Language-Conditioned Reinforcement Learning to Solve Misunderstandings with Action Corrections

Anonymous Author(s)
Affiliation
Address
email

Abstract

Human-to-human conversation is not just talking and listening. It is an incremental process where participants continually establish a common understanding to rule out misunderstandings. Current language understanding methods for intelligent robots do not consider this. There exist numerous approaches considering non-understandings, but they ignore the incremental process of resolving misunderstandings. In this article, we present a first formalization and experimental validation of incremental action-repair for robotic instruction-following based on reinforcement learning. To evaluate our approach, we propose a collection of benchmark environments for action correction in language-conditioned reinforcement learning, utilizing a synthetic instructor to generate language goals and their corresponding corrections. We show that a reinforcement learning agent can successfully learn to understand incremental corrections of misunderstood instructions.

1 Introduction

In real-world human-to-human communications, misunderstandings happen very frequently. Ambiguity, a lack of common ground, and incorrect statements are three examples of situations that cause misunderstandings. For example, you may ask your colleague to "give me the item on the right", referring to the object on *her* right, while your conversation partner assumes you refer to *your* right. Humans are extremely proficient in resolving such ambiguities by performing conversational repair, e.g., "No, not to my right, I mean your right!". Such repair commands enable conversation partners to incrementally establish a common understanding of what was said.

The state-of-the-art in robotic language understanding considers non-understandings [Bohus and Rudnicky, 2008, Bordes et al., 2017], but it entirely ignores the incremental resolution of misunderstandings with conversational repair.

The concept of action correction is not only important when the system is being deployed into the real world but should be, as we propose in this article, a fundamental part of the learning procedure. However, even flagship approaches like SayCan struggle with ambiguities and negations [Ahn et al., 2022], despite using large language models with a rich semantic knowledge of the world.

Recent advances in controlling robots with language achieved remarkable results in simulation [Lynch and Sermanet, 2021, Huang et al., 2022] and also made considerable progress in real-world applications [Shridhar et al., 2021]. Especially reinforcement learning (RL) turns out to be a suitable framework for training robots paired with large language models (LLM) [Ahn et al., 2022]. But it remains unclear how we can build an intelligent robot capable of resolving ambiguities, as we depict in Figure 1.

Submitted to the 2nd Workshop on Language in Reinforcement Learning, (NeurIPS 2022). Do not distribute.

In this article, we propose an extension of goal-conditioned RL, where goals are specified as word embeddings that can be dynamically extended with repair commands while a robot is executing a certain task. We hypothesize that our approach enables robotic agents to react appropriately to the following three yet unexplored cases of misunderstandings:

Ambiguity Instructions are ambiguous if they are underspecified. As an example, consider Figure 1. Here, the instruction *"grasp the cube"* is imprecise as there are two cubes and the instruction is referring to a categorical feature of the present objects that are identical for two entities.

Lack of common ground When operator and robot have a lack of common ground, the robot misunderstands the instruction due to unknown words used (out-of-distribution) or insufficient training. For example, if the robot's object detection method misclassifies a red apple for a tomato, the operator needs to use action corrections to interrupt the robot's interactions with the apple and guide it to the tomato.

Instruction Correction It can happen that the error is completely on the operator's side, e.g., when she/he accidentally utters an unintended instruction. For example, the operator might say *"reach the red cube"*, while having intended to instruct the robot to reach the green cube. This requires an action correction like "No, sorry, I mean the green cube".

In the following, we first highlight that the state of the art in robotic language understanding lacks methods to address such misunderstandings. Then we briefly describe our method, and, finally, we present our environment extension. We conclude the article with our experiments and results, showing how our method successfully fills this gap.

