CCL23-Eval任务4系统报告：基于深度学习的空间语义理解

谭臣坤
复旦大学
计算机科学技术学院
19307100058@fudan.edu.cn

胡先念
复旦大学
计算机科学技术学院
21210240194@m.fudan.edu.cn

邱锡鹏
复旦大学
计算机科学技术学院
xpqiu@fudan.edu.cn

摘要

本文介绍了参赛系统在第三届中文空间语义理解评测（SpaCE2023）采用的技术路线：面向空间语义异常识别任务提出了抽取方法，并结合生成器进一步完成了空间语义角色标注任务，空间场景异同判断任务则使用了大语言模型生成。本文进一步探索了大语言模型在评测数据集上的应用，发现指令设计是未来工作的重点和难点。参赛系统的代码和模型见https://github.com/ShacklesLay/Space2023。

关键词： 空间语义理解 ； 深度学习

System Report for CCL23-Eval Task4:Spatial Semantic Understanding Based on Deep Learning.

Chenkun Tan
Fudan University
School of Computer Science
19307100058@fudan.edu.cn

Xiannian Hu
Fudan University
School of Computer Science
21210240194@m.fudan.edu.cn

Xipeng Qiu
Fudan University
School of Computer Science
xpqiu@fudan.edu.cn

Abstract

This article introduces the technical approach adopted by the participating system in the 3rd Chinese Spatial Semantic Understanding Evaluation (SpaCE2023). For the spatial semantic anomaly recognition task, an extraction method was proposed, and combined with a generator to further complete the spatial semantic role labeling task. The spatial scene similarity judgment task used a large language model for generation. This article further explores the application of large language models on the evaluation dataset and finds that instruction design is a key and challenging area for future work. The code and models of the participating system can be found at https://github.com/ShacklesLay/Space2023.

Keywords: Spatial semantic understanding , Deep learning

1 引言

空间范畴是人类认知中重要的基础范畴。理解文本中的空间信息不仅需要掌握词汇、句法语义知识，还需要用到常识或背景知识，调动认知能力来构建空间场景。空间语义理解在自然

语言处理领域是一个热门的研究方向。近年来，随着大数据和深度学习技术的发展，越来越多的研究者开始关注如何让机器能够像人类一样理解自然语言中的空间信息。空间语义理解不仅仅是为了实现导航和文景转换等应用，更是为了探索人类语言理解的本质和人类认知的规律。因此，评测机器的空间语义理解能力并推进空间范畴的认知计算建模研究具有重要意义。

为了推进空间语义理解的研究，北京大学主办了第三届中文空间语义理解评测（SpaCE2023），并提出了空间语义异常识别、空间语义角色标注和空间场景异同判断三个子任务。这些子任务涵盖了空间语义理解的不同方面，旨在考察机器在理解自然语言中的空间信息方面的能力，为研究者提供一个公开、标准的评测平台。本文针对空间语义异常识别任务提出抽取方法，即微调模型来抽取出含有空间语义异常的文本片段；针对空间语义角色标注任务，本文提出先抽取再生成的两阶段方法，先抽取含有关键的空间语义角色的文本，再生成剩余的角色标注；针对空间场景异同判断任务，本文使用大语言模型和精心设计的指令来判断空间场景的异同并生成判断理由。同时，本文还进一步探索了大语言模型在评测数据集上的应用，发现指令设计是未来工作的重点和难点。

2 相关工作介绍

在空间语义理解的研究中，已经有许多研究者探索了如何让机器能够理解自然语言中的空间信息。这个方向的研究有助于机器更好地理解人类的语言交流，并在各种实际场景中得到应用。而在空间语义理解的评测方面，SpaCE系列评测则提供了一个重要的平台，为研究者提供了一个更为完善和具有挑战性的评测环境。因此，本文将分别介绍空间语义理解方向的进展和SpaCE系列评测的进展，以展示这个领域的最新研究动态。

