**复旦大学本科生毕业论文（设计）开题报告**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 姓名 | 黄呈松 | | 学号 | 19302010004 |
| 所在院系 | 软件学院 | | 专业 | 软件工程 |
| 指导教师 | 郑骁庆 | | 职称 | 副教授 |
| 校外指导教师  及其所属单位 | 无 | | 职称 | 无 |
| 论文（设计）题目 | 基于语言模型的日常任务规划研究 | | | |
| **开题报告：建议包含以下内容（可另附页）：**  1. 选题的目的和意义；  2. 国内外相关研究状况综述（列出相应的参考文献）；  3. 主要研究内容与基本思路，详细技术路线，并分析可行性、难点和创新点；  4. 预期成果及形式。  见附页 | | | | |
| **研究进度及具体时间安排（不够写可加行）** | | | | |
| 起止日期 | | 主要研究内容 | | |
| 2023.02.07-2023.03.03 | | 阅读国内外相关文献，撰写开题报告 | | |
| 2023.03.04-2023.03.31 | | 编写实验代码，记录实验数据 | | |
| 2023.04.01-2023.04.21 | | 分析实验数据，验证实验结果 | | |
| 2023.04.22-2023.05.05 | | 撰写论文，定稿提交 | | |
| **指导教师对课题报告的意见：**  **1．对选题依据、基本思路或技术路线的可行性、创新性的评价；**  **2．存在的主要问题和改进建议。**  指导教师签名： 年 月 日 | | | | |

# 开题报告

研究背景

日常机器人是一个正在发展的领域。随着人口老龄化的加剧，更加需要可以在家中完成一系列日常任务的机器人，来缓解护理人员不足的问题。这类机器人可以完成例如自动化打扫卫生，完成早饭准备等不需要专业知识的工作。而将人工智能应用在这类机器人中也是十分有前景的方向。已有一些工作通过人工智能算法来操控机器人在房间内完成这些工作，这些算法主要通过摄像头看到的图片与文本指令来执行相关动作。

在这一过程中，任务规划是至关重要的。以我们给机器人下令“泡一杯茶”为例，我们期望机器人自动的完成“找到茶杯茶叶，找到热水，泡好茶，端上来”这整个过程，而不是需要用户分四次布置这样的指令。这要求我们可以通过“泡一杯茶”这一全局指令来生成详细的四步规划。这类规划不同于简单的路线规划，不仅需要对于周围环境的认识，同时也需要常识知识的帮助，例如在这个例子中，需要了解泡茶需要热水这一常识性的知识才能做出正确的规划。

预训练语言模型在不同的任务中都取得了不错的性能，例如问题回答和常识推理等。这证明了预训练语言模型拥有一些解决这些问题所需要的常识知识，例如微波炉可以加热食物。更进一步，已有一些工作研究预训练语言模型是否可以将这些知识用于日常任务的规划中。但受限于预训练语言模型没有观察外部世界和与外部世界交互的能力，先前的工作并没有将环境信息纳入考虑。使得这些任务的规划中缺少例如“向左转”类似的导航规划。而这些规划是机器人可以适应新的环境的关键因素。

研究内容

通过已有ALFRED数据集构建一个从任务命令加环境信息到具体步骤的数据集，通过预训练语言模型来完成具体的规划过程，并通过表格的方法将环境信息加入输入，试图使得语言模型可以生成与环境相关的规划。同时提出每次只生成下一个步骤的迭代式规划方法。最后证明这两种方法都可以提升日常任务的规划性能。

研究意义

规划算法是家政机器人智能化的重要组成部分。通过研究机器人规划算法，可以使家政机器人能够自主的进行决策和规划，以实现更加复杂的任务，如清扫，洗衣，烹饪等。这些规划依赖于大量的常识知识，而已有的预训练语言模型被证明拥有了常识推理的能力，因此本文期望将预训练语言模型用于家政机器人的规划算法。同时完成的新数据集也可以帮助之后的研究者在这一任务上更好的进行研究。

