中文空间语义理解评测数据集质量评估研究 *

岳朋雪[1]　王诚文[1]　孙春晖[2]　詹卫东[2]　穗志方[1]

（1. 北京大学计算语言学教育部重点实验室　北京　100871；
2. 北京大学中国语言学研究中心 / 中文系　北京　100871）

[摘要] 中文空间语义理解能力评测（SpaCE2021）可以看作类人机器语言能力评测的一项重要尝试，其通过空间方位义词语替换的办法生成大量富含空间义信息的语料，构建了中文空间语义理解评测数据集。本文从生成题目的原句情况、可替换词与替换词的结构类型、测试集题目答案的正误分布及空间类型等四个方面分析了中文空间语义理解能力评测数据集的特点，进而通过人类被试和参赛系统的比较，详细分析了机器在不同类型空间词汇上的性能表现，总结了机器空间语义理解的一般规律，并从题目的无偏性和延续性两方面入手，为构建高质量中文评测数据集提出了具体建议。以上工作有助于提升空间语义评测数据集的质量，从而更好地提升相关评测任务的准确性和可靠性。

[关键词] 中文空间语义理解；数据集；机器表现；NLP 评测

[中图分类号] H087 [文献标识码] A [文章编号] 1003–5397（2023）01–0101–13

DOI:10.16499/j.cnki.1003-5397.2023.01.006

Research on Quality Evaluation of Chinese Spatial Semantic Understanding Evaluation Dataset

YUE Pengxue, WANG Chengwen, SUN Chunhui, ZHAN Weidong, SUI Zhifang

Abstract: Chinese Spatial Semantic Comprehension Ability Evaluation (SpaCE2021) can be regarded as an essential attempt to evaluate human-like machine language ability. We generate a large amount of corpus rich in spatial meaning by replacing words with spatial orientation meanings and constructs a Chinese spatial semantic understanding evaluation dataset. In this paper, we analyze the characteristics of our proposed dataset from four aspects: the original sentences of the generated test samples, the structure types of replaceable words and replaceable words, the spatial orientation types of the test set samples, and the correct and wrong distribution of the samples. Then, through the comparison of human and participating machine systems, the performance of machines on different types of spatial vocabulary is analyzed in detail, and the general rules of machine spatial semantic

[收稿日期] 2022–03–11

[作者简介] 岳朋雪，北京大学计算语言学研究所博士后，主要研究计算语言学、社会语言学；王诚文，北京大学计算语言学研究所博士后，主要研究计算语言学、语言工程；孙春晖，北京大学中文系博士生，主要研究计算语言学、语言知识工程、自然语言处理技术评测；詹卫东，北京大学中文系教授，主要研究计算语言学、现代汉语语法和语言知识工程；穗志方（通讯作者），北京大学计算机学院教授，主要研究自然语言处理。

* 本研究得到国家科技创新 2030 “新一代人工智能”重大项目（2020AAA0106701）的资助。

understanding are summarized. A high-quality Chinese spatial semantic evaluation data-set makes specific recommendations. The above work helps improve the quality of spatial semantic evaluation datasets, thereby improving the accuracy and reliability of related evaluation tasks.

Keywords: Chinese spatial semantic comprehension; data set; machine performance; NLP evaluation

一 引言

自图灵测试（Turing test）以来，人们开始探索如何提升机器的智能水平。尽管现在的语言模型能力越来越强，机器在一些复杂的语义理解、因果推理和常识知识方面仍表现不佳。基于认知神经科学的发展，神经人工智能（Neuro AI）领域提出扩展的“具身图灵测试（The Embodied Turing Test）”挑战，以实现真正的类人甚至超人的人工智能体（Zador et al., 2022）。空间认知能力是人类智力的基础能力和重要组成部分，面向计算机的空间语义理解任务不仅是 NLP 领域的基础课题，更是人工智能跨越计算智能、感知智能和认知智能三个阶段的综合性任务，可以看作是具身图灵测试的一种尝试。

然而，从数据规模、空间语言表达的真实度和自然度、空间语义理解能力的层次性等多方面来看，已有的空间语义理解任务还存在一定的局限性，且已有评测多集中在英文领域，面向中文文本的空间语义理解能力评测任务还不多见（詹卫东等，2022）。为了更好地对照人类语言认知，提升机器理解中文深层语义的能力，我们提出了面向机器的深层次语义理解任务——中文空间语义理解能力评测（SpaCE2021）[①]。作为类人机器语言能力评测的一项重要尝试，其通过空间方位义词语替换的办法生成大量富含空间义信息的测试语料，构建了中文空间语义评测数据集，并以判断替换句空间语义异常与否及异常归因的任务形式来考察机器空间语义理解能力，具体任务形式如表 1 所示。

表 1 SpaCE2021 子任务

子任务	解释	例题
1. 中文空间语义正误判断	判断给定中文文本是否存在空间关系异常	context: 孙萍不在办公室。这让孔太平感到有些束手无策。本来可以马上回到车上，但他在楼后多待了一会，才出来…… judge: false
2. 中文空间语义异常归因合理性判断	判断给定归因（搭配不当、语义冲突、信息冲突、不符合常识）是否可以用来解释给定中文文本中所存在的空间关系异常	context: 何俊英不知从哪儿钻了出来，连唱带笑跑到儿子跟前转一圈，就开始上人群侧跳舞去了。 reason: “人群”和“侧”不宜搭配 judge: true
3. 中文空间语义判断与归因联合任务	首先判断给定中文文本是否存在空间关系异常，若存在异常，则继续判断给定归因是否可以用来解释该异常	context: 老三想把欢迎会弄得热热闹闹的，一个劲往里让着街坊：“进去吧，外面请，到院子里头喝一盅。” judge：false reason: “外面请”和“到院子里头”语义冲突

