CCL24-Eval任务3系统报告：基于上下文学习的空间语义理解

武洪艳
广东外语外贸大学
信息科学与技术学院
2754976781@qq.com

林楠铠✉
广东工业大学
计算机学院
neakail@outlook.com

曾培健
广东工业大学
计算机学院
lil_ken@163.com

郑伟雄
广东工业大学
计算机学院
1473391854@qq.com

蒋盛益
广州商学院
信息技术与工程学院
jiangshengyi@163.com

阳爱民
广东工业大学
计算机学院
amyang18@163.com

摘要

空间语义理解任务致力于使语言模型能够准确解析和理解文本中描述的物体间的空间方位关系，这一能力对于深入理解自然语言并支持复杂的空间推理至关重要。本文聚焦于探索大模型的上下文学习策略在空间语义理解任务上的有效性，提出了一种基于选项相似度与空间语义理解能力相似度的样本选择策略。本文将上下文学习与高效微调融合对开源模型进行微调，以提高大模型的空间语义理解能力。此外，本文尝试结合开源模型和闭源模型的能力处理不同类型的样本。实验结果显示，本文所采用的策略有效地提高了大模型在空间语义理解任务上的性能。

关键词： 上下文学习 ； 高效微调 ； 样本选择

System Report for CCL24-Eval Task 3: Spatial Cognition Evaluation Based on In-context Learning

Hongyan Wu
Guangdong University of Foreign Studies
School of Information Science and Technology
2754976781@qq.com

Nankai Lin✉
Guangdong University of Technology
School of Computer Science and Technology
neakail@outlook.com

Peijian Zeng
Guangdong University of Technology
School of Computer Science and Technology
lil_ken@163.com

Weixiong Zheng
Guangdong University of Technology
School of Computer Science and Technology
1473391854@qq.com

Shegyi Jiang
Guangzhou College of Commerce
School of Information Technology and Engineering
jiangshengyi@163.com

Aimin Yang
Guangdong University of Technology
School of Computer Science and Technology
amyang18@163.com

Abstract

The Spatial Cognition Evaluation task aims to enable language models to parse and understand spatial relationships between objects described in the text, which is crucial for a deeper understanding of natural language and complex spatial reasoning. In this paper, we investigate the effectiveness of in-context learning for large language

第二十三届中国计算语言学大会论文集，第113页-第121页，太原，中国，2024年7月25日至28日。
卷3：评测报告

models on the spatial cognition evaluation task and propose a demonstration example selection strategy based on option similarity and spatial capability similarity. In this paper, we infuse in-context learning with parameter-efficient fine-tuning to fine-tune the open-source model, improving the spatial cognition capability of the large language model. Moreover, this paper attempts to combine the capabilities of open-source and closed-source models to tackle different types of samples. Experimental results reveal that our proposed strategy effectively improves the performance of the large language model on spatial cognition evaluation tasks.

Keywords: In-context Learning , Parameter-efficient Fine-tuning , Demonstration Example Selection

1 引言

空间范畴是人类认知的一个核心基础,大量空间信息存在于自然语言文本中。在通往人工智能的道路上，空间语义理解是不可绕开的一环。要准确理解文本中表达的空间语义，不仅需要语言知识，还需要使用空间认知能力，构建空间场景，并基于世界知识进行空间方位信息相关的推理。空间语义理解已成为自然语言处理领域中一个热门的研究主题，这在需要理解空间关系的导航系统、问答系统中具有关键的作用。近年来，随着大数据和深度学习技术的不断进步，越来越多的研究者开始研究如何使机器像人类一样能够准确理解自然语言中的空间信息。空间信息提取是理解文本中空间信息的关键，目前的研究主要致力于提取空间元素和空间关系去提升机器的空间语义理解能力。许多自然语言处理技术和机器学习方法已被应用于空间信息提取。例如，条件随机场(CRF)模型(Lafferty et al., 2001)被用于空间元素提取，支持向量机(SVM) (Suykens and Vandewalle, 1999; Roberts and Harabagiu, 2012a)和卷积神经网络(CNN) (Mazalov et al., 2015)模型被用于空间关系提取。各种语言资源，如GloVe (Pennington et al., 2014)、WordNet (Salaberri et al., 2015)和PropBank (Salaberri et al., 2015)也被用于空间信息提取。以ELMO(Peters et al., 2018)和GPT (Yang et al., 2019)为代表的语言模型在建模文本语义关系方面展现出优异的性能，被开发用于命名实体识别和语义角色标注，这使得预训练模型可以很容易应用到空间信息的提取。Shin等人(2020)提出了一个基于BERT的空间信息提取模型用于空间元素提取和空间关系提取。

