CCL23-Eval任务4总结报告：第三届中文空间语义理解评测

肖力铭[1,2,‡] 詹卫东[1,2,3,†,*] 穗志方[3,†] 秦宇航[1,2,‡] 孙春晖[1,2,†]

邢丹[1,2,‡] 李楠[1,2,†] 祝方韦[3,‡] 王培懿[3,‡]

[1]北京大学 中文系

[2]北京大学 中国语言学研究中心

[3]北京大学 计算语言学教育部重点实验室

[†]{zwd,szf,sch,linan2017}@pku.edu.cn

[‡]{lmxiao,hezonglianheng,xingdan,zhufangwei2022,wangpeiyi}@stu.pku.edu.cn

摘要

第三届中文空间语义理解评测任务（SpaCE2023）旨在测试机器的空间语义理解能力，包括三个子任务：（1）空间信息异常识别任务；（2）空间语义角色标注任务；（3）空间场景异同判断任务。本届评测在SpaCE2022的基础上，优化了子任务一和子任务二的任务设计，并提出了子任务三这一全新的评测任务。最终有1支队伍提交参赛结果，并且在子任务一上的成绩超过了基线模型。本文还报告了大语言模型ChatGPT在SpaCE2023三个子任务上的表现，结合问题提出指令设计可改进的方向。

关键词： 中文空间语义理解；评测任务；大语言模型

Overview of CCL23-Eval Task 4: The 3rd Chinese Spatial Cognition Evaluation

Liming Xiao[1,2,‡] Weidong Zhan[1,2,3,†,*] Zhifang Sui[3,†] Yuhang Qin[1,2,†]

Chunhui Sun[1,2,†] Dan Xing[1,2,‡] Nan Li[1,2,†] Fangwei Zhu[3,‡] Peiyi Wang[3,‡]

[1]Department of Chinese Language and Literature, Peking University

[2]Center for Chinese Linguistics, Peking University

[3]MOE Key Laboratory of Computational Linguistics, Peking University

[†]{zwd,szf,sch,linan2017}@pku.edu.cn

[‡]{lmxiao,hezonglianheng,xingdan,zhufangwei2022,wangpeiyi}@stu.pku.edu.cn

Abstract

The 3rd Chinese Spatial Cognition Evaluation Task (SpaCE2023) aims to test the machine's spatial semantic understanding capabilities, including three subtasks: (1) to detect and identify spatial semantic anomaly in a sentence; (2) to label the spatial roles and relations for spatial entities in a sentence; (3) to judge whether two spatial scenes in a pair of sentences are the same or different, then give a reason. Based on SpaCE2022, this year's evaluation optimized the task designs of subtasks 1 and 2, and proposed subtask 3 as a brand new evaluation task. 1 team submitted the results of test set and exceeded the baseline model on subtask 1. This paper also reports the performance of the large language model ChatGPT on SpaCE2023 dataset, and provides directions for improving prompt design combined with the issues raised.

Keywords: Chinese Spatial Cognition Evaluation , Evaluation Task , Large Language Model

1 引言

项目资助：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）；2022年度教育部人文社会科学重点研究基地重大项目（22JJD740004）

空间范畴是人类认知中重要的基础范畴，大量空间信息存在于自然语言文本中。在通往类人智能的道路上，空间语义理解是不可绕开的一环。为了评测机器的空间语义理解能力，推进空间范畴的认知计算建模研究，我们连续两年举办了**中文空间语义理解评测任务**（Spatial Cognition Evaluation）来考察机器的空间语义理解能力(詹卫东et al., 2022)。在第一届评测（SpaCE2021）中，我们通过替换方位词的方式生成了空间信息异常的文本，并设计异常判断任务，要求机器判断文本是否存在空间信息异常以及异常的归因类型。第二届评测（SpaCE2022）扩大了任务类型，增加了信息标注任务，要求机器在空间信息异常的文本中识别异常片段，在空间信息正常的文本中进行细粒度的语义角色标注。综合SpaCE2022中4个系统的成绩，空间信息正误判断任务的准确率（Accuracy）最高为0.7992，异常归因与识别联合任务的F1值最高为0.6748，语义角色标注任务的F1值最高为0.5069，说明机器捕捉异常片段的能力以及结构化空间信息的能力还有待提高。[0]