可行性，难点和创新点

可行性：前人已在表格理解，常识推理等任务上有了一定的进展。通过预训练语言模型进行的日常任务规划也已经成为研究的热点。本文期望结合这些工作的方法与已有的模型，将方法应用到全新的日常任务规划领域，因此本文提出的方法具有可行性。

难点：本文存在两个难点，一是如何在语言模型没有摄像头等输入设备的情况下将环境信息输入给语言模型，第二则是如何评价语言模型所给出的规划是否满足要求。本文第一次尝试通过表格的方式将环境信息进行编码来作为预训练语言模型的输入。此外本文也是第一次期望进行文本级别的规划评价。

创新点：本文首次将环境信息通过表格的形式进行编码，是的预训练语言模型有了解环境信息的能力，进而利用语言模型内涵的常识知识和推理能力进行日常任务的任务规划。本文同时新建了全新的该任务上的数据集和提出了全新的评价方法，给科研社区在该问题上的研究打下基础。

**预期成果**

预期成果主要分为三部分，首先是全新的基于ALFRED数据集生成的从任务命令到具体步骤的数据集，第二部分是对于生成质量的全新评价指标，最后是可以完成日常任务规划的语言模型。

相关研究综述

**常识推理**

常识推理（commonsense reasoning）是指通过每个人都有的普遍基本的知识或者经验进行推理，来推断出其他未知的信息。这一能力是人类认知过程的核心组成部分，同时也是构建自然语言理解系统或者人工智能系统的核心任务[[[1]](#endnote-0)][[[2]](#endnote-1)]。传统方法（深度学习）通过知识表示和知识图谱解决常识推理的相关任务[[[3]](#endnote-2)]。随着深度学习的发展，通过预训练语言模型来解决常识推理任务成为了主流方法。

不同于人类通过周围环境了解常识信息[[[4]](#endnote-3)],先前的工作证明了预训练语言模型通过在大规模无标签语言库的训练过程中可以获得常识相关的知识[[[5]](#endnote-4)]。因此预训练语言模型开始在没有外联知识库的情况下被经常的用于解决常识推理的问题[[[6]](#endnote-5)]。有许多任务从各个角度考察了与训练语言模型对于尝试知识的应用例如常识抽取[[[7]](#endnote-6)]，下一状态预测[[[8]](#endnote-7)]，文化和社会认知[[[9]](#endnote-8)]等。

而已有一些工作将预训练语言模型用于日常任务规划当中。例如ALFWorld[[[10]](#endnote-9)]构建了一个文字游戏的场景，在这种场景下检测语言模型在日常任务中的常识推理能力。语言模型可以根据任务目标和当前状态来生成下一步的步骤。ScienceWorld[[[11]](#endnote-10)]同样通过文字游戏的方式来检验模型的常识推理能力，不同于ALFWorld在日常任务中完成，ScienceWorld主要着眼于小学生科学实验常识例如热胀冷缩等。SayCan[[[12]](#endnote-11)]通过预训练语言模型生成一系列语义上可行的规划，之后通过强化学习算法来构建价值函数（value function）来选择在当前环境下的操作方法。同样GPT-3等大规模语言模型也被用于从文本中总结计划等[[[13]](#endnote-12)]。

**预训练语言模型与表格预训练**

预训练加微调是近几年来常用的将深度学习应用在下游任务的方法。预训练模型主要通过大规模语料进行无监督训练。预训练的编码器-解码器模型例如BART[[[14]](#endnote-13)]和T5[[[15]](#endnote-14)]在各种自然语言生成任务上都取得了不错的效果。尤其实在一些被认为跟常识推理相关的任务比如CommonsenseQA[[[16]](#endnote-15)]的优秀效果更是说明了这类预训练语言模型具有很强的常识推理能力，可能在这次的任务上获得不错的效果。

