Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6232–6239
Marseille, 20-25 June 2022

Task-Driven and Experience-Based Question Answering Corpus for In-Home Robot Application in the House3D Virtual Environment

Zhuoqun Xu[1], **Liubo Ouyang**[1], **Yang Liu**[2]
[1]Hunan University, [2]Samsung Research Center Beijing
{zhuoqunxu, oylb}@hnu.edu.cn
yang9004.liu@samsung.com

Abstract

At present, more and more work has begun to pay attention to the long-term housekeeping robot scene. Naturally, we wonder whether the robot can answer the questions raised by the owner according to the actual situation at home. These questions usually do not have a clear text context, are directly related to the actual scene, and it is difficult to find the answer from the general knowledge base (such as Wikipedia). Therefore, the experience accumulated from the task seems to be a more natural choice. We present a corpus called TEQA (task-driven and experience-based question answering) in the long-term household task. Based on a popular in-house virtual environment (AI2-THOR) and agent task experiences of ALFRED, we design six types of questions along with answering including 24 question templates, 37 answer templates, and nearly 10k different question answering pairs. Our corpus aims at investigating the ability of task experience understanding of agents for the daily question answering scenario on the ALFRED dataset.

Keywords: grounded QA corpus, interactive environment, task experience

1. Introduction

Question answering (QA) is a very important way to research language communication, and it is also an important part of natural language processing (NLP) (Yih et al., 2015; Wijmans et al., 2019) . Meanwhile, we notice that the trend of introducing the grounded language learning methodology is getting more attention (Wu et al., 2017; Nishida et al., 2019; Castro et al., 2020). The setting of grounded language learning is closer to the scene where natural language occurs, better presents the context where QA happens, and contains physical mapping (non-language). Assuming two men talk in a house, a default context of their dialog is the surrounding physical environment (their house), which is hardly described by the pure text. Environmental settings have emerged one after another, especially interactive environments for natural language research. Notably, ALFRED (Action Learning From Realistic Environments and Directives) (Shridhar et al., 2020), a benchmark for learning a mapping from natural language instructions to sequences of actions for domestic tasks.Naturally, we wonder whether the robot can answer the questions raised by the owner according to the actual situation at home. These questions usually do not have a clear text context (different from the current popular setting of reading comprehension), are directly related to the actual scene, and it is difficult to find the answer from the general knowledge base (such as Wikipedia). Therefore, the experience accumulated from the task seems to be a more natural choice.

For instance, the agent performs a task "*clean the lettuce*" in a kitchen scene. It will go through a series of actions: *pick up, walk, turn right, put, turn on*, and finally complete the task. After the task completes, we wonder if we can ask the agent a question *'Where do you clean the lettuce?'*, the agent should answer *'I clean the lettuce in the sink.'* , as shown in figure 1. Namely, it should have the ability to learn information from its experience of tasks like what humans do. This scenario of QA is rather essential for human daily conversations. To investigate such QA which is based on task experience, we require an interactive environment with rich task scenarios to a grounded QA corpus related to it.

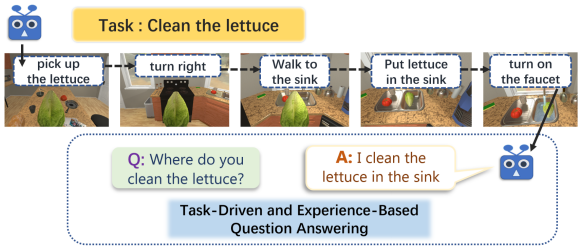

Figure 1: **An example of TEQA.** The agent tries to complete the task through the sequence of actions in the virtual environment and learns from the experience that can be used to answer the question in the process of the task or after the task is finished.

In this paper, we propose **TEQA (Task-Driven and Experience-Based Question Answering)**. Compared to other work, Visual Question Answering (VQA) (Antol et al., 2015) describes a scenario through a static image. While Embodied Question Answering (EQA) (Das et al., 2018) holds QA in a virtual environment, however, its platform lacks interaction which limits its task in navigation and interaction with objects. Besides, Interactive Question Answering (IQA) (Gordon et al., 2018) is deficient in changes of objects. We fo-

cus on tasks that require interaction and lead to state changes of environment and investigate whether the agent has the ability to understand the underlying information in task experiences via question-answering dialog. Specifically, using ALFRED as our benchmark, we create a QA corpus upon it, which contains 24 question templates and 37 answer templates, and covers 352 nouns(objects) and 33 verbs (actions/tasks). The question is divided into two classes with six types, including shallow three types of QA, which simply requires retrieving information to answer questions, and deep QA, which requires the agent to understand the underlying logic of experience. There are nearly 10k different questions generated.

