多模态同步语言神经影像数据集发布

作者：writer 来源：中国科学院

本文原地址：https://www.iikx.com/news/progress/20295.html

大脑在加工语言时，需要实时调动多个脑区的神经元进行协同工作。构建高时空分辨率的神经影像数据可以帮助我们更好地了解各个脑区以及脑区之间的协同合作，对于探索大脑的语言加工机制至关重要。当前已有的开源数据主要针对英文采集，只包括单一模态的神经影像数据，如高空间分辨率的功能核磁共振（fMRI）或高时间分辨率的脑磁图（MEG），且多使用1小时以内的实验材料，数据规模有限，无法借助数据需求量大的计算模型进行更全面、更深入的大脑语言加工机制探索。

中国科学院自动化研究所自然语言处理研究组历时近两年，采集处理完成了迄今为止国际上规模最大、包括信息最丰富的汉语同步多模态神经影像数据集，并于近日正式对外发布。相关研究成果发表在*Scientific Data*上。

该数据集是当前国际上最大规模的用于脑语言处理机制研究的多模态同步神经影像数据集，针对12个被试收听约6个小时故事时的功能核磁共振（fMRI）、脑磁图（MEG）、每个被试的T1/T2加权结构像、扩散磁共振成像（diffusion MRI）和静息态核磁共振（resting MRI）数据采集整理而成，采集流程如图1所示。为了便于利用计算模型进行脑语言处理机制的研究，所有故事材料都由人工标注了句法结构树，计算了文本中每个词汇对应的音频时间点、词频以及多种不同字和词汇的向量，如图2所示。所有测试指标均超越或可比拟已有的同类数据集，具有充分的质量保证。

该数据集的公开发布可为全方位研究大脑在真实场景下理解词汇、短语和句子时如何调动不同脑区以及不同脑区之间如何协同工作等科学问题提供重要支撑。该数据集覆盖了近万个汉语词汇，这对于探讨大脑理解汉语的认知机理具有重要意义，并将在探究自然语言计算模型与人脑语言处理机制之间的关系，以及如何利用神经影像数据提升现有语言计算模型的性能，从而构建新一代受脑启发的神经语言模型等系列工作中发挥作用。

论文链接

图1.神经影像实验数据采集流程

图2.实验材料对应的标注信息

研究团队单位：自动化研究所