SpaCE2022中文空间语义理解评测任务数据集分析报告

肖力铭[1,2,‡] 孙春晖[1,2,†] 詹卫东[1,2,3,†,*] 邢丹[1,2,‡] 李楠[1,2,†] 王诚文[3,†] 祝方韦[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
[†]{sch,zwd,linan2017,wangcw}@pku.edu.cn
[‡]{lmxiao,xingdan,zhufangwei2022}@stu.pku.edu.cn

摘要

第二届中文空间语义理解评测任务（SpaCE2022）旨在测试机器的空间语义理解能力，包括三个子任务：（1）中文空间语义正误判断任务；（2）中文空间语义异常归因与异常文本识别任务；（3）中文空间实体识别与空间方位关系标注任务。本文围绕SpaCE2022数据集介绍了标注规范和数据集制作流程，总结了改善数据集质量的方法，包括构建STEP标注体系，规范描述空间语义信息；基于语言学知识生成空间异常句子，提高数据多样性；采取双人标注、基于规则的实时质检、人工抽样审核等方式加强数据质量控制；分级管理标注数据，优选高质量数据进入数据集。通过考察数据集分布情况以及机器表现和人类表现，本文发现SpaCE2022数据集的标签分布存在明显偏差，而且正误判断任务和异常归因任务的主观性强，一致性低，这些问题有待在将来的SpaCE任务设计中做进一步优化。

关键词： 中文空间语义理解；评测基准数据集；质量控制；STEP标注规范

A Quality Assessment Report of the Chinese Spatial Cognition Evaluation Benchmark

Liming Xiao[1,2,‡] Chunhui Sun[1,2,†] Weidong Zhan[1,2,3,†,*] Dan Xing[1,2,‡]

Nan Li[1,2,†] Chengwen Wang[3,†] Fangwei Zhu[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
[†]{sch,zwd,linan2017,wangcw}@pku.edu.cn
[‡]{lmxiao,xingdan,zhufangwei2022}@stu.pku.edu.cn

Abstract

The Second Chinese Spatial Cognition Evaluation Task (SpaCE2022) aims to test the machine's spatial semantic understanding capabilities, including three subtasks: (1) to judge whether the spatial information in a sentence is correct or incorrect; (2) to determine what causes the abnormal spatial information in a sentence, and locate text fragments with wrong information in a sentence; (3) to label the spatial roles and relations for spatial entities in a sentence. This paper introduces the annotation specifications and the development of the SpaCE2022 dataset, summarizing four quality control methods used in the project. A STEP annotation specification is proposed to standardize the annotation of spatial information of a sentence. Under the guidance of linguistic knowledge, we used the method of replacing spatial semantic words in

*通讯作者
基金项目：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）

sentences to generate a large number of sentences that may contain spatial semantic anomalies. The sentences generated in this way are diverse in type. Quality control measures in corpus annotation include double-person annotation, rule-based real-time inspection, manual sampling review. According to different sources of annotated corpus, quality classification is carried out, and high-quality annotated data is selected into the final evaluation data set. By examining the dataset distribution as well as machine performance and human performance, this paper finds that the label distribution of the SpaCE2022 dataset exhibits significant bias. Both the correctness judgment task and spatial anomaly attribution task in SpaCE2022 are highly subjective and have low consistency, requiring further optimization in future version of SpaCE benchmark.

Keywords: Chinese Spatial Cognition Evaluation , Benchmark , Quality control , STEP annotation specification

1 引言

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，自然语言处理领域发布了多项评测任务，具体可以分为以下三类：（1）**空间信息标注任务**，要求机器根据给定的语义角色标注文本中的空间实体和空间关系，形式上与语义角色标注任务以及事件抽取任务相当，代表性的工作有SpRL任务(Kordjamshidi et al., 2012)；(Kolomiyets et al., 2013)和SpaceEval任务(Pustejovsky et al., 2015)；（2）**空间关系推理任务**，要求机器根据文本中已有的空间信息回答涉及空间关系推理的问题，代表性的工作有bAbI任务集(Weston et al., 2015)中的位置推理任务和路径推理任务，以及SpartQA任务(Mirzaee et al., 2021)；（3）**空间语义异常判断任务**，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，CCL 2021发布的中文空间语义理解评测任务（Spatial Cognition Evaluation 2021，简称SpaCE2021）(詹卫东et al., 2022)首次提出此类任务，认为如果机器能甄别错误的空间信息并进行正确的归因，就说明机器具有一定的空间语义理解能力[0]。

三类任务提供了评测机器空间语义理解能力的不同角度，但单从一个角度出发并不能全面评估机器的空间语义理解能力。标注类任务只要求对空间相关元素加以识别，因而无法进一步验证机器是否真正理解了这些元素的语义。在已有的推理类任务中，语料都是围绕人为设计的几何图形场景构造出的非真实文本，其空间表达方面的自然度与人类真实对话仍存在一定差距。判断类任务不要求机器定位造成空间语义异常的文本片段，不能测试机器提取空间信息的能力。