为了推进空间语义理解任务的发展，第四届中文空间语义理解评测关注于大语言模型的空间语义理解能力测试，SpaCE2024从空间信息实体识别、空间信息角色识别、空间信息异常识别、空间方位信息推理和空间异形同义识别五个层次测试机器中文空间语义理解的能力。这些测试语料来源于不同的领域，涵盖了空间语义理解的不同方面，旨在综合考察机器在理解自然语言中的空间信息方面的能力。本文基于中文空间语义理解任务探索了大模型的上下文学习在该任务上的有效性，我们基于选项相似度与空间语义理解能力相似度为测试样本选择示例。本文将上下文学习融入大模型的高效微调阶段，以提高大模型的空间语义理解能力。此外，本文尝试将开源模型和闭源模型融合用于处理不同难度的样本。实验结果显示，本文所采用的策略有效地提高了大模型在空间语义理解任务上的性能。

2 相关研究

目前，空间语义理解任务要求模型关注于不同层面的空间语义信息。SemEval 2012 (Kordjamshidi et al., 2012)，SemEval 2013和SemEval 2015提出了面向空间语义理解的多个评测，分别关注模型的静态空间和动态空间的语义角色标注能力。SemEval 2012引入了一个主要关注于静态空间关系的空间角色标注任务，SemEval 2013将空间关系扩展到动态，以捕获细粒度的语义。SpaCE中文空间语义理解评测则对模型提出了更高的空间语义理解要求。SpacE 2021提出了三个子任务，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。SpacE 2022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。SpacE 2023则在空间语义异常和空间语义角色标注的基础上增加了对空间场景的关注，考察机器对空间场景异同的判断。

