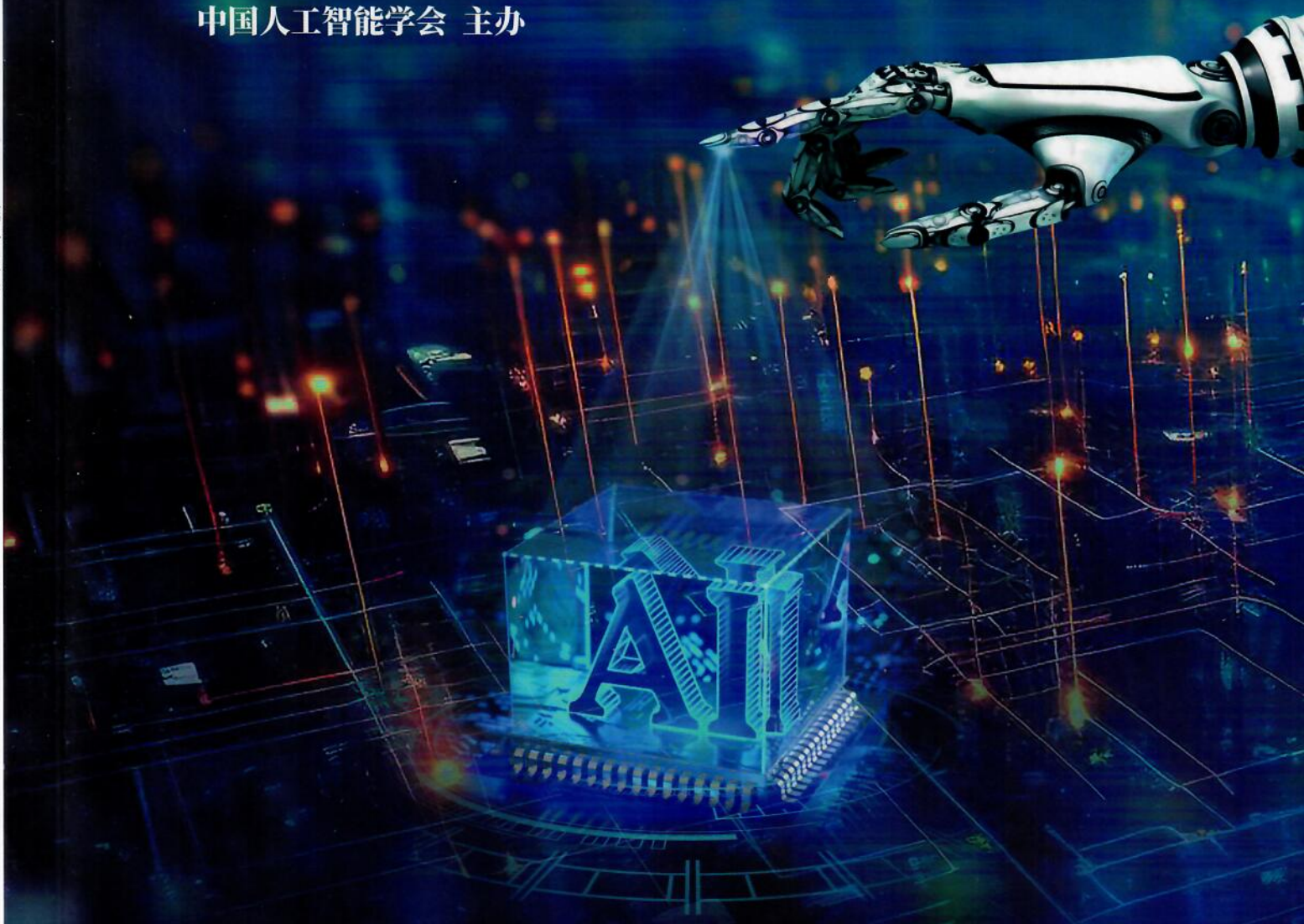
中国人工智能学会通讯

Chinese Association for Artificial Intelligence

第15卷 | 2025年 | 第5期

中国人工智能学会 主办



大模型知识机制探索

混元多模态大模型技术实践与思考

智能无人系统与具身智能

AIGS：全自主AI科学发现探索

CONTENTS 中国人工智能学会通讯 Chinese Association for Artificial Intelligence

目 录

大语言模型的空间理解能力评测
——知识驱动的合成数据研究

文 / 詹卫东

空间语义理解在自然语言处理（NLP）领域长期受到关注，是NLP评测的重要内容之一。然而值得注意的是，过往相关评测任务主要聚焦于语言中显性空间关系的分析，对于机器空间语义理解能力的评估不够全面，为此我们构建了中文空间语义理解评测任务（Spatial Cognition Evaluation，SpaCE），旨在系统评估机器对中文空间语义的深度认知的能力。该项目自2021年起连续四年推出SpaCE2021至SpaCE2024系列评测，形成业内首个持续性中文空间认知评测体系。本文着重解析该评测框架的设计方法论，并重点介绍基于知识驱动方法，合成空间常识推理数据集的研究。

1 机器空间语义理解能力评测概述

空间表达描述了物体之间的空间方位关系，是自然语言中的高频现象。准确理解文本中空间表达的语义不仅需要语言知识，还需要调用空间认知能力，构建空间场景，并基于世界知识进行空间方位信息相关的推理。

空间语义理解能力可以分为空间信息正误判别、异常空间信息识别、空间语义角色标注、空间表达异形同义判别、缺失参照成分找回和空间方位关系推理六个层次。