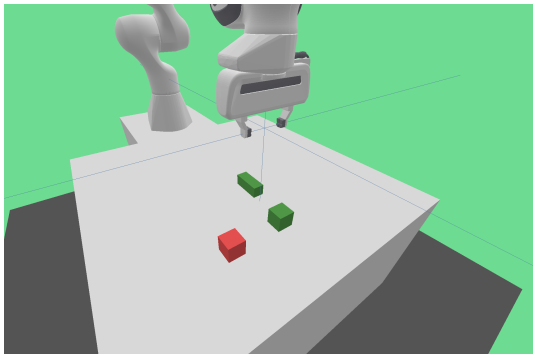

Figure 1: Consider the present scene with a green cube, a green cuboid, and a red cube. An instructor assigns the robot to *"grasp the cube"*. The robot might grasp an object by chance (let us assume the red cube). Now an action correction like *"actually, the green"* could resolve the robot's mistake caused by the ambiguity of the instruction. The robot should be still aware of the *"grasping"* task and initial shape word *"cube"*, to ideally ignore the green cuboid and approach the green cube.

2 Related Work

Natural language is the most important tool of communication, therefore researchers investigate its usage for robots for more than a decade [Steels, 2008] and it is still actively researched [Tellex et al., 2020]. Specifically grounded language is a key ingredient for true understanding because it is not always possible to transmit information without the background knowledge and considerations of other modalities [Bisk et al., 2020]. The notion of controlling intelligent agents with language has also found its way into deep RL research [Hill et al., 2019, Akakzia et al., 2021, Chevalier-Boisvert et al., 2019]. The review of Luketina et al. [2019] outlines the usage of language in RL where it is applied to assist or instruct an agent to solve a task. RL is an embodied setup and therefore provides the grounding of language through rewards [Hermann et al., 2017, Akakzia et al., 2021, Röder et al., 2022]. Furthermore, its compositional representation allows decomposing a task, e.g., *"remove the plate and clean the table"*, into two subtasks, namely *"remove plate"* and *"clean table"*. We assume that action corrections are of great benefit if such a compositional representation is utilized by the agent because a misunderstanding could originate in misinterpretation of one single word.

3 Method

We outline the general notion of action correction as dynamic goal extension, where we augment the episodic goal by the action correction. For a general example, consider the illustration in Figure 2. Here, the agent interacts with the world under the consideration of the instruction g_ℓ. The instructor is observing the environment state for interactions with wrong objects, on which an action correction is returned. The following goal is now a concatenation of the original goal and the action correction, $g_\ell \circ g_{ac}$.

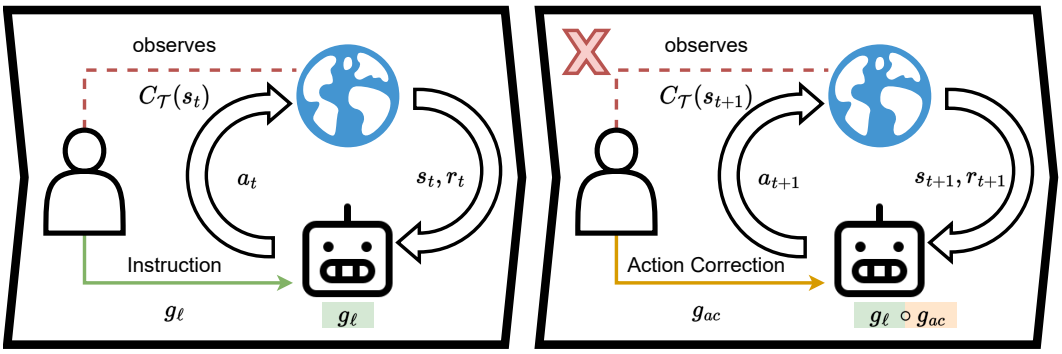

Figure 2: We illustrate the general case of an action correction. The episodic conversation starts with an instruction g_ℓ uttered by the instructor (left box). The agent, being in state s_{t-1} executes an action a_t to receive a scalar reward r_t and the next world state s_t. However, the instructor detects with the help of the task-specific condition function from Equation 1 a wrong behavior. Following, an action correction g_{ac} is returned, which is concatenated to the original goal $g_\ell \circ g_{ac}$ (right box). Under the consideration of the new linguistic goal, the agent attempts to correct its behavior.