表格是一种数据组织的方式，被广泛应用于记录，分析和呈现数据。可以更好的将大量数据集合在一起进行直观的展示。不同于常规的文本语句，表格带有一些具有语义的结构信息例如标题，行和列等，因此不能简单的将预训练语言模型直接应用在表格型数据上。在这样的背景下有许多先前的工作对于预训练模型进行了更多的表格优化的预训练方法。表格预训练主要目标在两个方向：生成更好的表格表示和将表格作为中间表示。

对于生成更好的表格表示，从TabNet[[[17]](#endnote-16)]使用采用序列关注（Sequential Attention）的方式，使用预测遮掩特征（Masked Features）作为无监督的与训练任务，得到更好的表格表示。TAPAS[[[18]](#endnote-17)]在BERT所使用的MLM（masked language model）的基础上加上了判断表格是否符合文本的与训练任务并且加上对于行列以及排名的embedding层来获得更好的表格cell表示。TAPEX[[[19]](#endnote-18)]通过大量表格结合SQL语句，通过SQL的执行结果进行大规模有监督训练。

**日常任务智能化**

机器人技术的发展使得机器人可以完成日常中的一些任务，然而这类家居机器人可以真正的进入日常生活需要机器人可以通过自然语言和摄像头了解到的知识来完成这样的任务，在这样的基础上AIFRED（Action Learning From Realistic Environments and Directives）数据集[[[20]](#endnote-19)]应运而生。任务集包括了在多个不同场景下的不同类型的任务。环境可以进行交互式操作等，以此为基础，前任提出了多种不同的解决方案用于操纵机器人在这类环境中的动作。

FILM[[[21]](#endnote-20)]（Following Instruction in Language with Modular Methods）证明了显式空间记忆（explicit spatial memory）和语义搜索策略（semantic search policy）可以在执行日常任务过程中为环境和指令提供更好的编码方式。同时证明了在没有专家轨迹和具体每一步指令的情况下，模型也可以完成一些任务，这说明了模型具有从高级指令分析出具体的每一步步骤的能力，作为支撑证明了本文的方法有很强的可行性。

LEBP[[[22]](#endnote-21)]（Language Expectation & Binding Policy）首先将每一步的step生成文本上的单个字的表示，之后通过绑定政策将这个表示对应到具体的操作，例如导航到某一点或对于面前的物体进行某种操作。该方法说明了，自然语言的步骤可以有效迁移到未见过的场景中，同时说明了分步骤的指令的重要价值，为本文研究提供了意义。