According to ALFRED's current environment and tasks, we build a series of templates to generate these question and answering instances. The core idea of our semi-automatic QA generation is that we can benefit from the advantage of the virtual environment. For example, in a scene (kitchen), the task for an agent is to wash the apple. Based on experience, the agent first finds the apple, then picks it up, puts it in the sink, and turns on the faucet. During such a process, we can easily access the ground truth information from the virtual environment of apple, including position, state, and other objects, such as a faucet which is seen during the task but not explicitly mentioned (in language) in the task. Thus, after the task is over, the question *"Why did you turn on the faucet ?"* is generated based on the corpus and context (task and sequence of actions). The agent ought to answer *"Because I need to wash the apple"* . This will prove that the agent relates the two actions (washing the apple and turning on the faucet), indicating that the agent can learn the relationship between the task and experience(information) to create the answer.

2. Related Work

QA is a common way to research the natural language such as VQA (Peng et al., 2016; Yu et al., 2019) that integrated vision and language. The machine obtains information from images and videos in order to answer questions. In essence, VQA is a static job that lacks behavior and state changes. There is no change or diversity in videos. Embodied Question Answering (Das et al., 2018) and Embodied AI (Smith and Gasser, 2014) are recent works that intelligence emerges in the interaction of an agent with an environment. After the question is posed, the agent deduces the answer by exploring the virtual environment. Compared with VQA, it has a more complex context and more diverse questions. The interactive environment (Kolve et al., 2017) has richer information that is grounded. The agent faces a room with multiple objects that demands navigation to explore the environment. EQA has aroused plenty of discussion and attention, th In the process of exploration, the agent needs to yield an answer through language grounding, visual sensing, and common-sense reasoning.e application of the virtual environment which makes agents have more space for it to improve learning ability.

Research in recent years has shown that language is more than just a connection between symbols, it also conveys functional meaning and includes the transfer of words to the grounding of physical concepts. It combines simple knowledge to describe more complex concepts and can reflect logic and purpose (Kottur et al., 2017). Therefore, grounded language learning becomes a hot topic in NLP, in which we encounter more and more challenges. EQA seems to be insufficient to explore more complex issues. For example, in ALFRED, where object interactions and state changes will occur. Thus, the new goal is how to make agents generate logic autonomously from experience. The specific environment in the ALFRED has a parallel interactive text world environment (Shridhar et al., 2020). This environment allows the agent to solve specific tasks by reasoning and learning high-level strategies in an abstract space. As a result, through a large number of expert demonstrations, ALFRED allows agents to learn how to complete household tasks. Combining language understanding, computer vision (CV), reinforcement learning, and navigation, it aims to transform language into interaction and action sequences (language-driven agents). Because its environment is a high simulation of the real world, which can also become a research platform for grounded language learning. In this direction, we hope to provide a schema to solve some issues.

3. Task-Driven and Experience-Based Question Answering

We propose the task-driven and experience-based question answering corpus based on the ALFRED benchmark. There are some premises before we discuss TEQA. Firstly, the virtual environment for us to build Q&A is interactive and fully observable, so semantics and information can be easily obtained. Secondly, the agent has been trained and has a certain ability to complete tasks. Finally, humans can complete tasks in a similar environment and utilize the experience to complete these latent questions. On this premise, we consider that TEQA is effective and meaningful.

3.1. Virtual Environment

We use AI2-THOR (A Near Photo-Realistic Interactable Framework for Embodied AI Agents) (Kolve et al., 2017), which is a controllable, well-designed simulation to the real-world in-house scenario. It contains lots of items and allows the agent to interact with these items to change the state of objects. For example, the apple can be sliced, and the potato can be cooked. Moreover, it also covers the various style of the home environment, including 120 different houses along with instances in a different style (e.g., different

sharp, textual cups in different locations). This supports our opinion that, in such a scenario, the agent cannot handle the question and answer with a static knowledge base. It has to dynamically adapt to the environment, which will constantly change. In other words, it has to learn to answer the QA with current experience and information summarized from previous experience. Meanwhile, the virtual environment usually provides an interface that contains all information of objects in the current environment. We can use it to generate legal QA.