Formally, the reward function R_g is determined by at least one or more task-specific conditions $C_\mathcal{T}$, that are evaluated at each step of the simulation

$$r_t = R_g(s_{t+1}, g_\ell) = \begin{cases} 0, & \text{if } C_\mathcal{T}(s_{t+1}, g_\ell) \\ -1, & \text{otherwise} \end{cases} \tag{1}$$

. The synthetic instructor uses the same function $C_\mathcal{T}$ from Equation 1 to also detect interactions with the non-goal objects. In case of a misunderstanding, the instructor returns an action correction, while having the full knowledge of the goal object and non-goal object's position and properties in the scene. Figure 1 illustrates an example of an ambiguity where the robot interacts with the wrong object because of a spatial setup (grasping the closer object) or by chance. Besides, the correction in this example is referring to a unique property of the actual goal object and is not making use of negations like *"not the red"*. We are aware of a recent experiment with negations in language-conditioned RL [Hill et al., 2020], but the authors are not utilizing them for action correction. They rather instruct an agent to go to an object that has not a specific property. However, we assume this to be learned in our scenarios when the corrections involve negations. In the following section, we give a brief description of our environments for action correction.

3.1 Learning Environment for Action Correction

We present an extension of the publicly available language-conditioned RL environment LANRO, published by Röder et al. [2022]. It provides a selection of 4 tasks, such as *reach*, *push*, *grasp*, and *lift*. An episode consists of two or three objects on a table (c.f. Figure 1) and a linguistic goal generated by a synthetic instructor. We provide the formal definition of the templates to generate the instructions in Subsection A.1. Like Röder et al. [2022], we make use of a condition-based reward function corresponding to the language goal (see Equation 1). We use the task-specific conditions to determine the moment we return an action correction for interacting with a wrong object (see Figure 2). From this point on, the original goal g_ℓ is extended by the action correction g_{ac}. In the end, we receive a concatenation of strings, hence $g_\ell \circ g_{ac}$. To bootstrap the learning of simple instruction-following, we challenge the agent to solve action corrections in only 50% of the episodes. Since we actually delegate the agent to fulfill the task of two different linguistic goals, the episode limit is increased to 100 steps.

4 Experiments

For our initial experiments, we make use of a language-conditioned Soft Actor-Critic baseline [Akakzia et al., 2021, Röder et al., 2022], that we pair like Röder et al. [2022] with a GRU as language encoder [Cho et al., 2014]. We conduct all the experiments with two and three objects on the table. The objects are sampled uniformly from all attribute combinations, but we put specific constraints on the overlapping features. Our goal object has at most 1 property in common with the non-goal objects. This allows us to create scenarios like the one in Figure 1, where the agent needs to consider every part of the language goal. Furthermore, it assures that there is a unique and valid solution after the instructor proposes the correction.

In Figure 3, we show the mean and the standard deviation of 3 random seeds in the *"reach"* and *"push"* tasks. Next to the default action corrections without negations (AC, red lines in Figure 3), we showcase the episodes with action corrections including negations (ACN, purple lines in Figure 3). As we sample 50% of the episodes with an action correction challenge, we provide additional plots for those episodes separately below the overall success rates plots (Figure 3e, Figure 3f, Figure 3g, and Figure 3h).