参考文献

1. [] Bhargava, P., & Ng, V. (2022). Commonsense Knowledge Reasoning and Generation with Pre-trained Language Models: A Survey. AAAI Conference on Artificial Intelligence. [↑](#endnote-ref-0)
2. [] Liu, H., & Singh, P. (2004). Commonsense Reasoning in and Over Natural Language. International Conference on Knowledge-Based Intelligent Information & Engineering Systems. [↑](#endnote-ref-1)
3. [] Zang, L., Cao, C., Cao, Y., Wu, Y., & Cao, C. (2013). A Survey of Commonsense Knowledge Acquisition. Journal of Computer Science and Technology, 28, 689 - 719. [↑](#endnote-ref-2)
4. [] Tincoff, R., & Jusczyk, P.W. (1999). Some Beginnings of Word Comprehension in 6-Month-Olds. *Psychological Science, 10*, 172 - 175. [↑](#endnote-ref-3)
5. [] Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A.H., & Riedel, S. (2019). Language Models as Knowledge Bases? *ArXiv, abs/1909.01066*. [↑](#endnote-ref-4)
6. [] Rajani, N., McCann, B., Xiong, C., & Socher, R. (2019). Explain Yourself! Leveraging Language Models for Commonsense Reasoning. *Annual Meeting of the Association for Computational Linguistics*. [↑](#endnote-ref-5)
7. [] Xu, F.F., Lin, B., & Zhu, K.Q. (2017). Automatic Extraction of Commonsense LocatedNear Knowledge. *Annual Meeting of the Association for Computational Linguistics*. [↑](#endnote-ref-6)
8. [] Zellers, R., Bisk, Y., Schwartz, R., & Choi, Y. (2018). SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. *Conference on Empirical Methods in Natural Language Processing*. [↑](#endnote-ref-7)
9. [] Lin, B., Xu, F.F., Zhu, K.Q., & Hwang, S. (2018). Mining Cross-Cultural Differences and Similarities in Social Media. *Annual Meeting of the Association for Computational Linguistics*. [↑](#endnote-ref-8)
10. [] Shridhar, M., Yuan, X., Côté, M., Bisk, Y., Trischler, A., & Hausknecht, M.J. (2020). ALFWorld: Aligning Text and Embodied Environments for Interactive Learning. *ArXiv, abs/2010.03768*. [↑](#endnote-ref-9)
11. [] Wang, R., Jansen, P.A., Côté, M., & Ammanabrolu, P. (2022). ScienceWorld: Is your Agent Smarter than a 5th Grader? *Conference on Empirical Methods in Natural Language Processing*. [↑](#endnote-ref-10)
12. [] Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Ho, D., Hsu, J., Ibarz, J., Ichter, B., Irpan, A., Jang, E., Ruano, R.J., Jeffrey, K., Jesmonth, S., Joshi, N.J., Julian, R.C., Kalashnikov, D., Kuang, Y., Lee, K., Levine, S., Lu, Y., Luu, L., Parada, C., Pastor, P., Quiambao, J., Rao, K., Rettinghouse, J., Reyes, D.M., Sermanet, P., Sievers, N., Tan, C., Toshev, A., Vanhoucke, V., Xia, F., Xiao, T., Xu, P., Xu, S., & Yan, M. (2022). Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. *ArXiv, abs/2204.01691*. [↑](#endnote-ref-11)
13. [] Olmo, A., Sreedharan, S., & Kambhampati, S. (2021). GPT3-to-plan: Extracting plans from text using GPT-3. *ArXiv, abs/2106.07131*. [↑](#endnote-ref-12)
14. [] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *Annual Meeting of the Association for Computational Linguistics*. [↑](#endnote-ref-13)
15. [] Raffel, C., Shazeer, N.M., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P.J. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *ArXiv, abs/1910.10683*. [↑](#endnote-ref-14)
16. [] Talmor, A., Herzig, J., Lourie, N., & Berant, J. (2019). CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. *ArXiv, abs/1811.00937*. [↑](#endnote-ref-15)
17. [] Arik, S.Ö., & Pfister, T. (2019). TabNet: Attentive Interpretable Tabular Learning. *ArXiv, abs/1908.07442*. [↑](#endnote-ref-16)
18. [] Herzig, J., Nowak, P.K., Müller, T., Piccinno, F., & Eisenschlos, J.M. (2020). TaPas: Weakly Supervised Table Parsing via Pre-training. *ArXiv, abs/2004.02349*. [↑](#endnote-ref-17)
19. [] Liu, Q., Chen, B., Guo, J., Lin, Z., & Lou, J. (2021). TAPEX: Table Pre-training via Learning a Neural SQL Executor. *ArXiv, abs/2107.07653*. [↑](#endnote-ref-18)
20. [] Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., Zettlemoyer, L., & Fox, D. (2019). ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10737-10746. [↑](#endnote-ref-19)
21. [] Min, S., Chaplot, D.S., Ravikumar, P., Bisk, Y., & Salakhutdinov, R. (2021). FILM: Following Instructions in Language with Modular Methods. *ArXiv, abs/2110.07342*. [↑](#endnote-ref-20)
22. [] Liu, H., Liu, Y., He, H., & Yang, H. (2022). LEBP - Language Expectation & Binding Policy: A Two-Stream Framework for Embodied Vision-and-Language Interaction Task Learning Agents. *ArXiv, abs/2203.04637*. [↑](#endnote-ref-21)