数据集作为评测任务的载体，是各种自然语言处理技术或方法得以进行评估的基石（Schlangen，2020；Madaan et al.，2021；Kiela et al.，2021），对评测的准确性和可靠性有着至关重要的影响（董青秀等，2021）。然而，现有研究更多重视如何提升模型，忽视分析数据集本身质量（Sambasivan et al.，2021；Bowman & Dahl，2021；Geiger et al.，2020），现有评测数据集还存在一些问题。Wang 等（2019）研究发现，CoNLL03 的命名实体识别数据集（测试集）中存在 5.38% 的标注错误，这些错误将会影响评测结果的可信性。同时，

在自然语言推理、常识推理和阅读理解的数据集中，研究者也发现大量偏差问题，严重影响了评测的可信性（Geva et al.，2019；Zhou & Bansal,2020；Saxon et al.，2021；Sugawara et al.，2020）。为了探索怎样的题目可以更好地评测机器空间语义理解能力，本文以 SpaCE2021 子任务 1（中文空间语义正误判断）的测试集作为研究对象，分析数据集本身特点以及参赛系统的表现，总结构建高质量中文空间语义理解评测数据集时应该注意的问题，以提升数据集质量，更好地提升相关评测任务的准确性和可靠性。

二　空间语义理解能力评测数据集（SpaCE2021）题目特点

（一）生成题目的原句数量

SpaCE2021 任务 1 测试集（以下简称“测试集”）共有 794 条题目，由分别替换 61 个原句[②]中的 154 个空间义词汇[③]构造而来。以表 2 中的原句为例，制作数据集时选定句中表示空间关系的“外面”和“后面”两词作为可替换词，分别将“外面”替换成 9 个形式相近的空间词，将“后面”替换成 6 个形式相近的空间词，每次只替换一个词，这个原句替换后共生成了 15 个替换句作为 SpaCE2021 任务 1 的测试题目，这些替换句的空间语义存在正常或异常两种情况。

表 2　替换示例

原句	可替换词	替换词	替换句数量
……一大群人已围在外面……我和哥哥走在后面，互相搀扶着，极力使自己镇静下来。	外面	附近、里面、旁边、前面、上面、四面、下面、右面、左面	15
	后面	附近、里面、上面、四面、外面、下面	

通常来说，原句中空间义可替换词越多，产生的替换句越多，而测试集原句中的可替换词数量与替换句数量并不是绝对的正比关系，如数据集中产生替换句数量最多的原句（共生成 46 个替换句），并不是空间义词语最多的原句，空间义词语最多的原句（共有 8 个可替换词），生成的替换句数量也并不是最多的。可见，替换句的生成数量除了受原句中空间义词汇数量多少的影响，还与原句中空间义词汇的类型有关，如常用的“上、下、中、外、里”等方位词和“出去、过来”等趋向动词，因有相对较多的形式相近、意义相关的替换词语而生成了较多的替换句。据统计，数据集中单次替换生成的句子数量最多的可替换词为“上”，同一原句同一位置的“上”被替换 16 次，生成 16 个替换句。

（二）题目答案的正误分布

测试集答案的正误分布如图 1 所示，题目的正误分布比例较为平衡，没有出现明显偏差。我们进一步统计了答案为“FALSE”的题目空间关系异常的几种类型，具体如表 3 所示。

表 3　答案为“FALSE”的题目空间异常类型

异常类型	数量	比例
方位词与名词搭配错误	217	0.505
动词位置使用了方位词	40	0.093
方位词汇表达的空间关系与上下文矛盾	173	0.402

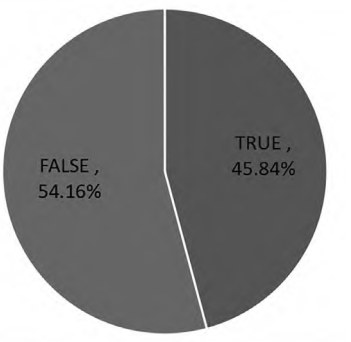

图 1　测试集题目答案的正误分布

测试集中，名词和方位词构成的空间类型题目有 381 道，几乎占到测试集题目的一半数量，其中 217 题有名词和方位词搭配不当问题，可见，名词与方位词的搭配不当是 SpaCE2021 任务 1 空间关系异常题目的主要类型。

现代汉语中，“上、下”既是动词也是方位词。数据集中由“上”替换产生的题目有 92 道，占题目总数的 11.59%，其中有 40 句由动词“上”替换成了方位词，这些替换句是明显的空间关系异常题目。如原句 1 中的动词“上”生成替换句时，不仅被替换为“回、进”等动词，还被替换成“内、侧、底、顶、前、后”等方位词，这些错误替换给参赛系统的机器表现了造成了噪声干扰。

原句 1：史婆婆回过头来，对白自在道：“你要是伤了我徒儿性命，我这就上碧螺山去，一辈子也不回来了。”白自在大怒，叫道：“你……你说去哪里？”

原句 2：那少女脸上微微一红，随即现出怒色，将瓷碗往桌上重重一放，转过身去，把铺在房角里的席子、薄被和枕头拿了起来，向房门走去。

原句 2 中，由“铺在房角里”可知“席子、薄被、枕头”位于一个较低的空间位置，因其后出现“拿”和“走”表示前后两个连续动作，“拿”后应该使用一个表示垂直方向的趋向动词“起来”，表示人拿着席子、薄被、枕头一起向房门方向走去。而测试题目中将“起来”替换成“过来”后，动作由垂直方向变成由远及近的动作轨迹，即将席子、薄被、枕头从房角拿到少女位置，与后面的“向房门走去”语义冲突。

