基于异形同义现象的机器空间语义理解能力评测研究*

詹卫东　秦宇航　肖力铭

摘　要　不同的方位义词语可以用于表达相同的空间场景。文章考察了汉语空间表达“异形同义”现象的不同类型，并据此制作测试题，作为“异形同义判别”子任务，成为中文空间语义理解能力 SpaCE 评测基准的一个组成部分。针对大语言模型的评测结果显示，大语言模型在“异形同义判别”任务上与人类水平尚有较大差距，且机器在不同试题上的表现特点也与人类表现有所不同。从空间认知图式的角度讲，大语言模型基于语符分布学习到的人类语言知识，还没有转化为类人的空间认知图式理解能力。

关键词　空间表达　空间认知　异形同义　机器语言能力评测　大语言模型

DOI:10.16134/j.cnki.cn31-1997/g2.2024.05.009

一、引　　言

语言中有的符号形式跟意义之间的对应关系相对固定，使用时对语境的依赖性相对较小，比如“汽车、学校、演奏……”；有的符号形式则需要在使用时结合语境才能确定其具体意义。比如表达物体之间空间方位关系的词语“上、下、前、后、上去、下去、这儿、那儿……”。图 1 中甲和乙在描述方块和圆球的位置关系时，就可能会出现“异形同义”的情况：甲说“圆球 K 在方块 Q 的前面”，乙说“圆球 K 在方块 Q 的后面”，两人说的句子形式不同（有一词之差异），但所描述的空间场景是相同的。[1]“前、后”等用于表达空间方位关系的词语，属于指示语（deixis）范畴，相比于形义对应关系相对固定的语言现象，跟指示范畴相关的形义对应关系更为复杂多样，会给计算机理解文本中的空间信息带来更大的挑战。

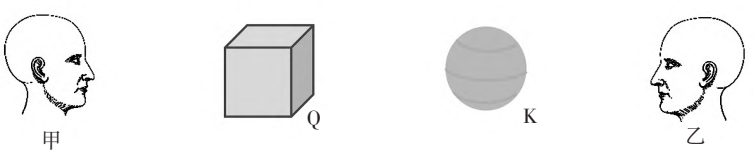

图 1　观察者视点对物体相对位置关系描述的影响示意

*　本文研究工作得到教育部人文社会科学重点研究基地重大项目“面向机器语言能力评测的综合型语言知识库研究”（项目编号 22JJD740004）的支持。北京大学中文系多位同学参与了本文工作，特别是邓思锐、李楠、邢丹、孙春晖、王佳骏、王希豪、胡楠、张子涵、崔香等在语料标注工作中贡献良多，特此一并致谢。

以深度学习方法训练的人工神经网络，通过观察海量文本中语言符号的分布模式，可以掌握类似于语言学研究所追求的“形式—意义”对应关系的知识。像 ChatGPT 这样的大语言模型表现出能与人流畅对话的能力，正是“意义即（形式）分布”这一抽象语言学原理的有效验证。不过，符号的意义是否完全等同于符号的形式分布呢？再进一步，训练语料的数据尽管是海量的，毕竟还是有限的。计算机在多大程度上，能从有限的语料（形式分布）中学习到具有无限可能的意义呢？

从理论上回答上述问题非常困难，信息处理领域的做法就是不断通过评测机器的语义理解水平来探索答案。在以往评测计算机空间信息处理能力的研究中，比较有代表性的评测任务是空间语义角色标注，如面向英语文本的空间语义角色标注任务 SpRL（2013，2015），多模态空间语义角色标注任务 mSpRL（2017）等。[2]语义角色标注任务是基于语言学理论对文本中的空间信息（包括实体和空间关系等）进行识别和分类，要求对文本进行细粒度结构化的综合分析，但这类任务侧重语言学专业知识，而不是诉诸普通人的语感。从“形式—意义”对应关系的角度看，自然语言的空间表达中有很多现象会超越符号通常的形义配对关系，呈现出不同程度的复杂性，对理解主体的认知加工能力提出了更高的要求。从这个角度考虑，我们尝试在评测任务设计时，实现从“语言学”到“语文学”的扩展（或者说某种程度的转向）：测试题的考察意图应该更直接体现机器对空间语言表达的直观理解能力，以更接近普通人语感的方式来回答问题。近四年我们依托中国计算语言学大会（CCL）的中文技术评测平台，组织了 SpaCE 系列评测大赛（SpaCE2021~2024），[3]先后设计了六项子任务：文本空间信息正误判别、文本异常空间信息识别、缺失参照成分找回、空间语义角色标注、空间表达异形同义判别、空间方位关系推理。[4]其中除空间语义角色标注、空间方位关系推理“专业色彩”较强外，其余四项任务都属于对一般人来说靠直觉就能回答的问题。比如例（1）是一个缺失参照成分找回任务[5]的例子：

（1）文本：他们五人推着自行车走到汽车旁，有两个人爬到了汽车上，接着就翻下来十筐苹果，下面三个人把筐盖掀开往他们自己的筐里倒。

问题：（　）下面三个人把筐盖掀开往他们自己的筐里倒？

显然，这样的任务就像是日常对话中的问答，不需要语言学专业知识参与。对人来说，回答这个问题需要理解整句话的语义，同时重点需要理解在特定场景中出现的实体以及实体间的空间方位关系。如果计算机也能像人一样做出正确的回答，就可以认为计算机也像人一样，是能够理解这句话中的空间信息的。

本文讨论 SpaCE 系列空间评测基准中的“异形同义判别任务”。第二部分分析“异形同义”现象的不同类型（对应不同复杂程度和不同认知难度的测试任务）；第三部分介绍评测数据集的制作方法；第四部分简要介绍大语言模型在这一任务上的表现；第五部分对比模型和人类被试在这一任务上的表现；第六部分对这项研究做一个总结。

二、空间“异形同义”现象的不同类型

空间范畴作为非常基础的语言认知概念，一直受到语言学界关注，研究成果非常丰

富。如果从“异形同义”的视角看汉语语法学界以往的工作，相关成果在三方面对本文工作有直接的启发。（1）注重区分不同空间方位参照类型，挖掘空间表达中影响说话人选取参照策略的不同因素。比如对“前、后、左、右”等方位词造成空间句异形同义现象的考察，可参见方经民（1987a，1987b）、林笛（1993）、郭锐（2004）等。（2）注重分析空间实体本身属性特征的影响，以及空间实体加上相关的动作等更复杂的语境信息，对空间表达异形同义的综合影响。比如对“上—里”可换用现象的分析，可参见高桥弥守彦（1992）、刘宁生（1994）；从实体属性角度讨论实体名词对其后方位词的选择限制，提出实体可居点特征分析框架，可参见储泽祥等（2008）。（3）对位移场景中的空间异形同义现象的考察，比如对“来—去、上来—下来”异形同义现象[“我马上就来 = 我马上就去”，“（登船场景中）跳上来 = 跳下来”]的分析，可参见齐沪扬（1996）、童小娥（2009）。

前人尚未对空间表达异形同义现象做系统全面的考察，[6]也未见有从机器空间理解能力评测的角度做相关语料数据的收集和标注工作。考虑到“异形同义”在真实语言使用中属于低频分布现象，为了在 SpaCE 评测基准中实现对机器的空间认知理解能力更为全面和深入的评测，我们设计了基于空间表达异形同义现象的异形同义判别任务。主要的思路是：穷尽性地考察汉语空间义词语的词对[7]（如“上—里、上—下、上—外、前—后、上来—过来、进来—下来……”），分析这些词对在表达空间场景时构成异形同义句对的可能性，以及造成异形同义的原因是什么。在收集到一定规模的句对语料基础上，就可以制作相应的试题（如判断题或选择题等），考察机器（或人）是否有能力判断：特定情境中空间方位义词语形式不同而其所指的空间场景却可能相同。