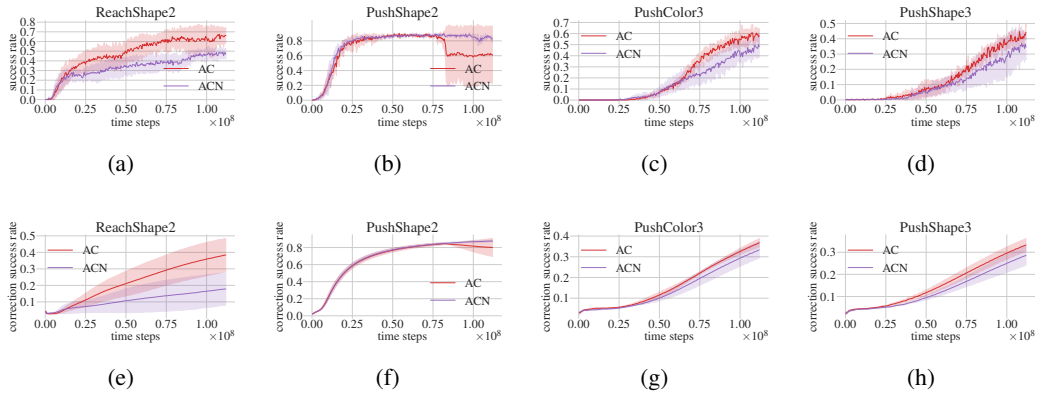

Figure 3: Our experiments showcase the success rate of a selection of *"reach"* and *"push"* experiments with different object configurations (Figure 3a, Figure 3b, Figure 3c, and Figure 3d). Below, we illustrate only the action correction success rate of the corresponding environments (Figure 3e, Figure 3f, Figure 3g, and Figure 3h).

As expected, learning action correction with negations is slightly more difficult than corrections created with editing terms and directly naming the desired goal object properties. Learning action corrections with just two objects in the scene is basically switching to the other object, after the action correction appears. However, it might still help with learning a robust representation because the agent still needs to grasp the nuances of the language. In Figure 3c and Figure 3d, the performance in the *"push"* task with 3 objects is shown. Although the success rate is increasing over time, the baseline requires millions of environment steps to surpass the threshold of only solving the single goal instructions shown in Figure 3c. We excluded the tasks that were too challenging for our baseline, but we expect addressing this with methods like hindsight learning (Subsection A.2), intrinsic motivation [Colas et al., 2020b, Akakzia et al., 2021] or hierarchical RL [Jiang et al., 2019].

In summary, it still takes additional effort to solve action correction challenges efficiently. However, we postulate our environment and initial experimental results to be valuable insight for future work.

5 Conclusion

This paper is the first to formalize and experimentally validate the use of incremental action correction for robotic instruction-following based on reinforcement learning. We present a novel collection of benchmark environments for which we show initial experimental results, using a common language-conditioned baseline algorithm.

References

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J. Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, and Mengyuan Yan. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances, 2022.

Ahmed Akakzia, Cédric Colas, Pierre-Yves Oudeyer, Mohamed Chetouani, and Olivier Sigaud. Grounding Language to Autonomously-Acquired Skills via Goal Generation. In *International Conference on Learning Representations*, Virtual (formerly Vienna, Austria), 2021. OpenReview.net.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. Experience Grounds Language. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Conference on Empirical Methods in Natural Language Processing*, pages 8718–8735, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.703.

Dan Bohus and Alexander I. Rudnicky. Sorry, I Didn't Catch That! In Laila Dybkjær and Wolfgang Minker, editors, *Recent Trends in Discourse and Dialogue*, volume 39, pages 123–154. Springer Netherlands, Dordrecht, 2008. ISBN 978-1-4020-6820-1 978-1-4020-6821-8. doi: 10.1007/978-1-4020-6821-8_6.

Antoine Bordes, Y. Lan Boureau, and Jason Weston. Learning End-to-End Goal-Oriented Dialog. In *International Conference on Learning Representations*, Toulon, France, 2017.

Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. BabyAI: A Platform to Study the Sample Efficiency of Grounded Language Learning. In *International Conference on Learning Representations*, New Orleans, Louisiana, USA, 2019. OpenReview.net.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, Doha, Qatar, 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179.

Cédric Colas, Ahmed Akakzia, Pierre-Yves Oudeyer, Mohamed Chetouani, and Olivier Sigaud. Language-Conditioned Goal Generation: A New Approach to Language Grounding in RL, 2020a.

