家地区用户输入内容的语言习惯、表达方式在理解和处理方面也可能存在较大的差异,从而进一步加大误解风险。

目前,我国人工智能语料库建设虽然取得一定发展成果,但总体水平还处于相对落后的状态,要全面推动语料库建设向深度、广度拓展,必须破解其发展中所面临的诸多瓶颈。

一些商业数据平台出于商业利益考虑,在长期发展过程中已经形成自己的数据壁垒,数据开放、共享共用难度较大,难以作为语料基础平台。国内不少商业大模型团队利用私有数据库建设数据集,例如百度的内容生态数据、腾讯的公众号数据、知乎的问答数据、阿里的电商和物流数据等,这些花费不菲的商业数据集难以作为开源资源免费提供给其他组织机构共享。此外,语料库内容还受制于知识产权保护,一些是由于语料库设计不合理导致,比如因收录大量全文而引起的版权问题,限制了语料库的对外开放,另一些问题则与版权保护制度有关。现实中语料内容的知识产权保护会对语料库建设和共享造成一定限制。

(中国社会科学院图书馆李栋)

让人工智能具备“汉语思维”

目前的生成式人工智能产品,多带有一定的“西方思维”色彩。因此,随着技术的不断发展,我国人工智能需要关注“汉语思维”,提供具有汉语意识的技术刻画方式和行之有效的技术实现路径。

汉语语言语法结构复杂,词汇量丰富。因此,在“汉语思维”同自然语言理解与自然语言生成相关联的过程中,需要关注汉语语法、词义、语义和语用等,需要深入理解汉语的语言和文化背景,尤其要考虑我国用户的文化习惯和心理需求。我们可以通过收集和整理不同领域和语境下的汉语文本进行语料库建设,然后依托语料库中的大量语言数据来训练模型,并将其应用到机器翻译、语音识别等领域。

在“汉语思维”方面,需要解决的问题包括汉语的音调、声调和语音变异等问题。此外,还需要考虑汉语的方言和口音等因素。比如,汉语中有许多词汇在不同语境下的发音和语调是不同的。建立一个汉语语音语料库,可以帮助语音识别系统更好地理解和分析汉语语音信号。此外,还可以在语音库中加入成语、俗语、习惯用语等固定搭配的语音样本,帮助系统更好地识别这些表达方式。

(同济大学外国语学院教授许文胜等)

本社社址:湖北省武汉市江岸区金桥大道113号新长江传媒大厦 | 邮政编码:430013 | 中闻集团武汉印务有限公司承印 | 印刷质量监督电话:027-85888888