今年依托第二十二届中国计算语言学大会CCL2023，我们举办了第三届中文空间语义理解评测任务（SpaCE2023），包括以下3个子任务：（1）**空间信息异常识别任务**，要求机器识别出文本中空间信息异常的片段；（2）**空间语义角色标注任务**，要求机器基于空间信息标注规范，对文本进行空间实体的识别与空间方位关系标注；（3）**空间场景异同判断任务**，要求机器判断两个在形式上相似的中文文本是否可以描述相同的空间场景，并说明判断理由。[1]

相较于SpaCE2022，本届评测优化了任务设计，不再设置人类表现一致性较低的异常判断任务，保留了人类表现一致性较高的异常识别任务和可操作性较强的语义角色标注任务。语义角色体系也得到了统一，从18个语义角色整合为15个。此外，SpaCE2023新定义了“空间场景异同判断”这一生成类任务，考察机器从文本中构建空间场景的的空间“想象”能力。比如，“站在电线杆下”和“站在电线杆旁”虽然使用了不同的方位词，但具备空间“想象”能力的机器应该能够发现这两个文本描述了同一个空间场景，即站在电线杆周围的地面上。除了任务的不同，SpaCE2023还修订了问题语料，提高了语料质量；扩大了参赛系统的范围，允许使用大模型参赛。最终有1支参赛队伍提交结果。

本文将报告SpaCE2023的总体情况。第2节介绍评测任务的基本内容，第3节介绍数据集的制作流程和基本信息，第4节展示基线系统、参赛系统以及大模型在各项任务上的表现，第5节总结评测概况，展望中文空间语义理解评测任务的发展。

2 评测任务

2.1 子任务一：空间信息异常识别

子任务一的输入是一个存在空间信息异常的文本，如“我清晨去公园散步时，总能看见**他站在电线杆里**，手里提着菜篮”中，“他”的空间方位存在异常。按照常识，人只能站在电线杆周围的区域，而不能在电线杆的内部。机器需要使用S-P-E标注法描述并输出异常文本片段，S指描述了空间方位信息的实体，P指空间实体的空间方位信息，可能涉及处所、起点、方向等信息，E指空间义相关事件，是动词性单位，表达了S之于P的方式、目的或原因，如[**S他，P在电线杆里，E站**]。文本片段的选取数量最多6个，即两个完整的S-P-E三元组；最少1个，即S-P-E中的单个元素。表1 展示了对应数量的标注情形和示例，S1、P1、E1、S2、P2、E2是具体使用的标签。

2.2 子任务二：空间语义角色标注

子任务二的输入是一段空间信息表述正常的文本，如“我清晨去公园散步时，总能看见他站在电线杆下，手里提着菜篮”中，描述了“我”、“他”、“菜篮”的空间信息。机器需要使用STEP标注体系描述并输出文本含有的所有空间信息，STEP标注体系在S-P-E标注法的基础上增加了时间信息（T）和空间事实性（F）的标注，描述的信息可概括为：“某空间实体在某时，经由某事件，处于某种空间方位关系，这一命题的事实性为真/假”。该体系共含15个元素，表2 是每个元素的含义。