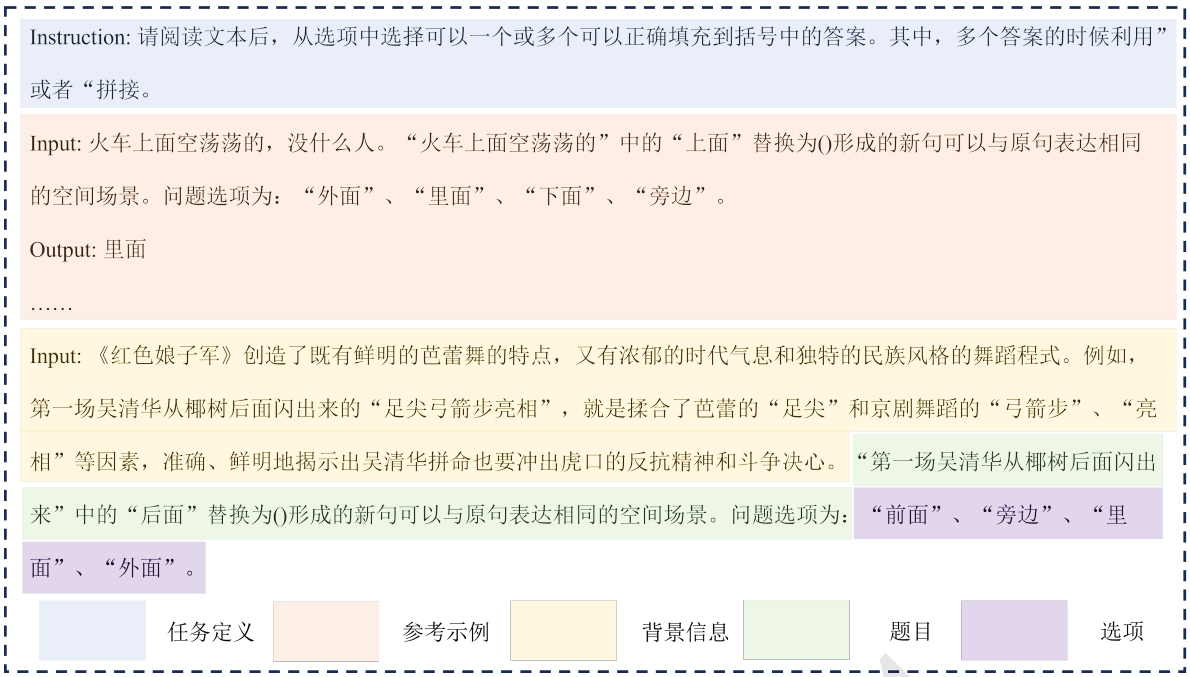

Figure 1: 大模型输入模板

早期的空间语义信息提取主要采用基于机器学习的方法。Nichols和Botros(2015)基于CRF和SVM提出了SpRL-CWW模型。该方法基于CRF使用多种输入特征来提取空间元素，例如使用GloVe的词嵌入、命名实体、词性和依存解析标签。SVM则用于从所有可能的三元组组合中过滤出正确的三元组得到空间关系。D'Souza和Ng(2012b)基于SVM提出了UTD-SpRL模型，该模型采用贪婪特征选择技术生成多个不同的特征并对空间关系相关的参数执行联合检测。一些研究者尝试将多种语言资源用于空间语义理解任务，以补充空间信息。Salaberri等人(2015)提出的X-Space模型使用WordNet中包含的地点、位置、方位等节点信息进行空间元素提取，同时使用PropBank中的参数信息对空间关系进行了分类。

随着深度学习的发展，基于神经网络的方法被开发用于空间信息的提取。Mazalov等人(2015)提出了一个基于卷积神经网络的语义角色标注系统来提取空间角色及其关系，成功地适应于空间信息提取。Dan等人(2020)使用BERT对给定图像的两个实体之间的空间关系进行预测。该模型包含一个由前馈网络实现的空间模型和一个BERT语言模型组成，其中语言模型被作为补充特征来预测图像中不可见的关系。尽管该方法使用了BERT作为语言模型，但仅限于对图像中给定的实体检测关系。Shin等人(2020)提出了BERT空间模型，使用BERT从原始文本中提取空间元素，确定它们对应的空间角色，进一步使用R-BERT对空间角色的关系进行提取，通过两个模块联合有效提高了空间信息提取的性能。本文则基于上下文学习和高效微调探索了大模型在空间语义理解任务上的综合能力，为未来的工作提供了新的研究思路。

3 基于上下文学习的空间语义理解

上下文学习使模型能够在不更新任何参数的情况下从特定任务的示例中学习，它将一些训练样本作为提示添加到推理阶段的测试样本之前。上下文学习的关键在于示例样本的选择，本文基于空间语义理解任务精心设计了样本选择策略，提出根据选项相似度和空间语义理解能力相似度为测试样本选择合适的示例样本。然后，本文将选择得到的示例样本作为前缀添加到测试样本之前构建指令提示，作为大模型的输入。对于开源大模型，本文在高效微调阶段和推理阶段均融合了示例样本作为模型的输入；而对于闭源大模型，本文仅在推理阶段使用示例样本。

3.1 示例样本选择

给定一个包含n个样本的数据集$D = \{(x_1, q_1, t_1, o_1, y_1), ...(x_n, q_n, t_n, o_n, y_n)\}$，对于第$i(i \in$

$[1,n]$)个样本，x_i，q_i 和t_i分别代表题目的背景信息、问题以及考察的空间语义理解能力类型，o_i和y_i表示该样本对应的选项和答案，其中，选项o_i包含四个不同的选项文本，即$o_i = \{o_i^1, o_i^2, o_i^3, o_i^4\}$。对于给定的样本$i$以及在数据集$D$中的另一个样本$j$，本文计算两个样本之间的选项相似度$s_o^{ij}$：

$$s_o^{ij} = \frac{|o_i \cap o_j|}{|o_i|}. \tag{1}$$

此外，对于样本i和样本j，如果两个样本考察的空间语义理解能力类型一样，两个样本之间的空间理解能力相似度则为1，否则，其空间能力相似度为0。具体地，两个样本之间的空间语义理解能力相似度s_c^{ij}定义如下：

$$s_c^{ij} == \begin{cases} 1 & t_i = t_j, \\ 0 & t_i \neq t_j. \end{cases} \tag{2}$$