2.1 空间语义理解方向

在空间语义理解的研究中，已经提出了各种方案来表示空间关系。其中，SpatialML (Mani et al., 2010) 提出了一种基于区域演算的方法，用于表征位置之间的方向和拓扑关系。而空间角色标注任务(Kordjamshidi et al., 2011) 则开发了一种语义角色标签方案，重点关注空间关系中的主要角色。此外，SemEval 2012 (Kordjamshidi et al., 2012)引入了空间语义角色标注任务，强调静态空间关系，SemEval 2013(Kolomiyets et al., 2013)将静态空间关系细颗粒化，并扩展到动态空间关系。而SemEval 2015 (Pustejovsky et al., 2015)则是第一个评估实现SpaceEval标注方案的系统的共享任务会议。SpaceEval标注方案也是当前通用的空间信息标注方案，许多的空间信息提取系统都是基于SpaceEval标注方案开发的。

空间关系提取任务可以分为传统的机器学习方法和神经网络方法。前者高度依赖于手动特征或显式句法结构。Nichols和Botros(2015)提出了SpRL-CWW模型，它使用CRF层来提取空间元素，然后引入SVM来分类空间关系。D'Souza和Ng(2015)提出了一种基于筛选的模型，通过贪心的特征选择技术生成各种手动特征。Salaberri等(2015)引入外部知识作为空间信息的补充，在此过程中，WordNet和PropBank提供了许多空间元素的信息。Kim和Lee(2016)提出了一种韩语空间关系提取模型，使用依赖关系来找到适合角色的合适元素。

随着神经网络的广泛应用，Ramrakhiyani等(2019)通过依存句法分析生成候选关系，并使用BiLSTM模型对候选关系进行分类。Shin等(2020)首先使用BERT-CRF提取空间角色，然后引入R-BERT(Wu and He, 2019)来提取空间关系。此外，一些研究关注于多模态空间关系提取。例如，Dan等人(2020)提出了一种空间BERT，它用两个实体以及包含这两个实体的图片来预测实体之间的空间关系。

2.2 SpaCE系列评测

SpaCE2021有三个子任务，分别是空间语义正误判断、空间语义异常归因合理性判断和空间语义判断与归因联合任务。它们的类型都是二元判断题，SpaCE课题组基于预训练模型BERT(Devlin et al., 2018)建立了一套基线系统，此次评测的所有参赛模型也都使用了主流大规模判别式预训练模型。

SpaCE2022有三个子任务，分别是空间语义正误判断、空间语义异常归因与异常文本识别和空间实体识别与空间方位关系标注任务。SpaCE课题组为评测建立了一套基线模型。子任务一使用预训练BERT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，寻找文本中能够出发事件抽取的关键词，然后根据触发词抽取其他元素。

3 模型与方法

3.1 子任务1

空间语义异常识别任务要求从给定中文文本中识别出具有空间语义异常的文本片段。每个文本片段包含3 个字段，分别是角色（role）、文本内容（text）和字序数组（idxes）。role的取值包括S1 P1 E1 S2 P2 E2，表示两个完整的“空间实体(S)-空间方位(P)-事件(E)”三元组。为了描述空间语义异常，最多可以选取6 个文本片段，最少可以选取1 个。

输入包括两个部分：数据编号“qid”和存在空间语义异常的文本“context”；输出也包括两个部分：数据编号“qid”和描述空间语义异常的文本片段“results”。输入输出样例如下所示：

子任务1数据示例

输入：
{ ”qid”:”1-train-626”,
”context”:”鲸每天都要睡觉。睡觉的时候，总是几头聚在一起，找一个比较安全的地方，头朝边，尾巴向外，围成一圈，静静地浮在海面中。如果听到什么声响，它们立即四散游开。” }

输出：
{ ”qid”:”1-train-626”, ”results”:[
[{ ”role”: ”P1”, ”text”: ”朝边”, ”idxes”: [35,36] }],
[{ ”role”: ”S1”, ”text”: ”鲸”, ”idxes”: [0] },
{ ”role”: ”P1”, ”text”: ”在海面中”, ”idxes”: [52,53,54,55] },
{ ”role”: ”E1”, ”text”: ”浮”, ”idxes”: [51] }]] }