在这样的背景下，第二十一届中国计算语言学大会（CCL 2022）发布了SpaCE2022技术评测任务[1]。SpaCE2022包含三个子任务：（1）**中文空间语义正误判断任务**，要求机器对文本的空间语义给出正常或异常的判断；（2）**中文空间语义异常归因与异常文本识别任务**，要求机器识别给定中文文本中空间信息异常的片段，并给出归因类型；（3）**中文空间实体识别与空间方位关系标注任务**，要求机器基于空间关系标注规范，对给定中文文本进行空间实体的识别与空间方位关系标注。

本文介绍SpaCE2022数据集的总体情况，并对参赛系统的表现和数据集质量进行评估和分析，指出数据集和机器模型存在的问题，以及这一任务相关的数据资源潜在的研究价值，为语言认知类评测任务的设计提供借鉴。下文第2节介绍数据集的标注规范、制作流程和数据分布情况，第3节展示基线模型和参赛队伍系统在各项任务上的表现，第4节通过人类表现来评估数据集质量，第5节总结数据集质量控制思路，展望语言认知类评测任务的发展。

[0]http://ccl.pku.edu.cn:8084/SpaCE2021/
[1]https://2030nlp.github.io/SpaCE2022/

第二十二届中国计算语言学大会论文集，第547页-第558页，哈尔滨，中国，2023年8月3日至5日。

2 SpaCE2022数据集总体情况

2.1 数据集标注规范

SpaCE2022数据集的标注面向三个子任务分为了三个层面[2]。一是面向空间语义正误判断任务，仅标注文本中是否存在异常空间信息，存在即标注“异常”，否则标注“正常”；二是面向空间语义异常归因任务，针对上一层标注中标签为“异常”的语料，进一步标注异常的类型以及存在异常的具体文本片段；三是面向空间实体识别与空间方位关系标注任务，精选了部分语料，基于STEP空间语义标注体系（S表示空间实体，T表示时间，E表示事件，P表示空间方位信息）进行精细的空间信息标注。

子任务一采用区间标注法，构造了一个空间语义完全正常到显然异常的连续区间，然后切分为四段子区间，分别对应四个标注选项，依次是：完全正常、尚能说通、比较牵强、显然异常。**完全正常**指句中实体的空间方位信息表达正确，毫无争议。**尚能说通**指实体的空间方位信息表达大致能成立，但表达的准确性和自然程度上存在一些问题，如“他听到一段钢琴声从茅屋边传出来”，用“茅屋里”表达感觉更自然，但“茅屋边”的表达也有出现的可能。**比较牵强**指实体的空间方位信息表达不大能成立，虽然这个句子的空间语义还没有到完全无法理解的程度，但空间语义信息罕见，如“他带着狗到森林旁去打猎”，这种情况基本不会出现，但从语义上看，“森林旁”所表示的空间关系仍然成立。**显然异常**指句中实体成分的空间方位信息表达错误，如“每辆自行车下座上都用一根扁担绑着两只很大的箩筐”，自行车只有“后座”，没有“下座”，与常识相悖，所以标注这个句子的空间信息为“显然异常”。

子任务二的标注需要选择造成空间信息异常的原因，并标注异常信息对应的文本片段。三个归因类型选项分别是：搭配不当、语义冲突、不符合常识或背景信息（以下简称不合常识），分别对应A、B、C三类标签。表1 是各选项的含义和标注示例。

选项	含义	示例
搭配不当	句中两个表示空间信息的语言成分，受到语法、韵律、语言习惯等因素的影响，不能组合。而且，这两个成分在其他语境中也都不能组合。	弯曲双膝，弯曲双肘，把头放在[**text1 地面**][**text2 边**]。
语义冲突	句中两处空间信息存在矛盾、冲突。这两处空间信息对应着两个事件。	池水照见了她的面容和身影；她笑，[**P1 池水里**]的[**S1 影子**]也向着她[**E1 笑**]；她假装生气，[**P2 池水外**]的[**S2 影子**]也向着她[**E2 生气**]。
不符合常识或背景信息	句中有一处文本片段的空间信息违反常识或者违反句子的背景信息。	那个苏联人孤零零地躺在那毫无遮掩的方场上，[**S 一只手臂**][**E 枕**][**P 在脑袋上面**]。

表 1: 子任务二的标注选项说明

在描述异常方面，搭配不当涉及语言使用中的搭配习惯问题，即两个相邻成分text1和text2是错误的组合。语义冲突和不合常识这两类的情况则更为复杂。为了更加规范地描述空间信息异常，子任务二引入空间语义三要素来描述句子的空间语义：空间实体（S）、空间方位（P）、空间义相关事件（E）。S指句中描述了空间方位信息的实体，P指空间实体在句中出现的空间方位信息，可能涉及处所、起点、终点等信息，E是动词性单位，表达了S位于P的方式、目的或原因。不合常识选择一组S-P-E三要素即可说明空间信息异常，而语义冲突需要两组S-P-E三要素来说明问题，见表1 中的示例。