[0]https://2030nlp.github.io/SpaCE2022/
[1]https://2030nlp.github.io/SpaCE2023/

第二十二届中国计算语言学大会论文集，第150页-第158页，哈尔滨，中国，2023年8月3日至5日。

数量	含义	示例
6	两个S-P-E元组存在信息冲突	她笑，[**P1 池水里**]的[**S1 影子**]向着她[**E1 笑**]；她假装生气，[**P2 池水外**]的[**S2 影子**]也向着她[**E2 生气**]。
5		[**S1 我们**][**P1 在树林里**][**E1 坐**]下来聊天，感觉秋天的[**P1 树林后**]已有些[**E1 微凉**]。
4		他一边跑，一边大叫:“[**S1 细马**][**P1 过去**]了![**S2 细马**][**P2 回来**]了！”
3	S-P-E元组描述的信息 无法构造空间场景	她又在衣袋里摸了半天，[**E1 摸**][**P1 进**][**S1 火柴**]来。
	S-P-E元组描述的信息 可以构造空间场景但不合常理	[**S1 小孩们**]特别爱逛庙会，为的是有机会[**P1 到市区**][**E1 看看野景**]
2	S-P-E元组描述的信息 无法构造空间场景	他[**E1 沉**][**P1 入水边**]，不见了踪影
	S-P-E元组描述的信息 可以构造空间场景但不合常理	她看着[**P1 橱窗边**]的[**S1 婚纱**]，心想自己穿上一定很好看。
1	元素所描述的实体、方位 或事件不存在	在[**S1 边脚掌**]接触地面的瞬间，快速做一个原地垫步跳。

表 1: 异常文本片段选取说明

序号	元素名称	所属	性质	含义
1	空间实体	S	片段	对应于被描述空间方位的空间实体。
2	参照实体	S	片段	对应于与1号元素形成距离关系的另一个空间实体。
3	事件	E	片段	与空间实体的空间方位直接关联的事件。
4	事实性	F	标签	如果空间方位命题是假的，则标签为”假”。
5	时间	T	片段	文中写明的与空间方位相关联事件的时间表述。
			标签	如果空间实体处于某种空间方位关系的时间在文中并未写明，但可以推断绝对时间，则选择标签：”说话时” / ”过去” / ”将来”。
			片段+标签	如果空间实体处于某种空间方位关系的时间在文中并未写明，但可以推断参照时间，则在文本中选择参照事件，并选择标签：”之时” / ”之前” / ”之后” / ”之间”。
6	处所	P	片段	描述静态空间实体相对某外部参照物的位置。
7	起点	P	片段	描述动态空间实体的方位发生变化的场景下，变化开始时实体的处所。
8	终点	P	片段	描述动态空间实体的方位发生变化的场景下，变化结束时实体的处所。
9	方向	P	片段	描述动态空间实体的位移方向。
10	朝向	P	片段	描述空间实体某一侧面所朝向的位置。
11	部件处所	P	片段	描述空间实体作为一个部件在整体中的位置。
12	部位	P	片段	描述了空间实体的某个部位。
13	形状	P	片段	描述了空间实体的形状。
14	路径	P	片段	描述了空间实体位移时经过的轨迹。
15	距离	P	片段	文中空间实体间定量距离的表述，与1号元素和2号元素共同描述。
			标签	空间实体间的定性距离，选择标签：”远” / ”近” / ”变远” / ”变近”。

表 2: 子任务二的元素含义

第二十二届中国计算语言学大会论文集，第150页-第158页，哈尔滨，中国，2023年8月3日至5日。

上述例句共有3条标注条目：①空间实体=“我”，时间=“清晨”，方向=“去公园”，事件=“散步”；②空间实体=“他”，时间=“我清晨去公园散步时”，处所=“在电线杆下”，事件=“站”；③空间实体=“菜篮”，处所=“手里”，事件=“提”。

2.3 子任务三：空间场景异同判断

子任务三的输入是两个包含空间场景描述的文本。两个文本在形式上存在差异，除去差异字符串，两个文本的其他部分完全相同，如“他站在电线杆下”和“他站在电线杆旁”只有方位词不同。机器需要判断两个文本是否可以描述相同的空间场景，并输出一段文本，内容包含列出差异字符串、说明空间场景相同或不同的理由。表3 是空间场景相同和不同的两个示例。

