AI为文学角色“测”人格

■本报记者 胡珉琦

在心理学中,要准确、客观测量一个人的性格,一般通过自我报告或填写性格测试来进行评价。可如果测量的对象是一个文学作品中的虚构人物呢?

传统文学分析方法只能基于定性研究,且与研究者的主观体验密切相关。如今,得益于生态化识别技术的人工智能分析法,科学家通过捕捉人物对话就能实现对文学角色的“人格测量”。

近期,中科院心理所和厦门大学科研团队的这一合作成果在线发表于《中文信息学报》和国际学术期刊 *Digital Scholarship in the Humanities*。

郭刚制图

文学人物定量分析难

1975年至1985年间,中国土地所有制改革开始实行,农村跟着进入了发展黄金期。作家路遥以一对兄弟——耿直、坚韧、质朴的孙少安和同样坚韧又更敏锐、灵活的孙少平的生存与生活故事,展示了中国西北农村一段荡气回肠的变迁史。

小说,是以刻画人物形象为中心的。《平凡的世界》作为当代文学史上的一部巨作,它所塑造的经典人物形象是许多文学批评家、评论家、作家研究和分析的对象。

而通过对小说人物性格的分析,可以帮助读者更准确、深刻地理解人物的语言、行动和内心世界,以及它所反映的现实背景。

论文通讯作者、中科院心理所研究员朱廷劭表示,单纯在文学界对小说人物的性格主要是定性分析,不过,这样的心理分析主要依靠的是研究者在研究过程中的主观体验和文学素养。比如,通常研究者会从微观方面将小说中的一个或多个人物的性格概括为主要的几个方面,然后对这几个方面的性格特点用文本中的有关描写进行佐证。

另一种思路是从语言学的角度出发,通过对情态动词和高频词的统计,分析小说人物性格。这是一种定量方法,但并不主流。

说到对人的内在分析,心理学是有系统的、科学的研究方法的。只是过去心理学的研究对象都是真实存在的,而非虚构人物。但引入心理学方法研究文学人物完全是可行的。

研究显示,目前,小说人物的心理分析主要依据心理学家提出的人格三层次模型。其中,最基础的是作出对人格特质最基本的描述。人格特质是个体行为的稳定的、显著性倾向;第二层是个人关注,是描述个人奋斗、生活任务、防御机制、应对策略等大量有关人格动机和策略等方面的建构,从而了解一个人在不同情境下的动机、关切和策略;第三层是生活叙事,需要去探究一个人从出生到成年的发展历程,也就是人生故事。

但朱廷劭指出,相较于后两层,通过最初的人物性格或人格特质分析对文学角色进行的讨论是最少的。原因就在于,这类分析必须依据基于自我报告的心理测量,显然,这在虚构人物身上很难实现。

▶以《平凡的世界》中的主人公孙少安、孙少平兄弟为例,分析两位主人公的总体性格与经历重大生活事件后的性格变化。结果如表1、2、3所示。

表1 主人公性格预测

	宜人性	尽责性	外向性	开放性	神经质
孙少安	67.91	84.23	94.13	81.97	75.61
孙少平	59.55	88.81	84.89	96.94	66.79

表2 结婚前后孙少安的性格变化

	宜人性	尽责性	外向性	开放性	神经质
结婚前	81.27	85.84	95.55	88.55	60.87
结婚后	60.69	83.38	93.36	78.38	83.59
差值	-20.58	-2.46	-2.19	-10.17	22.72
变化率(%)	-0.25	-0.03	-0.02	-0.12	0.37

表3 升学前后孙少平的性格变化

	宜人性	尽责性	外向性	开放性	神经质
升学前	50.89	82.70	53.25	89.94	77.85
升学后	62.18	90.67	94.47	99.03	63.44
差值	11.29	7.97	41.22	9.09	-14.41
变化率(%)	0.18	0.09	0.44	0.09	-0.23