从形成原因角度看，空间表达中的“异形同义”现象有不同情况，大致可以区分为六类：（A）两个方位义词语本身词义接近；（B）两个方位义词语的词义有包含关系；（C）两个空间义词语有多个义项，二者在某个义项上，对应的空间图景相近；（D）文本中方位词（f）所依附的参照物名词（N）缺失，可以有不同的补回方式，异形同义实际上是 N1+f1 跟 N2+f2 对应了相同的空间图景；（E）实体在文本中有投影物，异形同义实际上是 N1+f1+N2 和 N1+f2+N2 之间造成的所指实体跟其影像的“伪同指”；（F）空间关系固化语境中的“主宾可逆序”句型，即词序可逆而空间语义角色（关系）固定不变。下文分别讨论。

（一）方位义词语词义相近（A 类）

请看下面的例子：

（2）a. 每年开春，家里总是从地窖**里**把保存了一个冬季的地瓜种一筐筐运到上面。

b. 每年开春，家里总是从地窖**中**把保存了一个冬季的地瓜种一筐筐运到上面。

c. 每年开春，家里总是从地窖**内**把保存了一个冬季的地瓜种一筐筐运到上面。

例（2）中三个句子只有一个词的差异，即“里—中—内”在三句中不同，其余部分是完全相同的，三句构成最小对立的形式差异，同时，句子所表达的空间场景相同：“地瓜种冬季储藏在地窖里，开春时从地窖运到外面。”

词义相近的方位词对不多。类似的例子还有“旁—边”“一边—旁边”“旁边—附近”“旁边—侧面”等。

（二）方位义词语的词义有包含关系（B 类）

请看下面的例子。

（3）a. 在这个房间里，墙壁上挂着一幅画，画的是一片美丽的森林。画的**上端**是一片蓝天白云。

b. 在这个房间里，墙壁上挂着一幅画，画的是一片美丽的森林。画的**顶端**是一片蓝天白云。

（4）a. 阿姨将肉粽打开，粽叶放在**一边**备用。

b. 阿姨将肉粽打开，粽叶放在**右边**备用。

例（3）中两个句子只有一词之差：上端—顶端，从所指范围来说，“上端”指的区域包含了“顶端”，后者是前者的一部分。例（3）a 和例（3）b 整句所表达的空间场景基本可以看作是相同的。

例（4）中两个句子也是一词之差：一边—右边，从所指范围来说，“一边”指的区域既可以是“右边”，也可以是“左边”，后者是前者的一部分。如果不以精确传递信息为标准，例（4）a 和例（4）b 整句所表达的空间场景就也可以看作是相同的（类似于用“车”称呼小轿车）。

趋向动词之间也有类似的词义包含关系。请看例子：

（5）a. 三辆警车循着逃犯的逃跑路线，一路追到白石桥下，连日洪水的冲击，让平日能过大卡车的石桥看上去像是处在崩塌的边缘。中队长犹豫半晌，最终咬牙发出命令：开**过去**！

b. 三辆警车循着逃犯的逃跑路线，一路追到白石桥下，连日洪水的冲击，让平日能过大卡车的石桥看上去像是处在崩塌的边缘。中队长犹豫半晌，最终咬牙发出命令：开**上去**！

例（5）中两个句子只有一词之差：上去—过去，二者在句中都表示“警车向白石桥的方向移动”。《现代汉语词典》第 7 版对“上去”作为趋向动词用法的释义是“用在动词后，表示由低处向高处，或由近处向远处，或由主体向对象”；对“过去”作为趋向动词用法的释义是“用在动词后，表示离开或经过自己所在的地方”。可见，“上去”对位移特征的描述更具体，“过去”则更笼统，从这个意义上讲，“过去”跟“上去”的词义关系，类似于上面“上端—顶端”“一边—右边”的词义关系，也属于包含关系，即前者的空间方位特征相比于后者更为笼统，适用范围更大，后者相比于前者更为具体，适用范围更小。

词义有包含关系的方位词词对和趋向动词词对不多。前者主要有“上端—顶端、下端—底端、一边—右边、一边—左边、旁边—右边、旁边—左边”等；后者主要有“过去—上去、过去—下去、过去—进去、过去—出去、过来—上来、过来—下来、过来—进来、过来—出来”等。

（三）方位义词语表示的方位或方向重叠（C 类）

两个方位词词义之间即使没有相近或包含关系，但在特定上下文中，仍然可以使整句表示相同的空间场景。请看例子：

（6）a. 夜里打麻将，使她根本无法看书做作业，她只好搬个小木凳到小巷边的路灯**下**学习。

b. 夜里打麻将，使她根本无法看书做作业，她只好搬个小木凳到小巷边的路灯**旁**学习。

（7）a. 沿着木栈道，总书记步入林中。在一棵落叶松**下**，总书记还特地用手丈量了一番："长得很好，树干很直。"

b. 沿着木栈道，总书记步入林中。在一棵落叶松**前**，总书记还特地用手丈量了一番："长得很好，树干很直。"

例（6）中两个句子只有一词之差：下—旁，两句的空间场景相同，都是"她在路灯下面学习"，"路灯下 = 路灯旁"。尽管"下"跟"旁"的词典释义不同，但在借助路灯光线来学习的事件场景中，把"路灯"作为参照物，"下"和"旁"可以表示相对于参照物"路灯"而言相同的方位，即"路灯灯柱底端附近的位置"。

例（7）中两句的情况类似，"下"跟"前"的参照物是跟路灯类似的柱状物"落叶松"，"落叶松下 = 落叶松前"。图 2 是"下—旁—前"这三个方位词在参照物为柱状物时指向相同位置的示意图，S 是当前描述的空间实体，S 相对于参照物 R（柱状物）的位置，用"下、旁、前"描述，都指向相同的位置，即 R 底部的附近区域。

趋向动词之间也有类似方位词的这种"异形同义"现象，请看例子：

（8）a. 人的咽喉和食管同胃是相通的，喝**下去**的醋只会与鱼刺接触，醋的脱钙作用无法进行。因此，任你喝醋再多，也无济于事。

b. 人的咽喉和食管同胃是相通的，喝**进去**的醋只会与鱼刺接触，醋的脱钙作用无法进行。因此，任你喝醋再多，也无济于事。

例（8）中两句的差异是趋向动词"下去"和"进去"的对立。在食管这一垂直柱状容积物作为参照物实体的语境中，移动的物体（醋）从食管外进入到食管内部，同时也是从食管的顶部往下进入食管下方的位置。在这个场景中，下去（从高到低）= 进去（从外到里），不同的两个趋向动词，表达了这一场景中相同的位移方向。图 3 形象地描述了这一现象。