Cédric Colas, Tristan Karch, Nicolas Lair, Jean-Michel Dussoux, Clément Moulin-Frier, Peter Ford Dominey, and Pierre-Yves Oudeyer. Language as a Cognitive Tool to Imagine Goals in Curiosity-Driven Exploration. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3761–3774, Vancouver, Canada, 2020b. Curran Associates, Inc.

Karl Moritz Hermann, Felix Hill, Simon Green, Fumin Wang, Ryan Faulkner, Hubert Soyer, David Szepesvari, Wojciech Marian Czarnecki, Max Jaderberg, Denis Teplyashin, Marcus Wainwright, Chris Apps, Demis Hassabis, and Phil Blunsom. Grounded Language Learning in a Simulated 3D World. *arXiv:1706.06551 [cs, stat]*, 2017.

Felix Hill, Stephen Clark, Karl Moritz Hermann, and Phil Blunsom. Understanding Early Word Learning in Situated Artificial Agents. *arXiv:1710.09867 [cs]*, page online, 2019.

Felix Hill, Andrew Lampinen, Rosalia Schneider, Stephen Clark, Matthew Botvinick, James L. McClelland, and Adam Santoro. Environmental drivers of systematicity and generalization in a situated agent. In *International Conference on Learning Representations*, Virtual (formerly Addis Ababa, Ethiopia), 2020.

Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents. *arXiv:2201.07207 [cs]*, 2022.

YiDing Jiang, Shixiang (Shane) Gu, Kevin P Murphy, and Chelsea Finn. Language as an Abstraction for Hierarchical Deep Reinforcement Learning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 9419–9431, Vancouver, Canada, 2019. Curran Associates, Inc.

Jelena Luketina, Nantas Nardelli, Gregory Farquhar, Jakob N. Foerster, Jacob Andreas, Edward Grefenstette, Shimon Whiteson, and Tim Rocktäschel. A Survey of Reinforcement Learning Informed by Natural Language. In *International Joint Conference on Artificial Intelligence*, pages 6309–6317, Macao, China, 2019. International Joint Conferences on Artificial Intelligence Organization. ISBN 978-0-9992411-4-1. doi: 10.24963/ijcai.2019/880.

Corey Lynch and Pierre Sermanet. Language Conditioned Imitation Learning Over Unstructured Data. In Dylan A. Shell, Marc Toussaint, and M. Ani Hsieh, editors, *Robotics: Science and Systems*, Virtual, 2021. doi: 10.15607/RSS.2021.XVII.047.

Frank Röder, Manfred Eppe, and Stefan Wermter. Grounding Hindsight Instructions in Multi-Goal Reinforcement Learning for Robotics. In *International Conference on Development and Learning*, pages 170–177, London, UK, 2022. IEEE. ISBN 978-1-66541-310-7.

Mohit Shridhar, Lucas Manuelli, and Dieter Fox. CLIPort: What and Where Pathways for Robotic Manipulation. In Aleksandra Faust, editor, *Conference on Robot Learning*, Proceedings of Machine Learning Research, pages 894–906, London, UK, 2021. PMLR.

Luc Steels. The symbol grounding problem has been solved. So what's next? In *Symbols and Embodiment: Debates on Meaning and Cognition*, pages 223–244. Oxford University Press, New York, NY, US, first edition, 2008. ISBN 0-19-921727-0 978-0-19-921727-4.

Stefanie Tellex, Nakul Gopalan, Hadas Kress-Gazit, and Cynthia Matuszek. Robots That Use Language. *Annual Review of Control, Robotics, and Autonomous Systems*, 3(1):25–55, 2020. doi: 10.1146/annurev-control-101119-071628.

