用 AI 写出的第一本书面世：先进算法能对机器生成的内容负责吗？

人工智能2019-04-15

表的研究数量超过任何学者希望跟上的数量，但很快他们可能会依赖 AI 同伴阅读数千篇文章并从中提取摘要——这正是歌德大学的一个团队所做的。

学术出版商 Springer Nature 出版了第一本由机器学习生成的书籍——《锂离子电池：机器生成的当前研究摘要》，它概述了锂离子电池领域的最新研究成果，大约 250 页。

与电池研究一样有趣的是，它只与该项目的实际目的相关。人工智能的创造者，在本书的广泛而有趣的序言中，解释了他们的意图更多的是开始讨论机器生成的科学文献，从作者问题到技术和道德问题。

换句话说，他们的目的是产生问题，而不是答案。他们有丰富的问题：

谁是机器生成内容的创始人？算法的开发人员可以被视为作者吗？或者是从初始输入开始的人(例如“锂离子电池”作为术语)并调整各种参数？是否有指定的发起人？谁决定一台机器应该首先产生什么？从道德的角度来看，谁对机器生成的内容负责？

这里面用到的技术，是由 Springer Nature 和法拉克福歌德大学共同开发的一种先进算法：Beta Writer。它使用的是基于相似性的聚类分析，将海量的源文档排列成连贯的章节，然后创建文章的简洁摘要，同时，将文章内部加入超链接，这样利于读者进一步阅读原始的文章。

AI 这种创新化的结构化摘录成书，有利于研究人员更高效地管理海量信息，以及人们从海量内容里快速选择、使用和处理相关领域的文档。

他们之间已经进行了激烈的辩论，他们的同行以及与他们合作制作这本书的专家，研究人员清楚地知道这只是一个开始。但正如 Henning Schoenenberger 在序言中所写的那样，我们必须从某个地方开始，这就像任何地方一样好。

确实，我们已经成功地开发了第一个原型，这也表明还有很长的路要走：大型文本语料库的提取性总结仍然不完善，而且有时复述文本、语法和短语联想仍然显得笨拙。但是，由于我们要突出显示机器生成内容的当前状态和剩余边界，我们明确决定不对任何文本进行手动修改或复制编辑。

正如他们所说，这本书本身就是不完美和笨重的。但听起来自然的语言只是人工智能尝试的任务之一，因为它而忽略整体的成功是错误的。

人工智能在这个高度技术性的主题上分发了数千篇关于 1,086 篇论文，分析它们以找到关键词，参考文献，“代词回指”等等。然后根据他们的发现对论文进行聚类和组织，以便以逻辑的、基于章的方式呈现。

代表性的句子和摘要必须从论文中提取，然后重新制定以供审查，这既是出于版权的原因，也是因为原文的语法在新的背景下可能不起作用。（团队谈到的专家说，他们应该尽可能接近原文的意思，避免“创造性”的解释。）

想象一下，论文中最好的句子开头是“因此，正如 2014 年论文所建议的那样，它产生的绝缘系数提高了 24％。”

AI 必须很好地理解论文，它知道“它”是什么，并且在重构句子时，将“it”替换为该项，并且知道它可以取消“因此”和最后的旁注。

这必须完成数千次模拟，并且许多边缘情况会弹出模型不能正确处理或产生一些公认的笨拙的用语。例如：“这种研究的主要目的是获得具有优异性能的材料，如高容量、快速的锂离子扩散速率，易于操作和稳定的结构。”

最终，这本书具有可读性和可以想象的有用性，已经将大约一万页的研究归结为大约 250 页。但正如研究人员所说，这一承诺要大得多。

这本书里面包含了 2016-2018 年发表过的 150 多篇权威研究论文。仅在过去 3 年，关于锂电子电池的研究论文就发表了超出 53000 篇，这对试图学习这一领域的科学家是一个巨大的挑战，但 AI 的自动扫描和总结输出，能让科学家们把更多时间用在重要的研究上。

这里的目标似乎并不遥远，就是能够告诉一项服务“给我一份 50 页的生物工程最后 4 年的总结。

文本的灵活性意味着您也可以用西班牙语或韩语请求它。参数化意味着您可以轻松调整输出，强调区域和作者或排除关键字或无关紧要的主题。

可以预见，未来的学术出版以及各类书籍，将不再只是人为创造，而是有更多形式出现，包括了混合人机文本生成的书籍或完全由机器学习生成的文本。

- END -