Interactive Grounded Language Understanding in a Collaborative Environment: Retrospective on Iglu 2022 Competition

Julia Kiseleva[1] JULIA.KISELEVA@MICROSOFT.COM
Alexey Skrynnik[2] SKRYNNIKALEXEY@GMAIL.COM
Artem Zholus[3] ARTEM.ZHOLUS@GMAIL.COM
Shrestha Mohanty[1] SHRESTHA.MOHANTY@MICROSOFT.COM
Negar Arabzadeh[4] NARABZAD@UWATERLOO.COM
Marc-Alexandre Côté[1] MACOTE@MICROSOFT.COM
Mohammad Aliannejadi[5] M.ALIANNEJADI@UVA.NL
Milagro Teruel[1] MTERUEL@MICROSOFT.COM
Ziming Li[6] CSZIMINGLI@GMAIL.COM
Mikhail Burtsev[2,3] BURTSEV.M@GMAIL.COM
Maartje ter Hoeve[5] M.A.TERHOEVE@UVA.NL
Zoya Volovikova[3] VOLOVIKOVA.ZA@PHYSTECH.EDU
Aleksandr Panov[2,3] PANOV.AI@MIPT.RU
Yuxuan Sun[7] 53YUXUANS@FB.COM
Kavya Srinet[7] KSRINET@FB.COM
Arthur Szlam[8] ASZLAM@FB.COM
Ahmed Awadallah[1] HASSANAM@MICROSOFT.COM

Seungeun Rho[9] SEUNGEUN.RHO@KAKAOBRAIN.COM
Taehwan Kwon[9] TAEHWAN.KWON@KAKAOBRAIN.COM
Daniel Wontae Nam[9] DWTNAM@KAKAOBRAIN.COM

Felipe Bivort Haiek[10] FELIPEBIHAIEK@GMAIL.COM

Edwin Zhang[11] ETE@CS.UCSB.EDU
Linar Abdrazakov[3] ABDRAZAKOV.LR@PHYSTECH.EDU
Matthew Ho[11] MSHO@UCSB.EDU

Guo Qingyam[12] GQY22@MAILS.TSINGHUA.EDU.CN
Jason Zhang[12] ZHANGBEICHEN819@163.COM
Zhibin Guo[13] ZEBGOU@GMAIL.COM

[1]*Microsoft,* [2]*AIRI,* [3]*MIPT,* [4]*University of Waterloo,* [5]*University of Amsterdam,* [6]*Amazon Alexa,* [7]*Meta AI,* [8]*DeepMind,* [9]*Kakao Brain,* [10]*Universidad de Buenos Aires,* [11]*University of California, Santa Barbara,* [12]*Tsinghua University,* [13]*Renmin University of China*

Editors: Marco Ciccone, Gustavo Stolovitzky, Jacob Albrecht

Abstract

Human intelligence possesses the extraordinary ability to adapt rapidly to new tasks and multi-modal environments. This capacity emerges at an early age, as humans acquire new skills and learn to solve problems by imitating others or following natural language instructions. To facilitate research in this area, we recently hosted the second *IGLU: Interactive Grounded Language Understanding in a Collaborative Environment* competition. The primary objective of the competition is to address the challenge of creating interactive agents that can learn to solve complex tasks by receiving grounded natural language instructions in a collaborative environment. Given the complexity of this challenge, we divided it into two sub-tasks: first, deciding whether the provided grounded instruction requires clarification, and second, following a clear grounded instruction to complete the task description.

Keywords

Natural Language Understanding (NLU), Reinforcement Learning (RL), Grounded Learning, Interactive Learning, Embodied RL

1. Introduction

Humans possess an extraordinary ability to quickly adapt to new tasks and environments. From a young age, humans can acquire new skills and learn to solve new tasks either by imitating the behavior of others or by following natural language instructions provided to them (An, 1988; Council, 1999). Studies in developmental psychology have shown that natural language communication is an effective method for transmitting generic knowledge between individuals as young as infants Csibra and Gergely (2009). This form of learning can even accelerate the acquisition of new skills by avoiding trial-and-error when learning only from observations (Thomaz et al., 2019).

Inspired by these findings, the AI research community is attempting to develop grounded interactive *embodied agents* capable of engaging in natural back-and-forth dialogue with humans to assist them in completing real-world tasks (Winograd, 1971; Narayan-Chen et al., 2017; Levinson, 2019; Chen et al., 2020; Abramson et al., 2020). Importantly, the agent must understand when to initiate feedback requests if communication fails or instructions are unclear and must learn new domain-specific vocabulary (Aliannejadi et al., 2020, 2021; Rao and Daumé III, 2018; Narayan-Chen et al., 2019; Jayannavar et al., 2020; Arabzadeh et al., 2022). Despite significant efforts Wu et al. (2022); Kiseleva et al. (2022), the task is far from being solved. Therefore, we propose the second Interactive Grounded Language Understanding (IGLU) in a collaborative environment competition.