子任务三在S-P-E标注法的基础上增加了时间信息的标注，形成了STEP空间语义四要素标注体系，描述信息可概括为：“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

[2]了解标注规范详情请访问规范文档（https://2030nlp.github.io/Sp22AnnoOL/menu）和标注样例（https://2030nlp.github.io/Sp22AnnoOL/examples）。

第二十二届中国计算语言学大会论文集，第547页-第558页，哈尔滨，中国，2023年8月3日至5日。

命题的事实性为真/假”。事实性为假的空间信息用F要素来表示。按照该体系标注一段文本的空间信息，一条信息对应的数据为一个包含18项元素的元组，表2 是每个元素的含义。

序号	元素名称	所属要素	含义
1	空间实体	S	对应于被描述空间方位的空间实体。
2	参照实体	S	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	与空间实体的空间方位直接关联的事件。
4	原文时间	T	文中写明的与空间方位相关联事件的时间表述。
5	参照事件	T	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判断，则可通过此元素和6号元素共同描述。此元素描述了6号元素所参照的事件。
6	参照时间	T	当文中未出现描述空间方位关系的具体时间，且该时间可以判断时，通过此元素描述空间方位关系的时间。值有”说话时”/”过去”/”将来”/”之时”/”之前”/”之后”/”之间”
7	事实性	F	如果空间方位命题是假的，则该字段为”假”。
8	处所	P	描述静态空间实体相对某外部参照物的位置。
9	起点	P	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。
10	终点	P	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。
11	方向	P	描述动态空间实体的位移方向。（空间实体在动态中才有方向特征）
12	朝向	P	描述空间实体某一侧面所朝向的位置。
13	部件处所	P	描述空间实体作为一个部件在整体中的位置。
14	部位	P	描述了空间实体的某个部位。
15	形状	P	描述了空间实体的形状。
16	路径	P	描述了空间实体位移时经过的轨迹。
17	显式距离	P	文中写明的描述了空间实体间距离关系的表述，与1号元素和2号元素共同描述。
18	隐式距离	P	文中并未写明空间实体间的距离关系但可以推断，值为”远”/”近”/”变远”/”变近”。

表 2: 子任务三的元素含义

2.2 数据集制作流程概要

数据集的构建流程包括：筛选数据、生成替换句、开展标注工作、划分数据集。下面分别介绍:

1. 筛选合适的原始句子。SpaCE2022注重语料类型的丰富性，收集了报刊、文学作品、中小学语文课文、交通事故判决书、体育动作训练手册、地理百科全书等多领域的数百万字生语料，使用pkuseg(Luo et al., 2019)对语料进行分词和词性标注，通过定义空间语义表达加权值筛选得到待进入标注流程的原始文本，排除错别字和词性标注错误等问题后，得到6,643条语料（称为“原句”），进入后续标注流程。

2. 基于词语替换得到替换句语料。制作空间语义表达词表，并从可替换角度，对词表进行分组。根据这些分组，由程序遍历每个原始句子中的空间方位意义词语，将之替换为同组的其他词。6,643个“原句”经过替换，得到了47,920个“替换句”。每一个替换句都有1或2个词语与原始句不同。替换句的空间信息既可能异常，也可能正常。

与SpaCE2021相比，SpaCE2022的替换词表更具系统性，并且在进行替换操作时加入了过滤规则，替换效果更好。SpaCE2021替换词表不记录词性，在进行替换操作时直接根据字符匹配定位可替换词；SpaCE2022替换词表区分了方位词、处所词、趋向动词、介词、副词等词类，能够在替换时区分兼类词，如“上”字一词既有可能是方位词，也有可能是趋向动词，前

者对应替换词“下、中、前、后”等，后者对应替换词“下”。考虑到任务主要关注物理空间的空间语义，SpaCE2022不对抽象名词后面的方位词进行替换。此外，SpaCE2022在替换单个词的基础上还新增了同时替换两个方位词的替换模式。数据集的替换情况显示，平均每个原句生成的替换句数量从SpaCE2021的40.52下降到SpaCE2022的8.41，词表中每个词的平均替换频次从SpaCE2021的6.03下降到SpaCE2022的4.39，说明SpaCE2022替换句的多样性得到充分提升。

3. 就正误判断任务进行人工标注，并作筛选。共招募229名标注人员参与此项工作，标注了44,921个句子（皆为替换句）。标注时需判断句子的异常程度，标注界面如图1所示。每条语料由2名标注员标注，并有质检员进行抽查，以控制标注质量。标注完成后，根据双人标注的一致程度，对语料进行分流，其中2人一致标注为“完全正常”或一致标注为“显然异常”的句子被认为是最可靠的标注语料，共计15,747条，进入最终评测数据集。