输入文本	判断结果	输出文本
我清晨去公园散步时，总能看见他站在电线杆下，手里提着菜篮。 我清晨去公园散步时，总能看见他站在电线杆旁，手里提着菜篮。	相同	两段文本的形式差异在于“电线杆下”和“电线杆旁”。它们都描述了“他”相对于“电线杆”的位置：他站在电线杆周围的区域。因此，它们可以描述相同的空间场景。
市政府计划在火车站广场下开发一个大型美食城，解决往来旅客的休息和餐饮需求问题。 市政府计划在火车站广场旁开发一个大型美食城，解决往来旅客的休息和餐饮需求问题。	不同	两段文本的形式差异在于“广场下”和“广场旁”。前者描述的空间场景中：美食城在火车站广场下方，广场在地上，而美食城在地下。后者描述的空间场景中：美食城和火车站广场都在地上。因此，它们描述的空间场景不同。

表 3: 子任务三示例

3 数据集

子任务一共含7050条语料，每条语料平均约114个字符。子任务二共含2163条语料，每条语料平均约116个字符。表4 是两个任务的数据集规模。子任务三没有划分子集，标注阶段共形成了355条语料，每条语料含两个文本，每个文本平均约72个字符。共有10条语料作为示例公布，帮助参赛队伍理解任务目标，有100条语料用来测试，其余数据没有在本次评测中使用。

子任务	训练集	验证集	测试集	总计
1.空间语义异常识别	4962	700	1388	7050
2.空间语义角色标注	1529	207	427	2163

表 4: 子任务一和子任务二的数据集规模

在语料来源方面，子任务一和子任务二均涉及多种不同来源的语料。表5 是两个任务的语料来源在数据集的占比情况，“其他”包括“语义角色标注课题语料”和“语言学论文例句”。子任务三语料的一个重要来源是SpaCE2022中方位词替换后空间信息仍正常的句子，比较并分析这些替换句和原始句的空间场景，得到了一批空间场景相同和空间场景不同的语料。课题组成员再根据这批语料以及语言学研究成果，自拟形成其他语料。

子任务	人民日报	文学作品	语文课本	体育文本	交通判决书	地理百科	其他
1.空间语义异常识别	29	30	13	17	4	4	3
2.空间语义角色标注	34	22	20	-	11	7	6

表 5: 子任务一和子任务二的语料来源占比（%）

3.1 数据集制作流程

子任务一的训练集和验证集来自SpaCE2022的归因与识别联合任务数据集。该数据集设计了三个归因类型，“搭配不当”类型使用“text1”和“text2”标签来标注异常，其余两个类型使用S-P-E标注法。本届评测统一使用S-P-E标注法，我们根据“text1”和“text2”的词性搭配情况总结了44种转换模式，如名词后搭配方位词转换为P，方位词后搭配名词转换为S，趋向动词归入P，其他动词归入E等等。最后对转换结果进行人工检查，删除了4条问题语料，得到1455条成功转换为S-P-E标注法的语料。

子任务一的测试集包含868条重新修订的语料，它们原本在SpaCE2022数据集制作阶段因质量问题被剔除，经过修订后被质量筛选程序认为质量可靠。为接近7:1:2的子集分布比例，根据方位词缺省情况和语体占比，我们从SpaCE2022的测试集中选取了529条语料进行补充。最终，剔除17条搭配不当转换失败的语料后，形成了1388条语料的测试集。

子任务二的训练集和验证集来自SpaCE2022的语义角色标注数据集。测试集的语料来自去年因质量问题被排除的语料，共1029条。修订后，质量筛选程序认为有673条语料可靠。我们从中按语体占比选择了427条语料构建测试集，让数据集的子集分布比例接近7:1:2。

子任务三的语料制作流程包括自拟题目和交叉审查题目。自拟题目时，出题人根据给定的151个方位词替换对构造文本，并给文本打上标签，true代表正例，空间场景相同，false代表负例，空间场景不同。同时，列出与该空间场景有关的空间实体，然后从处所、朝向、终点、方向、路径和位置关系中选择1个作为判断空间场景异同所要涉及到的元素。审查阶段，出题人之间交叉验证构造的文本是否合乎语法、空间信息以及标签是否正确。最后，得到355条语料，231条（65%）正例，124条（35%）负例。