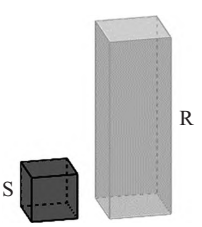

图 2 "下—旁—前"指相同位置示意

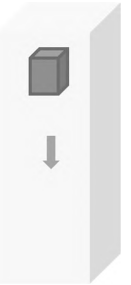

图 3 "下去—进去"指相同方向示意

像例（6）到例（8）这样的异形同义情况，不是由方位词或趋向动词自身的词义造成的，也不是由文本中相关的空间实体、参照物实体、位移事件等独立造成的，而是这些因

素共同作用的结果。如果把具体的复杂多样的空间场景抽象为有一定几何共性的示意图（如图 2、图 3 所示），在语言学中称为空间图式（Lakoff 1987; Talmy 2000）。从这个角度讲，可以说上述例句所呈现的异形同义现象，是因为不同方位词或趋向动词激活的“空间图式”有交集。

空间图式是从认知角度对物理意义上的空间场景所做的概念抽象，一个方位词或趋向动词可能对应一些典型的空间图式，但不太容易穷尽所有可能的空间图式，因为决定空间图式的因素比较多，而且有时不容易离析出来。例如：

（9）a. 昨晚，饭桌**上**，奶奶、爸爸和我争着同妈妈说话，直到我双手将妈妈的脸扳向我为止。

b. 昨晚，饭桌**旁**，奶奶、爸爸和我争着同妈妈说话，直到我双手将妈妈的脸扳向我为止。

例（9）中两句也是“异形同义”，虽然“上”跟“旁”是不同的方位词，但在例（9）的语境中，“饭桌上 = 饭桌旁”。如果要用空间图式来呈现这个场景中“上”和“旁”所指的位置，就不太容易，在很多场景中，“饭桌上”跟“饭桌旁”是不同的位置，比如“饭桌上的酒瓶”跟“饭桌旁的酒瓶”，肯定是指不同位置的酒瓶。但在多人围坐桌子吃饭的场景中，“在饭桌旁坐着的这些人”，也可以用“饭桌上的这些人”来指称。

实际上，我们收集的异形同义句对语料，多数都是“空间图式交集”类的（详见下文第三、第四部分），因为其他类别都有比较明确的形式特征条件或者词义特征条件，而“空间图式交集”这类情况，是方位词、空间实体、参照物实体、位移事件等因素综合作用的结果，尚无特别明晰、系统的界定标准，本文暂且把这类异形同义现象的成因归结为“空间图式交集”，更具体的原因分析，还有待将来进一步深入研究。

（四）方位词依附的参照物实体缺失（D 类）

前面三类方位词在句中都紧跟在其所依附的参照物实体名词之后。如果方位词所依附的参照物名词省略，在理解空间场景时，理论上就需要补出这个缺失的参照成分。这种情况下，也可能会造成异形同义现象。请看例子：

（10）a. 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**里面**的名师学习，享受优质的教育资源。”校长兴奋地说。

b. 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**外面**的名师学习，享受优质的教育资源。”校长兴奋地说。

c. 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。”校长兴奋地说。

（11）a. 至今菲律宾的土著居民在见面时，握过手后还要转身向**前**走几步，意思是向对方表明背后没有藏刀。

b. 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。

例（10）三个句子中只有一词之差：里面—外面—上面，这三个方位词依附的参照物名词没有跟方位词紧邻出现，其中例（10）a 的“里面”依附的参照物实体名词是“网课”（或“电子白板”），例（10）b 的“外面”依附的是“小县城”（或“教室”），例（10）c 的“上面”依附的是“电子白板”，但这三个句子所表达的空间场景可以说是完全相同的。

例（11）两句中也只有一词之差：前—后，这两个方位词依附的参照物名词没有跟方位词紧邻出现，两句表达的空间场景涉及位移动作：例（11）a 的“向前”指的是转身之后，人面向的前方；例（11）b 的“向后”指的是转身之前、人背向的后方。字面上，“前—后”两个方向相反，但在这两句所表达的空间场景中，实际上指向同一个绝对方向（比如“向东”），是相同的空间场景。区别仅仅在于，例（11）a“向前”方向的参照实体是转身之后的人（该人的面向）；例（11）b“向后”方向的参照实体是转身之前的人（该人的背向）。

例（10）和例（11）代表了两种参照物实体“缺失—找回”的情形。前者是在“同时”条件下，不同方位词参照了不同的空间实体；后者是在“历时”条件下，不同方位词参照了不同时点的同一个空间实体。二者都可以概括为：（N1）+f1=（N2）+f2，其中 N1 和 N2 是缺失的参照成分，可能是句中不同名称的空间实体，也可能是同一个空间实体在不同时点的变体。后一种情况出现的场景总是伴随着“转身、扭头”类转向动作。（孙陈亦待刊）

（五）空间实体在上下文中有投影（镜像）实体（E 类）

异形同义现象中，还有一类是实体在文本语境中有投影物，实体与投影物用同一个名词指称，即实体跟其影像“伪同指”，从而形成异形同义现象。请看例子：

（12）a. “笑一笑！”每次拍照前，摄影师都会对镜头**前**的人说这句话。甜甜的笑容挂在脸上，幸福感洋溢在镜头里。

b. “笑一笑！”每次拍照前，摄影师都会对镜头**里**的人说这句话。甜甜的笑容挂在脸上，幸福感洋溢在镜头里。

（13）a. 已经很多年没人这样叫李光头了，人们都是叫他“李总”，突然有人在后面叫他“李光头”，李光头心想是谁呀？回头一看是戴着口罩的宋钢，宋钢的眼睛在口罩上面的镜片**后**微笑。

b. 已经很多年没人这样叫李光头了，人们都是叫他“李总”，突然有人在后面叫他“李光头”，李光头心想是谁呀？回头一看是戴着口罩的宋钢，宋钢的眼睛在口罩上面的镜片**里**微笑。

例（12）和例（13）的共性是都有一个造成投影效果的“道具”，例（12）是通过“镜头”提供了投影；例（13）是通过“镜片”提供了投影。例（12）中，“镜头前的人”指真实物理世界中的实体人，“镜头里的人”指影像世界中的虚拟人，这两个实体具有一对一的投影关系。例（13）中，“宋钢的眼睛在镜片后”指真实物理世界的实体眼睛，“宋钢的眼睛在镜片里”指镜像世界中成像的眼睛，这两个实体也是一对一的投影关系。这种“伪同指”语境中造成的异形同义可以表示为：N1+f1+N2=N1+f2+N2’。因为 N_1（道具）的成像功能，使得 N_2 和 N_2’ 构成投影（镜像）关系，进而使得表面形式不同的“N1+f1”

和“N1+f2”约束构成镜像关系的两个名称相同的空间实体（N2=N2’）。在例（12）中，N2=N2’= 人；在例（13）中，N2=N2’= 宋钢的眼睛。

（六）主宾可逆序句型（F 类）

前五类异形同义都跟词汇语义有关。汉语中还有一类异形同义现象，跟特定构式有关。请看例子：

（14）a. 包好的包子在蒸锅里分三排摆放整齐后，她把**锅盖上锅盖**，然后打开计时器。

b. 包好的包子在蒸锅里分三排摆放整齐后，她把**锅盖盖上锅**，然后打开计时器。

（15）a. 我住在与福缘门隔着一条马路的娄斗桥，去北大食堂很方便。我常在那儿吃饭，**娄斗桥就正对着北大西门**。

b. 我住在与福缘门隔着一条马路的娄斗桥，去北大食堂很方便，我常在那儿吃饭，**北大西门就正对着娄斗桥**。

（16）a. 在吉林长春市一个繁忙路口附近，**一辆车前放着一个纸盒**，上面写着：口罩，环卫工人免费。……

b. 在吉林长春市一个繁忙路口附近，**一个纸盒放在一辆车前**，上面写着：口罩，环卫工人免费。……

上述例句在以往研究中属于“主宾可逆序句”这个话题。其特征是动词前后的主宾语可以调换位置，整句的命题语义基本相同，例（14）“锅盖上锅盖 = 锅盖盖上锅”，例（15）“北大西门正对着娄斗桥 = 娄斗桥正对着北大西门”，例（16）“一辆车前放着一个纸盒 = 一个纸盒放在一辆车前”。每个例子的 a、b 两句表面形式都有差异，但整句描述的空间场景相同。不过，跟前面五类不同，这一类的表面形式差异不是由替换一个词形成的最小对立。