（三）可替换词与替换词的类型

1. 可替换词的分布特点

图 2 统计了原句中 154 个可替换词的类型，从词类角度来看，指示代词和绝对方位词类可替换词较其他类型的空间词汇数量少。不仅不同类型空间义可替换词数量有所区别，同一类型的空间义词汇内部在数量分布上也存在差距。以单纯方位词为例，61 个原句中共有 15 个处于句子不同位置的方位词“中”被替换，而“后”和“外”分别只有 3 个被替换，也就是说，虽然同是方位词，“中”生成的替换句数量要远多于“后”和“外”的替换句。

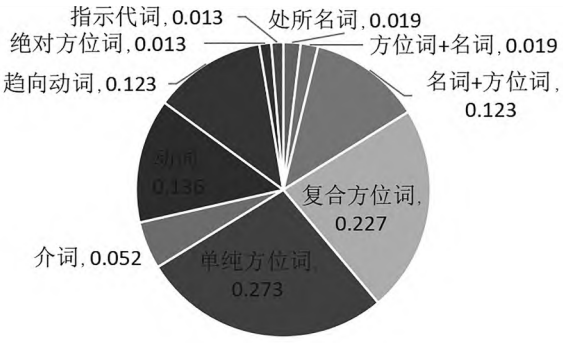

图 2　原句中的可替换词类型

2. 替换词的分布特点

图 3 显示了测试集题目中替换后的空间结构类型（以下简称“替换词”）。从替换词的空间方位类型来看，无论是方位词还是动词，均覆盖了水平、垂直和三维等常见空间维度；由指示代词构成的方位结构，覆盖到近指和远指两种不同的方位指称关系；由方位词和名词构成的空间结构中，名语素涉及了“山、江、水”等自然空间中的实体名词，“头、身、手”等身体部位名词，“桌、门”等物体名词以及“村、房”等生活名词，基本覆盖到了人类生活中常见的几种描述空间关系的参照名词类别。

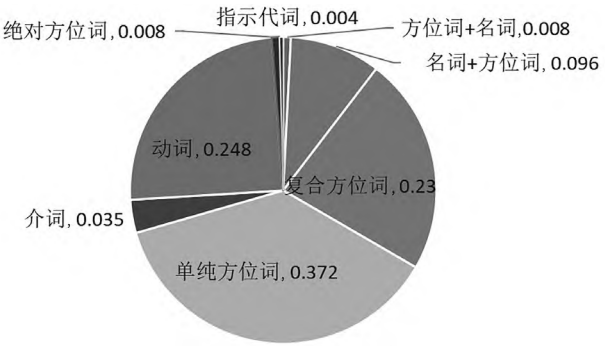

图 3　测试集题目替换词的类型

（四）测试集题目的空间义词汇搭配类型

汉语空间义词汇往往和相邻词语搭配构成空间语义表达结构，测试集中表示空间关系

的结构可以归纳为表 4 的几种类型。

表 4　测试集题目中的空间方位结构类型

空间方位结构	示例
方位（名）词 + 名词	前大门、身旁的两个汉子
方位（名）词 + 动词	四面只有一张小桌子、右臂伸出
名词 + 方位词	山门后、村前
名词 + 方位词 + 方位词	山外下
动词 + 方位（名）词	去四周、指着附近
动词 + 趋向动词	塌进去、摔上房外
介词 + 方位（名）词	于这边、到东行
指示词	这儿、那边

（五）中文空间语义正误判断数据集的特点

综合以上四个层面的分析可知，SpaCE2021 任务 1 通过替换句中空间词汇的方式生成了一定规模的替换句，运用较小成本构建了空间语义丰富的中文空间语义正误判断数据集，测试集题目在答案的正误分布上基本平衡，题目中的替换词虽然在具体的空间方位结构类型分布上有差别，但已基本覆盖了现代汉语词汇中常用的几种空间关系表达结构，作为测试题目能够较为全面地考查机器理解中文不同空间关系的能力。

三　空间语义理解能力评测中机器表现的特点

参加 SpaCE2021 评测任务的 8 个参赛系统全部采用了大规模预训练语言模型，6 个获奖系统很少使用预训练模型以外的技术或知识资源，只有一个参赛系统根据自行建立的方位词表对上下文进行了裁剪，从而降低模型过拟合率，但该做法并未带来明显提升。表 5 展示了人类被试和参赛系统在 SpaCE2021 任务 1 上的表现。人类被试的平均准确率更高，人类最高水平也明显超过参赛系统最高水平和基线系统，可见，机器在理解中文空间语义时与人类还存在一定差距。为更好地观察机器理解中文空间语义的能力，本文深入分析机器在测试集不同空间类型题目上的表现[④]，以更加细致地了解机器空间语义理解的特点。

表 5　人类和机器在测试集上的不同表现

类别	基线系统	机器水平			人类水平		
		最高	最低	平均	最高	最低	平均
准确率	0.673	0.734	0.631	0.695	0.849	0.600	0.715

为更细致地分析机器在不同类型空间义词汇上的表现，结合替换词表的分布，除介词和动词外，本文将名语素与方位语素组合而成的空间结构归为处所词，将“东南西北”等表示绝对方向的方位词归为“绝对方位词”，将“东南西北”以外的单音节方位词归为单纯方位词，将“东南西北”以外的由两个方位语素构成的方位词归为复合方位词，将包含指示代词“这、那”的词归为指示词。测试集中各类空间词汇题目分布比例及机器判断的平均正确率如表 6 所示。