Specifically, the goal of our competition is to approach the following scientific challenge: *How to build interactive embodied agents that learn to solve a task while provided with grounded natural language instructions in a collaborative environment?* By *'interactive agent'* we mean that the agent can: (1) follow the instructions correctly, (2) ask for clarification when needed, and (3) quickly adapt newly acquired skills. The IGLU challenge is naturally related to two fields of study that are highly relevant to the NeurIPS community: Natural Language Understanding and Generation (NLU / NLG) and Reinforcement Learning (RL).

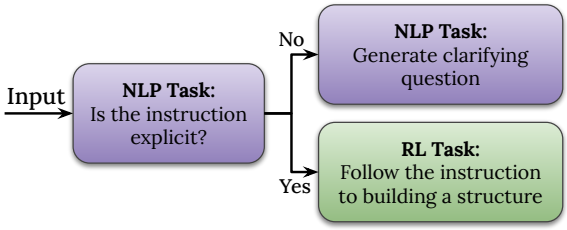

Figure 1: IGLU's general overview.

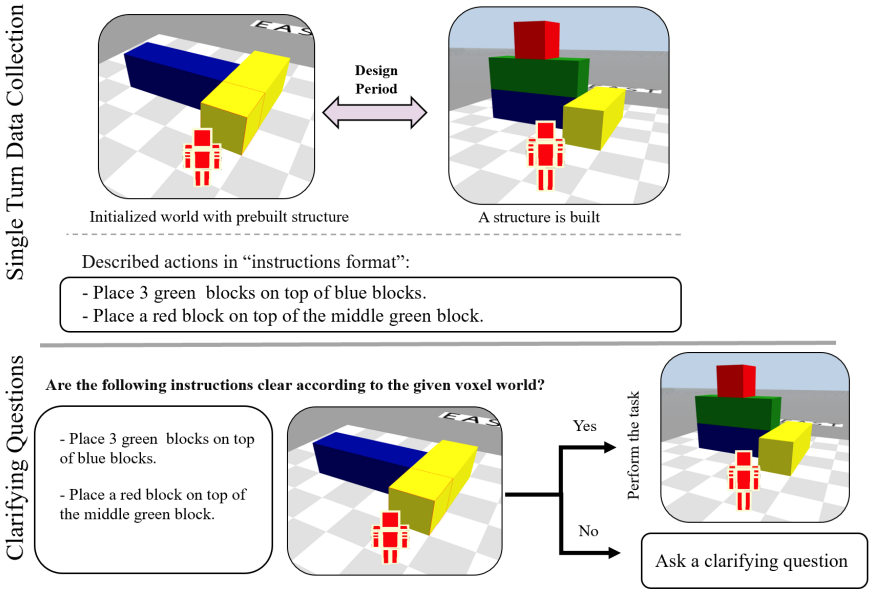

Figure 2: Overview of Data Collection

The challenge of developing an *'interactive agent'* naturally splits into two main tasks (Fig. 1):

- NLP-related: to decide if the provided grounded instruction needs to be clarified and generate clarifying questions if it's a case;
- RL-related: to follow a clear, grounded instruction to complete the described task.

Our vision suggests that the successful solutions to presented tasks can be combined into the end-to-end pipeline to develop desired interactive agents.

2. Data Collection

We leveraged and extended the previously collected multi-turn interactions dataset used in IGLU 2021 Kiseleva et al. (2022). We simplified the multi-turn dialogues interactions to single-turn interactions by removing the complexity of building a target structure. We instead have an annotator perform actions and provide instructions to another annotator. We also leverage the multi-turn interactions data to provide a starting point from which annotators can build. To elaborate, we design the following setup, as shown in Fig. 2, for collecting data:

- An interactive annotator or builder is dropped in the middle of a world where the structure is built partially. The partially completed world is retrieved from the multi-turn interactions dataset.
- The annotator is prompted to perform a sequence of building actions for a duration of one minute.
- The annotator then describes their performed set of actions in natural language, which will be displayed to another annotator as an instruction.
- The next annotator is shown the instruction and is asked to perform the steps mentioned in the instruction. If the instruction is not clear, they specify it as thus and ask clarification questions.

The data collection was performed in MTurk[1] where we integrate the extended CraftAssisft library Gray et al. (2019). This setup enables us to scale for participants quickly and collect a dataset consisting of natural language instructions, grid world states, actions performed based on those instructions, and a set of clarifying questions. More details on the tool, collected dataset, and its application are in Mohanty et al. (2022).