显然，上面六类的情况并不均衡，有的类界定标准相对清晰，内部相对匀质，比如 A、B、D、E 这四类；有的类内部情况不均匀，情况相对复杂，比如 C、F 这两类。除 F 类外，其他五类都跟方位词、趋向动词等空间语义功能标记成分直接相关，用于测试和评估机器的空间语义理解能力相对更合适一些。对于存在异形同义现象的句对，归入前五类中的哪一类，多数情况是比较清楚的。对于少数可能存在归类模糊的情形，我们在工作中明确一个优先序原则：A>B>D>E>C，即能归入前面一个类别，就不归入后面的类别。这个优先序主要考虑的是语义标准和形式标准的清晰性，即语义标准和形式标准越清楚，就越靠前（优先）。比如词义是否相近，最易判断，其次是词义之间是否有包含关系，再次看方位词在使用中是否有参照成分缺失现象，然后再看文本中是否存在有投影关系的实体，以上条件都不符合，最后就归入空间图式交集类。上文例（10）的语境中也涉及投影实体（电子白板），但从形式上看，方位词依附的参照成分缺失，因而优先归入 D 类而不归入 E 类。

三、语料的收集标注和数据集的制作

（一）语料与数据集制作流程

在第二部分对异形同义现象进行分类描写的基础上，我们可以制作试题，来测试机

器对异形同义现象的理解能力。试题制作分为两步：先是收集异形同义和异形异义（用于对照）的句对语料，在达到一定规模后，再将语料转换为试题形式。

1. 语料制作阶段的工作方式

语料来源主要是两个途径：一是来自我们制作的 SpaCE2022 中文空间语义正误判断任务数据集[8]中的句对；二是给出词对表，对表中的方位词对、趋向动词对，逐一由人工编写符合异形同义和异形异义条件的语料。

SpaCE2022 中有形如例（17）、例（18）这样的句对（为节省篇幅，替换词写在括号中）。

（17）1960 年 5 月 25 日凌晨，中国登山队员王富洲、贡布和屈银华首次从“不可逾越”的北坡登上了珠峰峰顶，首次在珠穆朗玛峰顶插**上（下）**五星红旗，创造了人类历史上第一次从北坡登上世界第一高峰的壮举。

（18）等大家都坐好，聂赫留朵夫也在他们**对面（中间）**坐下来，臂肘搁在桌上，面前摆着一张纸，他就根据纸上的提纲开始说明他的方案。

例（17）“插上”是原句用词，“插下”是替换后的语料，将句中一个趋向动词“上”替换为“下”后，语句依然合法，且并不改变整句所描述的空间场景，这个例子就构成一个“异形同义”句对。例（18）“对面”是原句用词，“中间”是替换“对面”后形成的新的语料，将句中的方位词“对面”替换为“中间”后，语句依然合法，但整句所描述的空间场景发生了改变，这个例子就构成一个“异形异义”句对。人工对例（17）标注“异形同义”，对例（18）标注“异形异义”，就完成了两条语料的收集工作。

可以想见，从自然语料中替换方位词或趋向动词形成的对比语料（句对），多数情况下，要么句子语法或语义异常，要么两句是异形异义的情况。对于很多词对，为得到数量均衡的“异形同义”和“异形异义”语料，就需要人工编写异形同义的句对。像例（18）中的“对面—中间”这个词对，要构造异形同义句对语料，就比较困难。下面是利用缺失参照物找回这个线索，为“对面—中间”构造的两条符合异形同义要求的语料示例：

（19）张飞一人立马在两军阵前。曹军阵营一字排开，距蜀军阵营也就百步之遥。阵前挂出三面将旗，分别写着“张”“许”“夏侯”字样，代表着曹魏军中战功赫赫名震一方的三员名将：张辽、许褚、夏侯杰。张飞挺矛直指正**对面（中间）**的许褚，厉声大喝：我乃燕人张翼德，谁敢跟我决一死战？

（20）铁路要经过很多高山，不得不开凿隧道，其中居庸关和八达岭两条隧道的工程最艰巨。居庸关山势高，岩层厚，詹天佑决定采用从两端同时向**对面（中间）**凿进的办法……把工期缩短了一半。

我们在 SpaCE2023 和 SpaCE2024 中都设置了“异形同义判别”任务，SpaCE2023 是首次尝试制作异形同义和异形异义句对语料，生成了 355 条语料。SpaCE2024 扩充了词对表，收集编写了更多语料，具体语料规模如表 1 所示。

表 1 SpaCE2024 语料标注期间人工编写异形同义、异形异义句对统计表[9]

工作组	词对示例	词对数量	句对数量	异形同义	异形异义
方位词—单音节组	上—下、上—里、上—中、上—内、上—外、上—旁、上—边、下—里、下—中、下—内、下—外……	40	208	109	99
方位词—双音节 1 组	上面—下面、上面—里面、上面—前面、上面—后面、上面—外面、上面—侧面、上面—旁边……	33	113	55	58
方位词—双音节 2 组	上边—下边、上边—里边、上边—前边、上边—后边、上边—外边、里边—旁边、外边—旁边……	20	62	34	28
方位词—双音节 3 组	对面—中间、对面—附近、对面—旁边、附近—旁边、旁边—中间、中间—附近	6	13	6	7
方位词—双音节 4 组	东边—右边、东边—前边、南边—前边、南边—后边、南边—右边、西边—左边、西边—前边、北边—后边、北边—左边	9	18	9	9
趋向动词—单音节组	上—下、上—进、上—出、上—起、下—进、下—出、进—出、来—去、出—来、出—去	10	29	14	15
趋向动词—双音节 1 组	上来—下来、上来—进来、上来—出来、上来—过来、上来—回来、上来—起来、下来—过来……	16	39	30	9
趋向动词—双音节 2 组	上去—下去、上去—进去、上去—出去、上去—过去、上去—回去、上去—起来、下去—过去……	17	62	35	27
总计		**151**	**544**	292	252

2. 从语料到试题的转换

SpaCE2023 任务[10]中，我们直接使用异形同义和异形异义句对语料，以判断题的形式来考察。一道试题给出两个对比文本 Context1 和 Context2（文本中有一对方位义词形成形式对立）。问题（答案）由两部分构成，先是判断（Judge），即 Context1 和 Context2 的关系属于“异形同义”还是“异形异义”，其次是释因（Reason），即给出判断的理由。试题样例如表 2 所示。

