“六书”多模态处理的形声表征以完善汉语语言模型

Li WEIGANG（李伟钢）[1]，Mayara C. MARINHO[1]，Denise L. LI[2]，Vitor V. DE OLIVEIRA[11]巴西利亚大学计算机科学系（CIC/UnB），巴西巴西利亚市，70910-900
[2]圣保罗大学经济管理会计审计学院（FEA/USP），巴西圣保罗市，05508-010

摘要：大型语言模型（LLMs）在自然语言处理中已取得显著成就，但在某些场景下，仍然面临解决中文语言处理复杂性的挑战。本文提出“六书”多模态处理（SWMP）框架，旨在考虑汉语形、声、音、像、意、会特性，便于中文语言多模态处理。在SWMP统一的理论框架下，提出“六书”形声编码（SWPC，简称“六书编码”）方法，使得对汉字的表达既能与语法有机结合，又反映汉语灵活应用的特点。文中设计的实验场景包括：（1）实验性建立汉字字根、偏旁（形部）和部件（声部）的图像和“六书”编码（SWPC）的数据库，实现汉语文字和图形的双模态处理；（2）表征若干汉词生成机制，建立提示性问/答模式，进行类比推理。使用SWPC处理中文形态关系数据集（CA8-Mor-10177）的所有问题，精度可达100%。（3）建立“六书”形声编码对词嵌入生成结果微调机制。对中文单词相似度数据集（COS960）中39.37%的问题，相似度计算与人工基础评估结果的平均相对误差低于25%。这些优于目前同类基准精度的结果表明，“六书编码”尝试体现汉语细腻的局部表征和整体关联等特点，可作为对现行汉语语言处理理论和技术的有效补充。

关键词：汉语语言模型；中文自然语言处理；生成式语言模型；多模态处理；六书
https://doi.org/10.1631/FITEE.2300384