3. NLP Task: Asking Clarifying Questions

Inspired by Aliannejadi et al. (2021); Dalton et al. (2020), we split the problem into the following research questions:

RQ1 When to ask clarifying questions?
Given the instruction from the Architect, a model needs to predict whether that instruction is sufficient to complete the described task or whether further clarification is needed. Simply put, here it is decided if we need to activate the Builder.

RQ2 What clarifying question to ask?
If the given instruction from the Architect is ambiguous, a clarifying question should be raised. In this research question, we are specifically interested in "what to ask" to clarify the given instruction. The original instruction and its clarification can be used as input for the Builder.

As a starting **baseline**, we use the implementation provided in Aliannejadi et al. (2021).

3.1. Solutions

To encourage participants to focus on both tasks: **When to ask** and **What to ask**, we divided the classification task (whether to ask clarifying questions) into buckets based on their F_1 scores: $[0-0.35, 0.35-0.5, 0.50-0.65, 0.65-0.75, 0.75-0.85, 0.85-0.90, 0.9-1]$. We then grouped the submissions into the corresponding ranges and evaluated the second task for each bucket. For example, if a classifier achieved an F1 score of 0.82, its binned F1 score would be 0.75. Another classifier with F1 score of 0.76 would also belong to the same bucket and the two classifiers would compete in the second task.

We present the performance of the best-performing submissions from all participating teams separately for both tasks in Fig. 3. The top-left plot in Fig. 3 shows the first task's performance in terms of F1 score for **when to ask** clarifying questions, with the baseline run highlighted in green. Similarly, we report performance of **what to ask** task in the top-right plot of Fig. 3 in terms of MRR@20. It is worth noting that while 19 teams were able to improve their performance on the first task, only 10 teams were able to make improvements on **what to ask** task. This indicates that it has higher difficulty.

Table 1: Results of the winners of NLP task

Team	**When:** F_1	**When:** MRR@20	# Submissions
craftsmanfly	0.751	0.596	74
try1try	0.766	0.592	39
FelipeB	0.761	0.550	54
Baseline	0.732	0.341	

Additionally, we highlight the results of the NLP task winners and compare them with the initial baseline in Tab. 1. More details about the baseline can be found in Mohanty et al. (2022). The first team, craftsmanfly, was able to obtain $F_1 = 0.751$ in the first task and MRR@20 of 0.596 in the second task.

1. https://www.mturk.com/

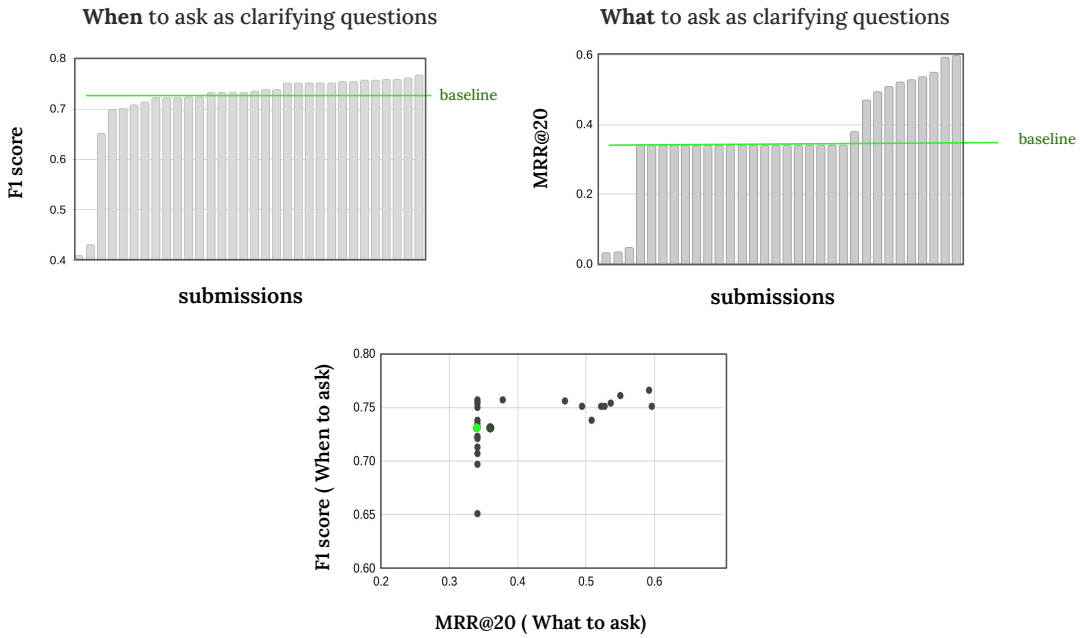

Figure 3: Distribution of achieved F1 scores for the when to ask as the classification task and MRR@20 for what to ask as the ranking task across the participants

3.2. NLP task: winning solutions

Next, we briefly explain the highlights of the solutions from the top-3 teams in NLP task.

