CCL24-Eval任务3系统报告：基于参数高效微调与半监督学习的空间语义理解

李晨阳[1,2]，张龙[1,2]，郑秋生[1,2]
[1]中原工学院 前沿信息技术研究院，河南 郑州 450007
[2]河南省网络舆情监测与智能分析重点实验室，河南 郑州 450007
2312826399@qq.com

摘要

本文介绍了我们在第二十三届中文计算语言大会的第四届中文空间语义理解评测任务中提交的参赛模型。该任务旨在测试机器的中文语义理解水平。现有研究显示，机器的中文语义理解水平与人类平均水平相比仍有较大差距。近年来，生成式大规模语言模型在自然语言处理任务中展现了出色的生成和泛化能力。在本次评测中，我们采用了对Qwen1.5-7b模型进行高效微调的方法，以端到端的形式实现空间语义的推理过程，并结合prompt优化和半监督学习提升推理表现。实验结果表明，我们的模型在该任务中取得了领先的效果。

关键词： 大语言模型 ； 高效微调 ； 半监督学习 ； prompt

System Report for CCL24-Eval Task 3: Spatial Semantic Understanding Based on Parameter-efficient Fine-tuning and Semi-supervised Learning

Chenyang Li[1,2], Long Zhang[1,2],Qiusheng Zheng[1,2]
[1]Frontier Information Technology Research Institute,
Zhongyuan University of Technology, Zhengzhou 450007 China
[2]Henan Key Laboratory on Public Opinion Intelligent Analysis, Zhengzhou China
2312826399@qq.com

Abstract

This paper introduces the models we submitted for the Fourth Chinese Spatial Semantic Understanding Evaluation Task at the 23rd Chinese Computational Linguistics Conference. This task aims to test the level of machine understanding of Chinese semantics. Existing research indicates that the level of machine understanding of Chinese semantics still lags significantly behind the average human level. In recent years, generative large-scale language models have demonstrated excellent generative and generalization capabilities in natural language processing tasks. For this evaluation, we employed an efficient fine-tuning method on the Qwen1.5-7b model to achieve the spatial semantic reasoning process in an end-to-end manner. We also combined prompt optimization and semi-supervised learning to enhance reasoning performance. Experimental results show that our model achieved leading performance in this task.

Keywords: Large Language Model , Efficient Fine-Tuning , Semi-Supervised Learning , prompt

1 引言

空间表达是自然语言中常见的现象，用来描绘物体之间的空间位置关系。空间范畴是人类认知总重要的基础范畴，大量空间信息存在于自然语言文本中。在通往人工智能的道路上，空间语义理解是不可绕开的关键步骤。著名认知语言学家Jackendoff(2004)在其概念语义学理论中指出，空间结构是语言系统的四种基本结构之一。理解文本中的空间表达语义，除了语言知识外，还需借助空间认知能力来构建空间场景，并依据世界知识进行有关空间方位信息的推理。通常认为，对文本中空间信息的理解，不仅需要掌握句段中词汇、句法语义知识，还需要具备一定的常识或背景知识，甚至是超出语言范畴的空间想象等认知能力，以此来构建空间场景。

在这样的背景下，第二十三届中国计算语言学大会发布了SpaCE2024技术评测。与之前相比，SpaCE2024不再划分子任务，而是以选择题的形式考察五个层次的空间语义理解能力：空间信息实体识别，要求从四个选项中选出文本信息的参照物。空间信息角色识别。要求从四个选项中选出文本信息的语义角色，或者选出与语义角色相对应的空间表达形式。空间信息异常识别，要求从四个选项中选出文本空间信息异常的语言表达。空间方位信息推理。要求基于文本给出的推理条件进行空间方位推理，从四个选项中选出推理结果。空间异形同义识别。要求从四个选项中选出能使两个文本异形同义或异义的空间义词语。

近期，随着预训练语言模型的参数量和数据量不断增大，大规模语言模型在人工智能领域取得了革命性进展，使得一些过去被视为仅限于人类能力范围的自然语言处理任务变得可行。相较于传统预训练模型，大模型能够记忆更广泛的世界知识，关注到更多的信息，并展现出传统模型不具备的涌现能力。通过指令微调和代码预训练等技术，大模型具备了理解人类指令、泛化到新任务、代码理解与生成、利用思维链推理以及处理长距离依赖等多项能力。基于此，我们采用了阿里发布的中文大模型Qwen1.5-7b进行端到端微调，以实现空间语义的推理过程。并结合prompt优化和由半监督学习产生的伪标签来提升推理表现。