基于对话的智能分析法

不过,在有了生态化识别技术之后,自我报告的测量方法就不再是虚构人物人格分析的障碍了。这一技术是基于生态化的行为数据,利用机器学习实现个体心理特征的自动识别的过程。

在此之前,中科院心理所计算网络心理实验室基于新媒体大数据和深度学习技术,开发了一款网络心理的研究工具——大五人格预测模型。简单说,研究人员可以利用社交媒体内容与大五人格量表的映射关系,对社交媒体使用者的人格进行自动识别,而无须通过量表进行测量。经过检验,这个模型预测值和量表测量结果之间的相关性已经达到了中等程度。

那么,小说人物能否使用大五人格预测模型进行分析?答案是肯定的。

社交媒体的内容一般是个人较为口语化的自我表达,也就是说,如果研究人员能把小说中与该人物相关的自我表达内容提取出来,录入模型,便可以得到预测结果。

在小说里,最符合人物自我表达属性的内容就是对话(“直接引语”)。以《平凡的世界》为例,研究人员首先按特定格式将小说文本中所有的对话提取出来,并以人物为分类条件拆分对话。每个人物的对话集就是系统分析的对象。

不过,角色有主次,分配到每个角色的对话体量也是不同的,体量过小必然会影响模型预测的有效性。孙少安、孙少平、田润叶和田晓霞是小说的4个主要人物,不仅具有人物代表性,得到作者的笔墨也是最多的。

接下去,研究人员需要对这4个人物各自的对话内容进行预处理。首先,他们利用中文分词工具,将所有自然语言进行拆分;然后,通过一种量化分析软件——中文心理分析词典,对分词得到的所有词汇进行词汇统计,得到完整的词类分布。该心理分析词典共有102个词类、6547个词,包括与尽责性相关的成就词、情绪性相关的焦虑词、外向性相关的朋友词等词类,词类之间可相互重叠,也包括对标点符号和词长的统计。

最后,根据词类统计的结果使用大五人格预测模型进行分析,得到该人物大五人格的预测分数,包括人物的宜人性、尽责性、外向性、开放性和神经质的分数。

该研究结果显示,从预测数值上看,开放性相对较强的孙少平和田晓霞是思想最超前的;尽责性较强的孙少平和田润叶是受传统道德观念影响较深的;外向性较强的孙少安和田润叶是年长且在社会上表现出良好交际性的;宜人性较强的是谦让、忍耐的孙少安和活泼大方的田晓霞;情绪性较强的孙少安、孙少平是因为贫穷而在生活中充满挫折和矛盾的。

朱廷劭表示,在与现有的文艺学文献对这一小说人物性格的分析研究成果进行比较之后,文学智能分析的结果是与前者相符合的,从而证明了预测是有效的。

人工智能与文学的联合

这是一次客观的、量化的科学分析方法与文学的结合,给文学人物分析提供了一种全新的思路,可以屏蔽一些文学主观评价带来的偏差。

“不过,仅将这种方法用于文学作品的个案分析意义有限。”朱廷劭告诉《中国科学报》记者。

他认为,这种具有客观性、可重复性和处理大型语料库优势的技术,可以同时用于大样本量的分析,试图去寻找不同历史时期、不同时代背景下的文学作品中,个体人格特质、命运与所处环境之间可能存在的独特的关系。

此外,文学领域存在大量的纪实作品,比如人物传记、口述史、书信等,文学智能分析还可以为那些与我们相隔久远的历史人物进行一次科学的人格测试,作为史料分析的新的客观手段,也不失为一种有价值的应用方向。

相关论文信息:*https://doi.org/10.1093/llc/fqy020*

阳光也能转化为燃料?