基于上述选项相似度和空间语义理解能力相似度两种不同的相似度，本文进一步将两种相似度相加作为两个样本总的相似度s^{ij}：

$$s^{ij} = s_o^{ij} + s_c^{ij}. \tag{3}$$

对于样本i，本文计算该样本与数据集D中其他的所有样本的相似度，然后选择相似度最大的Q个样本作为示例样本用于上下文学习。

3.2 提示设计

中文空间语义理解任务要求大模型根据输入的信息、题目以及选项，选择正确的答案。本文基于此精心设计了指令提示来引导大模型输出期望的回复。本文设计的提示模板如图1所示，包括五个部分：

(1)任务定义：该部分详细描述了大模型需要完成的具体任务。

(2)参考示例：这里提供了基于选项相似度和空间语义理解能力相似度所选择得到的Q个示例样本，提供了大模型执行空间语义理解任务时的输出示例。

(3)背景信息：这里对应题目的背景信息，机器需要阅读理解后回答问题。

(4)题目：机器需要回答的问题。形式上是一个句中有括号的陈述句。

(5)选项：这里对应题目的选项，在微调过程中作为输入的一部分。

3.3 高效微调

指令微调（Instruction tuning）训练语言模型遵循专门的自然语言指令，从而提高模型的性能和理解能力。这需要在特定任务的数据集上基于提示-回答的形式对模型进行训练。针对开源大模型，本文采用主流的高效微调方法LoRA微调。LoRA微调冻结预先训练的模型权重，并将可训练的低秩矩阵注入到每一层中，从而显著减少需要训练的参数量，从而实现使用更少的计算资源来训练大模型。具体来说，对于使用$W_0 \in \mathbb{R}^{d\times k}$ 作为权重矩阵的线性层，LoRA引入一组低秩矩阵进行训练，分别为$B \in \mathbb{R}^{d\times r}$ 和$A \in \mathbb{R}^{r\times k}$，其中，$k$表示输入的维度，$d$表示输出的维度，$r$是预定义的等级。对于输入$x$，前向传播过程的输出表示为：

$$h = W_0 x + \Delta W x = W_0 x + BAx. \tag{4}$$

在微调过程中，只有矩阵B 和A 被更新，而W_0 保持静态并且不接收梯度更新。同时，本文在微调过程中使用了因果语言建模（Causal Language Modeling, CLM）技术，该方法以自回归方式训练模型，根据给定的输入标记序列$(x_0, x_1, x_2, \ldots, x_{i-1})$ 预测下一个标记x_i，最终使用负对数似然函数来计算训练过程的损失：

$$L_{\mathrm{CLM}}(\Theta) = \mathbb{E}_{x\sim D}\left[-\sum_i \log p(x_i \mid x_0, x_1, \ldots, x_{i-1}; \Theta)\right], \tag{5}$$