- **Incorporating voxel-world state**: The winning solutions mostly relied on LLM-based question representations as the backbone. However, they additionally incorporated the voxel-world states as part of the input for the classifier. Here, the novelty of different teams lies in how they leveraged the information about the voxel world. For instance, some solutions encoded the state information, such as the colors and numbers of initialized blocks, in natural language format and concatenated it with the instruction, as illustrated in Fig. 3. Moreover, they took different sampling strategies to avoid redundancies and provide more balance in the training set. While converting the voxel-world information into natural language was commonly used between participants, a few teams take the representation of the world and create a 3D grid from it and pass the grid through CNN.

- **Data augmentation:** Data augmentation was popular among the top-performed solutions. Mostly, to balance the data distribution, Easy Data Augmentation (EDA) Wei and Zou (2019), synonym replacement, color replacement, and instruction segmentation were widely used.

- **Domain-Adaptive Fine-tuning:** Inspired by Gururangan et al. (2020), some teams perform domain-adaptive fine-tuning on datasets related to IGLU (e.g.,Narayan-Chen et al. (2019); Jayannavar et al. (2020) and IGLU 2022 Multi-turn Zholus et al. (2022))

- **Fast Gradient:** Due to the limited amount of training data, participants took different approaches to alleviate the overfitting problem. Among those, Fast Gradient Method (FGM) Goodfellow et al. (2014) which is a widely used regularization method in computer vision and NLP Miyato et al. (2016); Chen et al. (2019) was commonly used and showed superior performance.

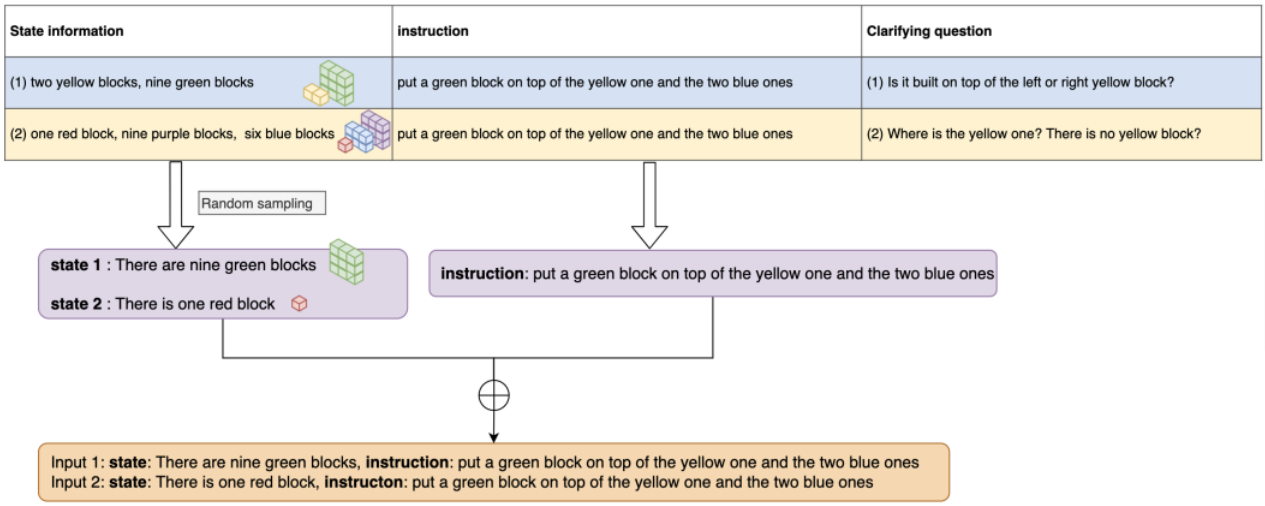

Figure 4: Example of encoding the voxel-world initial state.

4. RL Task: Building Structures

This study concerns an RL agent's ability to construct a target structure solely based on natural language instructions without any visual cues. The agent navigates and places colored blocks within a predetermined building area from a first-person perspective. The task involves two components: the context utterances, which specify the blocks previously placed, and the target utterances, which describe the remaining blocks to be placed. The Architect and the Builder's conversation provides these instructions. The RL agent receives a score at the end of each episode that reflects the degree of completeness of the constructed structure compared to the ground truth target structure.

4.1. IGLU GridWorld Environment

In our experience with IGLU 2021, we discovered the complexity of the builder task. The ideal builder must possess knowledge of building any structure. Given the vast state-space of possible block combinations, achieving this goal is challenging. To address this difficulty, we focused on optimizing the environment and implementing a more advanced baseline. As demonstrated by this year's results, these two directions were necessary to successfully solve the builder task.