近日,剑桥大学圣约翰学院成功地研制出一种使用半人工光合作用生产和储存太阳能的新方法。他们通过生物成分和人造技术,利用自然光将水转化为氢气和氧气。相关论文刊发在 *Nature Energy* 上。

该论文详细阐释了研究者如何利用他们的平台,来实现无辅助太阳能驱动的水分解。同时,他们还希望吸收比自然光合作用更多的太阳光。研究者称,此项研究可用于革新可再生能源生产的系统。

来自剑桥大学圣约翰学院的博士生Katarzyna Sokól说道:“自然光合作用效率并不高,因为它仅仅是为了生存,只需要少量的能量,即1%~2%的能量用以转换和存储。”

而人工光合作用虽然已存在数十年,但并没有成功地用于制造可再生能源。因为它需要使用价格昂贵且有毒的催化剂,这也就意味着人工光合作用的应用还无法扩大到工业水平。

基于此,研究人员试图通过使用酶产生的反应来完全解决人工光合作用的局限性。可喜的是,Sokól及其研究团队果然通过实验,不仅提高了人工光合作用产生和储存的能量,还重新激活了一种已在藻类中蛰伏数千年的生化反应过程。

“氢化酶是一种存在于藻类中的酶,能够将质子还原为氢气。在进化过程中,这一过程已被弃用,因为它不是生存所必需的。但是我们成功地绕过了这一过程,以达到我们想要的反应——将水分解为氢气和氧气。”Sokól希望,这一发现能够开发出用于太阳能转换的新型创新模型系统。

同时,她补充道:“令人兴奋的是,我们可以有选择性地挑选想要的工艺,并实现我们想要的化学反应。这为开发太阳能技术提供了一个很好的平台,同时,这种方法可以结合其他反应,通过实验的方法开发更强大的合成的太阳能技术。”

据了解,该模型是第一个成功使用氢化酶和光系统,来创建纯太阳能驱动的半人工光合作用的模型。因此,剑桥大学圣约翰学院Erwin Reisner博士将该研究描述为具有里程碑式的意义。他说道:“在该项研究之前,在将生物有机成分融合到无机材料中来组装半人工装置有种种困难,但是这项研究恰恰克服了这些困难,同时为未来开发太阳能转换系统开辟了新的道路。”(马晨)

相关论文信息:*DOI:10.1038/s41560-018-0232-y*

科学家向何处“流”

(上接第1版)

李江告诉记者,他拥有自己的ORCID号码——“每一位科研人员都可以在orcid.org网站上注册,并提交个人的学习工作经历,以及发表的论文清单”。据悉,目前还没有与ORCID功能类似的系统,“这一工具能在很大程度上解决科学家重名问题”。他介绍说,全球有大约280万科研人员注册了ORCID,其中西班牙和葡萄牙科学家人数较多,因为其资助机构要求科学家使用该系统。

“相比而言,中国的注册人数并不太多。前期只有英文期刊要求投稿人使用ORCID号码,但近期一些中文期刊也可以提出这种要求了。”

中国科学技术发展战略研究院研究员武夷山对于ORCID的作用表示肯定。他在采访中向《中国科学报》记者强调说,在大数据时代,首先要保证的就是数据准确,如果数据不准确,就是专业领域里所说的GIGO(garbage in, garbage out,即垃圾进,垃圾出),指错误数据的输入造成错误的输出结果,而ORCID可以很好地规避这个问题。

“在大数据时代,有很多很好的算法,但是如果数据不准确,这些算法并没有什么作用。”武夷山告诉记者,因为很多数据库曾经习惯于以论文作者的姓加上名字的缩写字母填充入数据库的“作者姓名”字段。比如“武夷山”就表示为“Y. WU”,“那全中国不知道有多少Y. WU啊!更别说一些常见名,比如李强,可能在一些大型科研机构就会有不止一个李强。那么,同机构内的某一常见姓名可能对应着不同的作者,你要区分谁是谁是很难的,给文献计量带来严重困扰。事实上,也确实有学术道德差劲的人钻这个空子,将别人的论文说成是自己的,因为另一个人的姓加上名字的缩写与其一致。有了ORCID,这种人就钻不了空子”。