空间信息正误判别是第一个任务，从最基础的开始，按照语言学分析语言现象思路，对真实语料中的空间信息进行词语替换形成正误对比。例如，“大客车沿新源路由北向南行驶至曹安公路路口处遇绿灯，遂右转弯由东向西行驶，适逢被害人李红英骑电动自行车沿新源路西侧非机动车道由北向南行驶，两车相撞。”这是一份上海市的交通事故认定书。把其中的“右转弯由东向西”改成“右转弯由西向东”，就构成了一个空间信息错误的文本。让机器对这两个文本的空间信息进行正误判别，如果能够识别错误文本，则进一步要求机器指出具体的错误信息是文本中哪个片段，这就是异常空间信息识别任务（即第二个任务）。例如，对于上面这段包含错误信息的文本“大客车沿新源路由北向南行驶至曹安公路路口处遇绿灯，遂右转弯由西向东行驶……”机器要能识别出信息1“大客车由北向南行驶”与信息2“大客车右转弯由西向东行驶”之间存在语义冲突。

第三个任务是空间语义角色标注。例如，“这就是我从水里捞出来的美国人手上戴着的戒指，里面刻着他的名字。”塞科高高地举起戒指说，“上面的名字是，”他一边看一边念出上面刻着的字：“约翰·罗伯森·邓纳姆。”我们把文本中的空间实体，以及空间信息，包括“处所”“起点”“终点”“方向”“朝向”“路径”等，概括了一共10个空间语义角色，同时与实体空间信息相关的的动词事件也要识别出来，这个任务与传统的NLP动词语义角色标注任务类似。

第四个任务空间异形同义判别是突出认知特色的任务。例如，“至今菲律宾的土著居民在见面时，

握过手后还要转身向后走几步，意思是向对方表明背后没有藏刀，是真诚地迎接对方。”我们把“转身向后走几步”替换成“转身向前走几步”，这两个文本的空间语义表达的空间场景是完全相同的。从语言学的角度分析，这个片段里，“向前”和“向后”方向信息是冗余的，重点其实都是转身走几步。这种异形同异的语言现象，我们一共归纳了 6 种类型，两个句子一对，形式上有方位词差异，空间语义上可能相同，也可能不同，编成判断题，即一个二分类任务，让机器来做判断。目前，这样的问题对于大模型也是具有挑战性的。