For this year's builder task, we recreated the RL environment[2]. The agent's goal is to complete a task expressed as an instruction written in natural language. The environment was implemented in pure Python, using a simplified version of an open-source Minecraft engine[3]. In the new

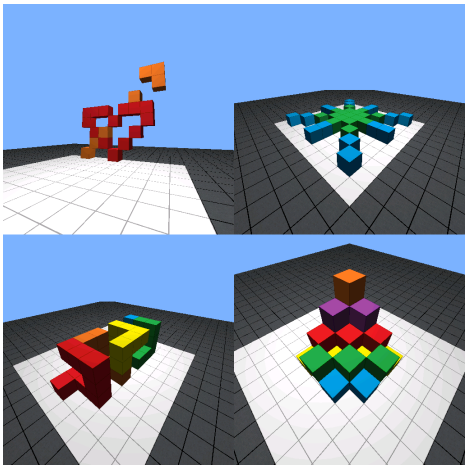

Figure 5: Examples of target structures rendered by the GridWorld environment

2. https://github.com/iglu-contest/gridworld
3. https://github.com/fogleman/Minecraft

version, the renderer is decoupled from the core environment, allowing it to run headless with GPU acceleration, that makes it fast and scalable for highly-loaded RL experiments.

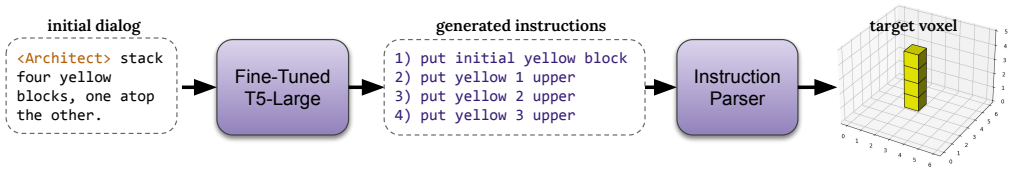

Figure 6: The Language module simultaneously predicts block coordinates and types using fine-tuned T5-large encoder-decoder transformer. The model relies solely on a dialogue between an Architect and a Builder as input.

The observation space consists of a point of view image of dimensions $(64, 64, 3)$, inventory item counts of size $(6,)$, a snapshot of the building zone with dimensions $(11, 11, 9)$, and the agent's position with pitch and yaw angles of size $(5,)$. The building zone is represented as a 3D tensor with block ids, where each block has a unique identification number (e.g., 0 for air, 1 for blue block, etc.). The agent can navigate over the building zone, place and destroy blocks, and switch between block types. Furthermore, a detailed description of the environment can be found in a separate paper Zholus et al. (2022).

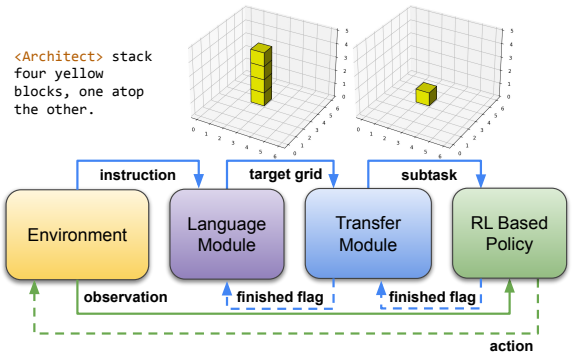

Figure 7: The general overview of the Language to Subtask Builder approach. The Transfer Module converts the voxel representation generated by the Language module into a sequence of simple subtasks that involve adding or removing one block at a time. The Transfer module heuristically orders the subtasks from bottom to top and left to right. The RL Based Policy Module, solves the tasks of navigating the agent and placing a block. The policy was trained on building random structures using the PPO approach.

4.2. Baseline

We propose an approach for training a general-purpose builder[4] that can solve structures not encountered during the training phase. Our *Language to Subtask Builder* agent converts natural language descriptions of a target structure into a grid representation that defines subtasks. These subtasks are then sequentially completed by the RL policy (Fig. 7). A detailed description of this solution is provided in Skrynnik et al. (2022).

The builder agent consists of three modules: (1) the Language module, which predicts the coordinates and types of blocks given text input (an example of which is shown in Fig.6), (2) the rule-based Transfer module, which iterates over the positions of the predicted blocks in a heuristically defined order, providing subtasks for (3) the RL-based Policy module, which solves the atomic subtasks of block placement or removal. The RL module operates using visual input, inventory, and compass observations provided by the environment, and a target block position provided by the Transfer module. The entire scheme of the presented approach is illustrated in Fig.7.

