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

詹卫东 [1]　孙春晖 [1]　岳朋雪 [2]　唐乾桐 [1]　秦梓巍 [2]

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

[**摘要**] 中文空间语义理解评测任务（SpaCE2021）以判断题的形式呈现，包括 3 个子任务：（1）中文空间语义正误判断；（2）中文空间语义异常归因合理性判断；（3）中文空间语义正误判断与异常归因合理性判断联合任务。SpaCE2021 数据集包含 18000 多道判断题，语料规模超过 200 万汉字。相较于传统空间语义角色标注及空间推理任务，SpaCE2021 兼顾了语料的真实性、数据集构建方法的便捷性以及空间语义理解的挑战性，是中文空间语义理解专项评测领域新的尝试。

[**关键词**] 空间语义理解；空间语义异常；分类任务；空间语义句集；自然语言处理评测
[**中图分类号**] H087 [**文献标识码**] A [**文章编号**] 1003–5397（2022）02–0099–12

SpaCE2021: A New Benchmark for Chinese Spatial Language Understanding

ZHAN Weidong, SUN Chunhui, YUE Pengxue, TANG Qiantong, QIN Ziwei

Abstract: This paper introduces SpaCE2021, a Chinese Spatial Language Understanding Evaluation benchmark. It contains three sub-tasks: (1) single sentence judgement, for classifying Chinese sentences that contain spatial expressions as acceptable or unacceptable according to their normal or abnormal spatial semantics; (2) error explanation and sentence pair judgement task, for judging whether a reason can explain an incorrect sentence with spatial semantic

[收稿日期] 2022–01–10

[作者简介] 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；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；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；唐乾桐，北京大学中文系硕士生，主要研究中文信息处理；秦梓巍，北京大学计算机学院硕士生，主要研究多模态内容理解。

* 本研究得到国家科技创新 2030 “新一代人工智能”重大项目（2020AAA0106701）、国家自然科学基金项目（62076008、61936012）资助。北京大学计算语言学研究所穗志方老师、常宝宝老师、李素建老师、中央民族大学曾立英老师，以及两校语言专业和计算机专业多位研究生参与了 SpaCE2021 评测方案的讨论、语料标注与评测数据集构建工作。复旦大学邱锡鹏老师在评测工作的组织、制定评价标准等方面给与了很多指导。SpaCE2021 评测大赛得到华为诺亚实验室赞助。在此一并致谢。

anomaly; (3) a joint task of the previous two sub-tasks. Space2021 data set contains more than 18000 judgment questions, and the corpus size is more than 2 million Chinese characters. Compared with traditional spatial semantic role tagging and spatial reasoning tasks, space2021 takes into account the authenticity of corpus, the convenience of data set construction method and the challenge of spatial semantic understanding. It is a new attempt in the special evaluation field of Chinese spatial language understanding.

Keywords: spatial language understanding; spatial semantic anomaly; classification task; spatial sentence dataset; NLP evaluation

一　引言

近年来，自然语言处理（NLP）评测任务不断推陈出新，大大推动了自然语言处理研究，也成为积累自然语言处理基础数据资源的重要方式。以 BERT（Devlin et al., 2019）、GPT-3（Floridi et al., 2020）等为代表的超大神经网络模型，在当下众多自然语言处理评测任务中取得了令人眼前一亮的优异成绩。不过，机器在自然语言处理评测中成绩的提高，是不是意味着自然语言理解能力真实水平的提高，还值得深入探究。已有一些研究就注意到模型可以从数据本身的偏置中学到答题捷径（董青秀等，2021；Trichelair et al., 2019）。为此，学术界越来越关注自然语言处理评测任务设计的合理性以及任务中数据集的质量。

本文介绍一种新型的中文空间语义理解评测任务（Spatial Cognition Evaluation，简称 SpaCE2021，网址：http://ccl.pku.edu.cn:8084/SpaCE2021/）。下文第二部分简介自然语言处理领域空间语义理解能力评测任务研究概况；第三部分详细介绍 SpaCE2021 的设计思路和数据集制作过程；第四部分对评测结果以及数据集的问题展开分析和讨论；最后第五部分是结语，对空间语义理解评测任务设计及其数据集提出改进思路，并展望发展前景。

二　空间语义理解能力评测概况

空间范畴是人类认知的基础范畴。计算机空间语义理解能力在自然语言处理领域一直都受到关注，是自然语言处理评测的重要内容之一。以往的具体评测任务主要有以下三类。

（一）空间语义信息标注

空间语义信息标注任务将句子中存在的空间实体和空间关系等空间语义要素标注出来，代表性的工作如：Kordjamshidi 等（2012）在国际语义评测工作坊（SemEval）首次提出空间角色标注（SpRL）任务，Kolomiyets 等（2013）进一步扩展了此任务中的角色和概念，并在次年的国际语义评测工作坊上发布。Pustejovsky 等（2015）提出了 SpaceEval（在 SemEval 2015 上发布），任务形式与 SpRL 基本相同，但改用 Pustejovsky 等（2011）提出的 ISO-Space 标注体系。这些评测任务在形式上与以 Propbank 为代表的语义角色标注任务以及命名实体识别任务相当，只是标注对象更为具体，限定在空间实体和空间关系方面。