流动,不一定非要跨国

科学家的自由流动类似于自然界的生态,阻碍科学家自由流动的行为破坏了学术生态。前文提到的Sugimoto认为,科学家的自由流动能产生多赢的效果,可以自由流动的科学家的影响力更大。

武夷山则对“流动更多的科学家影响力更大”这一观点持有不同观点,他分析说:“我关注过一些研究文献,根据其实证研究,迁徙过的作者的论文水平和被引次数确实可能要强于一直在一个地方呆着的作者。但这可能是个表面化的结论,似乎迁徙导致更多的交流,使科研人员开阔了视野,所以其论文的水平提高了;但往深里追究,你也许会发现,偏好迁徙流动的科学家本来就是能力较强的科学家。”

记者采访了多位有多国工作经验的科学家,他们大多明确表示确实不同国家的工作氛围有很大的不同,也有人提出,科学家的“流动”,不一定非要是跨国流动。

中国科学院高能物理研究所研究员张双南曾在美国工作十几年,他分析说,不同高校之间的风格也是引发科学家流动的原因之一。“哈佛可能就很喜欢挖人,而伯克利就很少,不同学校有不同的风格。像哈佛招聘一个助理教授,其实是对应一个终身教授名额的,但是等过几年,这个职位还是要公开招聘,很有可能不是这个助理教授获得。而这个助理教授在哈佛工作几年之后去到别的高校,也是很好的选择。每个学校的特点都很突出。”

科学家流动促进交流、打破知识边界、加强同领域或者跨领域合作的功能,是毋庸置疑的。但要实现这样的功能,却不一定非要是跨国界的流动。武夷山指出,对于欧洲科学家来说,从英国到法国再到丹麦,其实并不是一件多困难的事,“并不比中国科学家的省际流动更难”。

在武夷山看来,“流动”不妨从本单位跨专业合作开始。“很多人口头上总说要加强跨领域合作,但其实同一个单位不同专业之间的交流可能都很少。要解决这个问题,可尝试不同方式。像北欧一些高校,设置有专门的跨学科研究基金。申请条件很简单:不同系科的研究人员共同申请才行,这就促进了不同专业、不同系科之间的交流与合作。”

“蛟龙”号原第一副总设计师、万米级载人深潜器“彩虹鱼”项目负责人、上海交通大学教授崔维成曾在英国、丹麦、美国学习和工作,在采访中他也拓展了“流动”的定义:“国外很多项目的思路是,核心只有一两个人,面对非本专业的知识,自己去‘流动’、去学习,不断拓展自己的知识体系。”

相关论文信息:
https://doi.org/10.1016/j.apgeog.2018.04.017

热词

沙漠高铁

近日,在沙特设计最高时速为360公里的麦加至麦地那双线电气化高铁建成,这也是中国企业在海外参建的世界首条穿越沙漠地带高铁。

该高铁全长450.25公里,线路穿越阿拉伯沙漠,途经吉达、拉比格、阿卜杜拉国王经济城,其中麦加车站特大桥是全线的重点控制性工程,由中国企业独立承建。大桥总长1556米,横跨5条公路,桥梁最大宽度72.6米,为世界高速铁路桥梁之最。

据介绍,沙特地处地震带且年极端温度可达55摄氏度,为保证桥梁质量,中铁十八局通过技术创新,在梁体中埋设温度传感器和应变计,实时观察梁体的温度变化,及时调整施工工艺,满足了桥梁设计要求的地震安全性,以及震后快速通车的要求。

麦—麦高铁通车运营后,两地行车时间将由目前的4个小时缩短到2个小时,年客运量将突破1500万人次。

中国AI计算力报告

浪潮公司联合IDC近日在2018中国人工智能计算大会上发布了《2018中国AI计算力发展报告》。

算力方面,2017年AI硬件销售额同比增长235%,GPU、FPGA专用芯片和异构计算加速技术的快速发展。据IDC预测,到2025年,全球数据总量将达到163ZB,年均复合增长率将达到29%左右。