2 相关工作

为了评测机器的空间语义理解能力，自然语言处理领域的评测任务主要分为以下三类：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和时间抽取任务相似，代表性工作有SpRL任务(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，代表性工作又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隐类型，詹卫东等(2022)首次提出该任务，认为如果机器能够识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。

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等(2015)提出了一种基于筛选的模型，它使用多层来提取空间元素，然后引入分类器来分类空间关系。D'Souza等(2015)提出了一种基于筛选的模型，通过贪心的特征选择技术来生成各种手动特征。还有研究人员(Salaberri et al., 2015)引入外部知识作为空间信息的补充，在此过程中，提供了许多空间元素的信息。Kim等(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元素。

随着神经网络的广泛应用，Ramrakhiyani等(2019)提出了一种通过依存句法生成候选关系，并使用BiLSTM 模型对候选关系进行分类的方法。Shin等(2020)提出了一种使用CRF 进行空间关系分类的模型，Wu等(2019)则提出了基于AR-BERT 的方法来提取空间关系。此外，一些研究关注于多模态空间关系提取。例如，Dan等(2020)提出了一种空间BERT，它同时从文本以及包含实体的图片中来预测实体之间的空间关系。

3 实现方法

如图1所示，我们先对文本内容进行格式化处理，然后对主流大模型采用qlora方法(Dettmers et al., 2024)进行端到端微调，实现语义空间的推理过程。接着，选出效果较好的模型作为基线模型，最后采用prompt优化和半监督学习提升推理表现。

3.1 数据预处理

为了使得本评测任务的数据集能直接应用于大模型微调，在数据预处理阶段，我们编写

图 1: 模型的数据预处理、预测以及后处理过程

了代码对原始数据集进行了处理。如图2所示，首先识别原数据的qid字段，并将识别后的能力代号、题目类别加入到文本序列中，各项信息用“#”进行分隔，正确答案也作为一个序列用“#”进行分隔。我们还尝试了除此之外的其他prompt模板，以优化模型输入格式。


```
{
  "qid": "4-dev-m-1079",
  "text": "赵云、曹操、关羽、孙坚四人来到火锅店吃火锅，选了四人卡座坐下。卡座分列一张长方形桌子长边两侧，每排卡座上坐两人。面对面而坐。已知：赵云在关羽右边坐，曹操在孙坚同侧左边。",
  "question": "关羽和()不是并排坐的。",
  "option": {
    "A": "孙坚",
    "B": "赵云",
    "C": "曹操",
    "D": "以上选项都不是"
  },
  "answer": [
    "A",
    "C"
  ]
}
```

```
{
  "messages": [
    {
      "role": "user",
      "content": "赵云、曹操、关羽、孙坚四人来到火锅店吃火锅，选了四人卡座坐下。卡座分列一张长方形桌子长边两侧，每排卡座上坐两人。面对面而坐。已知：赵云在关羽右边坐，曹操在孙坚同侧左边。#空间推理#多选题#关羽和()不是并排坐的。#A:孙坚 B:赵云 C:曹操 D:以上选项都不是"
    },
    {
      "role": "assistant",
      "content": "A#C"
    }
  ]
}
```


图 2: 左为原格式数据，右为拼接后格式化数据

3.2 大模型微调

Qwen是一个全能的语言模型系列(Bai et al., 2023)，其使用了高达3万亿个token的数据进行预训练，涵盖多个类型、领域和任务，不仅包括基本的语言能力，还包括算术、编码和逻辑推理等高级技能。同时使用了复杂的流程进行数据清晰和质量控制。

在微调阶段，我们采用了qlora方法，这是一种高效的微调方法，可减少内存使用量。构造的输入数据包括任务指令和利用prompt拼接后的原句，标签则是对应的选项答案。我们把该任务视为序列生成任务，而不是直接生成符合本任务的数组形式，这样也是为了避免模型生成不符合格式的答案。我们发现，参数秩设置为64和96时效果最佳。如图3所示，针对于输入的问题，语言模型会将输入问题的上文作为参考进行信息的搜索，并以此来进行空间位置的推理，从而得出正确答案。