（二）空间关系推理

空间关系推理指根据已有空间关系信息推断出更多的空间关系信息，代表性的工作如：Weston 等（2015）提出了（bAbI）系列任务，其中的“位置推理”（Positional Reasoning），要求机器根据两个简单句子推理至多 3 个物体之间的 4 种方位关系。Mirzaee 等（2021）提出了 SpartQA 任务，基于 Suhr 等（2017）的 NLVR 数据集构建了更长的空间

故事文本，配合类型丰富、语句多样化和更多推理步骤的问题，形成了难度更高的空间推理任务。这些推理任务的语料都是以人工设计的简单几何图形为基础构造的，所涉及的空间实体类型及空间关系类型比较有限，语言的自然程度与人类真实语言环境中的空间语言表达存在较大差距。

（三）文景转换

文景转换指根据场景的语言描述生成相应的场景图片或视频（Coyne et al.，2011）。场景描述中一般包含若干空间实体及其空间关系代表性的工作，如：Liu 等（2021）提出了基于能量的机器学习模型（EBM），以因子分解的方式对多重空间关系进行分解再重组，可以对文本描述的空间实体（有一定的形状、颜色约束，数量 1 ~ 5 个）的多重空间关系进行场景解析并生成对应图片（项目展示网址 https://composevisualrelations.github.io/）。这类评测任务因其跨模态的性质，数据集制作成本较高，评测任务的实施也有较高的技术门槛（需兼有自然语言符号分析技术和图像分析技术）。

从数据制作成本及数据规模、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方面来看，已有任务存在一定的局限性。如何以较低成本制作较大规模的文本数据，用于机器对真实文本中空间语义理解能力的评测，并能在评测中体现理解能力的层次性，是本研究的主要考虑。同时，以往的任务设计均以英文为对象，中文文本的空间语义理解评测还不多见，正是在这样的背景下，本文提出了 SpaCE2021 空间语义理解评测任务。

三　中文空间语义理解评测 SpaCE2021 任务设计与数据集构建

现代形式语法学研究以语言表达形式的合语法性为首要问题。而合法性判断可以诉诸于普通人的语言直觉。SpaCE2021 受此启发，设计思路为：选取真实中文文本语料中富含空间信息的段落，采取由程序自动替换句中空间方位词语的方式，批量生成大量包含各类空间关系信息的语段。这些语段中，可能存在错误的空间关系信息，计算机需要甄别哪些语段的空间关系信息是正确的，哪些是错误的。通过这样的简单二分类任务，可以在一个比较粗的水平上，来评估机器的空间语义理解能力。这一设计最大的好处是以较低成本生产较大规模的数据集，兼顾了语料的真实性和数据集的构建效率和构建成本。

（一）SpaCE2021 的具体任务

SpaCE2021 评测共设计了 3 个子任务。

1. 中文空间语义正误判断任务

SpaCE2021 子任务 1 的主题是“中文空间语义正误判断”：要求机器判断给定的中文文本是否存在空间关系异常，经判定后不存在异常的句子标记为“True”，存在异常的句子标记为“False”。例如（句中划线部分是跟空间语义表达相关的主要成分）：

（1）信一送过来，他后了悔。他知道亲家的脾气多硬，多倔。要是钱先生见信后还不肯跟绑匪合作，那金三不就是把孩子往死里送了吗？　　（True）

（2）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，王实味满山遍野里找呀，喊呀，一直折腾到天黑，薄平就是不进去。　　（False）

2. 中文空间语义异常归因合理性判断任务

SpaCE2021 子任务 2 的主题是“中文空间语义异常归因判断”：要求机器判断给定的归因是否可以用来解释文本中存在的空间关系异常。异常原因有“不宜搭配”“语义冲突”“信息冲突”和“不符合常识”四类。如表 1 所示（text1、text2、commonsense 表示文本

字符串）。

表 1 SpaCE 2021 中的空间语义异常归因类型

类型	描述形式	具体含义
不宜搭配	“<text1>”和“<text2>”不宜搭配	主要是指二者因语法、韵律、习惯等因素，通常不搭配使用。
语义冲突	“<text1>”和“<text2>”语义冲突	主要是指二者的语义特征无法兼容。
信息冲突	“<text1>”与“<text2>”存在信息冲突	主要是指当前语境中的信息前后有矛盾。
不符合常识	“<text1>”不符合常识[：<commonsense>]	text1 所描述的内容不符合人类常识（由 commonsense 描述）。