第五个任务缺失参照成分找回，对空间语义理解的认知能力要求也较高。在文本中，方位词有时跟它的参照成分，即实体名词没有紧邻出现，就是参照成分缺失。例如，在一座小县城的一间教室里，工人们正在安装一块电子白板。“借助网课，我们的学生坐在教室里，就可以跟着上面的名师学习，享受优质的教育资源。”校长兴奋的说。这段话里，“上面”是相对于“白板”来说的，“白板”就是“上面”的参照成分。如果把“上面”替换为“外面”，“外面”指的就是“小县城外面”，参照成分就是“小县城”。对于方位词，如果没有跟参照名词相邻出现，人理解时会自动把实体名词补上。这个任务就是考察机器是否也具有这种理解能力。

第六个任务空间方位关系推理是考察机器对空间方位之间约束关系的理解。例如，“桌上有三块积木，红的在绿的上面，黄的在绿的下面。现在把最下面的拿到最上面来。移动之后，中间的积木是什么颜色的？”答案设置了四个选项：A. 红色、B. 黄色、C. 绿色、D. 无法确定。考察机器是否能推理得到答案 A。

所以，我们构造了上面这样的一系列任务，由易到难，不断升级来测试大模型机器空间理解能力到底是在什么层次，具有什么特点。

2 空间常识推理数据集制作方法

数据是大模型发展的一个重要且非常受关注的“瓶颈”问题。数据来源大体上有自然语料、标注语料、改写语料和生成语料等途径。依靠大模型的生成能力来生成数据、合成数据，是现在常见的一个方式。但是，从常识推理任务来讲，要生成高质量的推理数据，目前来说还有难度。我们调研了很多常识推理数据集，以及数据合成技术，决定采用编写专门的程序，基于人工知识库来合成推理数据，确保数据质量。

2.1 知识驱动的数据合成

目前，主流的数据合成技术很多都是基于大模型，而我们尝试让人类知识发挥更大的作用，基于知识库，用程序来控制，做高质量的比较精准可控的数据集。

如表 1 所示，我们收集了与方位词相关的词语表，基于词表，归纳表达空间方位关系的规则知识，形成知识库。对于数据合成程序来说，其输入是知识库，输出是题库。知识库包括知识（模板定义）、模板间逻辑关系、模板间推导规则和初始命题，最终生成的试题是选择题的形式。其流程是，知识库通过程序的推理器模块，生成事实库，再从事实库里抽取若干事实生成一个题目。然后再调用润色器模块，对题目的语言表达进行润色，最终得到比较符合自然语言表达习惯的逻辑推理题。

2.2 空间常识推理任务数据集的制作

这部分我们介绍从空间布局图式到空间关系知识库，再到最终生成空间推理题题库的过程。

空间常识推理是基于场景的，也就是推理任务要限定一个具体的空间范围，我们把这个场景称为空间布局。在一个布局下，既有显性表达出来的空间关系，也有隐含的没有明说的空间关系。比如四人卡座（见图 1），在卡座其中一排，两个人是左

表 1 表达空间方位信息的词语分类表

空间标记	词类	词语数量	例词
定位标记	方位词	236	上、上面、上边、上头、上端、左、右
	动词	7	在、有、是、位于、居于
	副词	3	到处、处处、四处
	介词	2	在、于
行程标记	动词	22	出发、启程、离开、撤退、抵达
	介词	11	到、至、从、自、由
趋向标记	动词	22	进、出、上、下、回
形态标记	形容词	19	高、低、矮、宽、窄
	动词	10	直、弯、屈、倾斜、弯曲
方向标记	动词	13	进、退、升、降、前进
	副词	6	正向、逆向、同向、同方向、反向
	介词	3	向、向着、往
朝向标记	动词	11	仰、俯、侧、扭、转
	介词	6	朝、朝着、向、向着、对
	区别词	2	向心、离心
距离标记	动词	11	贴、靠、靠近、接近、毗邻
	形容词	4	远、近、深、浅
	介词	2	距、离
合计		390	