该任务要求给定文本，输出若干异常语义片段并将输出结果以三元组的形式表示出来，相当于是给定文本和任务要求，然后抽取出特定的文本片段。由于抽取出的文本片段最多只有六个，对应六种不同的角色取值，因此我们将任务简化为对六个文本片段的抽取。具体来说，由于最多需要抽取六个文本片段（对应六个角色），因此我们构建一个长为12的数组，用来存储6个角色对应的文本片段在文本中出现的开头位置和结尾位置，作为该任务的预测目标。为了应对抽取数量不足六个的情况，我们在所有文本的开头添加一个标记，使得文本长度加1，当某个角色对应的文本片段不存在时，就将预测数组中表示它的开头位置和结尾位置的元素置为0。模型结构如图1所示。我们采用deberta-chinese-large(He et al., 2020)中文预训练模型对该抽取任务做微调。

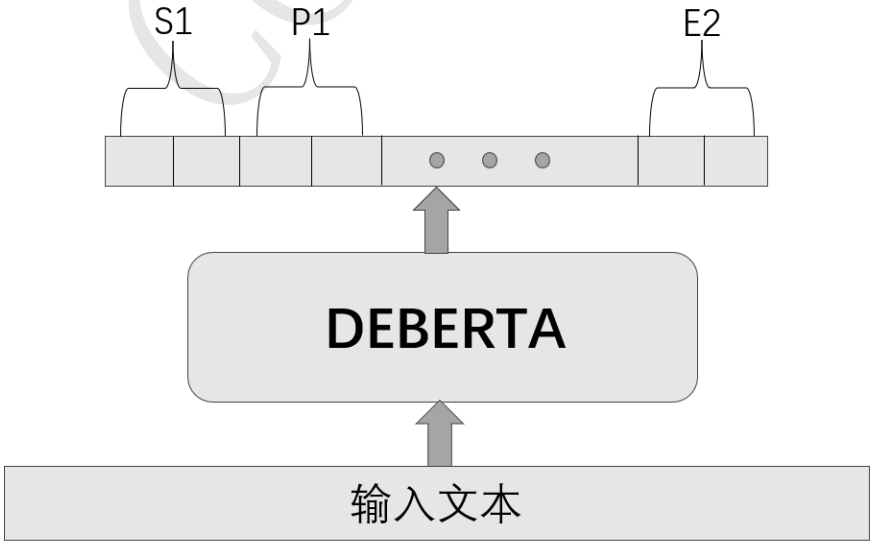

Figure 1: 子任务1模型结构示意图

模型输出的预测数组中可能会出现预测开头位置要大于结尾位置的情形，此时我们会交换开头和结尾的位置让它们符合规律。在得到存储有文本片段的开头位置和结尾位置的预测数组后，即可抽取出对应的异常文本片段的角色，文本和字序数组。

3.2 子任务2

空间语义角色标注任务要求对给定中文文本进行空间实体的识别与空间方位关系的关注。空间实体及其空间方位信息，描述了“某空间实体在某时，经由某事件，满足某种空间方位关系，这一命题的事实性为真/假”的信息。共有15 个语义角色（role）可供标注，15个语义角色分属于文本片段型角色(fragment)或标签型角色。文本片段型角色需要记录文本内容（text）和字序数组（idxes），标签型角色需要记录标签的值（label）。没有出现的角色不需要标注。

输入包括两个部分：数据编号“qid”和待标注的文本“context”；输出也包括两个部分：数据编号“qid”和标注的空间实体以及空间方位信息“results”。输入输出样例如下所示：

子任务2数据示例

输入：
{ ”qid”:”2-train-1192”,
”context”:”宋钢走的时候把五颗大白兔奶糖压在门前的石板下面，他说放在窗台上会被人拿走的。他走了几步又回来了，他说放在石板下面怕被蚯蚓吃了，他又去摘了两张梧桐树叶，把奶糖仔细包好了，重新放到石板下面。然后他的眼睛贴着门缝看看李光头，对李光头说：” }

输出：
{ ”qid”:”2-train-1192”,
”results”:[
[”role”:”空间实体”,”fragment”:”text”:”大白兔奶糖”,”idxes”:[9,10,11,12,13],
”role”:”事件”,”fragment”:”text”:”放”,”idxes”:[50],
”role”:”处所”,”fragment”:”text”:”在石板下面”,”idxes”:[51,52,53,54,55]]] }