（3a）庄先生从少年官回来之后，气呼呼一屁股坐在沙发上，目光上燃烧着痛苦的怒火。

（3b）只见谢烟客走进一个山洞前边，过了一会，洞中有黑烟冒出，却是在烹煮食物。

（3c）突然间轿中飞出一物，已罩住了他脑袋。那人登时眼前漆黑一片，大惊之下忙向后跃，再抓起罩在脚上之物，用力掷落，却是一顶官帽。

（3d）抬轿之人只要脚步稍慢，轿中软鞭挥出，刷刷几下，重重打在上面的轿夫背上，在前的轿夫不敢慢步，在后的轿夫也只得跟着飞奔，几名官差跟随在后。

例（3a）句存在异常，因为“目光”和“上”不宜搭配；例（3b）句的异常，是因为“山洞前面”和“走进”语义冲突；例（3c）句的异常，因为“罩住了他脑袋”与下文“罩在脚上之物”存在信息冲突；例（3d）句的异常，因为“上面的轿夫”不符合常识：轿夫不会在轿子上面抬轿。

3. 中文空间语义正误判断与异常归因联合任务

SpaCE2021 子任务 3 旨在考察机器联合处理前两个子任务的能力。机器要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正确。

（二）SpaCE2021 数据集的构建过程

数据集的构建过程包括以下 6 个步骤。

1. 收集生语料并进行分词和词性标注预处理。初始语料约 4200 万字，涵盖小说、散文、词典等多种文体。使用 pkuseg（Luo et al., 2021）进行分词和词性标注。

2. 筛选富含空间方位语义的语段。筛选标准为：（a）句中表达空间方位意义的词汇[①]较为密集；（b）句中出现至少 3 个实体词；（c）基于 BERT 的机器模型无法准确预测出句中的空间方位词语。基于上述标准，从 4200 万字原始语料中抽取得到 4021 个自然段落。

3. 基于词表批量生成更多语料。替换词表由 5600 组“原词—替换词”对构成，其中原词共 1200 个。根据该替换词表，对 4021 个段落进行替换操作，得到 16640 个“替换句”，每句都有一个方位词语与原始句不同，其空间语义可能正确，也可能异常，由人工做出判断。

4. 对句子逐条进行人工判断和标注。标注系统的界面如图 1 所示，句中标为红色的词语即程序自动替换的空间义词语（如图 1 中的“出来”“下来”），句中其他空间方位义词语以及实体词分别被标记为黄色和蓝色，方便标注人员区分。标注工作包括：（a）检查句中是否有分词、词性标注、错别字等方面的错误。存在这些情况的，直接将该句删除；（b）判断句子的可接受程度，包括“不成立”“勉强成立”“成立”三个选项，其中“不成立”和“勉强成立”皆被视为异常。（c）标注造成句子异常的原因[②]。为保证数据质量，最终数据集只保留了 2 名标注者标注一致的句子[③]。

在对 16640 个“替换句”进行人工标注后，得到有效标注结果 7858 句。其中，2702 句为无异常的句子，5156 句为有异常的错句，对每个错句，都标注了一个或多个归因，一共

图 1　SpaCE 2021 标注系统界面

构成了 8665 个“错句—归因对”（error – explanation sentence pair）。

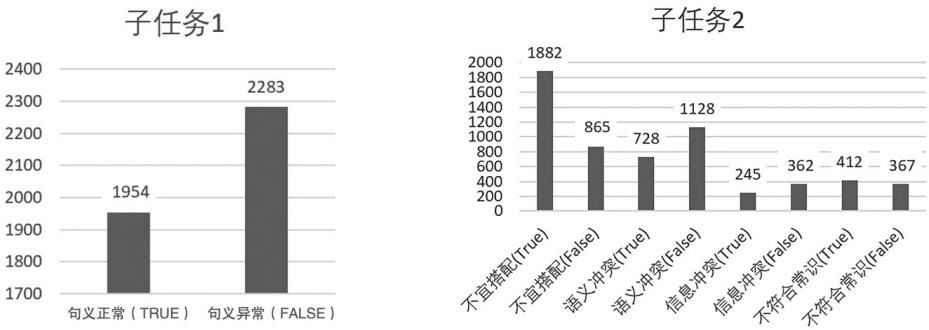
5. 基于规则生成错误归因。为了形成子任务 2 和子任务 3 的题目，除了需要通过人工标注得到的正确的“错句—归因对”，还需要再生成错误的“错句—归因对”。为此，针对各种归因类型，用程序自动生成了 74407 个可能错误的“错句—归因对”。经人工审核后，将确属错误的数据收录到最终的数据集④。

6. 构建数据集。将句子按照一定比例分配到训练集、开发集和测试集中。为了避免训练集或开发集中的句子与测试集中的句子来源于相同原始句子而导致的偏置，分配程序会确保不同数据集中的句子之间不存在同源关系，即不会由同一个原始句替换而来。