虽然子任务三对于生成的文本没有格式要求，但为了帮助参赛队伍理解任务目标，我们设计了参考模板并编写了10条示例语料的参考答案。表6 展示了参考模板以及填充实例。

标签	参考模板	填充实例
true	两段文本的形式差异在于<R1>和<R2>。两段文本都出现了以下空间实体：<R3>。尽管两段文本在描述<R4>上有形式差异，但实际上，<R1>和<R2>都描述了<R5>：<R6>。因此，这两段文本可以描述相同的空间场景。	两段文本的形式差异在于<电线杆下>和<电线杆旁>。两段文本都出现了以下空间实体：<他和电线杆>。尽管两段文本在描述<他站的位置>上有形式差异，但实际上，<电线杆下>和<电线杆旁>都描述了<他的处所>：<位于电线杆周围的区域>。因此，这两段文本可以描述相同的空间场景。
false	两段文本的形式差异在于<R1>和<R2>。两段文本都出现了以下空间实体：<R3>。两段文本在描述<R4>上有形式差异，表明<R1>和<R2>描述的<R5>是不同的：<R6>。因此，这两段文本不能描述相同的空间场景。	两段文本的形式差异在于<广场下>和<广场旁>。两段文本都出现了以下空间实体：<火车站广场和美食城>。两段文本在描述<美食城的位置>上有形式差异，表明<广场下>和<广场旁>描述的<美食城的处所>是不同的：<前者在火车站广场下方，后者和火车站广场都在地上>。因此，这两段文本不能描述相同的空间场景。

表 6: 子任务三的参考模板及填充实例

3.2 标签分布情况

子任务一有6个标签：S1、P1、E1、S2、P2、E2。异常文本片段的数量为1-3个时，可选取标签有S1、P1、E1。数量为4-6时，6个标签均可选取。由于标签分布与异常文本片段的选取数量密切相关，表7 统计了每种选取情况对应的标注条目数量[2]。

[2]一条语料可能有多条标注条目

片段数	1	2	3	4	5	6
标注条目数	613	1035	3779	192	48	785

表 7: 子任务一的文本片段选取情况

子任务二的标签是15个元素，一共出现了38663次，表8 是各标签在子任务二数据集中的出现次数。

标签	空间实体	事件	处所	方向	时间	终点	起点	路径
出现次数	12204	8527	7309	3388	2143	1828	950	486
标签	事实性	部位	朝向	参照实体	距离	形状	部件处所	
出现次数	463	447	250	218	218	130	102	

表 8: 子任务二的标签出现次数

子任务三的标签是true和false。100条测试语料中，有54条正例，46条负例，覆盖了87个替换对。元素占比约为：处所（59%）、方向（30%）、终点（4%）、位置关系（3%）、朝向（2%）、路径（2%），与355条总语料的分布一致。

4 评测情况

4.1 评测指标

子任务一设计了2种评价指标，分别是①文本识别准确性、②标签识别准确性。指标①是参考指标，只要字符正确则视为正确，以F1值的形式考察参赛系统对异常文本进行定位的能力；指标②是排名指标，要求字符和标签都正确，以F1值的形式考察参赛系统掌握S-P-E标注法的情况。

子任务二的数据在评测时组织为元组形式，每个元组对应一条空间信息标注，每个句子对应若干个元组。元组不定长，每个元组都记录了标注所使用的标签。评分程序会对参考答案和机器答案中的元组进行两两比较，对于每个参考元组和机器答案元组，程序计算其中每个标签的得分，求和得到该元组的得分，以及该题的总分。最后，根据每题得分计算所有题目的F1值，作为最终得分。

