Mind the Error! Detection and Localization of Instruction Errors in Vision-and-Language Navigation

Francesco Taioli[1,4*] Stefano Rosa[2] Alberto Castellini[1] Lorenzo Natale[2]
Alessio Del Bue[2] Alessandro Farinelli[1] Marco Cristani[1] Yiming Wang[3]

[1] University of Verona [2] Istituto Italiano di Tecnologia (IIT) [3] Fondazione Bruno Kessler [4] Polytechnic of Turin

Abstract

In Vision-and-Language Navigation in Continuous Environments (VLN-CE), agents have to navigate towards a target goal by executing a set of low-level actions, following a series of natural language instructions. All VLN-CE methods in the literature assume that language instructions are exact. However, in practice, instructions given by humans can contain errors. For the first time, we propose a novel benchmark dataset that introduces various types of instruction errors considering potential human causes, providing valuable insight into the robustness of VLN-CE agents. Moreover, we formally define the task of Instruction Error Detection and Localization, and propose a method that achieves best performances compared to baselines. Project page at https://intelligolabs.github.io/R2RIE-CE

1. Introduction

The emerging research on Vision-and-Language Navigation (VLN) [3, 6] aims to develop embodied agents that, following a given instruction in the format of natural language, can reach a target destination in a 3D environment. To facilitate the study of VLN, many benchmark datasets have been proposed [3, 5, 7, 8, 10]. All the previous benchmarks, however, only consider *correct* language instructions. This consideration can be brittle in reality as human often gives instructions that are approximate or ambiguous, or even prone to error as based on their memory. In this work, we first formally define the types of errors that may occur in language instructions for the VLN task in indoor environments, including *Direction, Room, Object, Room&Object* and a combination of *All* types of errors. Based on these definitions, we propose a novel benchmark in continuous environments built on top of the R2R-CE dataset [7]. Then, we propose a method based on a cross-modal transformer, fusing together the language features of the instruction with the observations of the agent, achieving competitive performance in solving the task of Detection and Localization of Instruction Errors, compared to a CLIP Alignment baseline. Our contributions are summarized below:

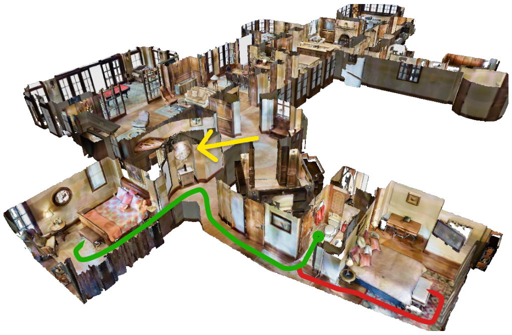

Figure 1. Changing *"right"* to *"left"* in *"Exit the bathroom and go left (✓right), then turn left at the big clock and go into the bedroom"* leads the agent to terminate exploration in the wrong location, disregarding the unseen *"big clock"* along the path.

- We establish the first benchmark *R2R with Instruction Errors in Continuous Environments* (R2RIE-CE), with a novel dataset and an evaluation protocol.
- We show that state-of-the-art VLN-CE methods are not robust to instruction errors using our proposed benchmark, necessitating the study of instruction errors in VLN.
- We propose a novel task *Instruction Error Detection and Localization*, with an effective baseline, Instruction Error Detector & Localizer (*IEDL*), based on a novel Instruction-Trajectory compatibility model.

2. R2RIE-CE Dataset

Our R2RIE-CE is constructed by artificially injecting various types of Instruction Errors into the given natural language instructions in R2R-CE, carefully considering error priors induced by human causes, including inaccurate scene memory and direction confusion. For each type of error, *i.e., Direction*, *Object*, *Room*, *Room&Object* and *All*, we create

[*] francesco.taioli@polito.it

a corresponding validation set based on the Val Unseen validation split of R2R-CE. Prior to dataset construction we have a filtering stage, *i.e.,* episodes must have Direction words for the validation set with *Direction Error*, or episodes must contain Direction, Object, and Room words for the validation set with *All Errors*. Following this filtering stage, we obtain a set of correct episodes $\mathcal{E}_\mathcal{C}$ for each type of Instruction Error. Then, for each episode $e_i \in \mathcal{E}_\mathcal{C}$, we create a corresponding erroneous episode in which we perturbed the instruction with the respective error. Specifically, *(i)* for the *Direction Error*, we swap directional words with their antonym, *e.g., left/right*, *go down/go up*, *into/out of*, etc. *(ii)* For the *Object Error*, we first identify a set of common object categories $\mathcal{C}$ that frequently appear in a set of language instructions, excluding synonyms. Each object class $c_i \in \mathcal{C}$ is associated with a set of object classes $\mathcal{C}_i$ comprised of the classes that often co-locate in the same room. We introduce an *Object Error* by swapping the ground-truth object class c_i to a random class $c_j \in \mathcal{C}_i$; *(iii)* For the *Room Error*, we consider the following set of rooms $\mathcal{R}$ that are common in indoor environments: *kitchen, archway, bathroom, bedroom, gym, lounge, hallway, living room, office, dining room, laundry, restroom*. Each room $r_i \in \mathcal{R}$ is associated with a set of rooms $\mathcal{R}_i$ that are often adjacent to r_i. We introduce a *Room Error* by swapping the ground-truth room r_i to a random room $r_j \in \mathcal{R}_i$; *(iv)* For the *Room&Object Error*, we introduce in an instruction both *Room* and *Object Error*; *(v)* For the *All Error*, we introduce in an instruction both *Direction* and *Room&Object Error*. The perturbed episodes will then be stored in the corresponding set of perturbed episodes, $\mathcal{E}_\mathcal{P}$. For each perturbed episode $e_i \in \mathcal{E}_\mathcal{P}$, both the error type and the position of the perturbed word are stored as metadata. Eventually, for each type of instruction error, we obtain sets $\mathcal{E}_\mathcal{C}$ and $\mathcal{E}_\mathcal{P}$.