表 6　测试集中不同空间方位类型题目的机器表现

词类	处所词	绝对方位词	单纯方位词	复合方位词	指示词	动词	介词
题目占比	0.103	0.008	0.372	0.231	0.004	0.249	0.071
机器平均正确率	0.617	0.945	0.779	0.657	0.630	0.641	0.540

整体来看，参赛系统在判断绝对方位词题目的空间语义时表现最好，单纯方位词次之。在判断介词题目的空间语义时表现最差，平均正确率刚刚达到半数，在判断处所词、复合方位词、指示词和动词的题目时，机器表现差距不大。总之，机器的表现会受到题目数量、词语的空间关系复杂程度等多方面影响，下文将具体分析机器在各类空间义词汇上的表现。

（一）在绝对方位词题目上的表现

机器在判断包含绝对方位词题目的空间语义时，表现出非常高的准确率。测试集中共有 6 个题目的替换词为绝对方位词，它们由原句 3 和原句 4 生成，分别由句中的“西”和“东南”替换成“东、北、南”和“东北、西北、西南”而来。因“东南西北”类方位词表示的方向是自然界中固定的，不会随着时间、地点等发生变化，即使句中空间实体的位置不明确，也不影响人们理解其所表示的空间关系。“东西南北”等绝对方位词作为汉语空间系统的基础词汇，在中文文本中高频出现，以预训练为主的模型容易从大规模文本中学到相关知识，对该类词汇的空间语义理解较为准确。

原句 3：这时四下里呼哨声均已止歇，……各人凝气屏息之中，只听得一个人喀、喀、喀的皮靴声，从西边沿着大街响将过来。

原句 4：入夜之后，小舟转向东南。在海中航行了三日，小船中只有些干粮清水，石破天和那船夫分食……

整体来看，绝对方位词内部替换后形成全新的空间关系，虽然替换后的句子空间语义与原句明显不同，但空间关系并无异常，因此 6 个题目的答案均为“TRUE”。如原句 3 和原句 4 中的绝对方位词被替换后，空间实体的位移方向虽发生了与原句不同的明显变化，但表义依然完整和准确，句子的空间语义成立，机器几乎都能做出正确判断。

结合参赛系统在 SpaCE2021 任务 1 上的表现可以看出，机器在判断绝对方位词题目的空间语义时呈现以下特点：机器在理解“东南西北”等绝对方位词的空间语义时，正确率很高。这与此类方位词表示的空间关系固定不变、不受空间主体的影响有关。

（二）在其他方位词题目上的表现

作为测试集中出现最多的空间义词类，由方位词替换生成的题目占比超过了 60%。整体来看，单纯方位词题目的机器平均正确率高达 0.779，是除绝对方位词以外机器平均正确率最高的一类词。

从题目来看，方位词的机器平均正确率高可能与机器识别语言结构“搭配不当”能力较强有关。构建数据集时，因题目由替换相似语境中的空间方位词得来，替换过程中生成了一些错误的名词和方位词结构，如“寓所下、底大门”等结构中的名词本身不具备方位词所表示空间关系，或如“石壁左、凌霄城左、小船左”等结构中的单音节方位词应该替换为双音节结构才可以正常搭配，无论是人类还是参赛系统，都可以对这类方位词题目准确给出“FALSE”的判断。

“方向”是现代汉语的重要空间系统之一，汉语中一般使用方位词表示方向，与之搭配的名词则为方向的参照点。为更好地观察机器理解汉语方位词所表示的空间语义关系，本文继续分析了参赛系统判断不同空间维度方位词的情况。表 7 可见，参赛系统在判断表示垂直方向的方位词题目时平均正确率最高，略高于水平方向方位词的正确率，而系统在判断更为复杂的三维空间方位词所在题目时表现最差。

结合参赛系统在 SpaCE2021 任务 1 上的表现可以看出，机器在判断方位词题目（非绝对方位词）的空间语义时呈现以下特点：机器能够很好地判断名词和方位词搭配不当题目的空间异常；从方位词内部来看，机器在判断表示垂直和水平等维度简单的方位词题目时

表 7　测试集中方位词题目及机器表现

空间维度	垂直方向	水平方向	三维空间
方位词	上、下、底、顶、上边、上面、上下、下边、下面、之下	边、侧、前、后、左、右、之左、之右、左边、左面、右边、右侧、右面、旁边、前边、前侧、前后、前面、两边、隔壁、后边、后面、后侧、边上、侧方、侧面	中、外、内、附近、里边、里面、内外、四面、四周、外边、外面、正中、之内
平均正确率	0.744	0.714	0.665

正确率高于三维空间等复杂空间方位题目。可见，机器的正确率与方位词的空间语义复杂程度相关，词语表示的空间语义越复杂，机器越难理解。

（三）在介词题目上的表现

“到、朝、向、经、从”等介词与名词性结构形成了“介词 +NP”式的动态空间关系，其中 NP 可表示位移的起点、终点或途经点，形成了不同方向的位移路径。SpaCE2021 中，参赛系统在替换词为介词的题目上表现较差，因由动词演化而来的介词语义比较虚化，逐渐表示范围、空间等抽象语义，与动词之间的语义有一定模糊性，加大了机器理解的难度。介词题目答案的正误分布及机器平均正确率如表 8 所示，机器在答案为“FALSE”的题目上正确率要低于答案为“TRUE”的题目。