子任务三不公布测试语料，参赛队伍需要提交模型或指令（prompt），由课题组运行程序，生成100道测试语料的结果。评价时，首先看异同判断的标签是否正确，如果错误，该题得0分；如果正确，采用人工评价对机器生成的文本进行解释有效性的打分。分数越高，表示判断空间场景异同的理由解释得越清楚。每题的分数划为0-5共6个等级，表9列出了等级含义。两名评分员均是课题组成员，他们会根据语料对机器生成的文本进行独立打分。最后取两名评分员的均分，转换为100分制，作为最终得分。

分数	等级含义
5分	调用文本以外的常识和世界知识，对空间场景进行重述。
4分	改写差异字符串，对空间场景进行重述。
3分	直接将差异字符串作为理由，未进行改写，没有对空间场景进行重述。
2分	所作解释与空间义有关，但与差异字符串所在的空间场景无关。
1分	所作解释与空间义无关。
0分	未作解释，没有结合文本答题。

表 9: 子任务三的人工评分标准

4.2 评测结果

SpaCE2023共有12支队伍报名，最终1支队伍提交了测试结果。课题组也开发了子任务一和子任务二的基线系统[3]，并测试了大语言模型GPT-3.5版本的应用ChatGPT在三个任务上的表

[3]https://github.com/2030NLP/SPaCE_Baseline_2023

现，评测结果见表10。

系统	子任务一（F1） 文本准确性	标签准确性	子任务二（F1）	子任务三（百分制）
复旦大学	**0.6526**	**0.5782**	0.4739	37.40
基线	0.6236	0.5547	**0.4803**	-
ChatGPT	0.4639	0.2521	0.1378	**40.40**

表 10: 评测结果

4.2.1 子任务一的模型与方法

复旦大学的参赛系统[4]使用阅读理解任务的范式来完成子任务一，将任务简化为对六个文本片段的抽取，每个文本片段对应于一个标签。他们采用了deberta-chinese-large(He et al., 2020)这个中文预训练模型，并对该抽取任务做微调，预测每个异常文本片段的开头和结尾。当抽取的片段数量不足六个时，模型会把位置指向文本开头设置的标记。基线系统使用了预训练模型BERT(Devlin et al., 2019)，采用了序列标注任务的范式，设置了一个序列标注层来判断每个词所属的标签。

复旦系统的F1值略高于基线系统。具体到七种语料来源上，二者在人民日报、文学作品、语文课本、体育文本以及其他上的表现仅相差1-4个百分点，但在地理百科上，复旦系统比基线系统高出约10个百分点，在交通判决书上高出约50个百分点。以“被告人宋某某驾驶牌号为沪C9XXXX的小型轿车沿本区甘德路由东向西行驶至辰塔路向东右转”为例，抽取式的复旦系统正确识别出[**S1轿车，P1由东向西，E1行驶**]和[**S2轿车，P2向东，E2右转**]的冲突之处，而序列标注式的基线系统可能在子句较长的情况下，逐词标注会受到非目标词的干扰，标注了“9X”、“轿车沿本”、“区甘”等不成词的片段。另外，在4-6个片段的数量上，基线系统也远少于复旦系统。以6个片段为例，基线系统中仅出现了6次，而复旦系统出现了212次。不过，序列标注式模型面对文本片段不连续的情况可能有一定的优势，如“塞到那个女学生座位四面”的异常片段是“到四面”，复旦系统选取的区间包括了“那个女学生座位”，基线系统给出了“到生四面”，更为接近答案。

两个系统的文本准确性和标签准确性都仅约0.6，仍有待提高。我们考察了130道两个系统的标签准确性都为0分的题目后，总结了机器表现较差的三个方面：①**介词异常**，如“骑自（向）”、“由（从）头到脚”；②**不常见但空间义正常的搭配**，如“趴在窗户边”、“站在田埂边”，机器认为这些搭配存在异常；③**结合语境和常识才能发现的异常**。比如，在上下文描述房间内部环境的文本中，“房外又热又闷”存在异常。再如，“在蜂箱里忙碌的姚生”违反了人不能在蜂箱内部的常识。这些异常机器都没有发现。