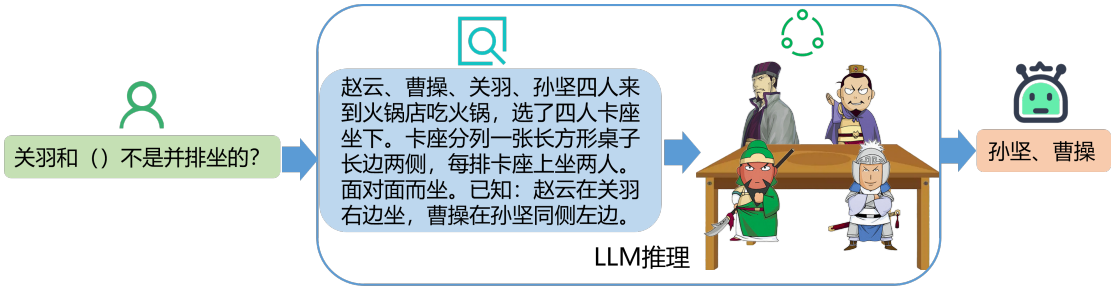

图 3: 大模型语义空间的推理过程

3.3 数据分析

为了进行后续分数的提升，我们对5种层次题型类别的数量和每个类别的准确率进行了统计，如图4所示，我们发现不论在训练集还是验证集中，都存在数据分布不平均，部分标签对应的数量较少的问题(Li et al., 2023)，如在训练集中，空间方位信息推理类型数量达到了1210条，其对应的准确率也较高。而空间异形同义识别类型的训练集只有5条，其对应的准确率也较低。为了提升该类的数据量，我们尝试采用增加数据集的方式来使得模型在训练时学习到更多的相关特征，我们首先采用同义替换、随机词插入等方法对该类的数据进行了数据增强，在模型上进行微调后，验证集分数有所提升，但在测试集上效果却较差，出现了过拟合的现象，我们猜测可能是数据增强引入了过多错误的噪声，进而破坏了数据原始的分布。于是我们采用了同样能增加训练集数量的伪标签方法。