表 2 SpaCE2023“异形同义判别”任务试题样例

异形同义题	Context1	一张微微泛黄的旧照片中，小伙子一身白色西装，脖子**上**系着领带，头发梳得整齐，与身旁衣着朴素的小女孩形成反差。	
	Context2	一张微微泛黄的旧照片中，小伙子一身白色西装，脖子**下**系着领带，头发梳得整齐，与身旁衣着朴素的小女孩形成反差。	
	Answer	Judge	同义
		Reason	两段文本的形式差异在于“脖子上”和“脖子下”。它们都描述了“领带”相对于“脖子”的位置：领带位于脖子表面和胸前。因此，这两段文本可以描述相同的空间场景。
异形异义题	Context1	兰兰惊奇地站在潜水桥**上**，透过玻璃看见大大小小的鱼游来游去，各种各样的船只从桥顶上驶过来划过去。	
	Context2	兰兰惊奇地站在潜水桥**下**，透过玻璃看见大大小小的鱼游来游去，各种各样的船只从桥顶上驶过来划过去。	
	Answer	Judge	异义
		Reason	两段文本的形式差异在于“潜水桥上”和“潜水桥下”。它们在描述“兰兰”相对于“潜水桥”的位置上存在差异，前者位于桥的上方，后者位于桥的下方。因此，这两段文本不能描述相同的空间场景。

以判断题的形式出题比较直观，但要求机器在判断异同之外，还要解释判断的理由。这些理由需要人工评分，成本较高。原因是虽然事先给了 Reason 的模板，仅要求机器填写表 2 中阴影部分的文本内容（相当于多个填空），但机器在生成文本时有可能没有严格遵循指令，生成的文本不符合模板要求，导致难以依靠程序自动评分。

SpaCE2024 的所有任务统一采用选择题形式命题，“异形同义判别”任务也改为选择题形式。试题样例详见表 3。

表 3 SpaCE2024“异形同义判别”任务试题样例

异形同义题	Text	眼看着水位越涨越高，易治林和其他老师们纷纷脱了鞋袜，光脚蹚进了漫水的走廊，抓紧搬运教学设施。**出来**时，水已经漫到易治林的腰部。		
	Question	“出来时”中的“出来”替换为（　）形成的新句可以与原句表达相同的空间场景。		
	Option	A. 下来　B. 进来　C. **回来**　D. 上来	Answer	C
异形异义题	Text	眼看着水位越涨越高，易治林和其他老师们纷纷脱了鞋袜，光脚蹚进了漫水的走廊，抓紧搬运教学设施。**出来**时，水已经漫到易治林的腰部。		
	Question	“出来时”中的“出来”替换为（　　）形成的新句也能描述一种空间场景（可以是常见的，也可以是不常见的），但明显与原句描述的空间场景不同。		
	Option	A. 下来　B. **进来**　C. 回来　D. 上来	Answer	B

改为选择题形式的好处是，语料的利用率相对更高。在一道选择题中，因为对比选项的增加（从判断题的 1 ∶ 1 对比变为选择题 1 ∶ 4 对比），替换对比项后形成的句子要么存在语法或语义错误，要么跟原句具有异形同义或异形异义关系，因而可以同时考察对语义正误的理解和对形义关系的判断。另外，异形同义现象的判断涉及比较复杂的认知因素，作为判断题，是二选一，有可能不同人对一个句对的理解差异也会比较大，但如果是选择题的形式，其他选项（非答案，干扰作用）可能对正确答案选项起到了一定程度的衬托作用。以表 3 的异形同义题为例，“出来”替换为“回来”，两句同义的条件是：先进再出 = 先进再回，要求文中“光脚蹚进了漫水的走廊”跟“出来”是相反的位移方向，这样，“出来”才能替换为“回来”而不改变空间场景。如果“光脚蹚进了漫水的走廊”中的“进”的位移方向跟“出来”是相同的位移方向，则“出来”跟“回来”就更倾向于理解为对立的方向，不是描述相同的（位移）空间场景。

（二）数据集的整体情况

SpaCE2023 中异形同义判别任务是判断题形式，我们从 355 个句对中选取了 100 个语料质量较好的句对，制作了 100 道判断题，其中 54 题为异形同义，46 题为异形异义，包含了上一节介绍的全部类型，不过总体数据规模比较小，主要是 C 类题（81 题），其余几类加起来共 19 题，A、E、F 三类一共才 7 道题。限于篇幅，这里不再展开介绍。

SpaCE2024 数据集扩充到 710 道选择题，按照机器评测的惯例，这些试题分为 3 份，其中训练集 5 道题，提供给机器学习，让机器熟悉题目形式；验证集 55 题，相当于人类考试中的模拟考试，用于评估机器的学习效果，改进学习策略；测试集 650 题，相当于人类考试中的正式考试。表 4 展示了 SpaCE2024“异形同义判别”任务数据集的语料字数规模概况；表 5 展示了“异形同义判别”任务测试集中单选题及多选题的数量和比例。

表 4　SpaCE2024“异形同义判别”任务数据集概况

项目	训练集	验证集	测试集	合计
题量	5	55	650	**710**
字数	1042	9852	114282	**125176**
最短句长	41	14	9	**9**
最长句长	173	179	220	**220**
平均句长	104.20	89.56	87.91	**88.15**
标准差	54.59	48.12	41.45	**42.05**

表 5　SpaCE2024“异形同义判别”任务测试集中单选题及多选题的数量和比例

类型	单选题		多选题		合计	
异形同义题	391	60.15%	51	7.85%	**442**	**68.00%**
异形异义题	126	19.38%	82	12.62%	**208**	**32.00%**
合计	**517**	**79.54%**	**133**	**20.46%**	**650**	**100.00%**

下文表 6 展示了 SpaCE2024 “异形同义判别” 任务测试集中涉及词对数量及对应的题量，并按照上文第二部分提出的类型体系分类计数。因 F 类（主宾可逆序句）的性质跟其他五类差异较大，且收集的这部分语料数量较少，故没有收入 SpaCE2024 数据集中。测试集中 C 类题最多，表 6 中进一步细分为 C1 类（方位词空间图式交集）和 C2 类（趋向动词空间图式交集）。相对而言，C1 类的异形同义题和异形异义题数量较为均衡，其他类别异形同义题的数量都明显多于异形异义题，显然，在分布均衡性方面，数据集还需要做进一步的改进。比如 A、D、E 三类异形异义题为 0，其中 A 类是词义相近词对，难以构造异形异义题，属于正常的偏置分布，D、E 则可以而且需要构造数量相当的异形异义对照题。此外，不同词对在测试题中分布平衡性还存在较大问题，数据集规模还有待提高。上文表 1 统计了目前数据集中词对类型（type）数为 151 对，表 6 统计的词对实例（token）数为 820 对，即每个词对平均在数据集中出现 5.43 次，以测试集 650 题为单位计，每个词对平均出现在 4.3 题中。实际上，出题数量达到 4 题以上的词对仅 43 个（占 28.5%），更多的词对（108 个）仅出现在 1 到 3 题中。即便是出题达到 4 题以上的词对，在异形同义题和异形异义题的比例上也很不均衡，比如题目频次前 5 的词对：上—里（16：2）、下—里（15：2）、上 $_{f}$—下 $_{f}$（8：8）、下面—里面（12：1）、上—中（12：1），只有 1 个词对两类题比例均衡。而在出现 4 题以上的全部 43 个词对中，也仅有 8 个词对（18.6%）的异形同义题和异形异义题比例相对均衡：上 $_{f}$—下 $_{f}$（8：8）、上—旁（3：4）、上去—下去（4：3）、后面—外面（3：3）、下面—外面（3：2）、中间—对面（2：3）、内—前（2：2），后边——外边（2：2）。以上情况表明：SpaCE2024 “异形同义判别任务” 数据集在数据规模和试题分布均衡性方面都还存在明显不足，还有待改进。