经过上述流程，得到最终发布的 SpaCE2021 数据集，其各子集构成如表 2 所示。其中子任务 1 和子任务 2 训练集数据的各类型题量分布如图 2 所示。

表 2　SpaCE2021 数据集的构成

子任务	训练集	验证集	测试集	总计	备注
1. 中文空间语义正误判断	4237	806	794	5837	各数据集之间不存在同源句子，下同。
2. 中文空间语义异常归因合理性判断	5989	2088	1952	10029	子任务 2 的全部数据集所使用的句子与子任务 1 的验证集和测试集无交集；子任务 2 训练集使用的句子与子任务 1 的训练集有交集。
3. 中文空间语义判断与归因联合任务	0	1203	1167	2370	子任务 3 不提供训练集。子任务 3 的验证集和测试集中使用的句子与子任务 1 相同。



（a）　　（b）

图 2　SpaCE2021 子任务 1 和子任务 2 训练数据集题量分布图

最终发布的数据集共计 18236 道题，题干文本共计 2034101 字（平均句长 111.54 字，

标准差 54.07），归因文本共计 209311 字（平均句长 16.88 字，标准差 7.30）。

（三）SpaCE2021 的评价指标

子任务 1 和子任务 2 以准确率（Acc，Accuracy）作为指标，公式为：

$$Acc = \text{命中正确答案的题数} / \text{题目总数}$$

子任务 3 以 F1 值作为指标，公式为：$F1 = 2 \times P \times R / (P + R)$

其中 P 和 R 分别代表准确率（Precision）和召回率（Recall），公式分别为：

$$P = (TP_2 + TN_2) / (TN_1 + FN_1)$$

$$R = (TP_2 + TN_2) / (TN_1 + FP_1)$$

上述公式中，TP、TN、FP、FN 分别代表命中数量、正确拒绝数量、误报数量、漏报数量；下标表示判断所属的阶段，“1”对应句子判断阶段，“2”对应归因判断阶段。

参加测试的机器系统可根据以上公式计算分值，并在各个子任务之中进行排名。多个参评系统综合排名计算方式如下：

$$Z_i = (\text{原始分—所有参赛系统原始分均值}) / \text{所有参赛系统原始分标准差} \quad （公式 1）$$

$$Z_{mean} = (Z_1 + Z_2 + Z_3) / 3 \quad （公式 2）$$

公式（1）将各个系统在每个任务中的得分换算为标准分数（Z-score）；公式（2）对一个系统在 3 个任务下的标准分数求均值（Z_{mean}），作为排名依据。

四　SpaCE2021 评测数据分析

（一）SpaCE2021 基线及机器总体表现

1. 基线系统（baseline system）

为考察主流模型表现，并为评测提供基线，课题组基于主流预训练模型 BERT 建立了一套基线系统[⑤]。该系统运行环境为 Ubuntu 18.04 操作系统，基于 CUDA 10.2 平台，通过 Python 3.6.x 编程。基线系统在 3 个任务上的实验结果如表 3 所示。

表 3　基于 BERT 的基线系统在 SpaCE2021 开发集及测试集上的表现

	子任务 1 准确率	子任务 2 准确率	子任务 3F1 值
开发集	0.6526	0.6769	0.5041
测试集	0.6725	0.7290	0.5267

在比赛任务发布后，课题组将该基线系统提供给参赛者，作为调试程序的比较基准参考。

2. 人类表现（human performance）

课题组组织了两次人类评测实验，招募了 7 名被试参与，其中被试 1 曾参与过语料标注工作。实验 1 从子任务 2 和子任务 3 的测试集中分别随机抽取 200 道和 117 道题目进行测验，最终获得 6 名被试的有效数据。实验 2 分别从 3 项子任务的测试集中随机抽取了 50、200、123 道题目，最终获得 4 名被试的有效数据[⑥]。结果如表 4 所示。

整体来看，人类被试的答题表现并不理想，一方面反映出空间语义理解有一定的主观性，另一方面也说明 SpaCE2021 任务中判别标准的清晰性、语料标注的规范性还有待加强。

3. 机器表现（machine performance）

SpaCE2021 在第 20 届中国计算语言学大会（CCL2021）上发布，共有 13 支队伍报名，最终有 8 支队伍提交了测试结果。所有系统均使用主流大规模预训练模型。几乎所有系统在 3 个子任务上的表现都超过了基线系统，具体成绩如表 5 所示。在子任务 2 中，超过