其中，Θ表示模型的参数。

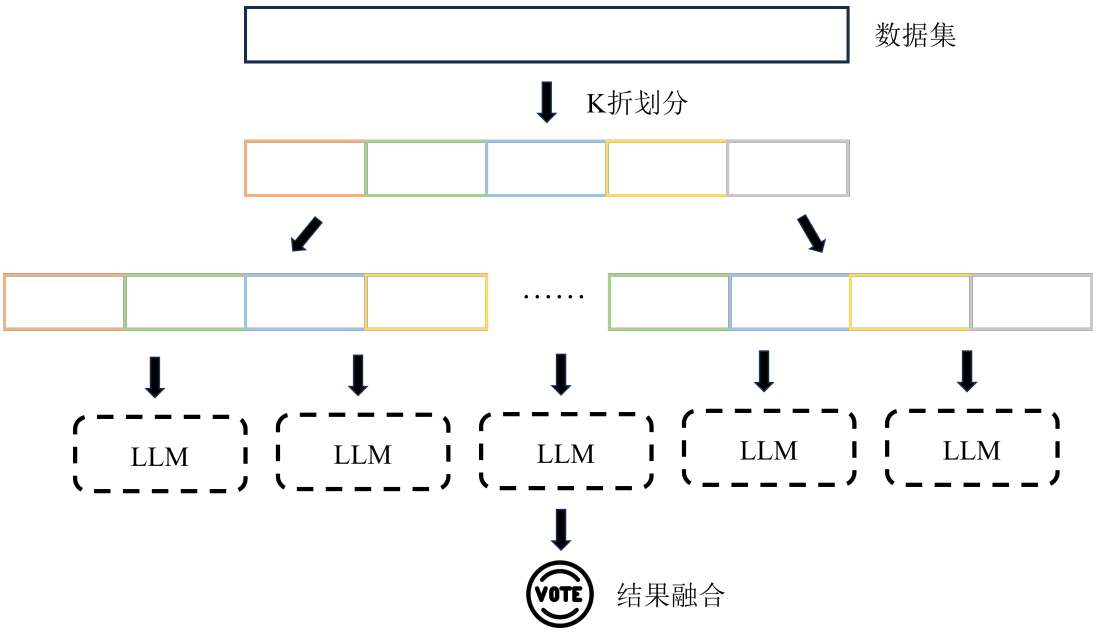

Figure 2: 模型融合流程

4 基于K折数据划分的模型融合

如图2所示，本文采用了K折交叉验证策略来增强模型的泛化能力，具体地，我们首先将整个数据集均匀划分为K个子集。在此过程中，本文选取$K = 5$作为交叉验证的折数。对于划分的K折数据，我们从中随机选取一折数据作为验证集，而其余$K - 1$折的数据则用作训练集。通过这种方式，我们分别训练了K个独立的模型。在模型评估阶段，每个模型都独立地对最终的测试集进行预测。为了得到更为稳健的预测结果，我们采用了集成学习中的投票机制来融合这K个模型的输出，通过投票法确定最终的预测结果。这种方法不仅可以减少模型在特定数据子集上的过拟合风险，同时能够有效提升模型在未知数据上的预测准确度。

空间信息实体识别数据示例：

输入：

{
“**qid**”：“1-train-s-913”，
“**text**”：“上海市浦东新区人民检察院指控，2021年1月7日11时45分许，被告人王某2驾驶牌号为沪CEXXXX的小型面包车沿本区临港大道由东向西行驶至祥凯路东约100米处时(时速大于67公里)，面包车左偏方向驶入中央绿化带，前部撞击路灯杆后向左侧翻，造成面包车前排乘员胡运送医途中死亡，王某2及后排乘员李某受伤。”，
“**question**”：“()前部撞击路灯杆后向左侧翻。”，
“**option**”：{“A”：“路灯杆”，“B”：“绿化带”，“C”：“面包车”，“D”：“以上选项都不是”}
}

输出：

{“**qid**”：“1-train-s-913”，“**answer**”：[“C”]}

Figure 3: 空间信息实体识别数据示例

Table 1: 数据集分布

训练集	验证集	测试集	总计
4,483	1,210	4,680	10,373

Table 2: 主实验结果

大模型	上下文学习	K折融合	总分	实体识别	角色识别	异常识别	空间推理	同义识别
ChatGLM3	否	是	44.22	64.91	79.48	59.20	25.44	31.69
ChatGLM3	是	否	50.53	74.21	85.20	67.80	31.23	36.00
ChatGLM3	是	是	52.34	73.68	86.62	73.00	32.40	39.69
文心4	是	否	53.89	67.19	93.12	76.60	28.97	56.46
文心4	是	是	55.21	69.82	92.86	76.20	30.68	58.62
ChatGLM3 + 文心4	是	是	56.45	73.68	92.86	76.20	32.40	58.62

5 实验与分析

5.1 实验设置

本文基于RTX 8000 48G GPU完成了所有实验，基于PyTorch框架[0]的代码开发模型。在开源大模型的微调方面，我们采用了支持LoRA微调的LMFlow框架[1]。此外，本次实验中，我们使用的基座模型包括开源大模型ChatGLM3-base[2] (Du et al., 2022) 和闭源大模型文心4[3]。对ChatGLM3-base模型进行微调时，我们将批处理大小设定为1，训练了7个Epoch，学习率设置为5e-5。