3.2. Experience

For the agent, there are tasks (daily goals) in each specific scene, and it performs a series of actions to complete the goal. In the process, the agent obtains object information through vision and infers the relation between them. This can be seen as a process of acquiring knowledge. And we call these contents experience and use a six-tuple to describe the agent's experience. $\mathbf{T}$ is the set of all tasks. $\mathcal{T} \in \mathbf{T}$ is one of the tasks. The sequence of actions generated in the task is $\mathcal{A} = \{ \mathbf{a} \mid \mathbf{a} \in \mathbf{A} \}$, and $\mathcal{A} \subseteq \mathbf{A}$, $\mathbf{A}$ is the action space. Obj is the set of interacting objects in the behavior. $Obj \times \mathcal{A} \rightarrow \mathcal{S}$ stands for the position changes of the objects after the interaction. $Obj \times \mathcal{A} \rightarrow \mathcal{C}$ stands for the property changes of the objects. The result of the task is $\mathcal{R}$. When the agent task is successfully completed, $\mathcal{R}(\mathcal{T}, \mathcal{A}) = \mathbf{1}$, otherwise $\mathcal{R} = \mathbf{0}$ or $-\mathbf{1}$. The experience $\mathbb{E}$ obtained by the agent in the process of completing the task is

$$\mathbb{E} = \{ \mathcal{T}, \mathcal{A}, Obj , \mathcal{S}, \mathcal{C}, \mathcal{R}(\mathcal{T}, \mathcal{A}) \}$$

The agent learns the experience when the task is successful, ensuring that it can learn the correct associations. Agents continuously acquire knowledge and understand the relations between them. Agents can use this experience to answer questions about causality and relevant connection. The agent possessing this kind of experience is similar to logic, which is closer to human thinking.

3.3. Consideration of Question Answering

In ALFRED, objects, scenes, and actions are fully observable. We can directly obtain all the data at each moment in the process from its interfaces, such as the position coordinates, state of the objects, the tasks contents, and agent behaviors. Consequently, the distance between objects can be calculated by coordinates. With the action of the task, the environment produces many data changes (position coordinate changes), and the agent can obtain the information of the objects through vision. Subsequently, the agent retrieves information and saves them as memory. Thereupon, we designed a QA corpus based on these, as shown in Figure 2. Inspired by Terry Winograd's (Winograd, 1974) and previous work (Peng et al., 2016; Gordon et al., 2018), we consider that the logical QA is not owing to statistics and probability. It should be based on facts and reality as the expression of the agent's internal logic, what reflects the agent's semantic understanding and cognition, and the process of transforming data into natural language. So we design specific templates and vocabulary sets for the QA corpus. Moreover, we set up multiple types of questions that are suitable for a variety of common family scenes. As a result, the known information and corpus are used to generate answers. Through the agent's answer to understand whether the agent has logic and dynamic perception. The contents of the questions emphasize the interactive process. Questions are not limited to the surface (visual), including behavior, logic, right and wrong, meaning, purpose, the location of objects that are not in the field of current vision (seen at a certain time), and the relationship. For example, *"do you need a knife to cut potatoes?" "What else can be used besides the knife?" "What objects are in the refrigerator?" "what can I do with the knife?" "And where can I put it?"* The agent can answer better by gaining experience from the interaction.

Question Template.

- (Surface state) Ask the state of an object: 'What is the state of *object* ? Where is the *object* ?'
- (Causality) Ask the logical relationship: 'Why did you *verb* the *object* ? Why is the *object adjective /preposition* the *object* ?'
- (Detailed process) Understand the detailed behavior process: 'How do you *verb* the *object* ? What should you do for the task ?'

Corresponding Answer Template.

- The *object* is *adjective* .The *object* is *preposition* the *object* . There is an *object preposition* the *object*.
- Because the *object* is *preposition* the *object* . Because I *verb* . Because the *object* needs to*verb*.
- I *verb* the *object* . Then, I *verb* the *object* . Then, I *verb* the *object* . (recyclable)

When the agent cannot answer, there are some answering example sentences. "Sorry, I can't answer this question." "I will continue to learn how to reply to you." "I believe I will do better next time." "There is something in your question that I don't understand."

3.4. Grounded QA Corpus

The corpus is dynamic and expandable. We compile elementary sentence patterns as templates so that QA sentences can be automatically generated. The template is constructed manually (setting the prescribed format and annotations) to ensure the correctness and representatives of the corpus and has a strong correlation with related tasks and objects.

