SpaCE2024 异形同义任务报告

秦宇航

北京大学中国语言文学系

2024 年 5 月 29 日

目录

目录

语料划分

语料分为训练集、验证集、测试集。由表1可知，本数据集以测试功能为主。

训练集	验证集	测试集	总计
5	55	650	710

表 1: 数据集语料划分

总体情况

本数据集按照考察的空间语义类型分为如下 2 类：

1. 空间异形同义类：正确选项替换题目中所给词语后，两段语料的空间语义合理，且可以表示相同的空间语义；
2. 空间异形异义类：正确选项替换题目中所给词语后，两段语料的空间语义合理，但不能表示相同的空间语义。

此外，本数据集包含单选题和多选题。

总体情况

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

表 2: 数据集各类型题目数量及比例

如表2所示，数据集以异形同义、单选为主。

总体情况

数据集中包含各选项的题目数量如表3所示。由表3可知，数据集的选项分布相对均衡。

选项	A	B	C	D
包含该选项的试题数	207	207	**214**	192

表 3: 选项在试题中的分布

语料溯源

- 数据集语料存在 2 个来源：
  1. SpaCE2022 任务 1 中的语料
  2. 工作组成员自拟语料
- 不同来源的语料在形式上存在区分
  1. 第 1 部分语料记录 SpaCE2022 赋予的 origin_id 值
  2. 第 2 部分语料赋予“synthetic”（合成）标签

语料溯源

	合成语料		非合成语料	
	数量	占比	数量	占比
训练集	3	**60.00%**	2	40.00%
验证集	33	**60.00%**	22	40.00%
测试集	419	**64.46%**	231	35.54%
总计	455	**64.08%**	255	35.92%

表 4: 合成语料和非合成语料在数据集的分布

由表4可知，本数据集中语料以合成语料为主，需要考虑实现语料收集、改写的自动化。

语料分批

语料分为 4 批（batch）。
批次间主要区别在于对语料合理性的审核不同。

- 第 1 批语料由 1 人收集或构造，2 人审核
- 其他批次语料由 1 人收集或构造，1 人审核

语料分批

	批次 1		批次 2		批次 3		批次 4	
训	**4**	**80.00%**	0	0.00%	1	20.00%	0	0.00%
验	**36**	**65.45%**	5	9.09%	5	9.09%	9	16.36%
测	**518**	**79.69%**	50	7.69%	16	2.46%	66	10.15%
全	**558**	**78.59%**	22	3.10%	55	7.75%	75	10.56%

表 5: 各批次语料在数据集中的分布

目录

总体表现

本次机器结果分析取评测总分排名前 6 的队伍提供的结果进行分析。这些队伍在本次任务上的总体表现得分如表6所示。由表6可知，本任务上模型在单选题上的表现好于多选题。

	第一名	第二名	第三名	第四名	第五名	第六名	平均
得分	**0.676923**	0.563077	0.543077	0.492308	0.586154	0.569231	0.571795
单选题	**0.711799**	0.605416	0.603482	0.597679	0.659574	0.649903	0.637975
多选题	**0.526316**	0.195489	0.308271	0.082707	0.210526	0.255639	0.263158

表 6: 排名前 6 的队伍在本任务上的总体得分

试题表现

试题上的主要表现如下：

- 对所有被分析的参赛队伍，在每道题目上计算相对于标准答案的 f1-score，这些值求平均得到该题的 f1-score，再对全部试题的 f1-score 做平均，该平均值为 0.658884。
- 存在 22 道试题，其 f1-score 为 0（即 6 支队伍全部做错）。

试题表现

- 从空间语义类型的角度，由表2和7、8可知，模型在空间异形异义类上表现相对更差。
- 正确项为“以上选项均不正确”的所有试题的平均 f1-score 为 0.403703，且所有试题的 f1-score 均小于 1（即至少有 1 支队伍答错）。
- 由表9，推测选项位置可能回影响模型回答问题时的表现：选项越靠后，错误率越高。
- 低 f1 试题的替换对情况没有明显规律。