四、大语言模型测试结果初步分析

本节介绍参加 SpaCE2024 评测的参赛系统（均采用大语言模型作为基座）在异形同义判别子任务上的表现。上文表 5 显示了 SpaCE2024 数据集区分单选题和多选题，以单选题为主，这样设置，主要是从增加试题难度的角度考虑，如果机器在多选题上也达到较高的正确率，就有更大把握认为机器对空间语义有深度理解能力。单选题中还有 27 道题答案设置为 “D. 以上选项均不正确”（异形同义题 24 道，异形异义题 3 道）。这类单选题和多选题类似，对机器而言难度更大。12 支参赛队伍中总分排名前 6 的系统在异形同义判别任务上单选题平均正确率是 0.62，单选题中答案为 “D. 以上选项均不正确” 的题，平均正确率为 0.40。多选题平均正确率是 0.30，是单选题的一半。从这个角度看，大语言模型对异形同义判别任务，还没有达到真正理解的水平。

表 6 给出了这些系统（以系统 1、2……称名）在 6 类异形同义现象测试题上的分项计分结果。[11] SpaCE2024 数据集异形同义题跟异形异义题的比例不够均衡，因此表 6 中同时也给出了各系统在这两类题上的分项计分。

表 6 大语言模型在 SpaCE2024 异形同义判别任务测试集上的表现

类型	子类	词对量	题量	系统 1	系统 2	系统 3	系统 4	系统 5	系统 6	分项平均	总平均
A	同义	33	31	0.81	0.65	0.97	0.65	0.87	0.81	0.79	0.79
	异义	0	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
C2	同义	72	65	0.71	0.63	0.69	0.65	0.72	0.65	0.67	0.64
	异义	29	26	0.65	0.54	0.69	0.42	0.38	0.58	0.54	
B	同义	27	26	0.62	0.69	0.77	0.77	0.73	0.58	0.69	0.63
	异义	4	4	0.25	0.00	0.50	0.00	0.25	0.25	0.21	
C1	同义	323	280	0.57	0.61	0.61	0.56	0.73	0.63	0.62	0.53
	异义	291	178	0.46	0.42	0.42	0.33	0.33	0.40	0.39	
D	同义	35	34	0.56	0.44	0.50	0.29	0.35	0.65	0.47	0.47
	异义	0	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
E	同义	6	6	0.17	0.00	0.67	0.17	0.33	0.33	0.28	0.28
	异义	0	0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	
合计		**820**	**650**	**0.56**	**0.54**	**0.59**	**0.49**	**0.59**	**0.57**		**0.56**

大语言模型是黑盒模型，其推理过程不可见，很难知道模型对一道题的回答是如何做出的选择，仅从模型对一道题的作答，也难以确定模型是否掌握了相关词对的空间语义知识。而且大语言模型还存在比较明显的稳定性问题，[12]再加上题量不大，因而考察大语言模型在具体词对和题目上的答题情况，目前还难以形成规律性的认识。[13]这里仅对大语言模型整体上较为明显的特点做一些概括说明。表 6 的分项计分，基本上反映了当前大语言模型基于语言符号的形式分布来学习意义的特点，模型在 A 类测试题上的表现明显优于其他类别。因为 A 类异形同义现象的判别主要诉诸方位词自身的词义，这类异形同义相关的“形式—意义”配对关系制约条件单一，容易被模型捕捉到。模型在 D 类和 E 类测试题上表现远远低于 A 类，这两类异形同义相关的“形式—意义”配对关系制约因素复杂，而且在自然语料中属于低频分布，模型相对来说不容易学习到判别条件，表现较差，也就在情理之中了。

值得一提的是，SpaCE2023 的异形同义判别任务 100 道判断题异形同义和异形异义题数量相当，ChatGPT3.5 在这两类题上的表现存在这样的情况：在对 54 道异形同义题做判断时，做对了 43 题（77%）；对 46 道异形异义题做判断时，做对了 28 题（61%），仅从判断结果来说，大语言模型得分都在及格线以上。但在进一步解释原因时，对异形同义题的解释，得分为 35 分，对异形异义题的解释，得分为 47 分（由人类专家评分）。前者比后者低 12 个百分点。这也同样反映了自然语料中不同类型语言现象的分布模式对模型表现的显著影响。自然语料中，异形异义无疑是远远多于异形同义的更为高频的语言现象，在总体表现上，模型对异形异义题的理解（成绩）自然也就比对异形同义题更好。

五、人类表现与模型表现的对比

为进一步考察大语言模型在异形同义判别任务上表现的特性，我们从 SpaCE2024 数据集中抽取了 100 道异形同义判别任务试题，组织了一个小规模的人类测试。数据分类情况如表 7 所示。其中有 10 道题是“重复题”用于测试回答稳定性，此外，有 4 题正确答案为“D. 以上选项均不正确”。人类被试共 8 人，其中 2 名被试答题无效，另外 6 名被试在重复题上得分超过 0.9（是大语言模型得分的 2 倍），我们选取这 6 名被试的成绩用于跟大语言模型的表现进行对比。大语言模型在 4 道答案为“D. 以上选项均不正确”题上的平均分为 0.46，6 名人类被试的平均分为 0.71。表 7 展示了单选题和多选题上人机成绩的对比；表 8 展示了不同类型异形同义判别题上人机成绩的对比。

表 7　人类测试题集（100 题）中单选题及多选题的数量和人机表现对比

类型	异形同义	异形异义	合计	人类得分			机器得分		
				平均	最高	最低	平均	最高	最低
单选题	58	22	80	0.90	0.93	0.88	0.61	0.66	0.56
多选题	9	11	20	0.70	0.80	0.60	0.34	0.45	0.20
合计	67	33	100	**0.86**	0.88	0.84	**0.56**	0.60	0.49

表 8　人类测试题中不同类型异形同义判别题上人机表现对比

测试题分类		题量	人类得分			机器得分		
			平均	最高	最低	平均	最高	最低
B	异形同义	1	1.00	1.00	1.00	0.50	1.00	0.00
	异形异义	0	—	—	—	—	—	—
C2	异形同义	8	0.98	1.00	0.88	0.54	0.75	0.25
	异形异义	2	0.67	1.00	0.00	0.83	1.00	0.50
A	异形同义	5	0.97	1.00	0.80	0.73	1.00	0.40
	异形异义	0	—	—	—	—	—	—
C1	异形同义	46	0.91	0.93	0.87	0.67	0.80	0.54
	异形异义	31	0.80	0.84	0.74	0.41	0.48	0.26
D	异形同义	4	0.63	0.75	0.50	0.25	0.50	0.00
	异形异义	0	—	—	—	—	—	—
E	异形同义	3	0.56	1.00	0.33	0.22	0.67	0.00
	异形异义	0	—	—	—	—	—	—
合计	**异形同义**	**67**	**0.89**	0.91	0.88	**0.61**	0.67	0.55
	异形异义	**33**	**0.79**	0.85	0.76	**0.43**	0.52	0.27