表 4　人类被试的表现

被试编号	实验 1		实验 2			子任务 2 准确率均值	子任务 3 F1 均值
	子任务 2 准确率	子任务 3F1	子任务 1 准确率	子任务 2 准确率	子任务 3 F1 值		
被试 1	0.8300	0.6891	0.8400	0.8100	0.6866	0.8200	0.6878
被试 2	0.8350	0.6076	0.6600	0.7300	0.5464	0.7825	0.5770
被试 3	0.8300	0.6154	0.6000	0.8550	0.6000	0.8425	0.6077
被试 6	0.7700	0.7244	0.7600	0.7650	0.6667	0.7675	0.6955
被试 4	0.7850	0.6164	/	/	/	0.7850	0.6164
被试 5	0.7650	0.6615	/	/	/	0.7650	0.6615
最大值	0.8350	0.7244	0.8400	0.8550	0.6866	0.8425	0.6955
均值	0.8025	0.6524	0.7150	0.7900	0.6249	0.7938	0.6410
标准差	0.0298	0.0434	0.0921	0.0470	0.0555	0.0282	0.0436

半数的系统表现超过了人类平均成绩；而在子任务 1 和子任务 3 中，也有部分队伍超过了人类平均成绩；但所有系统都没有达到人类最高成绩[⑦]。通过赛后对 SpaCE2021 数据集的分析，机器成绩接近人类水平，最主要的原因可能是数据集中的问题难度分布不均衡，总体来说测试难度偏低。比如子任务 2 中，“不宜搭配”类的试题数量远多于“违反常识”类的试题，前者对机器来说容易区分，后者对机器来说更难判断（可参见表 7 数据）。

表 5　SpaCE2021 测试结果

系统	子任务 1 准确率	子任务 2 准确率	子任务 3F1 值	Z_{mean}
队伍 1	0.734257	0.841189	0.647574	1.328
队伍 2	0.720403	0.813012	0.622041	0.709
队伍 3	0.729219	0.752561	0.657812	0.451
队伍 4	0.680101	0.768955	0.660972	0.103
队伍 5	0.692695	0.797643	0.584027	-0.021
队伍 6	0.678841	0.804816	0.554873	-0.308
队伍 7	0.630982	0.806352	0.616521	-0.373
队伍 8	0.692695	0.764857	0.557329	-0.543
基线系统	0.672544	0.728996	0.526651	-1.346
人类最高水平	0.8400	0.8425	0.6955	/
人类平均水平	0.7150	0.7938	0.6410	/

注：人类测试是随机抽取 SpaCE2021 测试集部分题目测得的结果，参见表 4。

（二）机器在 SpaCE2021 子任务 1 上表现的特点

基于人工神经网络构建的机器模型具有“黑箱”性质，目前还难以探究模型在空间语义理解上的内在过程和机制特点。不过，通过观察机器在 SpaCE2021 任务某些试题上的测试结果（外在表现），也可以在一定程度上了解机器空间语义理解能力的部分特点，为今后优化评测任务、提高数据集质量提供依据。

因多个系统参与了任务 1 的测试，不同系统如果对同一个题目给出了相同的答案，则可反映机器在判定空间语义信息时呈现的一些规律。为此，本节着重分析机器在任务 1 上答案一致性高的题目（见表 6），按结果对错可以划分为两组，表 6 中 A、D 类题为一组，这是所有系统都答对的题目，共 272 题，占任务 1 总题数（794 题）的 34.26%，占机器答题结果一致性高的题目总数（358 题）的 75.98%；表 6 中 B、C 类题为一组，这组包括了所有系

统和绝大多数系统都答错(只有一个系统答对)的题目，共 86 题，占任务 1 总题数的 10.83%，占机器答题结果一致性高的题目总数的 24.02%。

表 6 机器答案一致性高的题目数量分布情况

答案类型	标准答案 True	标准答案 False	合计
机器答案 True	(A) 104	(B) 51	155
机器答案 False	(C) 35	(D)168	203
合计	139	219	358

通过考察题目的语句内容，可以对机器易对题和易错题一些相对突出的特点做出大致区分：第一，A、D 类机器易对题中，主要涉及空间方位语义的表达式容易在局部语段做出判断，而对上下文信息的依赖较少，包括(1)表示绝对方向的方位词；(2)实际语料中少见的“名词 + 方位词”组配形式。第二，B、C 类机器易错题中，情况则相反，句中的空间方位语义要么有很强的相对性，要么不能仅根据局部语段做出判断，而是对上下文有很高的依赖或者需要调用常识知识。下面看一些具体的例子。先看机器易答对的 A、D 类题的情况。

(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……

例 4 中的方位词“东南”替换为“东北、西北、西南”等后，形成的句子语义虽不同，但就各句独立来看(脱离了更大语境),其空间语义都是成立的。这是 A 类机器易对题的情况。