表 8　测试集中介词题目分布及机器表现

题目答案	TRUE	FALSE
介词题目占比	0.75	0.25
平均正确率	0.68	0.46

为更好地了解机器在哪些介词题目上表现较好或较差，图 4 逐个统计了机器在不同介词题目上的正确率。可见，参赛系统在判断“距、朝、往、向”等引出终点（目的地）的介词题目时，正确率明显高于“于、从”等引出起点（出发点）的介词题目。

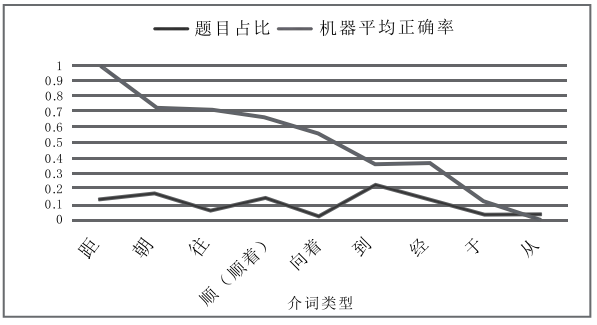

图 4　测试集中不同介词题目的机器表现

结合参赛系统在 SpaCE 2021 任务 1 上的表现可以看出，机器在判断介词题目的空间语义时呈现以下特点：机器理解介词空间语义时表现较差，并在表示不同位移方向的介词题目内部表现出明显差异；机器在判断引出终点的介词题目时，正确率明显高于引出起点的题目。出现这种情况可能有两个原因：一是机器能够较好地识别题目中的“终点”角色，但不能很好地识别“起点”角色；二是数据集中“从、于”等引出起点角色的介词题目数量少，现有结果还需要更多测试题目去验证。

（四）在空间指示词题目上的表现

指示代词“这、那”后加“儿、里、边”等词表示处所，表示空间关系时需要以说话人为参照点去判断方位关系。“这”类词表示距离说话人较近的位置，“那”类词表示距离说话人较远的位置。如果句中说话人位置信息缺失，“这”和“那”互换后通常不影响语义表达。如题 1 中的“那边”可替换为“这边”，因为不确定说话人位置和观测点，替换后空间关系成立，答案为“TRUE”，机器平均正确率为 1，即全部参赛系统都给出“TRUE”的判断。题 2 也如此，因他、龙岛主和说话人的位置都不确定，“他”在“这里”或是“那里”在句中都成立，二者可能指同一个位置，也可能是不同位置，并不影响语句的正常表达，因此答案也为“TRUE”，绝大多数参赛系统给出了“TRUE”的判断。而题 3 中，“我”去的地方应该是远离说话地，需使用“那儿”，不能替换为“这儿”，因此题 3 中的空间关系不成立，答案为“FALSE”，但机器判断正确率仅为 0.111，即绝大多数系统都给出了“TRUE”的错误判断。虽然测试集中指示词题目很少，但从仅有的三个指示词题目来看，参赛系统在判断指示词替换产生的题目时，倾向于给出“TRUE”的答案。可见，机器还未能准确把握句中的视角信息。

题 1：马老五坐在朱四边上的一桌，一直冷冷地朝那边看着，看了一阵子之后就开始站了起来……

题 2：他在这里自怨自艾，龙岛主以后的话就没怎么听进耳中。龙岛主说的是：“四十年前，我和木兄弟订交，意气相投，本想联手江湖……”

题 3：“……要是钱先生又让人给逮了去，日本人准会把明月留在庙里当诱饵，好逮老三和别的人。我上这儿去很不方便，你敢不敢去走一趟？”

结合参赛系统在 SpaCE 2021 任务 1 上的表现可以看出，机器在判断指示词题目的空间语义时呈现以下特点：机器在判断指示词题目时倾向给出单一答案，还未“拥有”句中空间主体和说话人的视角，不能准确判断指示词所指代的方向。

（五）在空间义动词题目上的表现

测试集中的空间义动词多为趋向动词，包括“进、出、上、下、来、去”等单音节动词以及由其扩展而来的双音节动词，表示实体的位移动作或方向。因“上、下”等具备动词和方位词的双重属性，导致题目制作过程中出现了部分方位词位置的“上”和“下”被替换为动词性质的“过、回、进”等，因这些替换后的词与名词构成的空间方位结构在语言交际中不存在，参赛系统的平均正确率高达 0.940，剔除这类明显异常的题目后，参赛系统判定动词题目空间语义正误的平均正确率为 0.600。表 9 显示了机器在判断正确替换后的动词题目上表现的特点。

表 9　不同类型动词题目的机器表现

动词题目类型		平均正确率	例题
音节	单音节动词	0.591	他们先去南方广州，到农村后，爹爹不是光听汇报，他执意要亲自到农村去，一直走过农民的家中，和农民直接对话。
	双音节动词	0.612	金三爷也在台阶上坐了出来。他忍住气，静下心来思索。
语用	单独作谓语的动词	0.585	曹文生坐了 23 天班房后，提着一只灰色旧包裹回到曹家桥家里。妻子、两个女儿一见文生上去了，以为是在梦里……
	与趋向动词搭配使用的动词	0.618	史婆婆回过头来，对白自在道：“你要是伤了我徒儿性命，我这就上碧螺山去，一辈子也不过去了。”
方向	由近及远	0.587	这句意想不到的话似一幢楼塌上去压在我们头上，把我们全压垮了。
	由远及近	0.640	里面危险，别回来。

