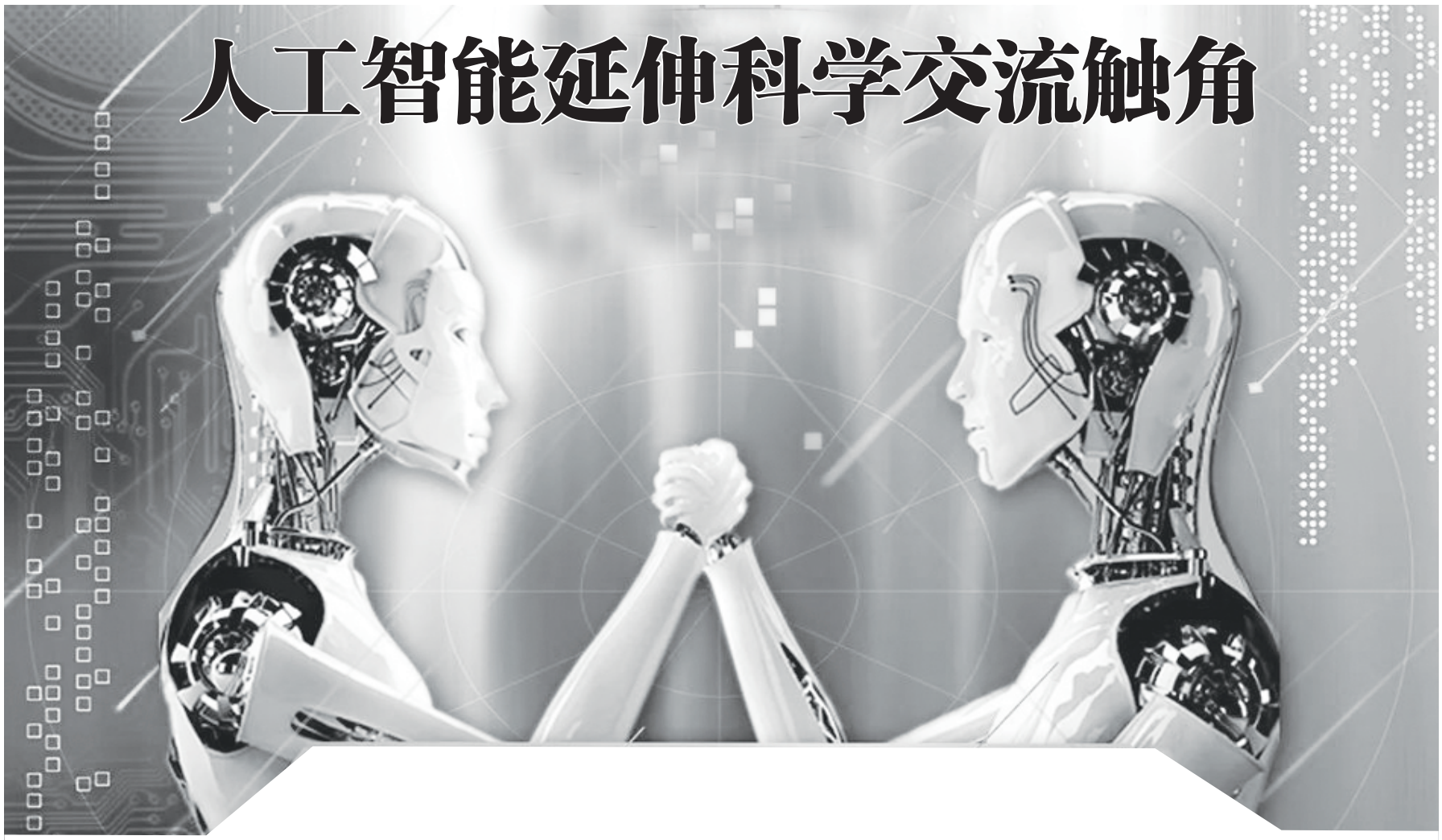
人工智能延伸科学交流触角



近日，一款看起来挺有文化的写稿机器人上线了。它叫小柯，由中国科学报社和北京大学科研团队共同研发。

小柯写的不是普通的稿子，而是中文科学新闻。据介绍，运用自然语言处理技术，小柯以英文论文摘要为基础，能够快速写出中文科学新闻底稿，然后由专业人士和报社的编辑进行把关和信息完善，帮助科学家以中文方式快速获取全球高水平英文论文中的最新科研进展。

目前小柯的作品已经上线。人工智能的触角，也在伸向各个领域。

小柯：一个尽职的摘要翻译转写者

记者发现，7月5日，小柯机器人发出第一篇稿子，截至8月22日记者统计时，小柯机器人共发稿415篇。初期更新时间距论文发表时间间隔一个月左右，现在可以做到当天或隔天更新，每天更新几篇到二十几篇不等。所选论文来自生命科学等领域，涉及《自然》《细胞》《新英格兰医学杂志》等期刊。

记者对照分析了小柯作品《单细胞测序揭示冠状动脉疾病保护机制》及其英文原文。新闻中，小柯先对论文主题、研究单位以及发表期刊进行简单介绍，后接英文原文摘要的翻译，大致反映原文内容；翻译时会对原文进行适当的语句简化，同时在对专业词语的翻译上也使用了如“血管平滑肌细胞”“保护性纤维帽”等专业表述。