5.2 数据集

本次中文空间语义理解任务数据集共10,373条样本，一共约100万字符，训练集、验证集和测试集规模如表1所示。所有样本均以选择题的形式呈现，以空间信息实体识别类型数据为例，如图3所示，每条样本包含qid、text、question、option和answer五个字段，其中qid代表题目编号，采用“能力代号-子集类别-题目类别-题目号”的策略。对于能力代号的表示，1代表实体识别，2代表角色识别，3代表异常识别，4代表空间推理，5代表同义识别。对于子集类别的表示，train代表训练集，dev代表开发集，test代表测试集。对于题目类别的表示，使用s代表答案只有1个的单选题，m代表答案有2个及以上的多选题。text代表该样本的文本材料，即机器需要阅读的文本。question代表机器需要回答的问题，形式上是一个有括号的陈述句。option代表选择题的选项，采用“选项字母-选项内容”键值对的形式，共有四个键-值对。answer代表选择题的答案。

5.3 评估指标

SpaCE2024使用准确率（Accuracy）作为评价指标，对于每个待检查的预测结果，如果与标准答案一致，则认为预测正确。中文空间语义理解任务中的准确率被定义为命中正确答案的题目数量$N_{correct}$占总的测试样本数量N_{total}的比例，计算过程如下：

$$Accuracy = \frac{N_{correct}}{N_{total}}. \tag{6}$$

5.4 实验结果

主实验结果 本文将评测提供的原始训练集和验证集合并后用于模型微调，微调过程进行五折划分与模型融合，实验结果如表2所示。可以发现，基于ChatGLM3融合上下文学习进行高效微调时，模型的性能相比仅使用指令微调时取得了显著提升，从44.22提升至50.53。同时，融合示例样本进行微调后的ChatGLM3在实体识别和角色识别两个类型上取得了优异的性能，分别

[0]https://github.com/pytorch/pytorch
[1]https://github.com/OptimalScale/LMFlow
[2]https://github.com/THUDM/ChatGLM3
[3]https://yiyan.baidu.com/

Table 3: 参数探究结果

Q	交叉验证集	测试（K折融合）
-	45.83	44.22
3	48.21	49.47
5	48.61	49.93

Table 4: 不同规模数据下的实验结果

数据	总分	实体识别	角色识别	异常识别	空间推理	同义识别
原始训练集	49.93	71.75	85.84	72.80	28.04	39.38
原始训练集+ 验证集	52.34	73.68	86.62	73.00	32.40	39.69

达到了74.21和85.20。但是，ChatGLM3总体上在空间推理和同义识别两个类别上表现较差，这种现象可能是由于空间推理往往涉及多层次的理解和推断，如物体的相对位置、运动路径等，对模型的推理能力有更高的要求。而同义识别需要对文本的丰富语义有深刻理解，依赖于其语境和上下文信息，涉及细粒度的语义分析，在通用语料训练的大模型则很难捕获细粒度的语义信息。此外，实验结果表明文心4模型在整体表现上比ChatGLM3性能更好，但是在实体识别能力与空间推理能力上表现不如ChatGLM3，因此，本文将文心4和ChatGLM3两个大模型的结果进行融合，分别处理不同空间语义理解能力的题目，以提高系统在中文空间语义理解任务上的表现。总体上，ChatGLM3和文心4在空间推理类型上均表现不佳，表明了大模型在推理任务上仍然面临挑战。

融合策略有效性验证 如表2所示，无论是在开源大模型ChatGLM3还是闭源大模型文心4上，利用五折划分分别构建不同的模型进行推理后融合的效果，均高于不采用融合策略的模型。使用了融合策略后，ChatGLM3的总体评分从50.53提升至52.34，而文心4的总体评分则从53.89提升至55.21，表明我们设计的融合策略有效提升了模型的空间语义理解能力。

参数探究结果 本文以ChatGLM3为实验模型，探究了不同的示例样本数量对于模型性能的影响，结果如表3所示。表中的实验仅采用原始训练集进行五折交叉验证与模型融合，可以发现，不同的示例样本数量Q对于模型的性能具有一定的影响，当选择5个参考示例时的模型性能优于选择3个参考示例时的模型。同时，考虑到随着参考示例的数量增加，所需的训练成本也同步增加，因此，本文采用示例样本数量Q为5作为主要实验设置。