试题表现

f1 均值	异形同义类	异形异义类
$f1 = 0$	7	15
$f1 > 0.17$	35	32

表 7: 模型低 f1 均值试题类型

语义类型	f1 均值
异形同义类	0.681394
异形异义类	0.61105

表 8: 模型各类型试题 f1 均值

试题表现

选项	A	B	C	D
$f1 = 0$ 试题数	4	5	**8**	5
$f1 < 0.17$ 试题数	13	18	16	**20**

表 9: 选项在低 f1 试题中的分布

词对表现

定义：命中及命中率

1. 命中：每道题目中至少存在 1 组正确的替换词对（“以上选项均不正确”视为特殊的词）。答者（含模型及人类被试）给出的答案中，若选中了这道题目中的特定替换词对，则称为这个词对的一个命中。
2. 命中率：某答者某词对的命中率定义为

$$\frac{\text{某答者该词对的命中数}}{\text{该词对为正解之一的题目总数}}$$

词对表现

第一名	第二名	第三名	第四名	第五名	第六名
上-下	上-下	上-下	上-下	上-里	上-下
上-里	上-里	上-里	上-里	上-下	上-中
上-中	上-中	上-中	上-中	上-中	下-里
下-里	上-前	下-里	下-里	下-里	上-里
上面-外面	下-里	下面-里面	上面-外面	中-里	中-里
中-里	上面-外面	上-前	中-里	下面-里面	下面-里面
下面-里面	上-外	上面-外面	下面-里面	上面-外面	下-旁
上-前	中-里	中-里	上面-里面	上面-里面	上-前
上-外	下面-里面	下-旁	上-前	下-前	上面-外面
下-旁	前-里	上-旁	上-外	下-旁	下-边

表 10: 各模型命中率最高的替换词对

词对表现

第一名	第二名	第三名	第四名	第五名	第六名
中-里	旁-边	旁-边	上面-里面	下-里	中-里
上面-外面	中-内	上-前	旁-边	中-里	下-旁
前-外	下去-进去	下去-进去	中-内	下面-里面	旁-边
旁-边	内-里	过来-进来	内-里	上面-里面	上-内
中-内	过来-进来	上面-侧面	前边-外边	上-内	外面-旁边
下去-进去	上来-回来	上去-进去	边-里	外面-旁边	中-内
内-里	前-里	下面-外面	下边-里边	中-内	下去-进去
过来-进来	上面-下面	前-后	上去-过去	内-里	内-里
下-外	外边-旁边	下来-出来	过去-进去	上来-过来	前-里
前面-外面	过去-进去	上去-过去	下去-过去	上去-下去	边-里

表 11: 各模型正例命中率最高的替换词对

词对表现

第一名	第二名	第三名	第四名	第五名	第六名
上-中	上-下	上-下	前-里	上-中	上-下
外-里	上-前	上面-后面	上面-下面	外-里	上-中
后面-里面	前-里	上面-下面	旁-里	上面-旁边	上-外
中-旁	上面-后面	外-里	上-里	上边-下边	上面-外面
外边-里边	中-边	中间-对面	下面-里面	内-外	上面-下面
上-里	上面-下面	上-里	上-后	中-外	上-内
下-旁	外-里	内-外	边-里	前边-里边	外-里
下面-里面	旁-里	中-外	内-外	中间-外面	上面-旁边
内-外	中-旁	旁边-里面	中-外	下面-侧面	上-里
前面-里面	上边-下边	前边-后边	前边-里边	下-外	下-旁

表 12: 各模型负例命中率最高的替换词对

词对表现

由表10、表11及表12可知：

- 模型高命中率替换词对具有强一致性。结合表13中试题数量与命中数的极强相关性，猜测其与词语在语言系统的频次有关。

	第一名	第二名	第三名	第四名	第五名	第六名
全部类型	**0.973347**	0.962107	0.951056	0.95911	0.918503	0.949012
正例	**0.961865**	0.945562	0.937423	0.921094	0.945175	0.901272
负例	**0.970186**	0.964506	0.926937	0.96114	0.875208	0.951152

表 13: 各类型试题上包含各替换对的试题数量与各模型的命中数的相关系数

词对表现

- 由表11和表12，正、负例命中率高的替换词对区别较大，正、负例命中率高的替换词对之间的表层语义分别呈相近、相反关系。结合表14中的弱一致性，模型主要理解方位义词语的表面语义，深层空间语义理解能力较弱。

第一名	第二名	第三名	第四名	第五名	第六名
-0.26241	-0.44257	-0.27329	-0.32206	-0.35663	-0.3179

表 14: 数据集中出现次数前 100 的替换对上各替换对的正例命中率、负例命中率的相关系数

两两一致

由表15，模型间的一致率不高。

	第一名	第二名	第三名	第四名	第五名	第六名
第一名	1	0.633216	0.583194	0.580044	0.590308	0.644667
第二名	0.633216	1	0.650476	0.535487	0.557692	0.598029
第三名	0.583194	0.650476	1	0.579194	0.582154	0.601582
第四名	0.580044	0.535487	0.579194	1	0.610256	0.598725
第五名	0.590308	0.557692	0.582154	0.610256	1	0.608103
第六名	0.644667	0.598029	0.601582	0.598725	0.608103	1