A Appendix

A.1 Grammar Templates

A.1.1 Instruction Templates

$$
\begin{aligned}
\langle\text{INSTRUCTION}\rangle &\models \langle\text{TASKVERB}\rangle\ \langle\text{ARTICLE}\rangle\ \langle\text{OBJECT}\rangle \\
\langle\text{OBJECT}\rangle &\models \langle\text{COLOR}\rangle\ \langle\text{SHAPE}\rangle \mid \langle\text{SHAPE}\rangle\ \texttt{object} \mid \langle\text{COLOR}\rangle\ \texttt{object} \\
\langle\text{SHAPE}\rangle &\models \langle\text{CUBE}\rangle \mid \langle\text{CUBOID}\rangle \mid \langle\text{RECTANGLE}\rangle \\
\langle\text{COLOR}\rangle &\models \texttt{red} \mid \texttt{green} \mid \texttt{blue} \mid \texttt{yellow} \mid \texttt{purple} \mid \texttt{orange} \mid \texttt{pink} \mid \texttt{cyan} \mid \texttt{brown} \\
\langle\text{TASKVERB}\rangle &\models (\texttt{reach} \mid \texttt{touch} \mid \texttt{contact}) \mid (\texttt{push} \mid \texttt{move} \mid \texttt{shift}) \\
\langle\text{CUBE}\rangle &\models \texttt{cube} \mid \texttt{box} \mid \texttt{block} \\
\langle\text{CUBOID}\rangle &\models \texttt{cuboid} \mid \texttt{brick} \mid \texttt{oblong} \\
\langle\text{CYLINDER}\rangle &\models \texttt{cylinder} \mid \texttt{barrel} \mid \texttt{tophat} \\
\langle\text{ARTICLE}\rangle &\models \texttt{the}
\end{aligned}
$$

Figure 4: Backus normal form (BNF) for our instruction set inspired by Chevalier-Boisvert et al. [2019]. Here we show the verbs for the *"reach"* and *"push"* task only.

A.1.2 Action Correction Templates

$$
\begin{aligned}
\langle\text{EXTENDED INSTRUCTION}\rangle &\models \langle\text{INSTRUCTION}\rangle\ \langle\text{CORRECTION}\rangle \\
\langle\text{CORRECTION}\rangle &\models \langle\text{BEGINNING}\rangle\ \langle\text{ARTICLE}\rangle\ \langle\text{OBJECT}\rangle \\
\langle\text{BEGINNING}\rangle &\models \langle\text{EXCUSE}\rangle \mid \langle\text{NEGATION}\rangle \mid \langle\text{EDIT}\rangle \\
\langle\text{EXCUSE}\rangle &\models \texttt{sorry} \mid \texttt{excuse me} \mid \texttt{no i meant} \mid \texttt{pardon} \\
\langle\text{NEGATION}\rangle &\models \texttt{not} \\
\langle\text{EDIT}\rangle &\models \texttt{actually}
\end{aligned}
$$

Figure 5: The extended BNF grammar for our action corrections paired with the primal instruction. These definitions apply to all tasks.

A.2 Hindsight Learning for Action Correction

We argue that language-conditioned RL has many properties with goal-conditioned RL in common and could benefit from its insights. In Colas et al. [2020a], the authors even speak for goal-agnostic agents where the goal originates in modalities like images, Cartesian Coordinates, language or other representations.

The challenge of correcting a behavior is to interpret the action correction in context of the past actions and initial instruction. Because one can think of the initial instruction as abstract goal description, the action correction acquaints about an alternation of the episodic goal.

Especially methods of hindsight learning are valuable to apply to language-conditioned RL [Jiang et al., 2019, Colas et al., 2020b, Akakzia et al., 2021, Röder et al., 2022]. However, for this article we keep it a question for future work on how to apply this for action correction.

A.3 Hyperparameters

The hyperparameters are a slight alternation of the ones used by Akakzia et al. [2021] and Röder et al. [2022].

Table 1

Parameter.	Values.
batch size	256
hidden size of MLPs	256
learning rate actor, critic and entropy tuning	1e−3
number of workers	16
trade-off coefficient α	0.2
buffer size	1e+6
word embedding size	32
discount γ	0.98
number of critics	1
polyak ρ	0.95