4. https://github.com/iglu-contest/iglu-2022-rl-baseline

4.3. Evaluation

To assess the RL agent's performance, we conduct numerous environmental episodes for each subtask of the hold-out part of the IGLU dataset. Each subtask initiates the world with a particular starting and target grid. The F_1 score is the primary evaluation metric, where the blocks added or removed serve as the ground truth, and the prediction is the difference between the initial world and the snapshot of the building zone at the end of the episode, represented as a 3D tensor. The episode ends either when the structure is complete, or when the time limit is reached. We permit the agent to choose when to end the episode as a separate action. For each task in the evaluation set, we perform several episodes and determine the weighted average of the task's F_1 scores, with the weights equal to the total number of blocks to add or remove.

4.4. Winning Solutions

Table 2: Results of the winners of RL task.

Team	F_1	Prec	Recall	Ep. Length	# of Submissions
Happy Iglu	0.254	0.331	0.264	391	89
FelipeB	0.178	0.335	0.153	283	18
Chuang	0.156	0.303	0.138	294	31
Baseline	0.150	0.256	0.134	281	

Tab. 2 presents the results of the winners of the RL task and compares them with the proposed baseline. The Happy Iglu team won by a significant margin, offering a multimodal end-to-end solution. FelipeB and the Chuang team improved the NLP part of the baseline to arrive at their solutions. A more comprehensive overview of the solutions is provided below.

4.4.1. FIRST PLACE: HAPPY IGLU TEAM

They developed an end-to-end RL approach to tackle the challenges in the IGLU environment. The approach outlines four directions to deal with challenges. Firstly, a reward function has been designed to incorporate task-specific rewards and penalties, as well as the F_1 score and "grid" reconstruction loss, which encourage better utilization of state information. Secondly, representation learning techniques have been employed to distill task-relevant information from high-dimensional input observations, including voxels of "grid" and "target_grid" input for the value function and hand-crafted features such as compass and color count. Thirdly, the team has addressed partial observability by processing the trajectory of past observations history using the TRXL transformer architecture. Lastly, COCOLM-LARGE has been used to create an embed vector for each instruction. These techniques have been successfully combined, resulting in high performance scores in the IGLU environment. The model (Fig. 8) was trained using BRAIN AGENT[5] distributed RL framework.

4.4.2. SECOND PLACE: FELIPEB

The solution to improve the NLP part of the baseline involved using a different model as the T5 model underperformed with a low BLEU score. To address this issue, a model with summarization pretraining such as PEGASUS Zhang et al. (2020) was used to translate utterances from *architect* into commands. The PEGASUS-LARGE model was trained using the same augmentations as the original

5. https://github.com/kakaobrain/brain-agent

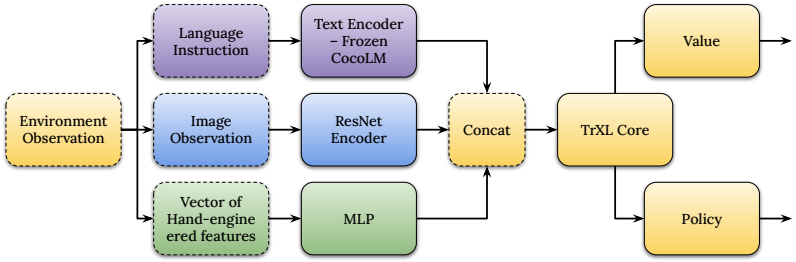

Figure 8: The network architecture of the approach submitted by the Happy Iglu team.

baseline. With the correct selection of hyperparameters and model replacement, the BLEU score improved from 0.3 to 0.95, and the F_1 score of the whole pipeline improved from 0.15 to 0.178.

During training, the building episodes were split to ensure that no event, regardless of color permutations, appeared in both the training and validation sets. However, the model had inconsistent BLEU score improvements. Even though augmented examples from the train set were included in the validation set, the model that resulted from the improper separation was the best performing in the online test. Finally, the best combination of training and inference history was achieved by training with a history of 10 *architect* utterances and inferring with all the possible available history.

4.4.3. THIRD PLACE: TEAM CHUANG

This solution utilized a provided baseline and also focused on enhancing the NLP component. To achieve this, the team reformulated the problem of generating the target grid as a text-to-video task, where the temporal dimension in the video corresponds to the third dimension in the voxel grid. The team employed an open-source implementation of video diffusion models and modified the model using context prompting. This involved adding the initial starting grid as a prefix to each language instruction to facilitate contextual generation. This encoding strategy enabled the diffusion model to effectively remove noise and generate the target delta grid.

Using contextual 3D diffusion models for the IGLU task yielded several benefits, including improved performance over the T5 model on local evaluation. However, the results were not as impressive when tested on a hold-out dataset, possibly due to difficulties in reconstructing the initial grid. Nevertheless, this approach enabled the generation of all blocks simultaneously, allowing for the capture of long-range dependencies and the enforcement of global consistency.