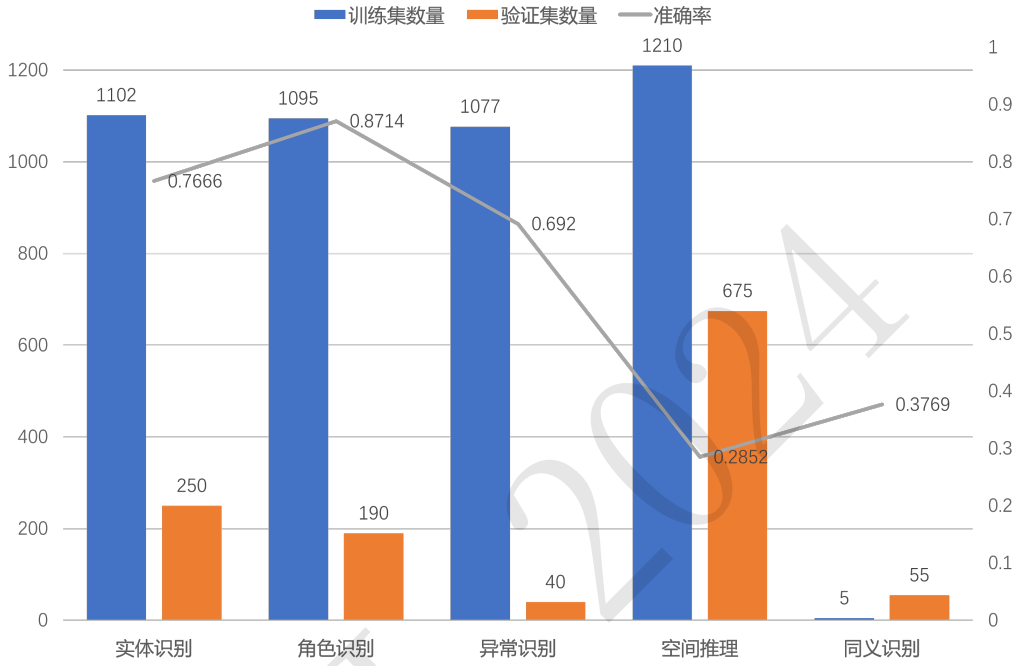

图 4: 大模型语义空间的推理过程

3.4 基于半监督学习的伪标签生成

伪标签方法来自于半监督学习，其核心思想是借助无标签的数据来提升有监督过程中的模型性能(Rizve et al., 2021; Berry et al., 2019)。由半监督学习生成伪标签的过程如图5所示，其主要是将模型对无标签的测试数据的预测结果加入到训练集中，从而增大数据量以提升模型效果。该方法适用于模型精度比较高的情况。由于我们未采用传统分类模型进行训练，从而无法得到预测类别的概率值。我们采用多个预测结果较高的文件，并取预测标签相同的部分作为伪标签加入到训练集中重新训练。经过统计，每次得到的伪标签数量都接近3k条，这表明我们的数据集在原本数量的基础上又增加了3k条数据。经过多轮的伪标签训练后，筛选出的伪标签会越来越接近，最终模型达到了拟合的状态，此时在进行后续的伪标签训练已经无法进一步再提升测试集的准确率。

3.5 模型融合

关于模型融合，周志华(2021)教授在其《机器学习》一书中提到：模型融合要好而不同，即模型差异性越大，融合效果越好。我们从两个方面来增加差异化，一是使用不同的多个模型，Qwen和ChatGLM4，二是重新划分训练集和验证集来改变模型输入。在使用这两个模型进行预测后，选择出每个层次类别中较好的部分进行结果的融合。

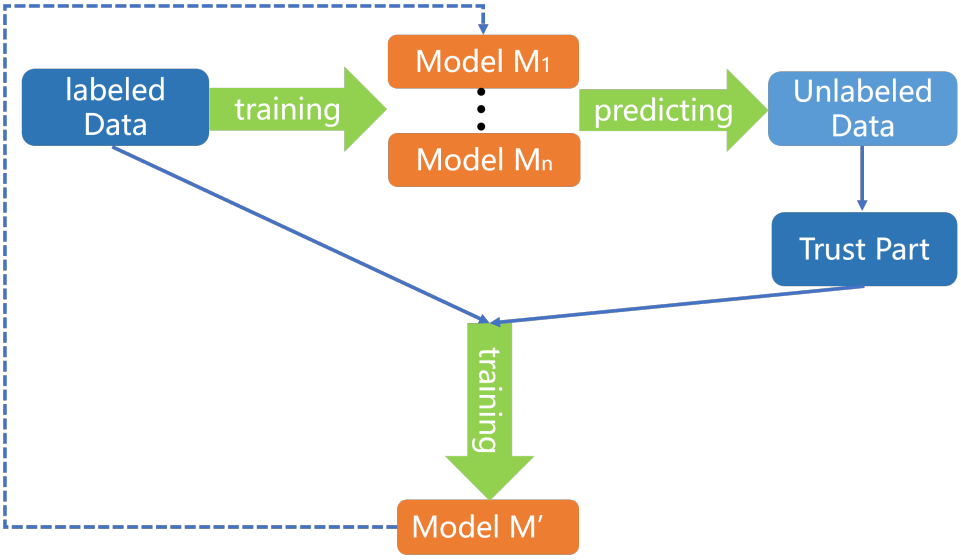

图 5: 基于半监督学习的伪标签生成

3.6 prompt模板构造

如图6所示，我们构造了两种prompt模板，并用不同颜色区分原数据的信息。尽管第二种模板比第一种更加流畅，但在实验中发现，在使用模型微调情况下，这两种prompt模板对结果的影响基本可以忽略不计，然而，由于第二种模板增加了一些额外的文本信息，导致其最大长度增加，从而增加了训练时的额外开销。相反，在调用大模型接口时，prompt的质量对结果影响非常显著。

他扶着移动的扶手，四下张望。右边有一个迷宫，里面满是叽叽喳喳叫嚷着的鹦鹉，左边是一家日杂品店，里面到处闪着铬的光芒。#实体识别#单选题#()右边有一个迷宫。#A:他 B:扶手 C:铬 D:鹦鹉"