表 15: 模型间的 f1 均值矩阵

本章小结

本章主要发现，机器在本任务上的表现具有如下特征：

1. 模型在单选题上的表现好于多选题；
2. 模型在空间异形异义类试题和正确项为“以上选项均不正确”的试题上表现较差，选项位置可能影响回答准确率；
3. 模型以方位词的字面义理解空间表达，缺乏理解深层空间语义的能力；
4. 模型间的一致率有待提高。

目录

总体表现

我们从数据集的测试集中抽取了 100 道试题，招募 8 名汉语母语者被试进行问卷测试。数据集的总体情况如表16所示。

	单选题	多选题	合计
异形同义类	58	9	67
异形异义类	22	11	33
合计	80	20	100

表 16: 人类测试题总体情况

总体表现

由表17可知，除被试 2 外，其余被试均表现较好，且单选题表现好于多选题。后续分析排除被试 2。

	被试 1	被试 2	被试 3	被试 4	被试 5	被试 6	被试 7	被试 8
正解题数	87	66	**88**	84	79	85	86	84
正解比例	0.87	0.66	**0.88**	0.84	0.79	0.85	0.86	0.84
正解单选题数	72	58	**74**	72	67	70	71	71
正解单选比例	0.9	0.725	**0.925**	0.9	0.8375	0.875	0.8875	0.8875
正解多选题数	**15**	8	14	12	12	**15**	**15**	13
正解多选比例	**0.75**	0.4	0.7	0.6	0.6	**0.75**	**0.75**	0.65
重测一致性	**1**	0.6	0.9	0.9	0.7	0.9	0.9	0.9

表 17: 人类被试总体表现

试题表现

试题上的主要表现如下：

- 总体 f1-score 平均值为 0.892810.
- f1-score 平均值为 0 的试题 1 道，经检查发现答案有误，修改答案后 7 名被试回答全部正确。
- f1-score 平均值为 0.428571（即 3 人同意标准答案）的试题有 2 道；f1-score 平均值为 0.571429 的试题有 5 道。

试题表现

在 f1-score 平均值较低的试题中：

- 有 1 题答案不合适，需要排除；
- 其余试题问题主要集中在 2 对方位词“上面-侧面”“中/内-前”上：
  - 2 月 4 日，贾娜医生的核酸检测结果显示为阴性，她痊愈了。隔离 7 天后，贾娜重新穿上了那身熟悉的“战袍”。看着镜子**中**的自己，她暗自打气：病毒不可怕，一起战胜它！（有 3 位被试选择“后”）
  - 校长办公桌上面有一个方形的鱼缸。饲养说明死死地粘在鱼缸的**上面**。（有 3 名被试选择“侧面”，有 1 名被试选择“里面”）

词对表现

人类的“命中”“命中率”概念与模型评估部分相同。由表18、表19及表20知，相比于模型，人类正、负例命中率高的替换词对区别不大，机器在达到人类理解深层空间语义方面尚有欠缺。

词对表现

被试 1	被试 3	被试 4	被试 5	被试 6	被试 7	被试 8
上面-外面	上面-外面	上面-外面	上面-外面	上面-外面	上面-外面	上面-外面
上-下	上-下	上-下	上-下	上-下	上-下	上-下
下-旁	下-旁	下-旁	下-旁	下-旁	下-旁	下-旁
上-中	上-中	下-里	上-中	上-中	上-中	上-边
上-边	下-里	上-边	下-里	下-里	下-里	上-中
上-内	上-内	上-中	上-内	上-内	上-边	上-内
下-里	上-边	内-前	上-边	上-边	中-里	下-里
内-前	内-前	中-里	中-里	中-里	上-内	中-里
中-里	中-里	上面-里面	内-前	上-后	上-后	上面-里面
上面-里面	上-后	上-后	上面-里面	后面-外面	后面-外面	后面-外面