表 7 和表 8 统计数据显示机器成绩显著低于人类水平，[14]说明空间异形同义判别任务对于大语言模型仍然属于高挑战任务。人类与机器得分的共性是：在 D、E 类任务上的表现明显低于 A、B、C 类任务。这一方面可能是 D、E 类试题数量少且试题质量不高造成了统计偏差，另一方面也提示：D、E 这类相对低频的语言现象，对人类而言，认知加工的难度和个体差异性也可能更大。对此，还需要在改进试题质量和规模后，做更进一步的对比研究。值得一提的是，在 A、B、C 三类异形同义题上人类被试超过 0.9 分，且被试之间一致性相对更好。而机器在这三类异形同义题上的表现，虽然整体相对其他类表现更好，但不同模型之间仍存在较大差异。请看下面两例：

（21）明美的速度慢于同组的其他同学。其他同学足足等了她半个小时，才等到从半山腰的观景台走**上来**的她。

“才等到从半山腰的观景台走上来的她”中的“上来”替换为（　　）形成的新句可以与原句表达相同的空间场景。

A. 起来　　B. 下来　　C. **过来**　　D. 进去

（22）凶手进入房间，杀害了房间内包括罗森堡在内的三人。每个人的头部都中了三枪。罗森堡的头部取出两颗子弹，枕头**里**又找到一颗。

“枕头里又找到一颗”中的“里”替换为（　　）形成的新句可以与原句表达相同的空间场景。

A. **中**　　B. 上　　C. 边　　D. 以上选项均不正确

例（21）考察“过来—上来”这对趋向动词，二者属于词义包含关系（B 类题），人类被试全部正确选择了答案 C，但机器 6 个系统中只有一半选 C，另外一半选了 B“下来”，而后者显然在这道题的语境中跟“上来”构成异形异义关系。

例（22）考察“里—中”这对方位词，二者属于词义相近关系（A 类题），人类被试全部正确选择了答案 A，但机器 6 个系统中有 4 个选 A，总成绩第一和第二的两个系统选择了 B。

上文举过的例（17）也在这 100 题中，属于 C2 类，选项设置为“A. 去、B. 下、C. 来、D. 回”，人类被试全部正确选择了答案 B，但机器 6 个系统中只有 2 个选了 B，另外 4 个系统选 A、C、D 的都有（分别是 2、1、1 次）。

以上情况显示，即便是形式和意义对应关系相对清楚，判别条件容易学习和掌握的空间义词对，机器目前的理解总体水平也较低，且跟人类表现特点有明显差异。

六、结　　语

本文研究了汉语中的空间表达“异形同义”现象（即**两个句子形式不同，仅有一词之差，而可用于描述相同的空间场景**），针对在传统自然语言处理任务上表现优异的大语言模型，本文首次提出了对机器更具挑战性的“空间异形同义判别任务”，并主要以人工编写方式制作了“异形同义”和“异形异义”句对语料，并转换为选择题，形成了空间异形同义判别任务测试数据集。我们分别在 SpaCE2023 和 SpaCE2024 评测大赛中，加入了这部分测试数据，进行了大语言模型测试和人类测试。测试结果显示：

（1）在测试数据集设计的全部可比项目，比如单选题、多选题、重复题等从纯形式角度设置的考察项目上，以及从原因角度对空间异形同义现象所区分的6个类型上，大语言模型的表现均显著低于人类平均水平，且大语言模型自身的内部一致性（稳定性）欠佳。

（2）大语言模型对自然语言意义的理解，更为显著地受到语言符号分布形式层面的影响，比如对出现频次更高的异形同义现象的理解能力要优于出现频次较低的同类现象；对“形式—意义”对应关系制约条件少的异形同义现象（上文的A、B类），理解能力优于制约条件多、需要更深认知能力的异形同义现象（C、D、E类）。

从初步结果来看，这项任务对考察大语言模型的“空间认知”能力，是有效的。不过，这项高认知难度的任务，对数据集的质量和规模，也提出了很高的要求。要让各个考察项目上的题量更具统计意义，让不同类别的题目分布更均衡，还需要针对空间异形同义现象，做进一步更细致的理论研究工作（尤其是对C类和D类异形同义现象做深入研究），同时在机器辅助生成语料、设计更好的试题形式、提高数据合成效率方面，也还需要更多探索。

附　注

［1］假如以地图模式的绝对方位“上北下南左西右东”来说，图1中K在Q的东边。

［2］SpRL是“Spatial Role Labeling”（空间角色标注）的缩写；mSpRL是“Multimodal Spatial Role Labeling”（多模态空间角色标注）的缩写。

［3］SpaCE是“Spatial Cognition Evaluation”（空间认知能力评估）的缩写，有关SpaCE评测基准（Benchmark）的情况介绍，可访问SpaCE2024网页查询：https://2030nlp.github.io/SpaCE2024/。

［4］关于文本空间信息正误判别，可参看詹卫东等（2022），关于后五项任务的介绍，可参考SpaCE2024网站。另外，空间推理任务相关研究，还可参看针对英文的SpartQA（2021）。SpartQA是“Spatial Reasoning on Textual Question Answering”（空间推理文本问答）的缩写。

［5］这个网页展示了大语言模型完成参照成分找回任务的测试示例：https://github.com/d0ubtfire/LLM_Evaulation/tree/main/ 对比大模型 / 空间信息理解 / 缺失参照成分找回。

［6］除调研大量期刊论文和学位论文外，我们也考察了相关权威辞书中对空间表达异形同义现象的描写情况。主要是《现代汉语词典》以及像吕叔湘（1999）《现代汉语八百词》、侯学超（1998）《现代汉语虚词词典》、张斌（2001）《现代汉语虚词词典》等描写虚词（语法功能词）类的词典。这些辞书基本上没有从异形同义这个角度对方位义词语进行描写分析。吕叔湘（1999）描写了常用方位词和趋向动词的用法，没有收录“左、右、东、南、西、北”；收录了“旁”，没有收录“边”。张斌（2001）收方位词“上、下、前、后、里、内、中、外”，但没有收“左、右”。也没有收趋向动词。侯学超（1998）没有收录方位词和趋向动词等表方位义词语。

［7］我们整理了一个汉语空间方位义词语表，详见https://github.com/2030NLP/SpaCE2024/tree/main/data。

［8］关于该数据集制作情况，可参看https://2030nlp.github.io/Sp22AnnoOL/task1_guide.html。

［9］值得补充说明的是，在151个词对中，方位词词对108个；趋向动词词对43个。只编写出“异形同义”语料的词对23个（如“里—内、中—内、旁—外、前面—旁边、对面—附近、下—出、上来—进来、上来—出来、进来—回来……”），只编写出“异形异义”语料的词对20个（如“里—外、

里—旁、中—外、内—外、上面—旁边、前面—侧面、中间—附近……”）。前者的典型词对是表空间义词语中的“同义词”，很难构造异形异义句对语料；后者的典型词对是“反义词”，很难构造异形同义句对语料。

［10］可参看 https://2030nlp.github.io/SpaCE2023/。

［11］查看全部参赛系统成绩榜，可访问网页：https://2030nlp.github.io/SpaCE2024/leaderboard.html。

［12］大语言模型对同一道题，生成的答案具有一定随机性。我们在 SpaCE2024 基准的每个子任务中都加入了 30 道“重复题”（包括题目和选项完全重复、题目不变但选项换序等形式），用于评估大模型的稳定性。在“异形同义判别”子任务上，排名前 6 的大语言模型，在“重复题”上的平均稳定性为 0.59（可以理解为 100 道重复题，只在其中 59 道题上，大语言模型的答案，无论对错，都保持稳定不变）。

［13］我们尝试考察大语言模型在不同词对题上的表现差异及可能的影响因素，但基于现有的题目和数据量，很难得出可靠的结论。从前 6 名系统的测试结果中，我们抽取了在异形同义题和异形异义题上表现均相对较好（平均正确率大于 0.6）的词对，分别是 23 个和 13 个，其中交集词对有 6 个：“上面—里面（9 题）、上 v—下 v（7 题）、下面—里面（13 题）、上—中（13 题）、上—里（18 题）、上去—下去（7 题）”，观察模型在这些词对题上的具体表现，并没有发现明显的规律。比如尽管“上 v—下 v”总成绩相对靠前，但对于上文例（17）的题，6 个模型中只有 2 个答对（正确率 33.33%），很难说大模型对“上 v—下 v”这对趋向动词的用法和语义理解掌握得比其他词对更好或更差。

［14］C2 类 2 道异形异义题是“例外”，人类成绩低于机器成绩，其中 1 名被试全部答错，得 0 分。

参考文献

1. 储泽祥，王寅 . 空间实体的可居点与后置方位词的选择 . 语言研究，2008（4）：50-62.