该任务要求给定文本，输出对语义角色的标注三元组，包含role，text，idxes三个部分。我们提出一阶段的方法和两阶段的方法。一阶段的方法是将该任务转换为生成任务。具体来说，我们将标注中所有三元组的text和role提取出来组成一个长文本作为新的生成目标，微调生成模型使它能够基于输入文本生成我们构建的作为新标注的长文本。我们实验了两个生成模型bart-large(Lewis et al., 2019)和CPT(Shao et al., 2021)，结果如表1所示。可以看出，CPT的生成效果要比BART的生成效果稍好，但是结果仍不是很理想。我们推测有两个原因，一方面是因为模型不太容易建模比较长的文本，另一方面是因为该任务的评测要求对“空间实体”和“参照实体”的生成不能全部错误，而生成出来的句子很容易出现这种问题。

两阶段的方法是使用抽取加生成的方法。我们训练两个模型，分别是用来抽取role为“空间实体”的三元组的抽取器，和用来生成剩余的三元组的生成器。具体来说，我们基于原始数据集的标注三元组，构建用于训练抽取器的抽取数据集和用于训练生成器的生成数据集。抽取数据集的输入数据是原始文本，标注数据是长度与输入文本相同的数组，如果标注三元组的role为“空间实体”，就根据该三元组的idxes将数组上的对应位置标为1，其余位置标0。构建抽取数据集之后，再用它微调标注器。生成数据集的输入数据是将原始文本和role为“空间实体”的三元组的text拼接得到的文本，标注数据是除role为“空间实体”的三元组以外的所有三元组的role和text组成的长文本。构建生成数据集之后，再用它微调生成器。我们使用Deberta模型作为标注器，由于之前的实验结果表明CPT的生成效果优于BART，因此我们使用CPT模型作为生成器。

推理过程如图2 所示。我们先将原始文本输入抽取器得到含有“空间实体”位置的输出数组，根据它抽取出role为“空间实体”的文本片段，然后将抽取出的文本片段与原始文本拼接起来再输入生成器，生成出除role为“空间实体”的三元组以外的剩余三元组的role和text，最后将它与role为“空间实体”的三元组一起格式化为与标注数据一致的格式。

表1展示了该任务采用的不同方法在验证集上的F1分数。可以看出，抽取加生成的方法极

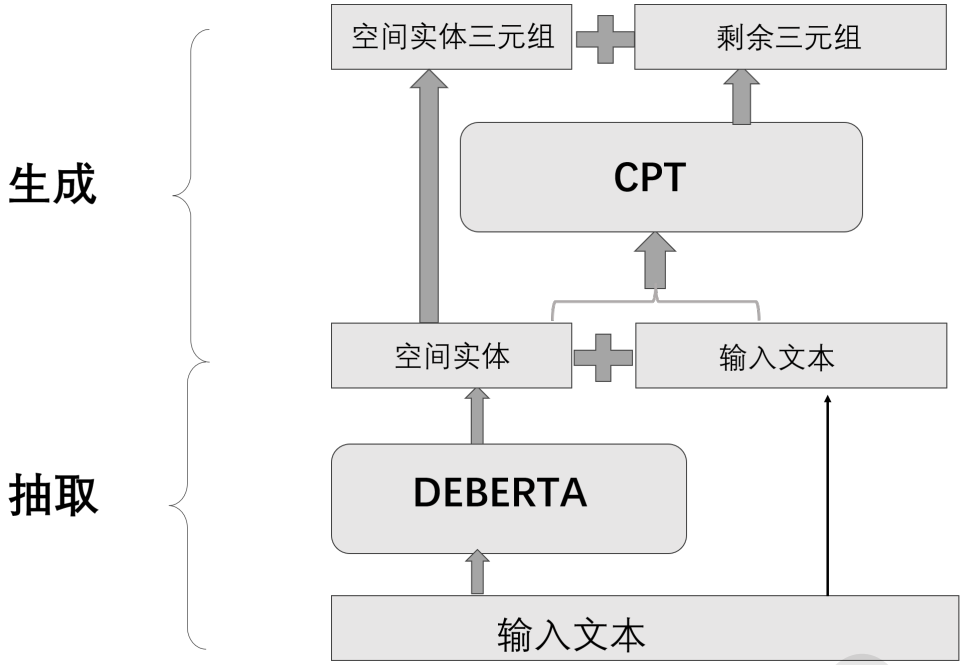

Figure 2: 子任务2模型结构示意图

大地提升了模型性能。

模型	F1
BART	35.33
CPT	37.55
Deberta+CPT	52.67

Table 1: 不同模型架构在验证集上的F1分数

3.3 子任务3

空间场景异同判断任务要求判断两个相似的中文文本是否描述相同的空间场景，并说明判断的理由。相似文本context1 和context2 存在差异文本C1 和C2，它们在形式上存在差异。C1 和C2 都是连续字符串，是合法的语言单位（词、词组、子句等），意义清晰且相对完整。context1 去除C1 后剩下的部分，和context2 去除C2 后剩下的部分，在形式上完全相同。