从表 9 来看，系统在判定表示由远及近的“来”和由近及远的“去”两类不同趋向动词题目时，平均正确率差别不是特别明显。但机器在判断“来”和“去”两词互相替换产生的句子时，表现出高度一致性，即无论句中“来”或“去”的使用是否符合参照系要求，参赛系统都倾向判断其空间语义正确。如题 4 所示，说话人与“老总”处于同一位置，“我们”在“上面防空洞那里”，说话人使用“我们”一词，表示以上面防空洞的视角来判定“老总”的位移方向，因此使用“来”空间语义也成立，答案为“TRUE”，参赛系统也给出“TRUE”的判断。而题 5 中，前句可以判断出说话人位置与“别人”一致，与“彭总”不同，“参谋”的位置与“彭总”也不同，应该是从说话人的位置出发去叫“彭总”，因此应当用“去”而非“来”，句子答案为“FALSE”，但参赛系统却给出了“TRUE”的判断。可见，机器还未完全学习到说话人的视角位置。

题 4：“老总，拿到上面防空洞里去了，我们都在那里等着你来研究作战方案呢。”

题 5：别人都进了洞子，就是彭总不来。参谋来叫了几次，他还是不肯出来。再去叫，说不定他又要发火了。邓华说：“老洪，彭总老和你开玩笑，你去叫吧。”

题 6：正在山上对着白云唱歌的薄平看到远远的山坡上走来的王实味，慌忙躲进一个山洞里，王实味漫山遍野里找呀，喊呀，一直折腾到天黑，薄平就是不过去。

复合趋向动词可以直接表示主体的动作，也可以构成“V+ 趋向动词”格式，表示主体动作方向，其空间语义的判断以“说话人位置 + 事物位置”作为参考位置。在“V+ 趋向动词”格式中，动词的语义特征决定了主体位移的方向，限制了其后趋向动词的范围。如动词“塌”的义项为“下陷”，其路径方向是从较高的位置向较低位置移动，因此，“塌”后出现的趋向动词多为表示垂直方向移动的“下”类词。机器在判定这类词时虽然一致性较高但正确率不稳定，如参赛系统一致判定“这句意想不到的话似一幢楼塌进去压在我们头上”的空间语义为“FALSE”，正确率很高，但在判定“这句意想不到的话似一幢楼塌出来压在我们头上”的空间语义时一致判定为“TRUE”，正确率又很低。又如，当趋向动词直接表示主体的动作，单独作谓语或谓语中心时，其空间关系涉及动作主体和路径的起点、终点等相对复杂的参照系，系统在判定句子空间语义时表现不稳定。如题 6 中“过去”一词所在位置的趋向动词需要根据上下文中“王实味”和“薄平”所在位置，即山洞的空间属性来选择，而脱离前文语境单独说“薄平就是不过去”时，其空间语义是成立的，但是在题 6 中就不能使用“过去”，而应该使用“出来”。可见，目前机器还未能正确判断这种比较复杂的空间关系。

结合参赛系统在 SpaCE2021 任务 1 上的表现可以看出，机器在判断动词题目的空间语义时呈现以下特点：机器在判断动词题目时表现较差，在不同类型动词上的表现差别不是特别明显。表示空间主体位移方向的“来”和“去”类趋向动词，受到主体视角的影响和动词的语义限制，对机器来说理解难度较大，是机器理解空间语义的难点，也是提升机器在空间语义理解能力上达到类人水平的重要方面。

（六）在处所词题目上的表现

前文提到，表示处所的空间结构经常出现在人们的日常交际中，它们脱离上下文单独使用时所表示的空间关系比较简单，容易理解，一些明显异常的空间关系也很容易被人识别出来，但对机器来说要判断这些常识信息还比较困难。如，题 7 中的“脚上”表示“一条辫子”“垂”的处所，根据常识可知，“辫子”“垂”的位置通常是“头、背”等部位，而非“脚”的位置，因此该句的空间关系不成立，答案为“FALSE”，但是参赛系统判断的正确率很低。又如题 8 中，“门”和“青石板路”是两个不同的空间实体，二者的位置关系不

应该是“青石板路”位于“门上”，应该是相离关系，因此该题目的答案为“FALSE”，但系统却判定为“TRUE”，可见目前机器还未能对类似的常识性空间知识做出正确判断。

题 7：在第一次剧务会议上，说到影片里中国青年的造型时，副导演拿出一幅钢笔画给大家看。纸上画着一个中国男子，长衫布鞋，头戴瓜皮小帽，脚上垂着一条辫子。

题 8：……他无牵无挂，任意漫游，走到傍晚，前面树林中露出一角黄墙，行到近处，见是一所寺观，屋宇宏伟，门上铺着一条宽阔平整的青石板路，山门中走出两个身负长剑的黄冠道人来。

结合参赛系统在 SpaCE2021 任务 1 上的表现可以看出，机器在判断处所词题目的空间语义时呈现以下特点：机器在判断一些明显不符合常识的处所词题目空间异常时表现较差，因为机器还未能掌握与空间相关的常识信息，而这类常识信息正是人类所熟悉的，很容易做出判断，但机器判断的平均正确率很低。因此，这类富含常识信息的空间语义题目也是机器理解空间语义的难点，要想真正提高机器的类人水平，需要探索如何提高机器理解人类生活常识和自然界常识问题的水平。

（七）在不同类型空间词汇题目上的表现

综合上文分析，虽然测试集题目在不同类型空间词汇的分布上有所差异，但从参赛系统在现有题目的平均正确率上还是可以观察到当前阶段机器空间语义理解的特点，具体如表 10 所示。