D 类机器易对题中大多是方位词和名词搭配不当(语料中无此搭配或低频搭配)。比如：

(5a) 房底、夜晚左、背内、寓所下、厅回、石壁过、小字顶、山角出、枪口侧、影片右、手左、耳左、耳右……

(5b) 市集左、市集右、厅堂左、小船侧、房左、寓所左、石壁左、凌霄城左、石室右、小船左……

(5c) 山外下、山上下、山里下

(5d) 顶大门、底大门

例 5 中的方位词和名词组配都存在问题。(5a) 中名词所代表的实体对方位有选择限制，比如“房”通常搭配“内、顶”等，不搭配“底”，“背”搭配“上”，不搭配“内”。(5b) 中名词所代表的实体从空间概念上可以选择后面方位词所代表的方位，但从韵律和用法习惯上，没有这样的组配方式，比如“小船一侧”可以，但“小船侧”则不合法。(5c) 中是方位词连用，“外 + 下、上 + 下、里 + 下”都是不合语法的组合形式。(5d) 是“方位词 + 名词”组合，也不合语法格式。这些语言表达式在真实语料中是低频分布，机器相对来说容易识别判断。

下面再看机器易错的 B、C 类题的情况。先看标准答案为“False”，机器判为“True”的 B 类题。

(6) 绑匪准会把明月留在庙里当诱饵，……我上这儿去很不方便，你敢不敢去走一趟？

(7) 她问我去哪里出差？我说西北。[...] 她兴奋极了，她不让我动手，要亲自给我理行李，把原来替我准备的来西北的东西一样一样重新摆到箱子里。

(8) 他耷拉着脑袋走近小庙，打眼角往外边瞅。

(9) 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。

(10) 有人大叫：“逃啊！”抢先到地道中奔入，余人也都跟了进去。石破天叫道：“里面危险，别进去！”却又有谁来听他的话？

上面例子的共性是，从句中加下划线成分所在局部语段看，其空间语义都是能成立的，但从更大的上下文来看，整句的空间语义却不成立。例 6 中的“这儿”、例 7 的“来”都涉及到空间方位的相对性，需要根据说话人当前处所跟言谈中有关处所的位置关系来确定。

从上下文可知，例 6 中说话人应该是去距离说话人所在地较远的处所，需要用“那儿”指代，不能用“这儿”。例 7 也是类似的问题，说话人“我”是离开某地前往西北，应该用“去西北”表达，而不是“来西北”。例 8 从第一个小句可知，“他”位于小庙的外边，只能是“往里面瞅”，不能再“往外边瞅”（跟第一个小句语义矛盾）。例 9 涉及到常识知识，“辫子”只能垂在脑后或背后，垂在“脚上”违背常识。例 10 涉及到多项词语的组配，其中表达空间方位义的介词组“到地道中”本身没问题，但“到地道中奔入”无法对应清晰的动态空间场景，应该是“向地道中奔入”才能得到正确的空间语义发生的场所。

再看标准答案是“True”，机器判为“False”的 C 类题。这类题在表 7 的四类题中数量最少，但并没有突出的共性。事实上，C 类题跟 A、B 、D 三类题都有所不同。就像一般对于合语法的句子，人们不太会问为什么合法，只有不合语法的句子，才会去问不合法的原因，C 类题都是正确（True）的句子，而机器判为“False”，如果要追问原因，这背后的可能性就很多，既可能是机器没有识别和理解句中的空间方位语义，也可能是因为机器“察觉”了句子中其他的语言表达存在问题，或者是二者的综合。也正因为如此，在不清楚机器分类算法的依据标准和运行过程情况下，仅从观察 C 类题的句子本身，难以发现和概括其共性特征。下面略举几例：

（11）从那图旁所注的小字中细加参详，得悉图中所绘的无名荒岛之左，藏有一份惊天动地的武功秘诀……

（12）信一送进来，他后了悔。他知道亲家的脾气多硬，多倔。

（13）当我们到达医院时，一大群人已围在四面。

上例下划线中的“之左”“进来”“四面”等方位表达词语是替换词。原词分别是“之上”“出去”“外面”。仅从个人语感体会，上面 3 例中的与空间方位语义相关的表达形式，都是相对低频的形式：“之左”在现代白话文中罕见；“送出去”常见而“送进来”少见；“围在里面”常见而“围在四面”罕见。或许是这个特征使机器将这几句都归入“False”类。当然，要搞清楚机器在对空间方位表达做出正误判断时的真实凭据，还需要更多的测试数据以及更丰富的标注信息。此外，例 11 标准答案为“True”，但这是数据标注人员的判断，像“之左”“之右”这样文言色彩过浓的语言形式，可能会影响人的语感判断，而机器学习的实际语料样本中这类表达式如果分布过少，就可能导致机器倾向于把这类句子归入“False”类。对此，在下一步改进 SpaCE2021 数据质量时，应注意尽量收入常见的空间方位相关表达形式，而避免包含像“之左”这样文言形式的句子。

（三）机器在 SpaCE2021 子任务 2 上表现的特点

子任务 1 的语料基本可以算是自然文本，训练集内部数据类型划分也相对平衡一些。子任务 2 的语料则包含了更多的人工构造性质。从数据量的角度说，子任务 2 的训练数据总体相对偏少，此外，内部各类的分布均衡性不佳（见图 2）。这些因素会影响测试结果的参考价值。表 7 是机器在任务 2 中不同归因类型及答案类型上的平均表现。