右并排坐的，这里就会有“最右侧”的隐含信息，最右侧的人，其相邻座位就只有在左侧，右边没有其他人了。在一个空间布局确定下来后，可以把该空间中全部实体的位置和相对方位关系作为“事实”全部列举出来，推理题的构建方式就是将一组事实中的一条事实删除（隐藏），让机器恢复出来。图 2 是 6 个人向心围坐的布局，我们称之为向心六角布局；如果 6 个人背对而坐，就形成一个新的离心六角布局。此外，为了考察上下方位关系，我们还设计了三层两列置物架的布局（见图 3）。

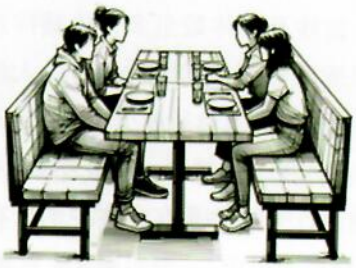


图 1 四人卡座

图 2 六人围坐（向心）

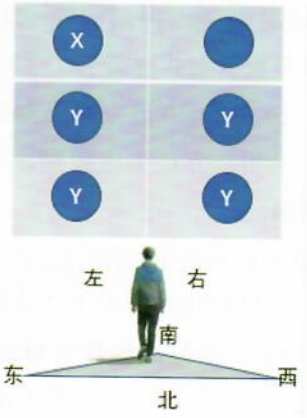


图 3 三层两列置物架布局

2.3 从空间布局到空间关系知识库

空间关系知识库是在空间布局约束下，罗列空间实体的方位关系事实陈述，以及方位关系之间的推导规则。如图 3 所示，知识库中描述 x 在三层东侧，x 和 y 不在同一层，则可以推导出 y 有 4 个可选位置。通过增加更多的关系（事实）描述，可以逐步通过 x（和其他实体）的位置信息，推导出 y 的位置信息。在知识库中，通过方位关系描述模板，可以描述一个空间布局中全部实体的位置信息。模板之间的逻辑关系包括等价、蕴含、包含、冲突、互反五类基本关系。通过逻辑关系，以及模板之间的推导规则，可以在给定事实基础上，扩展出更多的事实集合，进而形成试题的文本描述。

例如，在置物架上放置茉莉、君子兰、天竺葵、月季、郁金香、水仙 6 盆花，基于空间关系知识库里描述的三层两列方位关系事实，用这些花作为道具就可以由程序自动推理，最终编制出如图 4 所示的逻辑推理题。如果大模型能够按照题面描述正确推断出来所有花在置物架上的位置，就能回答“二层东侧是什么”这个问题。这道试题中包含 6 个实体，但在题面上只描述了 4 个实体位置信息。可以通过调节题面中实体个数，所包含事实陈述信息的条数等来控制推理难度。