4.5. Special Awards

We are awarding two teams with research prizes for their outstanding contributions to solving the builder task. The first prize is awarded to the HAPPY IGLU team for developing an end-to-end RL solution that utilizes the latest improvements and speed-ups introduced environment. Their model effectively extracts task-relevant information from high-dimensional RGB inputs by employing an auxiliary loss. The second prize is awarded to the team CHUANG for their innovative approach to rethinking the target generation module of the provided baseline. They introduced a diffusion model that incorporates classified free language guidance, and their agent utilizes an exploration policy to restore the environment's state.

5. Participation Statistics

A total of 54 teams, comprising 398 participants, participated in the competition. The RL Task had 15 active teams that made 382 submissions, while the NLP Task attracted even more attention and received 696 submissions. Our findings indicate that the competition was successful in engaging a diverse group of participants and promoting interest in the fields of RL and NLP.

We attribute the increase in the number of participants primarily to the work carried out to simplify the entry threshold to the competition. The RL environment was updated, allowing the use of modern distributed large-scale approaches. A baseline for the RL Task was developed to attract new participants who were less experienced in RL but could improve its other components (e.g., its NLP part). Furthermore, the NLP task was redesigned to be focused on clarifying questions to forsee an idea of designing an interactive agent.

6. Lessons learned

This year, we enforced a competition rule[6] that prohibited the use of any external fields or special properties of the environment during the evaluation. However, any information, including the agent's position and the grid state, was allowed during training machine learning models. Unfortunately, two RL task submissions were disqualified as they used a special feature of the simulator. Instead of utilizing the RL component of the provided baseline, these submissions employed a hand-crafted policy. The policy placed the agent in one of the corners of the environment, taking advantage of the limited area available for the agent's location. Afterward, it built blocks by performing programmed deterministic actions, disregarding the environment's observations.

7. Conclusion

We organized the 2022 IGLU competition with the aim of promoting the development of interactive embodied agents capable of learning to solve complex tasks by receiving instructions in natural language. The task of making agents interactive in a meaningful way has even more importance now since the recent advance in LLM. The participants in this year's competition demonstrated impressive results and significantly enhanced the agents' ability to understand and perform interactive grounded tasks. To achieve this, we reformulated the NLP task to include the generation of clarifying questions, expanded the builder dataset with additional data, created a fast RL environment, and established a strong baseline for comparison. This paper summarizes the participants' submissions and analyzes their performance.

Acknowledgments

We thank Microsoft for providing computational and financial support, and Amazon and Google for providing financial support. We would like to thank our advisory board: Tim Rocktäschel, Julia Hockenmaier, Bill Dolan, Ryen W. White, Maarten de Rijke and Oleg Rokhlenko.

6. The full competition rules is available at https://www.aicrowd.com/challenges/neurips-2022-iglu-challenge/challenge_rules

References

Josh Abramson, Arun Ahuja, Iain Barr, Arthur Brussee, Federico Carnevale, Mary Cassin, Rachita Chhaparia, Stephen Clark, Bogdan Damoc, Andrew Dudzik, et al. Imitating interactive intelligence. *arXiv preprint arXiv:2012.05672*, 2020.

Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. Convai3: Generating clarifying questions for open-domain dialogue systems (clariq). 2020.

Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeffrey Dalton, and Mikhail Burtsev. Building and evaluating open-domain dialogue corpora with clarifying questions. *arXiv preprint arXiv:2109.05794*, 2021.

Meltzoff An. Imitation, Objects, Tools, and the Rudiments of Language in Human Ontogeny, February 1988. URL https://pubmed.ncbi.nlm.nih.gov/23997403/. ISSN: 0393-9375 Issue: 1-2 Publisher: Hum Evol Volume: 3.

Negar Arabzadeh, Ali Ahmadvand, Julia Kiseleva, Yang Liu, Ahmed Hassan Awadallah, Ming Zhong, and Milad Shokouhi. Preme: Preference-based meeting exploration through an interactive questionnaire. *arXiv preprint arXiv:2205.02370*, 2022.

Valerie Chen, Abhinav Gupta, and Kenneth Marino. Ask your humans: Using human instructions to improve generalization in reinforcement learning. *arXiv preprint arXiv:2011.00517*, 2020.

Zhu Chen, Cheng Yu, Gan Zhe, Sun Siqi, Goldstein Thomas, and Liu Jingjing. Freelb: Enhanced adversarial training for language understanding. *arXiv preprint arXiv: 1909.11764*, 2019.

National Research Council. *How People Learn: Brain, Mind, Experience, and School: Expanded Edition*. August 1999. ISBN 978-0-309-07036-2. doi: 10.17226/9853. URL https://www.nap.edu/catalog/9853/how-people-learn-brain-mind-experience-and-school-expanded-edition.