表 18: 人类命中率最高的替换词对

词对表现

被试 1	被试 3	被试 4	被试 5	被试 6	被试 7	被试 8
下-旁	上面-外面	上面-外面	上面-外面	上面-外面	上面-外面	上面-外面
上-中	上-下	下-旁	下-旁	上-下	上-下	上-下
上-内	下-旁	中-里	上-中	下-旁	下-旁	下-旁
中-里	上-中	下-里	上-内	上-中	上-中	上-中
上-边	上-内	上-边	中-里	上-内	中-里	上-内
上面-里面	中-里	上面-里面	下-里	中-里	下-里	中-里
内-前	下-里	内-前	上-边	下-里	上-边	上-边
外边-旁边	内-前	下-外	上面-里面	上-边	外边-旁边	上面-里面
中-内	中-内	下面-前面	外边-旁边	外边-旁边	中-内	外边-旁边
下-外	下-外	上面-中间	中-内	中-内	下-外	中-内

表 19: 人类正例命中率最高的替换词对

词对表现

被试 1	被试 3	被试 4	被试 5	被试 6	被试 7	被试 8
上-下	上-下	上-下	上-下	上-下	上-下	上-下
上面-外面	上面-外面	上面-外面	上面-外面	上面-外面	上面-外面	上面-外面
后面-外面	上-后	上-后	上-后	上-后	上-后	后面-外面
中间-前面	后面-外面	外面-里面	外面-里面	后面-外面	后面-外面	外面-里面
下-旁	外面-里面	中间-前面	下-旁	外面-里面	外面-里面	中间-前面
上-中	中间-前面	下-旁	上-中	中间-前面	中间-前面	下-旁
上-里	上面-旁边	上-中	上-里	上面-旁边	上面-旁边	内-前
内-前	下-旁	上-里	内-前	下-旁	下-旁	中-前
中-前	上-中	内-前	中-前	上-中	上-中	中-后
中-后	上-里	中-前	中-后	上-里	上-里	上边-里边

表 20: 人类负例命中率最高的替换词对

两两一致

由表21，该矩阵非对角线元素的平均值为0.848721。这说明人类被试间的一致性较好，但仍有较大差异。

	被试 1	被试 3	被试 4	被试 5	被试 5	被试 6	被试 7
被试 1	1	0.869238	0.857	0.832333	0.838	0.84	0.84
被试 3	0.869238	1	0.864238	0.864333	0.880667	0.916333	0.854333
被试 4	0.857	0.864238	1	0.822333	0.813	0.845	0.823
被试 5	0.832333	0.864333	0.822333	1	0.821333	0.818	0.796
被试 6	0.838	0.880667	0.813	0.821333	1	0.884	0.876
被试 7	0.84	0.916333	0.845	0.818	0.884	1	0.868
被试 8	0.84	0.854333	0.823	0.796	0.876	0.868	1

表 21: 人类被试间的 f1 均值矩阵

人机对比

由表22来看，模型在人类测试题上的表现和模型在总体测试题上的表现相差不大，与人类的总体表现还有较大差距。

	第一名	第二名	第三名	第四名	第五名	第六名	平均
人类	**0.71**	0.49	0.58	0.49	0.53	0.58	0.563333
总体	**0.676923**	0.563077	0.543077	0.492308	0.586154	0.569231	0.571795

表 22: 模型在人类测试题上的表现

本章小结

本章主要考察了人类被试的表现以评估任务质量、试题质量和机器表现。目前发现：

- 任务质量上，本任务属于高难度任务，模糊性和争议性相对较大，人类得分低于预期。
- 试题质量上，试题质量较高，人类测试题有较好的代表性。相对 SpaCE2023 而言，本任务试题制作难度进一步增大，针对试题错误和争议问题需要进一步改进。
- 机器表现上，目前机器表现与人类尚存较大差距。这一差距可能来源于机器对深层空间语义的理解不足，需要进一步评估。

后续问题

- 本数据集中存在“同文题”，而在人类测试数据集中这类题目的比例更高。
- 所谓“同文题”，指的是具有相同题干文本的 2 道试题。这些试题可能属于相同的空间语义类型或不同的空间语义类型。示例见下页。
- 后续需要针对模型和人类在同文题上的表现进行更细致的评价，以考察答者对于深层空间语义的理解能力。

后续问题

“同文题”示例：

- TEXT: 毕业典礼结束了。其他同学早已收拾好所有物品，离开这所学校。只有小茜站在教室外面，看着空荡荡的教室，心中感慨万千。
- QUESTIONS: “只有小茜站在教室外面”中的“外面”替换为 () 形成的新句可以与原句表达相同的空间场景/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。

后续问题

- OPTIONS:
  - 后面
  - 旁边
  - 里面
  - 上面
- ANSWERS: 旁边/后面 & 里面

请大家批评指正！