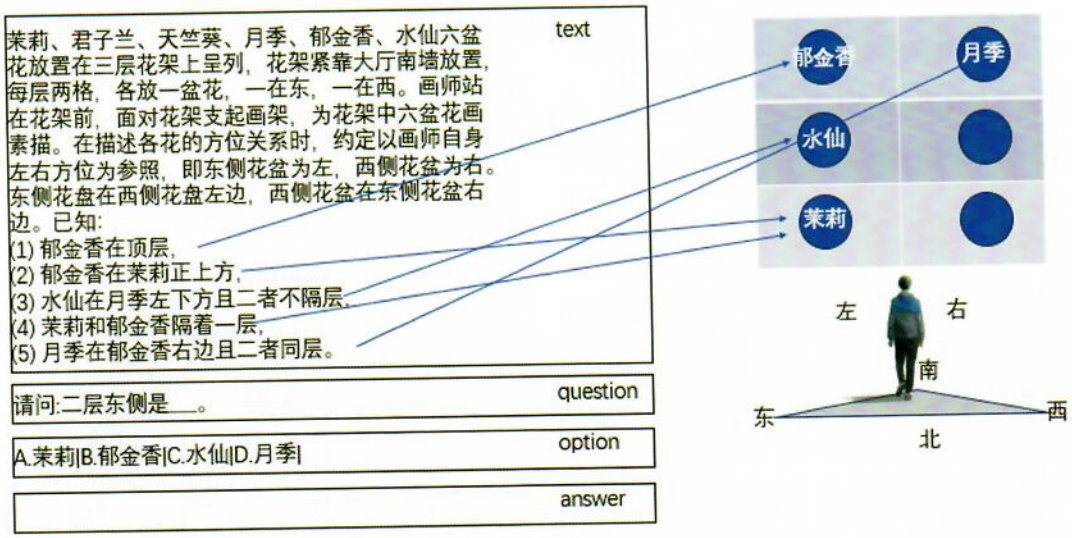


图 4 空间推理题示例

3 SpaCE2024 评测：空间关系推理任务

按照上述思路制作得到的 SpaCE2024 评测中的空间关系推理数据集共覆盖了 32 种空间关系、10 个空间布局，一共 2 040 道空间推理题。具体情况如表 2 所示。采用此方法生成数据，可以大批量生成上万道推理题。

SpaCE2024 获奖队伍的成绩显示，大模型在像“角色识别”这样的传统任务上，表现都非常好，成绩在 85 分以上，最好成绩甚至超过 90 分；但是在像“异形同义判别”这样需要较高认知能力的任务，目前还没有超过 60 分；在“空间方位推理”任务上，没有超过 40 分，最高分是 37.16。所以，空间推理任务对于目前的大模型，还是比较有测试价值的。

总之，初步测试的结果与我们的预期比较一致，即高认知难度的任务大模型处理得不太好。在空间推理任务上，涉及实体数量多，依赖推理已知信息的条数多和推理链的长度长的题目，大模型表现就差；反之，依赖的推理信息越少，模型表现就越好。另外，答案的选项个数越少，模型表现越好；反之，模型表现越差。空间关系在文本中的分布也存在差异，比如在自然语料中，“面对面”关系是最常见的，因为人的社会交际场景都是以面对面为主的，“背

表 2 SpaCE2024 评测数据集中空间关系在试题中的分布统计

空间关系	题量	空间关系	题量	空间关系	题量	空间关系	题量
右	514	正对面	56	底层	27	右上	16
左	491	背对	48	东南	26	左上	15
横向相邻	263	纵向相邻	48	右下	23	东北	14
横向不相邻	225	北	41	南	22	对侧	8
逆时针	116	上	37	西南	21	斜下	3
斜对面	113	下	35	同侧	20	斜上	3
顺时针	93	东	32	西北	18	不面对面	3
西	59	纵向不相邻	29	左下	17	身体不同向	3

对背”关系在文本中就少得多，在空间推理题中，涉及“背对背”方位关系的问题就答得比较差。

4 进一步拓展实验

空间关系推理题的制作方法可以拓展，融合其他常识知识。下面举两个例子说明，一个是空间常识和社会亲属关系知识的融合；一个是空间常识跟自然领域知识融合的拓展实验工作。

图 5 展示了空间常识和社会亲属关系知识的融合推理题。在四人卡座布局下，“罗小芬坐在靠窗的位置，脸上洋溢着幸福的笑容。她的父亲罗淮坐在她正对面。坐在罗淮斜对面的是他的女婿李夕明。罗小芬的母亲王玉坐在自己丈夫罗淮的左手边。”这段文本描述对应的空间场景如图 5（a）所示，人物关系如图 5（b）所示。针对这段文本的测试题是：（1）坐在王玉斜对面的是____。答案为罗小芬；（2）王玉斜对面的人坐在_____左边？答案是李夕明；（3）_____ 没挨着李夕明的岳父坐？答案是李夕明和罗小芬。为了制作这样的常识推理题，需要准备一个亲属关系知识库。