Gergely Csibra and György Gergely. Natural pedagogy. *Trends in cognitive sciences*, 13(4):148–153, 2009.

Jeff Dalton, Aleksandr Chuklin, Julia Kiseleva, and Mikhail Burtsev. Proceedings of the 5th international workshop on search-oriented conversational ai (scai). In *Proceedings of the 5th International Workshop on Search-Oriented Conversational AI (SCAI)*, 2020.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Jonathan Gray, Kavya Srinet, Yacine Jernite, Haonan Yu, Zhuoyuan Chen, Demi Guo, Siddharth Goyal, C. Lawrence Zitnick, and Arthur Szlam. CraftAssist: A Framework for Dialogue-enabled Interactive Agents. *arXiv:1907.08584 [cs]*, July 2019. URL http://arxiv.org/abs/1907.08584. arXiv: 1907.08584.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*, 2020.

Prashant Jayannavar, Anjali Narayan-Chen, and Julia Hockenmaier. Learning to execute instructions in a minecraft dialogue. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 2589–2602, 2020.

Julia Kiseleva, Ziming Li, Mohammad Aliannejadi, Shrestha Mohanty, Maartje ter Hoeve, Mikhail Burtsev, Alexey Skrynnik, Artem Zholus, Aleksandr Panov, Kavya Srinet, Arthur Szlam, Yuxuan Sun, Katja Hofmann, Marc-Alexandre Côté, Ahmed Awadallah, Linar Abdrazakov, Igor Churin, Putra Manggala, Kata Naszadi, Michiel van der Meer, and Taewoon Kim. Interactive grounded language understanding in a collaborative environment: Iglu 2021. In *NeurIPS 2021 Competitions and Demonstrations Track*, pages 146–161. PMLR, 2022.

Stephen C Levinson. Tom m. mitchell, simon garrod, john e. laird, stephen c. levinson, and kenneth r. koedinger. *Interactive Task Learning: Humans, Robots, and Agents Acquiring New Tasks through Natural Interactions*, 26:9, 2019.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*, 2016.

Shrestha Mohanty, Negar Arabzadeh, Milagro Teruel, Yuxuan Sun, Artem Zholus, Alexey Skrynnik, Mikhail Burtsev, Kavya Srinet, Aleksandr Panov, Arthur Szlam, Marc-Alexandre Côté, and Julia Kiseleva. Collecting interactive multi-modal datasets for grounded language understanding. *arXiv preprint arXiv:2211.06552*, 2022.

Anjali Narayan-Chen, Colin Graber, Mayukh Das, Md Rakibul Islam, Soham Dan, Sriraam Natarajan, Janardhan Rao Doppa, Julia Hockenmaier, Martha Palmer, and Dan Roth. Towards problem solving agents that communicate and learn. In *Proceedings of the First Workshop on Language Grounding for Robotics*, pages 95–103, 2017.

Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. Collaborative dialogue in minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5405–5415, 2019.

Sudha Rao and Hal Daumé III. Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. *arXiv preprint arXiv:1805.04655*, 2018.

Alexey Skrynnik, Zoya Volovikova, Marc-Alexandre Côté, Anton Voronov, Artem Zholus, Negar Arabzadeh, Shrestha Mohanty, Milagro Teruel, Ahmed Awadallah, Aleksandr Panov, Mikhail Burtsev, and Julia Kiseleva. Learning to solve voxel building embodied tasks from pixels and natural language instructions. *arXiv preprint arXiv:2211.00688*, 2022.

Andrea L Thomaz, Elena Lieven, Maya Cakmak, Joyce Y Chai, Simon Garrod, Wayne D Gray, Stephen C Levinson, Ana Paiva, and Nele Russwinkel. Interaction for task instruction and learning. In *Interactive task learning: Humans, robots, and agents acquiring new tasks through natural interactions*, pages 91–110. MIT Press, 2019.

Jason W. Wei and Kai Zou. EDA: easy data augmentation techniques for boosting performance on text classification tasks. *CoRR*, abs/1901.11196, 2019. URL http://arxiv.org/abs/1901.11196.

Terry Winograd. Procedures as a representation for data in a computer program for understanding natural language. Technical report, MASSACHUSETTS INST OF TECH CAMBRIDGE PROJECT MAC, 1971.

Anne Wu, Kianté Brantley, Noriyuki Kojima, and Yoav Artzi. lilgym: Natural language visual reasoning with reinforcement learning. *arXiv preprint arXiv:2211.01994*, 2022.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR, 2020.

Artem Zholus, Alexey Skrynnik, Shrestha Mohanty, Zoya Volovikova, Julia Kiseleva, Artur Szlam, Marc-Alexandre Coté, and Aleksandr I Panov. Iglu gridworld: Simple and fast environment for embodied dialog agents. *arXiv preprint arXiv:2206.00142*, 2022.