请阅读以下文本并回答问题：他扶着移动的扶手，四下张望。右边有一个迷宫，里面满是叽叽喳喳叫嚷着的鹦鹉，左边是一家日杂品店，里面到处闪着铬的光芒。题型：实体识别单选题。问题：()右边有一个迷宫。括号中应该填入什么？选项：A.他 B.扶手 C.铬 D.鹦鹉 请选择正确答案。

图 6: 左为prompt模板1，右为prompt模板2

4 实验

4.1 实验结果

如表4.1所示，我们列出了采用ChatGLM3、Qwen-7b、Qwen1.5-7b和调用ChatGLM4接口的测试集结果。在选择基线模型的过程中，我们发现Qwen1.5-7b模型要明显优于其他模型。但为了更好的使用半监督学习来生成高质量的伪标签，我们同样也使用了除Qwen1.5以外的模型进行训练，以融合出更高质量的伪标签。随着Qwen模型逐轮加入伪标签数据后，模型预测的准确率不断提升。同时我们发现，在使用微调时，prompt模板的优劣对结果不会有太明显的影响。而在调用大模型接口时，prompt模板的质量对结果影响非常明显。最终，通过多轮伪标签的应用，我们模型预测准确率达到了0.5516的分数，再通过多个模型结果的融合，最终达到了0.566的分数。如表4.1所示，实验结果证明，我们的模型取得了较为先进的效果，并在本任务中取得第4名的成绩。

模型	方法	准确率	实体识别	角色识别	异常识别	空间推理	同义识别
Baseline	微调	0.4792	0.7509	0.8818	0.6860	0.2196	0.4200
ChatGLM3-6b [0]	微调	0.4370	0.6720	0.7727	0.6320	0.2348	0.3185
Qwen-7b[1]	微调	0.5035	0.7667	0.8714	0.6920	0.2853	0.3769
	+1轮伪标签	0.5344	0.8175	0.9039	0.7580	0.3157	0.3630
	+2轮伪标签	0.5423	0.8316	0.9000	0.7920	0.3122	0.3954
	+3轮伪标签	0.5434	0.8421	0.9051	0.7980	0.3078	0.3970
Qwen1.5-7b[2]	微调	0.5100	0.7795	0.8810	0.6900	0.2966	0.4277
	+1轮伪标签	0.5465	0.8351	0.9103	0.7740	0.3181	0.4046
	+2轮伪标签	0.5485	0.8403	0.8961	0.7960	0.3196	0.4092
	+3轮伪标签	0.5516	0.8526	0.9078	0.7860	0.3210	0.4092
ChatGLM-4[3]	api+模板1	0.3870	0.5543	0.7987	0.4580	0.1838	0.3353
	api+模板2	0.4832	0.6070	0.9064	0.7080	0.2309	0.4923
-	结果融合	0.5660	0.8526	0.9142	0.7980	0.3210	0.4923

表 1: 方法结果

排名	队伍	准确率	实体识别	角色识别	异常识别	空间推理	同义识别
1	TeleAI	0.6024	0.8947	0.9364	0.8480	0.3471	0.5631
2	zyy	0.5969	0.8491	0.9143	0.8100	0.3716	0.5131
3	涛涛不绝	0.5949	0.7719	0.9429	0.7800	0.3711	0.5877
4	Prompt	0.5660	0.8526	0.9142	0.7980	0.3210	0.4923
5	猪门永存	0.5647	0.7368	0.9286	0.7620	0.3240	0.5862
6	龙年旺旺空间站	0.5620	0.7965	0.9429	0.7420	0.3064	0.5692
7	panda	0.5448	0.7509	0.9117	0.7540	0.3044	0.5231
8	一个短篇	0.5355	0.7333	0.9013	0.7960	0.2858	0.5123
9	欣崽全球研究生后援会	0.5199	0.8053	0.9000	0.7020	0.34076	0.2415
10	CPIC1	0.4865	0.7667	0.8610	0.6220	0.2603	0.4031
11	少小离家	0.4724	0.6719	0.8182	0.7000	0.2735	0.3369
12	Azur Promilia	0.3364	0.4947	0.6727	0.2160	0.2172	0.2662
-	Baseline	0.4792	0.7509	0.8818	0.6860	0.2196	0.4200