输入包括三个部分：数据编号“qid”、context1和context2；输出包含两个部分：数据编号“qid”和都两个文本空间场景异同的判断“judge”和理由“reason”。该任务提供两个答题模板，分别对应空间场景相同和不同的情况，以帮助及其生成更自然的文本。使用模板的数据样例如下所示，引号“”表示插入到模板插槽中的内容。

第二十二届中国计算语言学大会论文集，第139页-第149页，哈尔滨，中国，2023年8月3日至5日。

两段文本表示相同的空间场景

输入：
{”qid”：”3-1”
”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。”}

输出：
{”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。”}] }

两段文本表示不同的空间场景

输入：
{”qid”：”3-2”
”context1”：”兰兰惊奇地站在潜水桥上，透过玻璃看见大大小小的鱼游来游去，各种各样的船只从桥顶上驶过来划过去。”
”context2”：”兰兰惊奇地站在潜水桥下，透过玻璃看见大大小小的鱼游来游去，各种各样的船只从桥顶上驶过来划过去。”}

输出：
{”results”:[
{”judge”：”false”,
”reason”:”两段文本的形式差异在于“潜水桥上”和“潜水桥下”。两段文本中都出现了以下空间实体：“兰兰”和“潜水桥”。两段文本在描述“兰兰”站立的位置上存在形式差异，表明“兰兰站在潜水桥上”和“兰兰站在潜水桥下”描述“兰兰”的处所是不同的，前者位于桥的上方，后者位于桥的下方。因此，这两段文本不能描述相同的空间场景。”}] }

该任务要求给定两段文本，然后生成对两段文本空间场景异同的判断和理由。我们使用大语言模型进行生成。大语言模型生成的关键之处在于指令设计。对于该任务，我们采用了少样本加思维链(Wei et al., 2022)的方式设计指令。如下所示，思维链方法指的是把人类思考问题的过程，即所谓的思维链，用自然语言的形式显性地放在指令中。

样例——标准指令和思维链指令

标准指令:
问题：小明有5个网球，他又买了2个罐网球，每罐有3个网球。他有多少个网球?
答案：11个
问题：

思维链指令
问题：小明有5个网球，他又买了2个罐网球，每罐有3个网球。他有多少个网球?
答案：小明一开始有5个球，每罐有3个网球那么2罐有6个网球。5+6=11。11个
问题：

思维链适用于涉及推理的生成问题，相比于直接生成，思维链能通过一步步地引导来指示

模型推理出更好的结果。少样本则能够通过提供更多的样例，帮助模型理解输出格式以及思维链推理的过程。

我们在设计具体的指令时，先通过多轮对话的方式引导模型输出合适的结果，然后将多轮对话使用的指令改写为思维链形式的指令，再进行添加少样本，修饰指令等改动。我们使用的指令如下所示，省略号“.....”表示我们为了方便展示而省略的部分内容。

指令

context1 和context2 存在差异文本C1 和C2，它们在形式上存在差异。C1 和C2 都是连续字符串，是合法的语言单位（词、词组、子句等），意义清晰且相对完整，context1 去除C1 后剩下的部分，和context2 去除C2 后剩下的部分，在形式上完全相同。

指出它们的差异文本C1和C2;

给出包含C1的完整短语P1，包含C2的完整短语P2。完整短语指被符号分割开的部分句子，如果C1和C2已经是被符号分割开的句子，则P1和P2就是C1和C2;

指出P1和P2中包含的空间实体;

根据P1和P2判断它们所描述的空间实体所处的空间场景是否一致;

选择一个适合的模板进行输出;
模板一：两段文本表示相同的空间场景: .....
模板二：两段文本表示不同的空间场景: .....