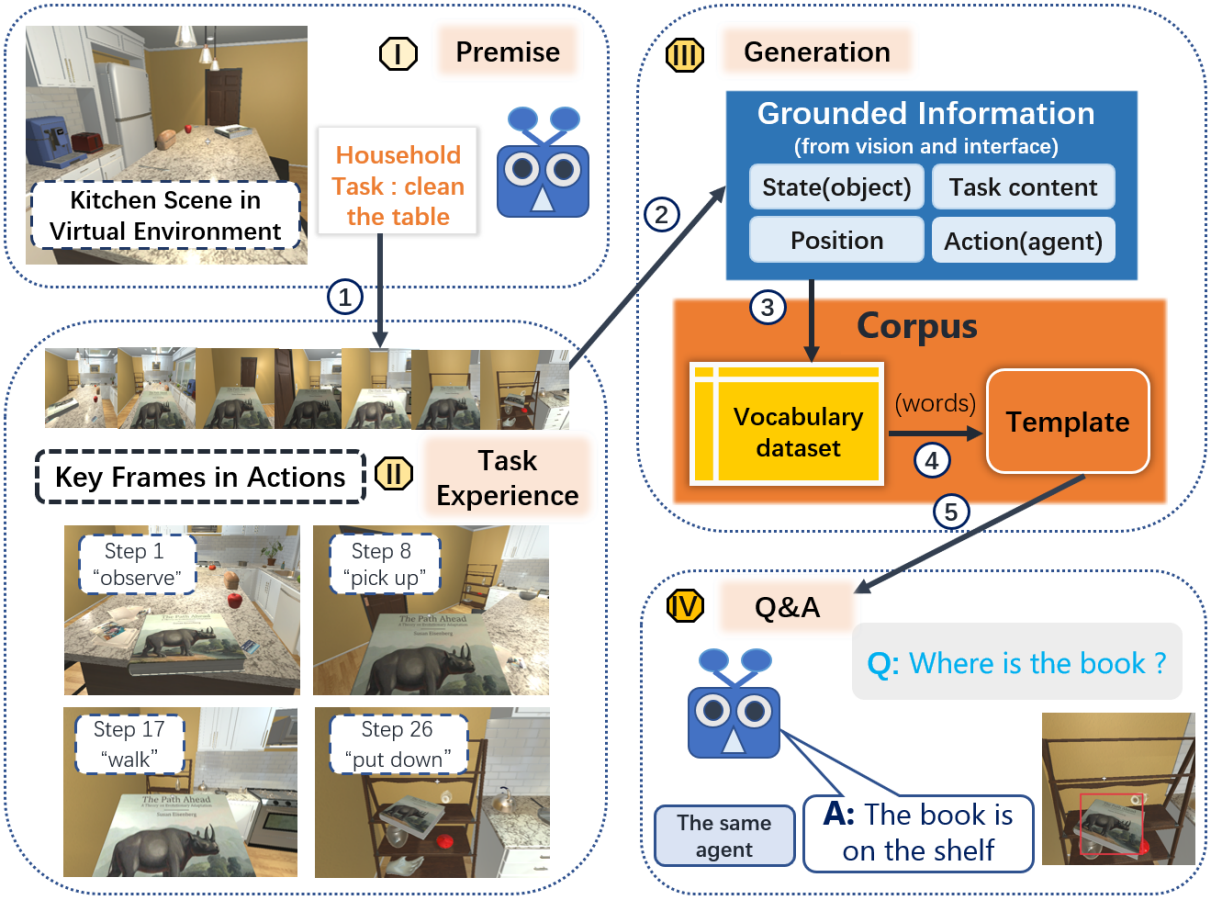

Figure 2: **Main workflow of TEQA.** A trained agent is required to complete a daily task in a virtual environment, which can learn experiences from action sequences. (1) The agent completes the goal through a series of behaviors. (2) With the generation of changes in the actions(attributes and positions of objects), the agent collects object information through vision. (3) QA is based on the content of the task, the objects in the current environment, and the action sequences of the agent. (4) The words in the dataset are filled in the template to generate a question, similarly, we can create massive questions. (5) The robot yields the answer in the same way.

Premise. The agent must have certain prior and background knowledge. In other words, it is necessary to set up some logical operation inside the agent, preset a knowledge base containing common sense (deriving answers), models for semantic understanding (tasks and questions) and visual processing (extracting features), as well as modules for data processing and information retrieved. Moreover, the agent is trained in advance (by expert demonstration), it can understand the task and produce actions for the goal. And these works have already yielded initial results in ALFRED.

Construction sequence: *planning, collecting corpus, inputting data and words, labeling, building templates, setting restrictions*. This is not only a specific work but also a methodological proposal. We will continue to update and maintain the corpus.

Vocabulary Dataset. It is constructed manually and combined with existing word corpus and synonyms, which can satisfy most scenes of the household task. The noun dataset contains tools and foods in the environment and daily lives, such as *apple, pot, bowl, table, etc*. The verb dataset includes actions in ALFRED and common housework, such as *put, pick up, open, etc*. The adjective dataset includes descriptive words for the states of objects and environment, such as *cooked, moveable, dirty, break, etc*. The preposition vocabulary dataset includes commonly used: *on, in, at, from, under, below, near, etc*.