不过，这也不全是小柯的功劳，因为稿件发出前，还有人工审校这一步骤。北京大学计算机科学技术研究所研究员万小军团队负责小柯的系统总体设计与联合技术攻关。他告诉科技日报记者，目前机器翻译系统的性能很大程度上依赖于其所使用的训练数据，即平行语料。目前的平行语料多为新闻语料，因此训练得到的机器翻译模型对于日常新闻的翻译效果较好。但学术文献（比如生物学术论文）与日常新闻在用词造句等方面都有较大差别，机器翻译系统对于学术文献翻译的效果并不理想。

这一次，他们通过融合领域知识进行语句智能筛选，选择适合大众理解的语句，并基于语句简化提升语句翻译质量。“英文学术论文摘要适合专业科研人员阅读，但摘要中的语句并不都适合写到科学新闻中面向大众传播，因此需要结合编辑提供的先验知识，采用计算机算法对语句进行筛选，保留适合进行大众新闻传播的语句。”万小军说。

自然语言处理技术不只能让机器人写稿

研发小柯用了半年时间，万小军表示，和一般写稿机器人相比，一个好的跨语言科技新闻写稿机器人需要进行两次重要的信息转换过程：一次是不同语言的转换，将英文文本转换为中文文本；另一次是语言风格的转换，将学术型文字表达转换为大众能够接受的通俗文字表达。“这两次转换都具有较大的挑战性，目前并没有完全解决。后续还需要进一步积累数据，调整算法模型，才能取得更好的效果。”万小军说。

接下来，团队还将继续优化小柯，让它写出的科学新闻内容更丰富，表达更生动。

当然，翻译撰写科技新闻稿件，只是自然语言处理等人工智能技术在学术交流中所能大显身手的领域之一。

“基本上，只要人类交流和工作过程中涉及到语言和文字的地方，自然语言处理技术都有可能发挥作用。”万小军说，在科研论文写作过程中，可以借助自然语言处理技术帮助推荐参考文献，并自动生成 related work 等章节的文字；业界也有基于自然语言处理技术自动编撰图书的尝试。“我个人也接触到很多很有意思也很有挑战的应用需求，但可惜的是不少需求都无法基于目前的自然语言处理技术进行实现。自然语言处理技术还需要进一步地发展和突破，我相信在未来将有更多的用武之地。”

中国知网常务副总经理张宏伟长期关注自然语言处理，大数据和人工智能方面的应用研究。他告诉科技日报记者，在数字出版和知识服务的全链条中，你都能看到人工智能和机器学习技术的身影。

人工智能可以对数字出版的选题策划、协同撰稿、内容编审进行赋能。大数据标注机器人则能对海量文献信息资源进行 OCR 文字识别，智能版面分析，知识元抽取，自动分类，自动标引主题，自动生成摘要，自动翻译，自动标注引用和参考文献。

人们熟悉的论文抄袭检测，同样需要智能技术。它不是简单的语句重复检测，而是要对文本内容（包括图片、公式、表格等）进行语义索引，“看你在思想上有没有抄袭别人”。如果存在不同语言之间的互抄，还需要动用“机器翻译”。张宏伟表示，初级的语义抄袭可以由机器揪出来，不过，如果足够有“心机”，完全用自己的语言“洗”了别人的思想，对人工智能的技术要求一下就提高了许多。目前已有利用神经网络模型对文本内容构建高维度语义索引等新技术出现，不管是中文还是英文，一律映射到一个统一的语义空间，实现真正基于内容理解的语义级全文比对检索。

知识库是智慧社会的基础设施

至于在学术研究中必不可少的资料索引，看似简单，也仍然具有技术含量。

张宏伟说，数字出版和数字图书馆的资源类型非常丰富，有大量文本、图像和音视频数据，且数据是非结构化的，若想对其进行深度的挖掘利用，难度不小。

就拿常见的信息检索来说，首先得做到结果要全，相关度要高；再进阶一步，能不能用自然语言交互的方式检索；升级一下难度，用智能问答的方式查找信息，能否直接给出答案？“要让检索功能变得更贴心，计算机要‘学会’阅读资料，总结、推理然后回答。它需要把海量的数据资源变成自己可以理解的知识库。”张宏伟说。

深度学习等统计方法严重依赖于大样本数据，然而，现实世界中，很多实际问题仅仅依靠统计方法是无法解决的，这就需要建立专门的计算机能理解的知识库，实现真正的人工智能。但构建知识库，本身是一项极其艰难且耗时漫长的工作。毕竟，机器和人对知识的理解方式大相径庭。

张宏伟说，像知网这样的机构正在致力于深度整合全球知识信息资源，建设世界知识大数据。也在让文本文献碎片化、网络化，依据知识使用的场景，采用半自动知识抽取算法来构建面向垂直领域的知识图谱。2019年知网陆续推出了一些基于知识图谱的行业智慧应用产品，如医疗领域的临床智能诊断，法律领域的智能量刑判案等。

“不过，我们在这些领域刚刚起步。我个人觉得，还是要少一点浮躁，踏踏实实做一些基础性的工作。没有知识的支撑，就谈不上‘智慧’。”在张宏伟看来，知识库和人工智能，本身就是互相促进、相互赋能的关系。构建知识库需要人工智能，而人工智能的发展，也离不开知识库。怎么将人类的知识库转换成计算机能理解的知识库是人工智能的核心问题，面临许多困难，需要学术界和产业界共同努力。

（据《科技日报》）