4.2.2 子任务二的模型与方法

子任务二中，复旦系统采用抽取加生成的方法，训练了一个deberta抽取器来抽取“空间实体”元素，再训练一个CPT生成器(Shao et al., 2021)来生成剩余的元素。基线系统将子任务视为事件抽取任务，先抽取触发词，即事件元素，再围绕触发词抽取与其相关的其他元素。

复旦系统的F1值接近但未超过基线系统。两个系统所采用方法的差别反映在空间实体的数量差异上，复旦系统以空间实体为中心抽取其他元素，识别了2468个空间实体，而基线系统以事件为中心抽取其他元素，不善于处理缺少触发词的元组，仅识别了1808个空间实体。测试集共标注了3029个空间实体，其中约30%的元组缺少事件元素，从这一点看，以空间实体为中心的抽取方法可能比较符合数据集的特点。

复旦系统仅考虑片段型元素，忽略标签型元素，没有输出事实性元素、距离标签和方向标签。基线系统则没有输出时间标签和与距离关系有关的参照实体和距离元素。未来可进一步研究标签型元素和低频元素的识别。

4.2.3 子任务三的指令与方法

复旦大学使用了大语言模型来完成子任务三，在设计指令时结合COT（chain of thought，思维链）(Wei et al., 2022)和few-shot方法(Brown et al., 2020)，通过COT引导模型推理出更好

[4]https://github.com/ShacklesLay/Space2023

的结果，然后提供1个空间场景相同的样例和1个空间场景不同的样例，帮助模型理解输出格式以及推理的过程。

COT指令包含5个步骤：1.找出两个文本的差异字符C1和C2；2.找出差异字符所在的完整短语P1和P2。有时仅凭差异文本不足以判断空间场景是否异同，所以扩展了判断空间；3.找出短语中包含的空间实体；4.判断空间场景异同；5.选择模板进行输出。模板是课题组提供的参考模板。

两个样例均包括输入、思维链和输出三个部分，表11 展示了指令中空间场景相同的样例。最后，课题组使用提交的指令，在ChatGPT上运行了100道测试语料的结果，共有65题判断正确，进入人工评分阶段，得分为37.40分。

输入	思维链	输出
”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。”	差异在于“脖子上”和“脖子下”。 根据提供的上下文，包含C1的完整短语P1可以是：”脖子上系着领带”；而包含C2的完整短语P2则可以是：”脖子下系着领带”。 P1和P2中包含的空间实体如下：“小伙子”、“脖子”、“领带”。 根据P1和P2的描述，它们所描述的空间实体”领带” 位于相同的位置，P1描述的是领带在脖子表面，而P2描述的是领带在脖子胸前。因此，它们所处的空间场景一致。	{”results”:[{ ”judge”:”true”, ”reason”:”两段文本的形式差异在于“脖子上”和“脖子下”。两段文本中都出现了以下空间实体：“小伙子”、“脖子”和“领带”。尽管两段文本在描述“领带”系着的位置上有形式差异，但实际上，“脖子上系着领带”和“脖子下系着领带”描述“领带”的处所是相同的，都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。”}]}

表 11: 复旦大学指令中的样例

4.2.4 ChaptGPT的表现

SpaCE2023测试了GPT3.5版本的ChatGPT在数据集上的表现。子任务一的指令说明了S-P-E标注法。受限于ChatGPT可接受的输入长度，我们提供了5条样例帮助模型理解任务目标和输出格式，这5条样例涵盖了5种语料来源和6种文本片段。子任务二的指令说明了15个元素的概念，提供了3条样例，涵盖3种语料来源和9个元素。子任务三则直接提供10个样例以供学习，每个样例包括语料文本、判断结果和按模板编写好的判断理由。样例涉及处所、方向、终点、位置关系这四个要素，含5条正例和5条负例。对于子任务一和子任务二的输出，我们发现文本片段的下标存在很多错误，所以用程序重新寻找了每个文本片段的下标。