Input:
”context1”: ”兰兰惊奇地站在潜水桥上，透过玻璃看见大大小小的鱼游来游去，各种各样的船只从桥顶上驶过来划过去。”
”context2”: ”兰兰惊奇地站在潜水桥下，透过玻璃看见大大小小的鱼游来游去，各种各样的船只从桥顶上驶过来划过去。”

Thought:
差异在于文本中描述兰兰所站的位置不同，一个是“潜水桥上”，另一个是“潜水桥下”。
根据提供的上下文，包含C1的完整短语P1可以是：”兰兰惊奇地站在潜水桥上”；而包含C2的完整短语P2则可以是：”兰兰惊奇地站在潜水桥下”。
P1和P2中包含的空间实体如下：”兰兰”、”潜水桥”。
根据P1和P2的描述，它们所描述的空间实体”兰兰” 分别位于不同的位置，P1描述的是兰兰站在潜水桥的上方，而P2描述的是兰兰站在潜水桥的下方。因此，它们所处的空间场景不一致。

Output:.....

（第二个样例）.....

Input:

我们将该问题的解决过程划分为以下五步，并将它们显示地放在指令中，指示模型使用思维链的方式生成:

1. 找出context1和context2的差异文本C1和C2;

2. 找出C1和C2所在的完整短语P1和P2。有时仅凭差异文本不足以判断是否空间场景异同，因此我们扩展了判断空间;

3. 找出P1和P2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。

在指示语之后，我们加入了两个样例，分别表示相同的空间场景和不同的空间场景两种情况（报告只展示了第二种），并在样例中手动添加了模型的思维链过程，即Thought部分。

4 评价指标与结果

4.1 子任务1的评价标准

子任务1 采用角色识别准确性作为评价指标，计算过程如下：

①对于每个待检查的results（称为“待检项”），与参考答案中的results（称为“参考项”）进行逐个比较。对于待检项中的每个role 中的每个字符（idxes 字段），仅当参考项的相同role中找到了该字符，才视为该字符是正确检出的字符；

②按上述标准计算待检项与参考项文本的F1 值，公式如下：

$F1 = (2 \times P \times R) / (P + R)$

$P =$ 正确检出字符数/ 待检项总字符数, $R =$ 正确检出字符数/ 参考项总字符数

③找到得分最高的待检项，以此项得分计为该题在角色识别准确性上的得分。

子任务1 同时提供文本识别准确性作为参考指标。该指标不限制字符所属的role，只要在参考项的任意role 中找到了该字符，则视为是正确检出的字符。

4.2 子任务2的评价标准

子任务2按照以下步骤计算得分，作为评价指标：

①分别从参考答案和提交答案中获取results 元组数组，称为参考数组和待检数组，其中的每个元组分别称为参考元组和待检元组，元组中的role 分别称为参考角色和待检角色；

②对待检元组和参考元组进行匹配。对于每个元组对，按照以下程序计算待检元组的得分：当role 是“空间实体”或“参照实体”时，计算idxes 字段的F1 值；对于其他文本片段类的角色，计算text 字段的F1 值；对于标签类的角色，相同计1 分，不同计0 分。时间角色（5 号角色）可能既有文本片段，也有标签，得分取二者的均值。当参考角色和待检角色都不出现，此角色不计分。当参考角色和待检角色有一个不出现，此角色也不计分。如果空间实体类角色（1-2 号角色）完全不匹配，则整个待检元组得0 分，否则对所有待检角色的得分求和，作为此待检元组的得分；

③对于所有待检数组和参考数组，遍历所有可能的匹配方式，使得待检数组中的所有待检元组能够得到最高总分。以此得分作为此题最终得分；

④对于数据集中所有题目计算F1 值，作为子任务2 的最终得分。

4.3 子任务3的评价标准

子任务3 首先根据judge 字段的结果评价异同判断是否正确，如果错误，该题得0 分；如果正确，采用人工评价对reason 字段的生成文本进行解释准确性的打分。分数越高，表示判断空间场景异同的理由解释得越清楚。共有两名评分员对100道题的结果进行人工评分。

4.4 结果

表2 展示了我们在各个子任务上取得的分数。

Table 2: 在测试集上得到的各个评价指标

task1 排名指标-角色准确性（F1）	task1 参考指标-文本准确性（F1）	task2（F1）	task3（百分制）
0.5782	0.6526	0.4739	37.40