2. 方经民 . 汉语“左”“右”方位参照中的主视和客视——兼与游顺钊先生讨论 . 语言教学与研究，1987a（3）：52-60，154.

3. 方经民 . 现代汉语方位参照聚合类型 . 语言研究，1987b（2）：3-13.

4. 高桥弥守彦 . 是用“上”还是用“里”. 语言教学与研究，1992（2）：47-60.

5. 郭锐 . 方位词“前、后、左、右”的参照策略 . // 黄正德主编 . 中国语言学论丛（第三辑）. 北京：北京语言大学出版社，2004：1-30.

6. 侯学超 . 现代汉语虚词词典 . 北京大学出版社，1998.

7. 李敏 . 现代汉语主宾可互易句的考察 . 语言教学与研究，1998（4）：51-59.

8. 廖秋忠 . 空间方位词和方位参考点 . 中国语文，1989（1）：9-19.

9. 林笛（李平）. 汉语空间方位词的语用考察 . // 北京大学汉语语言研究中心《语言学论丛》编委会编 . 语言学论丛（第十八辑）. 北京：商务印书馆，1993：3-37.

10. 刘宁生 . 汉语怎样表达物体的空间关系 . 中国语文，1994（3）：169-179.

11. 吕叔湘 . 现代汉语八百词 . 北京：商务印书馆，1999.

12. 齐沪扬 . 空间位移中主观参照“来 / 去”的语用含义 . 世界汉语教学，1996（4）：56-65.

13. 孙陈亦 . 是什么让“前”与“后”的对立消失，待刊 .

14. 童小娥 . 从事件的角度看补语“上来”和“下来”的对称与不对称 . 世界汉语教学，2009（4）：495-507.

15. 肖力铭,孙春晖,詹卫东,等 . SpaCE2022 中文空间语义理解评测任务数据集分析报告(A Quality Assessment Report of the Chinese Spatial Cognition Evaluation Benchmark). //*Proceedings of the 22nd Chinese National Conference on Computational Linguistics*. Harbin, China: Chinese Information Processing Society of China, 2023: 547-558.

16. 肖力铭,詹卫东,穗志方,等 . CCL23-Eval 任务 4 总结报告:第三届中文空间语义理解评测(Overview of CCL23-Eval Task 4: The 3rd Chinese Spatial Cognition Evaluation). //*Proceedings of the 22nd Chinese National Conference on Computational Linguistics* (*Vol 3: Evaluations*), 2023: 150-158.

17. 詹卫东,孙春晖,岳朋雪,等 . 空间语义理解能力评测任务设计的新思路—SpaCE2021 数据集的研制 . 语言文字应用, 2022(2): 99-110.

18. 张斌 . 现代汉语虚词词典 . 北京:商务印书馆, 2001.

19. 张其昀 . 运动义动词“上”、“下”用法考辨 . 语言研究, 1995(1): 37-43.

20. 中国社会科学院语言研究所词典编辑室编 . 现代汉语词典(第 7 版). 北京:商务印书馆, 2016.

21. Clark H H. Space, Time, Semantics and Child. // Moore T E.(ed.) *Cognitive Development and the Acquisition of Language*, New York: Academic Press, 1973: 27-62.

22. Herskovits A. *Language and Spatial Cognition: An Interdisciplinary Study of Prepositions in English*. Cambridge: Cambridge University Press, 1986.

23. Kolomiyets O, Kordjamshidi P, Bethard S, *et al*. Semeval-2013 Task 3: Spatial Role Labeling, *Proceedings of the Seventh International Workshop on Semantic Evaluation* (*SemEval 2013*), 2013: 255-262.

24. Kordjamshidi P, Rahgooy T, Marie-Francine M, *et al*. CLEF 2017: Multimodal Spatial Role Labeling (mSpRL) Task Overview. *International Conference of the Cross-Language Evaluation Forum for European Languages*, 2017.

25. Lakoff G. *Women, Fire and Dangerous Things: What Categories Reveal about the World.* Chicago: University of Chicago Press, 1987.

26. Mirzaee R, Faghihi H R, Ning Q, *et al*. SPARTQA: A Textual Question Answering Benchmark for Spatial Reasoning. //*Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021: 4582-4598.

27. Pustejovsky J, Kordjamshidi P, Moens M F, *et al*. SemEval-2015 task 8: SpaceEval. // *Proceedings of the 9th International Workshop on Semantic Evaluation*, 2015: 884-894.

28. Talmy L. *Toward a Cognitive Semantics: Concept Structuring Systems*. Cambridge: MIT Press, 2000.

29. Xiao Liming, Nan Hu, Weidong Zhan, *et al*. Overview of CCL24-Eval Task 3: The Fourth Evaluation on Chinese Spatial Cognition. https://github.com/2030NLP/SpaCE2024/tree/main/docs/Overview of SpaCE2024.pdf. 2024.

(北京大学中文系 北京 100871)

(责任编辑 刘 博)

CISHU YANJIU
LEXICOGRAPHICAL STUDIES
September, 2024

Abstracts of Major Papers in This Issue

Evaluating Machine Capability in Spatial Cognition Through Synonymous Variants

Zhan Weidong　Qin Yuhang　Xiao Liming

Abstract: In Chinese, different spatial locative expressions can describe the same spatial scene. This phenomenon of "synonymous variants", which are caused by different factors, can be categorized into various types. As part of the SpaCE benchmark series designed to evaluate machine spatial semantic understanding, a subtask known as "Synonymous Variants Discrimination" is created. Evaluation results on large language models show that there is still a significant gap between their performance on this task and human-level performance. Moreover, the characteristics of machine performance on different types of test items are noticeably different from those of humans. From the perspective of spatial cognitive schemas, the human language knowledge acquired by large language models based on symbol distribution has not yet transformed into human-like understanding capability for spatial information cognition.

Keywords: spatial expressions, spatial cognition, different forms with the same meaning, machine language ability evaluation, large language models

A Survey on Automatic Compilation of Chinese Dictionary Entries: A Case Study of ChatGPT

Zhang Yongwei　Liu Ting

Abstract: This article investigates the performance of large language models in the automatic compilation of Chinese dictionary entries, using ChatGPT as an example. The study selects 40 word samples covering different dimensions such as the part of speech, word length, and number of senses. ChatGPT is used to generate explanations, which are then compared and analyzed with entry definitions in the 7th edition of the *Modern Chinese*