4. 就异常归因任务进行人工标注，并作筛选。共178名标注人员参与，标注了10,614个句子（皆为替换句）。此步骤以正误判断任务中被标注为异常的句子为原材料，要求标注者判断异常的类型，并选出存在异常的具体文本片段。异常类型参见上文表1 说明。标注界面如图2所示。每个句子由1-2名人员标注，并定期审核标注内容。最终根据双人标注的一致性、审核意见以及标注员标注水平分级情况，选出其中7,068条相对可靠的标注句进入评测数据集。

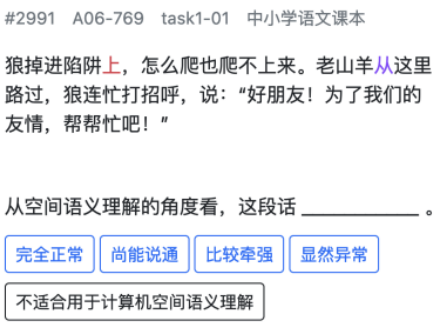

图 1: SpaCE2022正误判断标注系统工作界面

图 2: SpaCE2022异常归因标注系统工作界面

5. 使用STEP标注体系进行细粒度空间信息标注。共71名标注人员参与，标注了3,223个句子。此步骤要求标注者遵照前文介绍的STEP标注体系，将句子中涉及的空间信息详尽地标注出来。标注界面如图3所示。每个句子仅由1名人员标注。为保证质量，审核员对每名标注员的标注进行定期抽检；同时标注工具也提供了自动检查功能，将可能存在的问题实时地通过标注界面反馈给标注员，如被标注为空间方位P的片段首尾一般不为动词。标注完成后，通过程序剔除了怀疑有不符合标注规范情况的部分语料，最终选取了2,152个句子进入评测数据集。

划分训练集、验证集和测试集时，为避免机器通过比对同一原句生成的不同替换句学习到替换规律，规定某一原句和它生成的所有替换句只能出现在同一个数据集合中，即训练集、验证集或测试集三者之一。经过上述步骤，最终得到SpaCE2022数据集，共24,947条标注语料，数据集构成如表3所示。

子任务	训练集	验证集	测试集	总计
1.中文空间语义正误判断任务	10993	1602	3152	15747
2.中文空间语义异常归因与异常文本识别任务	4966	700	1402	7068
3.中文空间实体识别与空间方位关系标注任务	1529	207	396	2132

表 3: SpaCE2022数据集的构成

图 3: SpaCE2022空间语义角色标注系统工作界面

2.3 数据分布情况

SpaCE2022语料共计285万字符，每段语料字符数均值为114.23，标准差为49.57。语料涉及多种不同类型和来源，各类语料比例为：报刊（37%）、文学作品（25%）、中小学课本（20%）、交通事故判决书（9%），体育动作（6%）、地理百科（2%）、其他（1%）。下面分别考察各子任务的标签分布情况。

子任务一有两个标签："正常"和"异常"，下面用"正例"指标签为"正常"的语料，用"负例"指标签为"异常"的语料，表4是各子集的标签分布情况。正负例比重是正例数和负例数的比值，可以衡量二元标签的平衡性，越靠近1，正例和负例越平衡。从表中可看出，子任务一数据集以负例为主，说明替换空间义词语更有可能引发空间信息异常。测试集的正负例比例接近1，制作子任务一数据集时优先考虑了测试集的平衡性。

数据集	正例数	负例数	正负例之比
训练集	2677	8316	0.32
验证集	705	897	0.79
测试集	1695	1457	1.16
合计	5077	10670	0.48

表 4: 子任务一各子集的标签分布情况

子任务一中，替换词和替换对的分布呈现出了标签偏向。双音节处所词所在的语料倾向于空间信息正常，而有两个替换词（下文称为替换组）的语料则倾向于空间信息异常。近义词构成的替换对使用前后，空间义基本未发生改变，所以偏向正常，而反义词构成的替换对偏向异常，如"下来→上来"、"进去→出来"。此外，中心义词和外围义词的替换也容易导致空间信息异常，如"中→旁"、"中→外"。

子任务二共有三个标签，分别对应三种归因类型，具体参见上文表1。表5是子任务二数据集归因标签的频次。"&"代表联合归因。整个数据集中，三类标签的分布很不平衡，约58%的语料被归因为不符合常识或背景信息，其次是搭配不当和语义冲突。联合归因中，搭配不当经常和不符合常识或背景信息共现，说明这两个类型的区分并不显著。三类归因共现最少见。