子任务一的得分反映出ChatGPT仅找出了约46%与空间异常有关的片段，且并未有效掌握S-P-E标注法。输出结果主要存在两个问题：①**P信息的成分较为混乱**，填写了许多不表示方位信息的动词性成分和形容词性成分；②**大量使用4-6个片段来描述异常**，但填写的片段并没有信息冲突。在接下来的工作中，子任务一的指令设计将描述每个元素可填入的语法单位，规定每种文本片段的使用情况。

子任务二的F1值仅有0.1分。输出结果含有许多与空间语义无关的标注，如“他等父亲”标注了空间实体“他”和事件“等”，说明没有理解任务目标。同一个标签也在同一个元组中使用多次，如一个元组标注了两个及以上的空间实体，甚至出现“幻觉”，使用了规范以外的元素，如“数量”、“状态”、“情感”等。子任务二的指令需要采用新的设计思路，如使用思维链的方式让其找出所有空间实体，然后围绕每个空间实体寻找相关的元素。指令还应说明元素的使用限制，提供更多用例展示元素的使用条件。

子任务三的输出中，共有71题判断正确，较好地理解了任务目标。人工评分阶段获得40.40分，生成的文本主要存在两个问题：

（1）与空间义无关。如“等潮水涨**上来**时才登上小岛”和“等潮水涨**起来**时才登上小岛”描述的空间场景是相同的，ChatGPT给出的理由是“都描述了登上小岛的时间”，没有涉及潮水上涨的方向。

（2）直接将差异字符串作为解释。如“在名称**中**使用’示范区’字样”和“在名称**前**使用’示范区’字样”描述的空间场景是相同的，但所作解释是“企业可以在名称中或名称前使用’示范区’字样”，仍然没有说明相同的原因。

在改进子任务三的指令时，可将模板槽位拆解为多个问答题的形式，引导大语言模型做出更好地回答。

5 结论与未来工作

本文介绍了第三届中文空间语义理解评测的概况。SpaCE2023在提高语料质量的基础上，优化了异常识别任务、空间标注任务的底层设计，创新性地提出了空间场景异同判断任务，提供了考察空间语义理解能力的新角度。评测最终吸引了复旦大学队伍参赛，他们研发的抽取式模型在子任务一上取得0.5782的F1值，超过了基线模型。子任务二使用了抽取加生成的方法，虽然未能超过基线系统，但所采用的以空间实体为中心的抽取方法贴合任务设计。在接下来的工作中，如何有效区分异常的空间信息和正常的空间信息、如何不遗漏地提取空间元素可能是新的重点和难点。

本届评测还出现了大语言模型的身影。复旦大学参赛队使用思维链和提供少样本的方法引导大语言模型进行空间场景异同判断并生成理由，取得了37.40分的成绩。课题组测试了ChatGPT在SpaCE2023数据集上的表现，子任务一和子任务二的表现远不如参赛系统和基线系统，未来的指令设计应着力于解构任务，帮助大模型理解任务目标。ChatGPT对子任务三这一生成式任务的理解较好，仅提供2个示例就能对两个文本的空间场景进行比较并判断异同，但对于判断作出的解释仍不够到位，应继续引导它结合世界知识，对空间场景进行重述，比如将“车下”进一步描述为“车辆底盘与地面的空隙区域”。

整体来看，参赛系统、基线系统和ChatGPT在三个子任务上的得分均有待提高，反映出空间语义的复杂性。我们将继续提出新的评测角度，开展新的评测任务，以期推动空间语义理解任务乃至语言认知类评测任务的研究和发展。大语言模型的空间语义理解能力是未来重要的研究目标之一，我们将设计更具针对性的评测方法。

参考文献

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Hang Yan, Fei Yang, Li Zhe, Hujun Bao, and Xipeng Qiu. 2021. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. *arXiv preprint arXiv:2109.05729*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

詹卫东, 孙春晖, 岳朋雪, 唐乾桐, and 秦梓巍. 2022. 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, (02):99–110.