第二十二届中国计算语言学大会论文集，第139页-第149页，哈尔滨，中国，2023年8月3日至5日。

5 探索与改进

5.1 不足与改进

子任务1采用抽取区间的方式来识别异常文本并没有与标注的方式进行比对。由于该任务的数据集存在文本片段可能不连续的情况，而抽取区间的方式只能关注连续区间，因此抽取区间的方法面对不连续的区间表现会比较差。例如当待处理文本如下时：

> **伤心崖是戛洛山下的一座山峰**，像被一把利斧从中间剖开，从山底下的流沙河抬头往上看，宛如一线天。隔河对峙的两座山峰相距约六米左右，两座山都是笔直的绝壁。

正确抽取出的三元组应该为：

```
{"role": "S1", "text": "伤心崖", "idxes": [0, 1, 2]},
{"role": "P1", "text": "戛洛山下", "idxes": [4, 5, 6, 7]},
{"role": "E1", "text": "是一座山峰", "idxes": [3, 9, 10, 11, 12]}
```

该三元组是从第一句话抽取出来的，且存在文本片段不连续的情况。在该文本上使用抽取区间的方式得到的结果如下：

```
{"role": "S1", "text": "伤心崖是戛洛山下的一座山", "idxes": [0, 1, 2,3,4,5,6,7,8,9,10,11,12]},
{"role": "P1", "text": "下的一座山峰，像被一把利斧从中间剖开", "idxes": [7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26]},
{"role": "E1", "text": "剖开，从山底下的流沙河抬头往上看，宛如一线天。", "idxes": [23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45]}
```

由于抽取区间的方式只能关注连续区间，因此它不能将第一句话拆分成三个文本片段，而只能将它作为单个文本片段抽取出来，导致后续错误地抽取出了其它文本片段。未来可以考虑探索标注的方式，以及改善抽取区间的方式在不连续区间表现较差的问题。

子任务2提出全部生成的方法和抽取加生成的方法，但是并没有考虑全部抽取的方法，不足以证明抽取加生成的方法是最优的组合。同时，该任务还给出了语义角色之间的约束关系，如下所示，但我们并没有对约束进行建模。

子任务2的约束关系

1号角色“空间实体”是核心，必定存在。
2号角色“参照实体”只与15号角色“距离”同时出现。
6-14号角色不会与15号角色同时出现。

未来可以考虑补充全部抽取的方法，以及对模型输出的结果进行约束。

5.2 对大语言模型的探索

在前两个任务上，我们尝试使用大语言模型（主要是MOSS和ChatGPT）进行生成，实验过程中存在两个主要的问题。

首先是模型的生成效果很差。这主要体现在子任务1上，我们的模型能够生成出符合要求的输出格式，但是抽取出的文本片段基本没有体现出空间语义异常。生成效果主要跟两个因素有关：大语言模型本身和指令设计。无论是GPT系列模型，还是MOSS，在子任务3上都取得了不错的表现，考虑到本次测评的三个子任务有所关联，因此可以推测大语言模型有能力解决前两个子任务。由此我们推测问题主要出现在指令的设计上。在指令设计上，我们尝试了少样本和思维链独立使用以及组合使用等多种方法，发现使用少样本和简单的格式说明就能让大语言模型生成出明确的输出格式。但是使模型识别出空间语义异常十分困难。我们在使用指令进行实验时，多次出现大语言模型找不到空间语义异常的情况，此时它们往往输出与下列文本类似的内容：

第二十二届中国计算语言学大会论文集，第139页-第149页，哈尔滨，中国，2023年8月3日至5日。

在给定的文本中，没有空间语义异常片段。

而有时则是大语言模型没有理解什么是空间语义异常，而胡乱生成的情况。例如当输入以下文本时：

她又在衣袋里摸了半天，**摸进火柴**来。她把那大蜡烛插到坟堆的顶上，点了起来。这晚上没有风，蜡烛的火焰向上直升，一点也不摇晃。老妇人对着这烛光，坐在坟边，一动也不动，两臂交叉抱在胸前，披着那黑色的大围巾。