标签	A	B	C	A&B	A&C	B&C	A&B&C
频次	870	821	4102	77	848	317	33

表 5: 子任务二数据集的归因标签分布情况

第二十二届中国计算语言学大会论文集，第547页-第558页，哈尔滨，中国，2023年8月3日至5日。

替换词和替换组在归因类型标签的分布上没有明显偏向。在联合归因中，替换词和替换组之间也都表现出单因大于双因、大于三因的现象，未发现数据偏差。替换对中，原词和替换词的搭配能力有明显区别时，偏向**搭配不当**，比如“当地”一般修饰名词，而“原地”一般修饰动词。偏向**语义冲突**的替换对包含了绝对方位词，替换后的方向在上下文构建的空间场景不能成立，与其他绝对方位词相冲突。偏向**不合常识**的替换对以“两侧”作原词为主，替换成其他单侧的方位词后，不能满足空间实体的要求，违反了常识。

子任务三的STEP标注体系共使用了18个元素，训练集、验证集和测试集的元素总量分别为25224, 3848和5357。每个元素占子集元素总量的占比情况见附录A。每个元素在各子集中均有分布且占比接近，但元素之间的分布严重失衡。空间实体比事件高约10个百分点，说明不是所有标注都有事件信息出现。这个特点对于使用事件抽取范式的系统而言是一大挑战，因为缺少动词性单位意味着缺少事件抽取所需的触发词。

3 机器在SpaCE2022任务上的表现

3.1 评测指标

子任务一是正误判断任务，以准确率为指标。子任务二设计了3种评价指标，分别是①异常文本归因准确性，②异常文本识别准确性，③异常元素识别准确性。指标①以准确率的形式考察参赛系统进行异常归类的能力；指标②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③则以F1值的形式考察参赛系统对于异常信息所属的具体要素进行分类的能力。鉴于指标③实际上涵盖了前两项指标所考察的方面，我们使用指标③作为评测赛事的排名依据。

子任务三的数据在评测阶段组织为元组形式，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。每个元组含有18个槽位。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，程序根据一定标准计算其中每个元素的得分，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。

3.2 机器在各项任务上的表现

3.2.1 基线系统

SpaCE课题组为评测建立了一套基线系统。子任务一使用预训练模型BERT(Devlin et al., 2019)构建了一个二元分类器。子任务二设置了一个分类层预测归因类型，以及一个序列标注层判断每个词所属的元素，两个模块采用独立编码器。子任务三首先进行序列标注任务，寻找文本中能够触发事件抽取的触发词，然后根据触发词抽取其他元素。

3.2.2 参赛系统

SpaCE2022共有32支队伍报名，最终3支队伍提交了测试结果。参赛系统普遍使用不同的模型架构来分别完成三个子任务。子任务一中，参赛系统均使用了判别式预训练模型Electra(Clark et al., 2020)来完成二分类任务，队伍1和队伍3进一步使用集成模型的方法提升准确率，队伍2则构建了方位词表，通过计算方位词的最大替换概率来判断句子是否存在空间信息异常。

子任务二中，队伍1使用阅读理解任务的范式，针对三个归因类型分别训练了三个模型，模型会预测每个异常文本片段的开头和结尾。队伍2训练了一个序列标注器，给每个词打上S、P、E或O（表示非目标词）的标签，从而找到异常文本片段。队伍3利用w2ner(Li et al., 2022)架构可以抽取不连续实体的特性，同时抽取所有归因类型的异常文本片段。

子任务三中，队伍1采用抽取和生成两阶段的方法，抽出18元组的主语（空间实体）后，使用生成模型生成其他部分。队伍2使用信息抽取预训练模型UIE抽取出每个空间元素。队伍3将18元组拆分为多个3元组，使用关系抽取模型gplinker(苏剑林, 2022)抽出3元组后，再合并为18元组。

表6是参赛系统和基线系统在三个子任务上的表现。所有参赛系统在子任务一的表现都超过了基线模型，因为基线模型不擅长同时捕捉多种错误模式，而参赛系统通过集成模型等方式能改善此问题。子任务二中，所有系统在文本准确性指标上的表现都优于归因准确率指标，一方面说明所有系统都更擅长识别异常文本片段，另一方面可能说明三种归因类型的区分度不明显，归因难度较大。所有系统的元素准确性指标得分也都显著低于文本准确性指标，说明系

统在为异常文本片段标注SPE要素时遇到了困难。子任务三中，所有参赛系统的得分均低于基线系统，且所有系统的得分均低于0.6，主要的原因可能是元组和元素的抽取数量较多，而且约30%的元组没有触发词，这对事件抽取模型而言有较大的难度。

系统	子任务一准确率	子任务二			子任务三F1值
		归因准确率	文本准确性F1值	元素准确性F1值	
队伍1	0.7865	0.0036	**0.8075**	**0.6748**	0.4950
队伍2	**0.7992**	**0.5827**	0.6432	0.4877	0.3870
队伍3	0.7985	0.2268	0.3324	0.2822	0.4387
基线系统	0.5864	0.5599	0.5812	0.4403	**0.5069**