不难看到，子任务 2 中，机器更擅长判断词语语义冲突和词语不宜搭配类型的问题，而不太擅长判断上下文信息冲突类和常识相关类问题。表 7 统计数据显示，机器在常识类异常的判断上存在明显的偏误，表现为正例判断的正确率最高，负例判断的正确率最低，说明凡是遇到常识类判断题，机器总是更倾向于判断为“True”，由此可知：机器并没有真正掌握相应的常识。类似的情况也出现在不宜搭配类的判断上。对于不宜搭配类的判断题，机器也更倾向于判断为“True”，因此在正例上的正确率高，负例上的正确率低。

表 7　机器在 SpaCE2021 子任务 2 上的平均表现

归因类型	答案类型	分项		合计	
		题数	平均正确率	题数	平均正确率
不宜搭配	True	632	0.8671	871	0.8101
	False	239	0.6594		
语义冲突	True	173	0.8277	429	0.8615
	False	256	0.8844		
信息冲突	True	142	0.6521	355	0.7251
	False	213	0.7737		
不符合常识	True	103	0.9534	297	0.7273
	False	194	0.6072		

不过从数据中也可以看出，子任务 2 测试数据集中各类归因答案的分布存在较大的均衡性问题，如“不宜搭配”类的答案中，“True”与“False”的比例超过了 2.5:1；而“不符合常识”类的答案里，“True”与“False”的比例也接近 1:2，可能导致正负例的正确率差异被放大（可对比图 2 训练集中的数据分布情况）。

五　结语

SpaCE2021是面向中文文本的机器空间语义理解能力评测任务设计方面的初步尝试。大略算是对机器的空间语义理解能力做了表象层面的考察。从参评系统的表现来看，当前基于大数据预训练模型的人工神经网络系统在这一任务上的表现跟人类表现比较接近。但如果透过机器得分不算差的表象，深入到机器答题的具体数据，不难看出，机器在理解句子中空间语义相关常识，分析上下文中空间信息冲突方面，仍然存在明显的不足。

为了更为全面和准确地探究机器空间语义理解能力，并推动机器的相关能力在 NLP 应用系统中的落地，下一步的改进工作应着重考虑以下几个方面的问题。

第一，空间范畴作为人类基本认知范畴，人对文本中的空间语义信息理解也有明显的主观性。SpaCE2021 数据标注中人类标注员的一致性未达优秀级（参见本文附注 2）。人类被试在测试集上的小规模抽样答题结果平均成绩也仅在 70 ~ 80 分（参见表 4），这些都说明空间语义理解评测任务的设计还需要进一步优化。特别是测试题需要再简化，以人类判断标准来说，要尽可能有清晰的答案，而不宜在测试数据中容纳对错界限模糊的试题。

第二，语料选取更加注重分布均衡，且更多地考虑应用场景。比如 SpaCE2021 中有较大比例的武侠小说语料，距现在时间较远，且语言风格半文不白，并不适合用来测试机器的空间语义理解能力。在后续工作中，将选取特定领域的专项语料，比如体育教材中的人体动作描述类语料、交通事故现场描述类语料，等等。

第三，试题设计和测试内容选取方面更多考虑“人—机”对比（特别是认知维度上的对比）价值。比如参考儿童语言习得中空间方位认知和语言能力的发展顺序，参考留学生学习中文产生的中介语语料中的空间方位表达偏误类型来设计测试题，以利于将机器评测结果与人类表现进行更全方位的对照和研究。

第四，在语句正误标注的基础上，进一步考虑语句间空间语义表现相同和相异现象的标注，以便对机器空间语义理解能力做出更细粒度的评价和更具解释性的分析，同时使得机器空间语义理解能力评测数据集具有更高的语言学本体研究价值。这里略举两例说明：

（14a）她缓缓地回过头，朝着【面前】带着潮气的泥土，深深地吸了一口气……

（14b）她缓缓地回过头，朝着【身后】带着潮气的泥土，深深地吸了一口气……

（15a）【窗前】谁种芭蕉树？阴满中庭；阴满中庭，叶叶心心、舒卷有馀情。

（15b）病里得书如中甲，【窗前】览镜试新妆。

例 14 中，“面前”与“身后”是一对反义词，但人们去理解整句所表达的空间场景时，却对应出一幅相同的画面，即例（14a）和（14b）语义相同；例 15 中，都包含同一个方位短语“窗前”，但例（15a）的方位指“窗外”，例（15b）的方位指“窗内”（在试妆人的面前）。