Question Types. The goal of ALFRED is to develop a domestic robot assistant. In this scene, we need to treat the agent as a real human and ask it some common questions to understand its logic. Based on the above three question types, we continue to subdivide the content. We consider the possible actions and potential logic, design multiple templates for each question, and set a corresponding answer template for each question. question classification (the specific question is a variant of the following classification):

- The information of the object after the task (what)

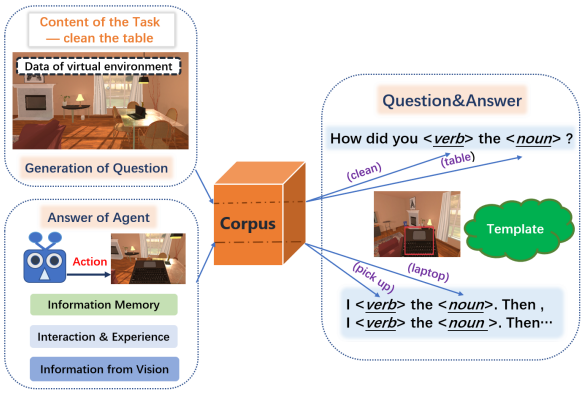

Figure 3: **Corpus uses context information to generate grounded QA.** The key information matching vocabularies are filled into the template to ask questions. And the agent uses its information and experience to answer the questions in the same way.

- The detailed process of completing the task (how)
- Relation between objects (what)
- The source and purpose of the object (what)
- Reasons for changes in the state of objects (why)
- The meaning of some actions (why)

Consider the possible actions and potential logic, design multiple templates for each question, and set corresponding answer templates as well.

Implementation. The corpus is designed in view of a rapid application for TEQA in a (virtual) home environment, which must be combined with the task completion model. The agent retrieves information while performing or completing tasks, hence it can conduct QA training. The agent conducts experiments in the order of *semantic understanding (task) → actions (collecting information and learning knowledge in the behavior) → asking questions → logical derivation → answering questions*.

Analysis. The process of using the corpus is shown in Figure 3. During the process or after the behavior occurs, according to the content of the task and environment, extract the nouns and add them to the question template to generate the question. We can choose any question type to ask the agent, due to the detailed information of the task being known in the behavior, the question generation method is fully feasible. As for the agent, it extracts characteristic information through vision and saves part of the information. After analyzing the semantics of the question, the agent retrieves the real-time data from its "memory". After a logical derivation process, it matches words in the vocabulary set and selects a template to produce the answer.

For specific tasks, there are fixed answers (manually set), which can be used as test sets for TEQA. Comparing the similarity and relevance between the agent's answer and the right answer can quantify the judgment

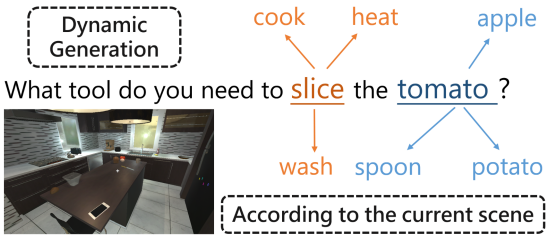

Figure 4: **Example of question.** The form and meaning of the question can be changed to understand different contents and logic about the agent in a specific scene.

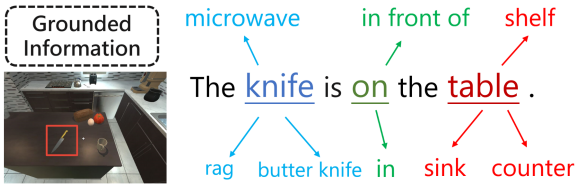

Figure 5: The answer must correspond to the question and be able to visualize the logical process of the agent.

of the TEQA result (score). In this way, we can achieve similar semi-supervised learning (reinforcement learning or self-supervised learning), or use TEQA as a benchmark for testing and exploring the internal logic of the agent. From the above, in a closed environment (daily scene), the agent is passable to utilize the corpus to complete automatic QA. Therefore, in TEQA, there may not necessarily be direct human participation and supervision. TEQA semanticizes and visualizes the "thinking" process of the machine that converts data into the natural language in the agent's self-question and self-answer. In the grounded language learning and process of studying the cognitive logic of the agent, this is an indispensable and meaningful part of the evaluation standard and reference.