表 10　中文空间语义理解的机器表现特点

空间题目类型	机器表现	影响因素
绝对方位词题目	好	空间语义固定、简单，不受外界影响
单纯方位词题目	较好	表示静态的空间关系，与名词搭配表示空间关系时，搭配不当问题较为突出，受语义复杂度影响
复合方位词题目	较差	表示静态的空间关系，受语义复杂度影响
处所词题目	较差	表示静态的空间关系，多涉及常识问题
介词题目	差	表示动态的空间关系，受空间主体的视角影响
动词题目	差	表示动态的空间关系，受空间主体视角影响和动词语义的限制
指示词题目	差	参与的空间主体较多，受空间主体视角、参照系等影响

整体来看，可以得出以下结论：（1）机器在表示静态的空间关系题目上的表现要好于表示动态位移关系的空间题目，而词汇本身表达的空间关系越简单，机器越容易理解，词汇表达的空间关系越复杂，机器越难理解。如涉及视角、参照系等复杂空间关系的指示词、动词和介词，机器判断还未能呈现出有序规律，表现出的理解能力还不够稳定。（2）在判断因搭配不当造成异常的空间题目时机器表现很好，但涉及常识问题，机器表现就变得较差。可见，机器在空间语义理解上还未能很好地达到类人水平。这就需要我们构建更好的测试集，设计更多有效的测试题目去探索和判断影响其空间语义理解的因素，总结机器中文空间语义理解的一般规律，以逐步提升机器在不同类型空间题目上的表现。

四　构建高质量评测数据集的一些建议

究竟什么样的题目能够真实考察机器的中文空间语义理解能力？而什么样的评测数据集可以更好地测试机器的类人能力？作为中文空间语义理解能力评测的初步探索，SpaCE2021 在构建数据集时进行了科学、大胆地尝试，以较小成本生成了包含丰富空间实

体和空间关系的替换语料（詹卫东等，2022），弥补了现阶段中文空间语义评测语料不足的情况，是专门用于中文空间语义评测的数据集。但从数据集题目构成和参赛系统的表现来看，现有数据集还不能全面考察机器的中文空间语义理解能力，其参赛系统表现还不能很好体现机器理解中文空间语义的真实水平。为此，我们以 SpaCE2021 评测数据集为基础，提出构建高质量中文空间语义评测数据集需要注意的问题，为语料规模小的中文机器评测数据集的构建提供有价值的参考。

（一）平衡题目类型分布，确保评测数据集的无偏性

评测数据集的无偏性要求真实、全面反映机器理解空间语义的能力，而不是在某些特定题目上的得分。以中文空间语义理解为例，构建数据集时应该注意考查机器细粒度空间语义理解题目的平衡。

从数据集的考察范围来看，高质量的中文空间语义评测数据集应覆盖现代汉语中所有的空间方位词汇，且数量要达到一定规模，避免因某些常用空间词类数量少而降低评测结果的可信度。某些空间词汇题目数量过少会使机器在判断该类词语空间语义时表现的特征不明显，或者虽在少量题目上表现出一定规律，但因题目数量有限而使结论的信度降低。同时，各类空间词汇内部具体词语的分布也要做到平衡。以 SpaCE2021 任务 1 测试集为例，数据集中包含指示词、介词和绝对方位词题目数量极少，仅占题目总量的 0.083。介词“从、向”是汉语中常用的引出起点和终点关系的介词，但均只在题目中出现一次，而同类的“到”则出现了 6 次。又如，替换词中单纯方位词虽然有 13 个，却并未涉及汉语中常用的单纯方位词“里”。为此，在数据集构建阶段，需要确保测试题目覆盖到更多、更全面的空间语义范畴。在选取语料时，须确保原句中的空间关系覆盖到所有常用的空间关系词类和具体的空间词语，力求可替换词中不同词类和词语的分布平衡。确定替换标准时，则需避免将可替换词替换为不同结构或词性的词，控制减少因搭配错误造成空间关系异常题目的数量。

无偏差的数据集还要求题目答案的正误分布平衡。对于中文空间语义评测来说，不仅要确保数据集答案的整体分布平衡，更要细化到确保不同空间方位类型题目答案分布均衡，避免出现某些空间方位词题目答案全部为“FALSE”或者“TRUE”的情况。如 SpaCE2021 任务 1 测试集中，某些替换对所对应的句子答案全部为“FALSE”或者“TRUE”，或某一答案类型的题目明显高于另一答案类型，有着明显的分布偏差，如表 11 所示。这就需要我们在构建数据集时应注意细粒度的答案分布平衡。

表 11　替换对答案正误分布偏差示例

替换对	TRUE	FALSE	总数
中→顶	0	26	26
中→底	3	18	21

此外，为保证数据集质量，选取测试语料时，应兼顾到不同题材范围，同时考虑语料的时代特征，选取年代较近的语料制做题目进行考察，避免出现一些文言色彩过浓或过旧的语料作为考察题目，使数据集更符合时下人类语言表达的特点。

（二）根据不同阶段评测任务特点，适时调整数据集题目难度

通常来说，评测任务具有延续性，因此用于评测的数据集需要有延续性，新一轮的评测数据集可以在现有数据集的基础上进行改进。中文空间语义评测是一个极具挑战性的评测任务，涉及到语言知识和常识知识，甚至视觉想象、运动规划等多模态的感知经验和认知经