表 2: 测试集排行榜[4]

4.2 结果分析

在使用大模型之前，我们也尝试使用了类似Bert的传统分类模型，然而，传统模型在测试集上的表现要远低于基线模型的结果。这一结果表明基座模型对于推理效果有着显著的影响，而大模型在预训练获得的世界知识和涌现能力对空间语义理解能力任务有着重要帮助。我们在对大模型微调时也面临着一个普遍问题，即幻觉现象(Huang et al., 2023)。当模型生成的文本不遵循原文或者不符合事实时，我们就认为模型出现了幻觉，尽管在我们训练集的选项中只有4个选项可选，但模型在结果预测时仍会产生4个选项以外的答案，为此，我们暂时只采用正则表达式来过滤掉这些无效答案。

[0]https://github.com/THUDM/ChatGLM3
[1]https://github.com/QwenLM/Qwen
[2]https://github.com/QwenLM/Qwen1.5
[3]https://chatglm.cn
[4]https://2030nlp.github.io/SpaCE2024/leaderboard.html

通过大量实验发现，数据增强方法在大多数任务中，尤其是小样本任务，通常会有不同程度的提升效果。然而，在本任务中，由于数据量规模并不小，采用类似随机词插入和随机词删除等通过添加噪声来实现数据增强的方法可能改变了原本的数据分布，反而没有显著提升效果。相反，使用半监督学习生成的伪标签可以增加数据集的规模，提升模型预测的准确率，并增加模型的泛化性。在使用多轮伪标签方法后，后续筛选得出的伪标签几乎不会有变化，导致模型的性能不再有提升，这时可以采用模型融合技术，取差异较大的多个模型，分别学习不同的输入，使得多个模型之间学到的知识特征尽量不同，这样使得多个模型可以更好的融合，提升性能。

5 总结

在本任务中，除了提供的训练集和测试集以外，任务组织者还额外提供了空间语义词表信息，但我们在仅使用了训练集和验证集的情况下就达到了超过基线模型的效果，后续并采用prompt优化和半监督学习的方式来进一步提升推理的表现。为了进一步优化结果，未来我们将针对使用训练集存在的过拟合问题，考虑在划分训练集和验证集时进行数据均衡。使用模型融合时，采用五折交叉验证法来训练多个模型，然后在对多个预测结果取平均，以取得更好的预测效果。

参考文献

Ray Jackendoff. Précis of foundations of language: Brain, meaning, grammar, evolution. *Behavioral and Brain Sciences*, 26(6):651–65; discussion 666–707, 2004.

Eric Nichols and Fadi Botros. Sprl-cww: Spatial relation classification with independent multi-class models. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 895–901, 2015.

Jennifer D'Souza and Vincent Ng. 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, 2015.

Haritz Salaberri, Olatz Arregi, and Beñat Zapirain. 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)*, pages 856–861, 2015.

Bogyum Kim and Jae Sung Lee. Extracting spatial entities and relations in korean text. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2389–2396, 2016.

Nitin Ramrakhiyani, Girish Palshikar, and Vasudeva Varma. A simple neural approach to spatial role labelling. In *Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part II 41*, pages 102–108. Springer, 2019.

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

Shanchan Wu and Yifan He. Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 2361–2364, 2019.

Soham Dan, Hangfeng He, and Dan Roth. Understanding spatial relations through multiple modalities. *arXiv preprint arXiv:2007.09551*, 2020.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.

Chenyang Li, Long Zhang, Qiusheng Zheng, Zhongjie Zhao, and Ziwei Chen. User preference prediction for online dialogue systems based on pre-trained large model. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 349–357. Springer, 2023.

Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*, 2021.

Michael W Berry, Azlinah Mohamed, and Bee Wah Yap. *Supervised and unsupervised learning for data science*. Springer, 2019.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

Oleksandr Kolomiyets, Parisa Kordjamshidi, Marie Francine Moens, and Steven Bethard. 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, 2013.

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

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

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

詹卫东, 孙春晖, 岳朋雪, 唐乾桐, and 秦梓巍. 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, pages 99–110, 2022.

Zhi-Hua Zhou. *Machine learning*. Springer nature, 2021.