Instance. It happens in the kitchen, and the agent is washing dishes to complete a task. And when you need to use a tool to slice the tomato, you ask the agent *"What tool do you need to slice the tomato?"* (certainly, replacing words in the template can generate different questions, as shown in Figure 4. The question can also be *"How do you cut potato?" " Why is the knife in the fridge?" "Why did the position of the pot change?" "Can you finish the task with a spoon?"*). Regarding this question, the agent has seen a knife on the kitchen table, and it used to use a knife to slice something. The agent tries to associate the butter knife with the task of 'slice' through experience. We hope that the agent will answer you *"The knife is on the table"* after retrieving the information like Figure 5. It may also generate these answers: *"The butter knife is in the sink." "The knife is on the shelf."* Of course, the subject may not be a knife (*"The cup is on the counter"*).

Instance	Question Answering
Q :	*"How do you wash towel ?"*
A :	*"I put the towel in the sink.*
	Then, I turn on the faucet."
Q :	*"Why do you put the pen in the drawer ?"*
A :	*"The desktop needs to clean."*
Q :	*"Why do you turn on the microwave ?"*
A :	*"The bread needs to heat."*
Q :	*"What tools do you need to cut tomato ?"*
A :	*"A knife."*
Q :	*"What is in the refrigerator ?"*
A :	*"There is lettuce in the refrigerator.*
	There is bread in the refrigerator
	There is an egg in the refrigerator."
Q :	*" Where does the shirt come from ? "*
A :	*" I take the shirt from the cabinet. "*

Table 1: **Some instances of question answering.** 'Q' is to ask the agent, 'A' is the answer of the agent.

This QA detects whether the agent can learn knowledge in the environment and actions. The following cites some TEQA in Table 1. After the interaction and behavior, we set up a series of questions that contain six types. Through the agent's answer, we can know its action process and logic rather than just the result. This will help improve the agent's task success rate, language understanding, and learning ability.

3.5. Statistics

The ALFRED dataset contains 25743 language instructions, corresponding to 8055 expert demonstrations, a total of 2685 task parameters. In 120 different indoor scenes: 30 kitchens, 30 bathrooms, 30 bedrooms, and 30 living rooms. It is worth noting that there are about two hundred object types in the virtual environment with more than two thousand different styles (one object class may have multiple styles). In particular, there are unique object classes (58 types) and receptacle object classes (26 types).In AI2-THOR framework, the agent can generate 13 actions (5 navigation operations: *move ahead, rotate right, rotate right, look up and look down*, and 7 interactive operations: *pick, place, open, close, toggle on, toggle off, and slice*). Tasks are divided into 7 types. Moreover, the target success rate is about 8% that has a comparatively large room for improvement. In a scene, the detailed information included in the dataset contains all objects in the virtual environment and their positions and states (Pickupable, Sliceable, Receptacle, etc.). At each frame, we can obtain the event metadata, including the position of the agent and each step of its actions. So in the experiment, we collected all types of objects in the virtual environment, although they may have many different appearances, as shown in Table 2. In other words, the content of the corpus will fully generalize the content of the ALFRED dataset. In addition, when we construct the vocabulary dataset, we add synonyms and generalizations of the objects to enhance the description of daily scenes. We have increased the types of verbs to reflect the actions of the agent in more detail to satisfy the demands of the household tasks. Adjectives (color and shape) are content that is not in the ALFRED dataset, which helps the agent to analyze and learn the attributes of the objects through visual analysis. We have marked the commonly used prepositions and prefer that the agent use prepositions to describe the spatial relationship between objects. In order to reduce the bias of the corpus, we have referred to open-source datasets of related scenes. Consequently, through the real-world mapping and virtual 3D environment reference, it is guaranteed that the corpus can cover daily life scenes. This also ensures that our corpus can complete the goal of TEQA.

Statistics of Corpus	Description
question template	24
answer template	37
noun(object)	352
verb (action)	33
adjective (state)	46
preposition (location)	17
example sentence	582
generation of different questions	9336

Table 2: **Corpus for TEQA.** Multiple permutations and grounded setting enrich the generation of questions.

Question Type	Content	Amount
surface state	color and quantity	3
information	position and distance	4
visual analysis	the state change of object	2
causality	reason of change or action	4
experience	how to do or tool be used	6
logical derivation	purpose of object or action	5

Table 3: **From the shallower to the deeper.** Questions may be asked at any time, and their content is uncertain. It relies on real-time data and has specific significance.

Since the QA is dynamically generated, the corpus is easy to expand. Owing to the arrangement of various words that can produce enough questions, the corpus is not limited by the information scale of the virtual environment. Question templates covering various aspects are given in Table 3. In particular, more templates will easily create extremely more problems.