(a) (b)

图 5 空间场景 + 人物关系描述示意图

图 6 展示了自然常识和空间常识融合场景的推理题。对应文本是：“老王家的仓库坐北朝南（仓库地基是一个正方形）。勤劳的老王在仓库的前后左右各开垦了一块田，分别种植了西瓜、火龙果、番茄、青椒这四种作物。已知：① 种植水果的两块田地相隔距离要尽可能远；② 老王在仓库西边田地里撒下了黑色的种子；③ 仓库门前田里的作物开黄色的花；④ 仓库东边和北边田里作物的果实表皮颜色接近。”通过以上信息，就可以推导出这些作物相对于仓库的空间方位。针对这段文本的测试题如：（1）老王家仓库东侧田地种的是____。答案是西瓜；（2）老王从仓库出来，走最

近的路去田里摘西瓜，他出门应该 ____ 。答案是左转。

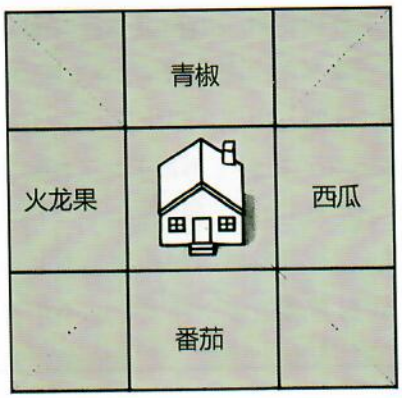


图 6 空间场景 + 自然物属性知识示意图

5 结束语

我们提出的合成数据制作方法是知识驱动的，对于常识推理任务来说，可以基于跨领域的知识库，自动生成大规模的测试数据。推理题基于统一的程序逻辑。采用自研的程序生成数据，相比于大模型生成的数据，质量更为可靠，而且覆盖知识点，对于知识分布的控制，对试题难度的控制，可以更灵活。整个数据合成的流程全透明可定制，通过这种方式，不仅可以对大模型的能力进行打分，还可以从多个维度进行评价，对模型能力进行解释，针对模型的不足，更有针对性的生成训练语料，帮助微调模型。

（参考文献略）

詹卫东

北京大学中文系教授，北京大学中国语言学研究中心副主任、计算语言学研究所副所长。主要从事现代汉语形式语法、语言知识工程与中文信息处理、语言文字应用方面的研究。代表性成果有《面向中文信息处理的现代汉语短语结构规则研究》《出版物上数字用法》《〈出版物上数字用法〉解读》。参编《计算语言学概论》《自然语言处理》《现代汉语》等多部教材。在国内外学术刊物发表论文多篇。近年来研究兴趣主要集中在现代汉语构式资源库建设，面向认知智能的机器语言理解能力评测等。

（上接第 05 页）

一个与电子羊一样的完全不一样的生物，那么人类能理解电子羊这样一个与我们完全不一样的智能吗？最后，由于这方面研究还非常初步，目前的研究结论呈现出碎片化，缺乏显著性和一致性的问题。因此，大模型知识机制的研究还处于一个盲人摸象的阶段，但是随着我们一点点的往前研究，终究会对大模型的知识机制形成深入理解。

（参考文献略）

韩先培

中国科学院软件研究所研究员，中文信息处理实验室副主任，国家优青获得者，入选中国科协青年人才托举计划及北京智源青年科学家；CIPS 理事及语言与知识计算专委会副主任。主要研究方向为大语言模型、知识计算及自然语言处理。承担中国科学院战略先导、科技创新 2030 课题，以及国家重点研发专项等课题 10 余项，发表论文 60 余篇，相关成果获 CIPS 汉王青年创新奖一等奖及科学技术奖一等奖。