表 6: 参赛系统和基线系统在任务上的表现

4 人类在SpaCE2022任务上的表现

课题组在数据集评测阶段进行了人类测试，包括用相同指标计算人类得分，以及进行一致性检验。人类得分反映了任务的难度，也为参赛系统提供人类基准。一致性检验关注不同人对同样的语料是否有一致的标注结果，反映人类对标注任务的理解是否趋同。一致性低，可能是评测任务的问题主观性比较强，不同人的看法不太容易趋同。如果存在上述问题，则有理由怀疑数据集的质量不可靠。下面分别介绍人类在三个子任务上的表现情况。

4.1 子任务一的人类表现

课题组招募了来自不同年级和专业的大学生共7名被试，进行子任务一的标注培训。培训合格后，他们需要独立完成100道测试题。这100道题目是根据语体占比随机抽取的，包括50条正例和50条负例。表7是所有被试的准确率，最高分为0.95，最低分为0.69，平均分为0.78，与参赛系统的得分相当。被试4和被试7属于异常值，最终取5名被试的结果进行了Kappa值的计算，衡量人类在子任务一的一致性，如表8所示。Kappa均值为0.53，在Kappa值分级标准下属于中等水平[3]。

人类	准确率
被试1	0.74
被试2	0.78
被试3	0.83
被试4	0.69
被试5	**0.95**
被试6	0.80
被试7	0.69
均值	0.78

表 7: 子任务一人类被试准确率

人类	被试1	被试2	被试3	被试5	被试6
被试1		0.53	0.57	0.49	0.42
被试2	0.53		0.51	0.56	**0.37**
被试3	0.57	0.51		**0.76**	0.48
被试5	0.49	0.56	0.76		0.59
被试6	0.42	0.37	0.48	0.59	
均值	0.43	0.47	0.52	0.54	0.41

表 8: 子任务一人类被试的Kappa值

进一步观察每一道题的答题情况，有76道题是至少4名被试做出了相同的判断，其中有38道题为5名被试的判断都相同，这说明这76道题的客观性比较强，被试有较为趋同的理解。包含近义词替换对的题目，被试倾向于做出空间信息正常的判断，因为可以构建出相似的空间场景。如“今晚在院子里坐着乘凉”，如果用“中”替换“里”，这个句子的空间信息依然正常。包含反义词替换对的题目，被试则倾向于做出空间信息异常的判断。如“口袋全挂在外面像是被抢劫了一样”，如果用“里面”替换“外面”，空间义与下文的“抢劫”相冲突。如果空间实体的维度形态在人类的认知中趋于一致，那么被试对实体可以或不可以搭配的方位词也有比较一致的认识，从而得出相同的答案。如“森林”是三维实体，当搭配强调二维平面的方位词“上”时，会出现搭配不当的情况，被试很容易发现异常。

剩余24道题的主观性较强，不同人的看法有所不同。一种情况是上下文缺少助于判断的线索，被试需要调用个人认知经验。比如“在宇航中心的食堂前吃了早饭”，有的被试认为可以构

[3]Kappa值在[0.41, 0.60]区间为中等的一致性。

建“食堂前”的场景并且在这里吃早饭，空间信息是正常的，而有的被试则认为基于常识，“食堂里”才是供应早饭并提供吃早饭的地方，所以空间信息异常。另一种情况是替换词为包含绝对方向词“东、南、西、北”等的交通文本，被试需要有较好的方位感和空间想象能力，否则容易判断错误。

4.2 子任务二的人类表现

子任务二共有9名被试参与人类一致性检验，经过培训后，他们需要独立标注50条随机抽取自子任务二测试集的语料。表9是所有被试的得分，文本准确率平均分为0.82，归因准确率均分为0.65。每名被试的文本准确性都比归因准确率高，被试2甚至高出约42个百分点，这说明被试能够较好地定位异常文本片段，但对归因类型的理解不到位，也反映出归因类型的划分主观性较强，有较大的改进空间。与参赛系统相比，人类在文本准确性上的表现显著优于队伍2和队伍3，类型准确率的表现显著优于队伍1和队伍3，但与表现最好的系统相当。

人类	文本准确性（F1）	归因准确率（Acc）
被试1	0.84	0.70
被试2	**0.78**	**0.36**
被试3	0.81	0.74
被试4	0.82	0.62
被试5	**0.87**	**0.76**
被试6	0.82	0.64
被试7	0.83	0.70
被试8	0.83	0.70
被试9	0.81	0.66
平均值	**0.82**	**0.65**