4. Discussion and Conclusion

We believe that the learning ability of the agent is the key factor to improve the task success rate. Therefore, we introduce TEQA to research the comprehension of agents. By the feedback of the agent, we comprehend the steps of its actions (not just the results) to further

improve its internal algorithms. For a natural language text, changing one of the words may make the meaning completely different from the previous one. Semantic comprehension should not only consider the connection between symbols but also the agent learns logic similar to humans. Thus, we consider that the agent should not acquire results from its surmise. As a result, we use grounded information to force the agent to understand knowledge(rather than guess). On this basis, we propose a corpus to accomplish the TEQA. The way of grounded language learning is more similar to the human language environment that presets the context of the conversation. The richness and authenticity of grounded information cannot be described by data and text. The finite objects have infinite random arrangements, and there are many possibilities for agent behavior. The virtual 3D environment is closer to the scene of daily life, the agent in which is like a baby, constantly exploring and learning. And TEQA is like its test papers and examinations. Within the scope of a problem domain and natural language rules, our QA corpus has sufficient feasibility and high efficiency, which will be beneficial to research on domestic robots and grounded language learning.

In fact, using the corpus can complete automatic task-driven and experience-based question answering to the agent. But there are also some shortcomings. TEQA can only be limited to family tasks, besides, the objects and behaviors are finite. Besides, the detail of learning knowledge from experience needs to be further explored. In the future, we will perfect our ALFRED task completion model and implement the TEQA corpus on it for exploring superior performance, and we will continue to optimize and improve the QA corpus, such as more sentence templates, richer nouns, and descriptive words, adding more adjectives (depicting the characteristics of objects), more types of QA, multiple languages and scenes, more detailed corpus tags, larger scale, and more accurate evaluation indicators. Certainly, we could make two agents with different experiences ask questions to each other, which may produce more interesting results. Within the context of the information boundary, the semantic understanding and logical derivation of the agent will become an important direction of NLP, which is the basis for the agent to have intelligence. No matter TEQA, grounded language learning, or ALFRED, it is a complex problem that requires combining multiple fields. Our work can help these, but it is not enough.

5. References

Winograd, T. . (1974). Understanding natural language. Leonardo, 3(1), 1-191.

Das, A. , Datta, S. , Gkioxari, G. , Lee, S. , Parikh, D. , & Batra, D. .(2018). Embodied Question Answering. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Shridhar, M. , Thomason, J. , Gordon, D. , Bisk, Y. , & Fox, D. . (2020). ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Antol S , Agrawal A , Lu J , et al. (2015). VQA: Visual Question Answering[J]. International Journal of Computer Vision, 2015, 123(1):4-31.

Szlam, A. , Gray, J. , Srinet, K. , Jernite, Y. , Joulin, A. ,& Sy Nn Aeve, G. , et al. (2019). Why build an assistant in minecraft?. arXiv preprint arXiv:1907.09273

Kottur, S. , Moura, J. , Lee, S. , & D Batra. (2017). Natural Language Does Not Emerge 'Naturally' in Multi-Agent Dialog. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP).

Eli Bingham, Piero Molino, Paul Szerlip, Fritz Obermeyer, Noah D. Goodman (2017). Characterizing how Visual Question Answering models scale with the world, NIPS 2017

Gordon, D. , Kembhavi, A. , Rastegari, M. , Redmon, J. , & Farhadi, A. . (2018). IQA: Visual Question Answering in Interactive Environments. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Kolve, E. , Mottaghi, R. , Gordon, D. , Zhu, Y. , & Farhadi, A. . (2017). Ai2-thor: an interactive 3d environment for visual ai. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Shridhar, M. , X Yuan, MA Cté, Y Bisk, & Hausknecht, M. . (2020). Alfworld: aligning text and embodied environments for interactive learning. Conference on Learning Representations (ICLR).

Das, A. , Agrawal, H. , Zitnick, L. , Parikh, D. , & Batra, D. . (2017). Human attention in visual question answering. Computer Vision and Image Understanding.

Clark, H. H. . (1996). Using language. Cambridge University Press: Cambridge), 952:274–296, 1996.

Peng, Z. , Goyal, Y. , Summers-Stay, D. , Batra, D. , & Parikh, D. . (2016). Yin and Yang: Balancing and Answering Binary Visual Questions. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Su, H. , Shen, X. , Zhao, S. , Zhou, X. , Hu, P. , & Zhong, R. , et al. (2020). Diversifying Dialogue Generation with Non-Conversational Text. In ACL 2020

Castro, Santiago , Azab, Mahmoud , Stroud, Jonathan and Noujaim, Cristina , Wang, Ruoyao , Deng, Jia & Mihalcea, Rada.LifeQA: A Real-life Dataset for Video Question Answering. In LREC 2020.