《报告》给出了中国AI计算力城市排名,前五位的城市是杭州、北京、深圳、上海、合肥,处于AI计算发展的第一阵营。成都、重庆、武汉、广州、贵阳位列AI发展的第二阵营。

Inception v3

纽约大学医学院的研究人员开发了一个新的机器学习程序,不仅能够以97%的准确率确定患者的肺癌类型,甚至还可以识别导致异常细胞生长的变异基因。

研究人员团队使用了来自Google的深度卷积神经网络Inception v3,并使用来自The Cancer Genome Atlas(TCGA)数据集的1634张图像对其进行了重新训练。TCGA是一个由美国国家癌症研究所(NCI)和美国国家人类基因组研究所(National Human Genome Research Institute,NHGRI)维护的公共数据集,包含了33种不同类型的癌症,以及每种癌症中存在的基因组变化数据。

在完成了对Inception v3的训练之后,研究人员开始使用该神经网络,来区分腺癌(LUAD)和鳞状细胞癌(LUSC),这两种癌症都是肺癌最常见的形式。结果显示,尽管样本中出现了在之前训练中从未出现的特征,比如血块、炎症、坏死区域和肺萎缩等等,Inception v3仍然可以正确识别绝大部分样本中的肺癌类型,正确率最高可达到97%。更加令人印象深刻的是,该模型在一台拥有单一图形处理器的电脑上运行时,平均计算时间只需短短20秒。

恐龙DNA

英国肯特大学生命科学学院的达伦·格里芬教授领导的研究团队重建了暴龙等恐龙在显微镜下可能呈现的基因组结构。为了重建这一消失了漫长时间的遗传密码,研究人员追踪了从爬行类祖先到现生动物的染色体在演化历史中的变化。

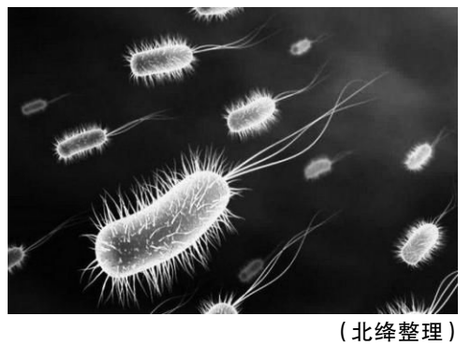
从现有资料推断2.6亿年前鸟类和海龟的共同祖先可能的基因组结构。大约过了2000万年后,第一批恐龙才开始出现在地球上。研究人员现在认为,恐龙可能具有约40对染色体。这与现生鸟类相似,并且几乎是人类染色体数的两倍。“我们认为这导致了分化。具有大量染色体使恐龙比其他动物类别更多地进行基因‘洗牌’,”格里芬教授说,“这种‘洗牌’意味着恐龙能更快地演化,以帮助它们即使在地球变化时也能生存下来。”

肠道细菌发电

科学家最新一项研究表明,人体肠道细菌能够“发电”。研究者之一、美国加州大学伯克利分校微生物学家丹尼尔·波特努瓦称,发电细菌并非一个新概念,在远离人类生活的环境中可以找到,例如湖泊底部。

但是之前科学家还不知道腐烂植物或者哺乳动物体内的细菌能以一种更简单的方式发电,尤其是农场牲畜。在实验室里,波特努瓦和研究小组首次培育了一批单核细胞增生李斯特菌,在日常饮食中,人们很容易吞入这种细菌,从而感染李斯特菌病。这种食物中毒对免疫系统低下、孕妇、新生儿和老年人群最危险。

通过将单核细胞李斯特菌放置在电化室里,能用电线或者电极捕获生成的电子。研究小组发现这些食源性细菌可以产生电流。



(北绛整理)