数据规模的影响 本文进一步在ChatGLM3上进行实验，分别探究仅采用原始训练集进行五折划分与模型融合，以及采用原始训练集和验证集进行五折划分和模型融合两种数据规模的效果。在测试集的实验结果如表4所示。可以观察到，增加训练数据量能显著提高模型性能。通过仅添加1000多条样本进行训练，模型的性能已从49.93提高至52.34。这表明，目前的空间语义理解数据集的规模还不足以完全发挥开源大模型在此任务上的潜力。因此，采用数据增强策略以进一步提升模型性能，是未来值得探索的一个研究方向。

6 总结

本文基于上下文学习策略探索了大模型在中文空间语义理解任务上的表现，揭示了大模型在该任务上的潜力。此外，通过基于选项相似度与空间语义理解能力相似度的策略去选择示例样本，研究结果显示，我们提出的样本选择方法能有效提高大模型在中文空间语义理解任务上的性能。实验结果表明，在高效微调阶段融合上下文学习策略能显著提升模型的空间语义理解能力，特别是在开源和闭源模型的综合使用上表现出色。此外，本文的研究不仅提高了模型的实际应用效果，也为未来在空间语义理解领域的研究提供了新的思路和框架。尽管取得了一定的成果，但模型在空间推理类型和同义识别类型上仍存在显著的差距，在接下来的工作中，如何有效提升模型的推理能力和同义区分能力可能是新的重点和难点。空间语义理解仍有待进一步深入研究。未来的工作可以在优化样本选择算法、数据增强策略、探索更多维度的语义关系，以及提升模型在更广泛自然语言处理任务中的泛化能力等方面进行。

致谢

本研究受国家社会科学基金项目（No.22BTQ045）资助。

参考文献

Soham Dan, Hangfeng He, and Dan Roth. 2020. Understanding spatial relations through multiple modalities. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 2368–2372. European Language Resources Association.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Parisa Kordjamshidi, Steven Bethard, and Marie-Francine Moens. 2012. SemEval-2012 task 3: Spatial role labeling. In Eneko Agirre, Johan Bos, Mona Diab, Suresh Manandhar, Yuval Marton, and Deniz Yuret, editors, **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)*, pages 365–373, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Carla E. Brodley and Andrea Pohoreckyj Danyluk, editors, *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289. Morgan Kaufmann.

Alexey Mazalov, Bruno Martins, and David Martins de Matos. 2015. Spatial role labeling with convolutional neural networks. In Ross S. Purves and Christopher B. Jones, editors, *Proceedings of the 9th Workshop on Geographic Information Retrieval, GIR 2015, Paris, France, November 26-27, 2015*, pages 12:1–12:7. ACM.

Eric Nichols and Fadi Botros. 2015. Sprl-cww: Spatial relation classification with independent multi-class models. In Daniel M. Cer, David Jurgens, Preslav Nakov, and Torsten Zesch, editors, *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*, pages 895–901. The Association for Computer Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In Marilyn A. Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Kirk Roberts and Sanda M. Harabagiu. 2012a. Utd-sprl: A joint approach to spatial role labeling. In Eneko Agirre, Johan Bos, and Mona T. Diab, editors, *Proceedings of the 6th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2012, Montréal, Canada, June 7-8, 2012*, pages 419–424. The Association for Computer Linguistics.

Kirk Roberts and Sanda M. Harabagiu. 2012b. Utd-sprl: A joint approach to spatial role labeling. In Eneko Agirre, Johan Bos, and Mona T. Diab, editors, *Proceedings of the 6th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2012, Montréal, Canada, June 7-8, 2012*, pages 419–424. The Association for Computer Linguistics.

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 Daniel M. Cer, David Jurgens, Preslav Nakov, and Torsten Zesch, editors, *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2015, Denver, Colorado, USA, June 4-5, 2015*, pages 856–861. The Association for Computer Linguistics.

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.

Johan A. K. Suykens and Joos Vandewalle. 1999. Least squares support vector machine classifiers. *Neural Process. Lett.*, 9(3):293–300.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.