Harnad, S. . (1990). The symbol grounding problem. Physica D Nonlinear Phenomena, 42.

Wang, P. , Wu, Q. , Shen, C. , Hengel, A. , & Dick, A. . (2015). Explicit knowledge-based reasoning for visual question answering. Computer Science.

Macmahon, M. , Stankiewicz, B. , & Kuipers, B. . (2006). Walk the Talk: Connecting Lan-

guage, Knowledge, and Action in Route Instructions. DBLP.

Yu, L. , Chen, X. , Gkioxari, G. , Bansal, M. , & Batra, D. . (2019). Multi-Target Embodied Question Answering. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Jorge, E. , M Kågebäck, Gustavsson, E. , & Johansson, F. D. . (2016). Learning to play guess who? and inventing a grounded language as a consequence. arXiv preprint arXiv:1611.03218 [cs.AI]

Lazaridou, A. , Peysakhovich, A. , & Baroni, M. . (2016). Multi-agent cooperation and the emergence of (natural) language. In ICLR 2017.

Austin, J. L. . (1966). How to do things with words. Analysis, 23(Suppl-1), 58-64.

Cangelosi, A. , Greco, A. , & Harnad, S. . (2002). In cangelosi a parisi d (eds) (2002). simulating the evolution of language. In: Hurford J.R. Studdert-Kennedy M. and Knight C. (eds), Approaches.

Yih, W. T. , Chang, M. W. , He, X. , & Gao, J. . (2015). Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing.

Côté, M., Kádár, Á., Yuan, X., Kybartas, B.A., Barnes, T., Fine, E., Moore, J., Hausknecht, M.J., Asri, L.E., Adada, M., Tay, W., & Trischler, A. (2018). TextWorld: A Learning Environment for Text-based Games. CGW@IJCAI.

Roma Patel, Roma Pavlick, and Stefanie Tellex. Learning to ground language to temporal logical form. In NAACL, 2019.

Lynch, C. , & Sermanet, P. . (2020). Grounding language in play, in RSS 2021.

Mehta, H. , Artzi, Y. , Baldridge, J. , Ie, E. , Mirowski, P. . (2020). Retouchdown: adding touchdown to streetlearn as a shareable resource for language grounding tasks in street view. arXiv:2001.03671

Wu, Q. , D Teney, Wang, P. , Shen, C. , Dick, A. , & Anton, V. . (2017). Visual question answering: a survey of methods and datasets. Computer Vision Image Understanding, S1077314217300772.

Savva, M. , Kadian, A. , Maksymets, O. , Zhao, Y. , Wijmans, E. , Jain, B. , et al. (2019). Habitat: a platform for embodied ai research. In ICCV.

Singh, K. P. , Bhambri, S. , Kim, B. , Mo Tt Aghi, R. , & Choi, J. . (2020). Moca: a modular object-centric approach for interactive instruction following. In ICCV.

Puig, X. , Ra, K. , Boben, M. , Li, J. , Torralba, A. . (2018). Virtualhome: simulating household activities via programs In CVPR 2018.

Nishida, K. , Nishida, K. , Nagata, M. , Otsuka, A., Saito, I. ,& Asano, H. , et al. (2019). Answering while Summarizing: Multi-task Learning for Multi-hop QA with Evidence Extraction. In ACL 2019.

Wijmans, E. , Datta, S. , Maksymets, O. , Das, A. , Gkioxari, G. , & Lee, S. , et al. (2019). Embodied question answering in photorealistic environments with point cloud perception (CVPR).

Smith, L. , & Gasser, M. . (2014). The development of embodied cognition: six lessons from babies. Artificial Life, 11(1-2), 13-29.

Colas, A. , Kim, S. , Dernoncourt, F. , Gupte, S. , & Kim, D. S. . (2019). Tutorialvqa: question answering dataset for tutorial videos. In LREC 2020.

Petroni, F. , Rocktschel, T. , Lewis, P. , Bakhtin, A., Wu, Y. ,& Miller, A. H. , et al. (2019). Language Models as Knowledge Bases?. In EMNLP 2019.

Liu, Y., Wang, S., Zhang, J., Zong, C. (2019). Experience-based causality learning for intelligent agents. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 18(4), 1-22.

Sydorova, A. , N Poerner, & Roth, B. . (2019). Interpretable question answering on knowledge bases and text. In ACL 2019.

Tran, K. Q., Nguyen, A. T., Le, A. T. H., Van Nguyen, K. (2021). ViVQA: Vietnamese Visual Question Answering. In Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation (pp. 546-554).