验。因此，中文空间语义评测任务题目的设置在兼顾各类中文空间语义表达类型的同时，也需要具有一定的区分度，并随着不同阶段评测任务的特点调整难度。

从语义理解上看，由指示词、介词和动词构成的空间方位关系复杂，且容易受到上下文语境的限制，理解时需要确定不同空间实体的视角和参照点，人类在判断这类题目的空间语义关系时也存在一定难度，且参赛系统在这类题目上的正确率较低，是评测任务中难度较大的题目，较简单的“名词 + 方位词”结构更能考察机器深层次空间语义理解能力。因此，后续评测任务构建高质量中文空间语义理解数据集时，可以继续深入考察这类题目的机器表现特点，将其作为难度较高的题目类型去考察机器理解深层次空间语义的能力。

五　结语

评测数据集的质量是保证评测科学、有效的基石。本文以首届中文空间语义理解能力评测（SpaCE2021）数据集与参赛系统的表现为研究对象，分析了数据集的细粒度分布特征以及参赛系统在细粒度空间词汇类型上的表现，认为后续的中文空间语义评测任务须平衡数据集中不同空间方位词类及具体词语的数量以保证评测的信度和效度，并在确保数据集答案整体分布平衡的基础上，保持不同类型题目正误分布平衡，在复用本次机器表现相对较差题目的同时，选择更能从语义角度深层次评测机器空间语义理解的题目进行深入考察。

中文空间语义理解评测任务可以看作具身图灵测试在特定领域的一种尝试，面向特定领域的机器评测在面临语料规模小和不自然、不符合人类表达习惯的情况下，可以借鉴 SpaCE2021 的数据集构建方法，并注重从认知角度去评价和分析机器表现，探索机器理解中文深层语义的能力。

[附　注]

① 本次评测任务的具体情况见 SpaCE2021 官方网站：https://github.com/2030NLP/SpaCE2021。

② 本文将生成 SpaCE2021 数据集的原始语料称为“原句”，将替换后作为测试集题目的句子称为“替换句”。将原句中被替换的空间词语称为“可替换词”，将替换句（题目）中替换后的空间词语称为“替换词”。

③ 空间义词汇指现代汉语中表示空间关系的词语，包括方位词、处所词、指示词、趋向动词、介词等，具体分类可查询词表链接：https://2030nlp.github.io/SpaCE2021/words。

④ 下文提到的机器表现皆为 8 个参赛队伍在任务中准确率的均值。

[参考文献]

[1] 董青秀，穗志方，詹卫东，常宝宝 . 自然语言处理评测中的问题与对策 [J]. 中文信息学报，2021，（6）.

[2] 詹卫东，孙春晖，岳朋雪，唐乾桐，秦梓巍 . 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 [J]. 语言文字应用，2022，（2）.

[3] Bowman,S. R. & Dahl,G. E. What will it take to fix benchmarking in natural language understanding?[J].arXiv preprint arXiv:2104.02145, 2021.

[4] Geiger,R. S.,Yu,K.,Yang,Y.,Dai,M.,Qiu,J.,Tang,R. & Huang,J. Garbage in,garbage out? Do machine

learning application papers in social computing report where human- labeled training data comes from? [C]. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020.

[5] Geva,M.,Goldberg,Y. & Berant,J. Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets[J].arXiv preprint arXiv:1908.07898, 2019.

[6] Kiela,D.,Bartolo,M.,Nie,Y.,Kaushik,D. & Williams,A. Dynabench:Rethinking benchmarking in NLP[J].arXiv preprint arXiv:2104.14337, 2021.

[7] Madaan,A.,Mcmillan-Major,A.,Parikh,A.,Bosselut,A.,Anuoluwapo,A.,Majumder,B.P.,Emezue,C.,Garbacea,C., Kumar,D. & Das,D.The GEM benchmark:Natural language generation,its evaluation and metrics [J].arXiv preprint arXiv:2102.01672, 2021.

[8] Sambasivan,N., Kapania,S., Highfill,H.,Akrong,D. & Aroyo,L.M. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI[C].Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021.

[9] Saxon, M., Wang, X. & Wang,W.Y. Automatically identifying semantic bias in crowdsourced natural language inference datasets[J].arXiv preprint arXiv:2112.09237, 2021.

[10] Schlangen,D. Targeting the benchmark: On methodology in current natural language processing research[J]. arXiv preprint arXiv:2007.04792, 2020.

[11] Sugawara, S., Stenetorp, P., Inui,K. & Aizawa, A. Assessing the benchmarking capacity of machine reading comprehension datasets[C]. Proceedings of the AAAI Conference on ArtiFIcial Intelligence, 2020.

[12] Wang,Z.,Shang,J.,Liu,L.,Lu,L.,Liu,J. & Han, J. Crossweigh: Training named entity tagger from imperfect annota- tions[J].arXiv preprint arXiv:1909.01441, 2019.

[13] Zhou,X. & Bansal,M. Towards robustifying NLI models against lexical dataset biases[J].arXiv preprint arXiv:2005.04732, 2020.

[14] Zador,A.,Escola,S.,Richards,B.,Ölveczky,B.,Bengio,Y.,Boahen,K.,Botvinick,M.,Chklovskii,D.,Churchland,A.,Clopath,C.,DiCarlo,J.,Ganguli,S.,Hawkins,J.,Köerding,K.,Köulakov,A.,LeCun,Y.,Lillicrap,T.,Marblestone,A.,Olshausen,B.,Pouget,A.,Savin,C.,Sejnowski,T.,Simoncelli,E.,Solla,S.,Sussillo,D.,Tolias.A.S. & Tsao,D.Toward next-generation artificial intelligence:Catalyzing the Neuro AI revolution[J]. arXiv preprint arXiv:2210.08340, 2022.

（责任编辑　陈丽湘）