表 9: 子任务二人类被试的文本准确性和归因准确率

4.3 子任务三的人类表现

课题组对子任务三开展了数据集抽检质量评估工作。评估人员需要检查题目的标准答案，对认为有误的答案进行增加、删除或改动等操作。课题组招募了4名审核员完成这项工作。根据语体占比和标签占比，课题组从测试集中随机抽取了70条数据分发给审核员。最后，以每一位审核员的结果为标准答案，使用子任务三的评测脚本计算其他审核员的F1值。如果F1值高，说明他们有相同的增删改操作，对任务有较为一致的理解和认识。表10是审核员两两比对的F1值，总平均值达0.88，具有较高的一致性。其中，审核1和审核4的一致性最高。

人类	**审核1**	**审核2**	**审核3**	**审核4**
审核1		0.89	0.89	**0.92**
审核2	0.89		**0.86**	0.87
审核3	0.89	0.86		0.88
审核4	0.92	0.87	0.88	
均值	0.90	0.87	0.88	0.89

表 10: 子任务三人类审核的一致性得分

进一步考察每一条标注数据的一致性，约24%的数据完全一致，一致率大于80%的数据占八成。通过分析13条增删改次数差别较大的语料，发现造成标注差异的因素有：

（1）空间隐喻可能造成对空间实体的标注有主观认识上的差异。没有实体的抽象概念在空间隐喻的作用下可能会被视为空间实体。如“那三个老者的讥笑一句也没听进耳中”，使用了空间隐喻来描绘话语的传递过程，3号审核员认为“讥笑”是空间实体，标注了终点“耳中”、方向“进”和事实性“假”；

（2）空间推理可能造成对隐性空间信息的标注有主观认识上的差异。对于文本中通过简单推理可以得出的隐性空间语义，任务三要求进行标注，但标注规范本身对此缺少明确规定，

标注员往往难以把控尺度。简单的推理能够增加标注的丰富度，不合适的推理可能会让标注变得过于复杂。如“两个孩子往里面张望，一辆汽车都没有”，可以推理出“汽车不在里面”的空间义，简单且合理。但对于“一个巨大的小行星或彗星撞击地球”，通过标注距离角色来表示彗星与地球越来越近，不同的标注员处理就可能不一致，因为这个事件并不强调二者之间的距离，推理性较弱；

（3）对可能性与事实性的区分有主观认识上的差异。“可能”、“也许”等词可能影响事实性的判断，如“垃圾可能掉入河道内”，三名审核员认为“垃圾”位于“河道内”的事实为假，有一人认为是真。另外，带虚拟语气的句子和假设关系的句子也常常出现事实性标注不一致的问题；

（4）空间方位描述中存在一词充任多个角色的情况，容易漏标。如“手里的瓜子和蚕豆越掉越多”，在“手里的瓜子和蚕豆”中，“手里”是“处所”。在整个句子中，而“掉”蕴含了位移信息，“手里”又可以作为“起点”理解。4名审核员都只标注了其中的一种情况，存在信息标注不完备的问题。

通过考察人类在SpaCE2022任务上的表现，子任务一，对文本中空间信息是正常还是异常的判断，一致性较差，作为评测任务，评价难度较大；子任务二，识别异常文本片段的可操作性较强，一致性较高，但异常归因则界限模糊，一致性较低；子任务三，人类标注员在标注时也存在一定的主观差异，且达到标注信息完备的难度较高，反映出该任务的信息丰富度和复杂性，这是参赛系统普遍表现不佳的原因之一。

5 结语

SpaCE2022在SpaCE2021的基础上扩充语料规模和语料类型，提出了覆盖面更广的任务设计，包括：考察机器对空间信息异常的判断能力、异常片段的识别能力和归因能力，以及提取结构化文本空间信息的能力。为了构建高质量的数据集，SpaCE2022课题组在语料准备阶段，利用词类、约束规则等语言学知识，提高了空间信息异常句的生成多样性；在任务设计阶段，提出了S-P-E标注法和STEP标注体系，对空间信息异常和结构化空间语义信息的表征做了明确而系统的规范；在标注阶段，采用规则自动查错、双人标注和抽样审核等方式控制标注质量；在构建数据集阶段，对标注数据的价值进行合理分级，优先选用高质量数据，并充分考虑了语料在训练、验证和测试数据集上分布的合理性。

尽管如此，机器和人类的表现仍反映出数据集的质量控制问题。人类在子任务一和子任务二上的表现反映出正误判断和异常归因的任务设计存在一定的主观性，不同人对空间异常和归因类型的理解不一定相同，这可能降低评测结论的信度（Reliability）。机器和人类在子任务二的文本准确性指标上的表现都显著优于归因准确率指标，说明S-P-E标注法既有助于人类选取趋同的文本片段，也有利于机器学习描绘异常的方法。另一方面，也说明归因分类任务的设计还需要进一步优化，增加约束条件，提高标注的一致性。