模型可能找出的存在空间语义异常的文本片段为：

她把那大蜡烛插到坟堆的顶上，点了起来。

其次是模型的生成速度的问题。前两个任务的测试集包含成百上千条文本，因此必须考虑大语言模型的生成速度。大语言模型的生成速度可以视作（每次生成的生成数量/推理延迟）。每次生成的生成数量可以通过在指令中加入多个输入文本来提升，然而，大语言模型的输入长度限制了指令能够插入的文本数量，并且插入多个输入文本时模型不一定还能具有插入单个文本时的推理能力，因此每次生成的文本数量仍然有限。而在推理延迟上，MOSS目前的开源版本进行单次推理所需要的时间受部署设备的影响，在1-10分钟不等，其量化版本的部署和推理时间更短，但是性能也会有所降低。而GPT3.5和GPT4的推理延迟更短，但需要考虑API的调用成本。

除了之前使用的直接推理生成的方式，我们还尝试了微调MOSS再推理生成的方法。由于微调MOSS需要的算力要求很高，我们主要尝试LoRA微调(Tang et al.,)和QLoRA微调(Dettmers et al.,)的方法。然而，微调MOSS时面临了与之前相似的难题，即如何设计微调使用的对话。微调使用的对话类似于指令，一个区别是使用少样本方法容易使得模型生成偏向少样本指令的内容。由于指令和微调使用的对话设计不佳，微调后的模型效果也不好。

需要注意的是，在任务三上我们的指令设计很顺利，一方面是因为任务目标非常清晰且提供了可用的输出模板，另一方面是输入是两段存在差异的文本，且场景异同的判断主要基于两段文本之间的差异文本，这使得模型能够明确差异所处的空间，并进行进一步地推理。

6 总结与展望

在本次SpaCE2023第三届中文空间语义理解评测中，我们使用抽取的方法和抽取加生成的方法解决前两个子任务，使用大语言模型生成的方法解决第三个子任务。前两个子任务的实验设计有着改良的空间，并且仍然可以探索大语言模型在这两个任务上的应用。指令设计是大语言模型应用的核心难点，如何设计出适配前两个子任务的指令以及适用于大多数任务的通用指令是我们未来主要关心并探索的研究方向。

参考文献

Soham Dan, Hangfeng He, and Dan Roth. 2020. Understanding spatial relations through multiple modalities. *arXiv: Computation and Language*, Jul.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jennifer D'Souza and Vincent Ng. 2015. Sieve-based spatial relation extraction with expanding parse trees. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 758–768.

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

Bogyum Kim and JaeSung Lee. 2016. Extracting spatial entities and relations in korean text. *International Conference on Computational Linguistics*, Dec.

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, Martijn Van Otterlo, and Marie-Francine Moens. 2011. Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Transactions on Speech and Language Processing (TSLP)*, 8(3):1–36.

Parisa Kordjamshidi, Steven Bethard, and Marie-Francine Moens. 2012. Semeval-2012 task 3: Spatial role labeling. *Joint Conference on Lexical and Computational Semantics*, Jun.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Inderjeet Mani, Christy Doran, Dave Harris, Janet Hitzeman, Rob Quimby, Justin Richer, Ben Wellner, Scott Mardis, and Seamus Clancy. 2010. Spatialml: annotation scheme, resources, and evaluation. *Language Resources and Evaluation*, page 263–280, Sep.

Eric Nichols and Fadi Botros. 2015. Sprl-cww: Spatial relation classification with independent multi-class models. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Jan.

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)*, Jan.

Nitin Ramrakhiyani, Girish Palshikar, and Vasudeva Varma, 2019. *A Simple Neural Approach to Spatial Role Labelling*, page 102–108. Jan.

Haritz Salaberri, Olatz Arregi, and Beñat Zapirain. 2015. Ixagroupehuspaceeval: (x-space) a wordnet-based approach towards the automatic recognition of spatial information following the iso-space annotation scheme. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Jan.

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*.

Hyeong Jin Shin, Jeong Yeon Park, Dae Bum Yuk, and Jae Sung Lee. 2020. Bert-based spatial information extraction. In *Proceedings of the Third International Workshop on Spatial Language Understanding*, pages 10–17.

Zhiwei Tang, Tsung-Hui Chang, Xiaojing Ye, and Hongyuan Zha. Low-rank matrix recovery with unknown correspondence.

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.

Shanchan Wu and Yifan He. 2019. Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, Nov.