目前 SpaCE2021 数据集中的句子（测试题）都是各自独立的，还没有像上面两例这样的对照句，还不能把句子关联起来考虑空间方位语义的理解问题。

综合来说，SpaCE2021 任务设计出发点是以较低成本生产较大规模的真实语料数据集，为此，课题组拟定了语句正误判断任务和错误归因任务作为数据制作的主要手段，从而可以通过众包方式快速产生大量数据。课题组从 2021 年 1 月底启动大规模数据标注工作，到 3 月底完成全部数据集制作并发布评测任务，说明本文提出的思路是切实可行的。不过，毋庸讳言，SpaCE2021 的评测目标是考察机器的空间语义理解能力，手段则是用分类任务来观察机器在中文句子正误判断上的表现。手段和目的之间存在一定的距离。如何让自然语言处理评测任务的测试手段（试题及指标）跟它的测试目的更为契合？实际上是考察机器智能真实水平时需要回答的核心问题。在 SpaCE2021 基础上，课题组接下来还要在扩大数据规模、提高数据质量、优化评测任务设计方面做更深入的研究，希望改进后的数据集更具学术价值和应用价值，能引起学界对空间语义理解问题的更多关注和研究兴趣。

［附　注］

① 可访问 https://2030nlp.github.io/SpaCE2021/words 查看空间方位分类词表。该词表以分组形式，把具有相似上下文的表示空间方位意义的词语归入一组。

② 若句子不存在异常，标注者需标注替换后的句子与原始句子是否发生了较大语义变化。这一信息未在数据集中使用。

③ 每个句子均由 2 名标注者标注，对于句子是否成立，2 名标注者 Kappa 值均值为 0.564，标准差为 0.145；在句子成立方面取得一致的情况下，2 名标注者对句子归因类型的 Kappa 值均值为 0.685，标准差为 0.123；总体来说一致性程度未达到优良水平，反映包含空间语义在内的句义理解存在较大个体差异，同时也存在标注任务定义不够清晰，标注流程还需要进一步规范。

④ 第 4 步人工标注有 40 人参与，第 5 步人工审核有 18 人参与。全部数据加工工作大约耗时 1 个月。

⑤ 该基线系统可在 https://github.com/2030NLP/SpaCE2021-Baseline 查看。

⑥ 当被试答题出现机械规律，或体现出其并未能理解任务要求时，其数据被视为无效。

⑦ 队伍 1 提交的系统在子任务 2 中达到了 0.8412 的准确率，与人类最高水平 0.8425 非常接近。

［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，2021，35（6）.

［2］Coyne，B. & Sproat，R. WordsEye: An automatic text-to-scene conversion system［C］. Proceedings of the 28th annual conference on Computer graphics and interactive techniques. 2001: 487-496.

[3] Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding[C].Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019: 4171-4186.

[4] Floridi, L. & Chiriatti, M. GPT-3: Its nature, scope, limits, and consequences[J]. *Minds and Machines*, 2020, 30（4）.

[5] Kordjamshidi, P. , Bethard, S. & Moens, M. F. SemEval-2012 task 3: Spatial role labeling[C]. Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012). 2012, 2: 365-373.

[6] Kolomiyets, O., Kordjamshidi, P., Bethard, S. & Moens, M. F. Semeval-2013 task 3: Spatial role labeling[C].Proceedings of the seventh international workshop on semantic evaluation (SemEval 2013). 2013: 255-262.

[7] Liu, N., Li, S., Du, Y., Tenenbaum, J. B. & Torralba, A. Tenenbaum, Antonio Torralba. Learning to Compose Visual Relations[J]. *Advances in Neural Information Processing Systems*, 2021,（34）.

[8] Luo, R., Xu, J., Zhang, Y., Ren, X., & Sun, X. Pkuseg: A toolkit for multi-domain chinese word segmentation[J]. arXiv preprint arXiv:1906.11455, 2019.

[9] Mirzaee, R., Faghihi, H. R., Ning, Q. & Kordjmashidi, P. SPARTQA: A Textual Question Answering Benchmark for Spatial Reasoning[C]. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 4582-4598.

[10] Pustejovsky, J., Kordjamshidi, P., Moens, M. F., Levine, A. & Yocum, Z. Semeval-2015 task 8: Spaceeval[C]. Proceedings of the 9th International Workshop on Semantic Evaluation. 2015: 884-894.

[11] Pustejovsky, J., Moszkowicz, J. L. & Verhagen, M. ISO-Space: The annotation of spatial information in language[C]. Proceedings of the Sixth Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation. 2011, 6: 1-9.

[12] Suhr, A., Lewis, M., Yeh, J. & Artzi, Y. A corpus of natural language for visual reasoning[C]. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2017: 217-223.

[13] Trichelair, P., Emami, A., Trischler, A., Suleman, K. & Cheung, J. C.K. How Reasonable are Common-Sense Reasoning Tasks: A Case-Study on the Winograd Schema Challenge and SWAG[C].Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019: 3382-3387.

[14] Weston, J., Bordes, A., Chopra, S., Mikolov, T., Rush, A. M. & Van Merrinboer, B. Towards ai-complete question answering: A set of prerequisite toy tasks[J]. arXiv preprint arXiv:1502.05698, 2015.

（责任编辑　常文斐）