数据集的质量控制还体现在标签的分布情况。标签分布失衡会降低数据集的效度（Validity），比如子任务一的替换组在标签分布上明显偏向负例，这可能使机器捕捉到替换词数量与标签分布的关系，无法有效反映机器真实的空间语义理解能力。再如，子任务三的标注规范虽然覆盖到了所有标签，但实际语料中某些标签数量过少，训练不充分，机器可能无法学习到相应的空间知识。标签分布失衡还在一定程度上反映了质量分级策略的局限性，分级规则可能让具有某一特征的可用数据被错误排除在外，如子任务三的标注规范要求P信息不以时体助词结尾，但分级规则没有区分时体助词“过”和表经过义的“过”，导致以经过义“过”为结尾的路径标签没有进入数据集。

为了更为全面和准确地探究机器空间语义理解能力，并推动语言认知类评测任务的发展，数据集的质量控制仍然是下一步待改进的重要问题。其中影响标注一致性的认知因素应研究更有效的规范和约束方法，而如果标注员经过多轮培训后，标注一致性仍不够理想，则需要考虑修改标注规范的体系设计。构建数据集时还应确保覆盖到尽可能多的评测对象，避免标签分布存在明显的偏差。

参考文献

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

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.

Oleksandr Kolomiyets, Parisa Kordjamshidi, Marie Francine Moens, and Steven Bethard. 2013. Semeval-2013 task 3: Spatial role labeling. In *Second joint conference on lexical and computational semantics (* SEM), Volume 2: Proceedings of the seventh international workshop on semantic evaluation (SemEval 2013)*, pages 255–262.

Parisa Kordjamshidi, Steven Bethard, and Marie-Francine Moens. 2012. Semeval-2012 task 3: Spatial role labeling. In {** SEM 2012*}*: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation* {*(SemEval 2012)*}, volume 2, pages 365–373. ACL.

Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10965–10973.

Ruixuan Luo, Jingjing Xu, Yi Zhang, Zhiyuan Zhang, Xuancheng Ren, and Xu Sun. 2019. Pkuseg: A toolkit for multi-domain chinese word segmentation. *CoRR*, abs/1906.11455.

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjmashidi. 2021. Spartqa:: A textual question answering benchmark for spatial reasoning. *arXiv preprint arXiv:2104.05832*.

James Pustejovsky, Parisa Kordjamshidi, Marie-Francine Moens, Aaron Levine, Seth Dworman, and Zachary Yocum. 2015. Semeval-2015 task 8: Spaceeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (semeval 2015)*, pages 884–894. ACL.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

苏剑林. 2022. Gplinker：基于globalpointer的实体关系联合抽取. https://kexue.fm/archives/8888.

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.

A 附录

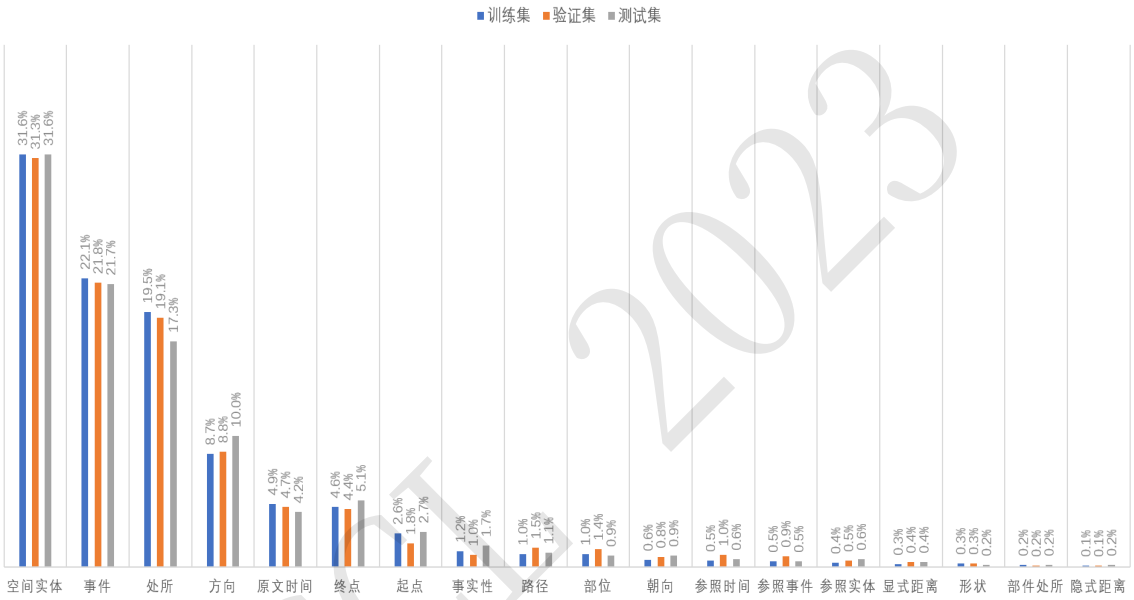

图 4: 子任务三数据集的标签分布图