3. R2RIE-CE Benchmark

With R2RIE-CE dataset, we can evaluate the robustness of VLN policies, and benchmark the novel task in terms of instruction error detection and localization, an important intermediate task for further addressing instruction errors in VLN policy learning.

Instruction Error Detection and Localization. Error detection aims to classify if an instruction contains errors, while localization aims to identify the positions of the error occurrences within the instruction. To address this task, we propose an effective method, short for IEDL, which takes visual observation embeddings Γ produced by the frozen policy [1]. A cross-modal multi-layer transformer fuses trajectory set Γ and instruction embeddings Υ. The enriched [CLS] token is fed into two classification heads, to produce alignment score $\sigma(a)$ and error location in the instruction, respectively. As the task is novel, we also construct a set of baselines for comparison: *(i) Random*: We randomly classify each instruction as correct or wrong, and randomly predict wrong token indices. *(ii) CLIP Alignment* (zero-shot): we extract the set of room and object tokens $\mathcal{K}$ from instruction $\mathcal{I}$ via an off-the-shelf POS tagger [4]. Errors are identified by comparing $\mathcal{K}$ with the observed objects and rooms $\mathcal{S}$ (labels are extracted using CLIP [9]) during navigation.

Performance Metrics. We use standard metrics for evaluating VLN performance as in prior works [2, 3], *i.e.,* Success Rate (SR), and Success rate weighted by Path Length (SPL). We evaluate *Detection of Instruction Errors* using the Area Under the ROC Curve (AUC, main metric) as in [11]. We then propose *Absolute Token Distance* (ATD), a novel metric to assess *Instruction Error Localization*, defined as the absolute difference between the predicted position of the perturbed token and the true position of the perturbed token.

Table 1. We show the performance (SR, SPL and ΔSR%) of [1] on our proposed benchmark. Then, we show the classification (AUC) and localization (ATD) performance of different methods.

Error type	Policy [1]			Random		CLIP Alignment		**IEDL**	
	SR ↑	**SPL** ↑	$\Delta_{SR}(\%)$	**AUC** ↑	**ATD** ↓	**AUC** ↑	**ATD** ↓	**AUC** ↑	**ATD** ↓
Direction	0.53	0.43	-18.64	0.50	10.54	0.50	11.05	0.59	7.74
Room	0.58	0.49	-6.66	0.50	11.03	0.57	9.63	0.79	7.85
Object	0.56	0.46	-8.47	0.51	10.94	0.59	8.76	0.74	10.14
Room&Object	0.57	0.47	-11.47	0.49	11.56	0.64	7.98	0.90	6.75
All	0.52	0.43	-30.64	0.51	12.22	0.63	8.68	0.93	5.68
Avg.	0.55	0.46	-15.17	0.50	11.26	0.59	9.22	**0.82**	**7.46**

Are VLN policies robust to instruction errors? We report the results in Tab. 1 under the SR, SPL and $\Delta_{SR}(\%)$ columns, where $\Delta_{SR}(\%) = SR(\mathcal{E}_\mathcal{C}) - SR(\mathcal{E}_\mathcal{P})$. *Direction* type of error has the largest effect on the navigation policy, with a -18.64% of SR. Following that, we have *Object* error type with a -8.47% and the *Room* error with a decrease of -6.66%. Interestingly, differently from [12], VLN-CE agents are more affected by perturbation of directional tokens than by object tokens (-18.64% *vs* -8.47%).

Can we detect and localize instruction errors? Tab. 1 reports the results regarding error detection by the AUC, and error localization by ATD. Regarding the detection performance, the random baseline is presented as a means to identify potential biases and establish a lower-bound, achieving an AUC of ~ 0.50, as expected. CLIP Alignment seems to be effective in the presence of *Object* and *Room* types of errors. We can see that *IEDL* achieves the best AUC in all the benchmarks by a large margin. A lower AUC of 0.59 for *IEDL* also seems to highlight the challenges of the *Direction* error type. Regarding the localization performance, CLIP Alignment baseline performs better for all error types compared to random. Our proposed *IEDL* achieves the best localization performance across all benchmarks, except for the *Object*, in which CLIP Alignment seems to be particularly effective. Finally, it is worth noticing that the mean ATD of *IEDL* is 7.46, which is close to the length of a typical sub-sentence within each instruction.

References

[1] Dong An, Yuankai Qi, Yangguang Li, Yan Huang, Liang Wang, Tieniu Tan, and Jing Shao. BEVBert: Multimodal Map Pre-training for Language-guided Navigation. *ICCV*, 2023. 2

[2] Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018. 2

[3] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sunderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-Language Navigation: Interpreting Visually-Grounded Navigation Instructions in Real Environments. In *CVPR*, 2018. 1, 2

[4] Steven Bird and Edward Loper. NLTK: the natural language toolkit. In *ACL*, 2004. 2

[5] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niebner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3D: Learning from RGB-D Data in Indoor Environments. In *3DV*, pages 667–676, CA, USA, 2017. 1

[6] Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Wang. Vision-and-Language Navigation: A Survey of Tasks, Methods, and Future Directions. In *Proc. of Association for Computational Linguistics*. ACL, 2022. 1

[7] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. *Beyond the Nav-Graph: Vision-and-Language Navigation in Continuous Environments*, page 104–120. ECCV, 2020. 1

[8] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-Across-Room: Multilingual Vision-and-Language Navigation with Dense Spatiotemporal Grounding. In *EMNLP*, 2020. 1

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 2

[10] Zun Wang, Jialu Li, Yicong Hong, Yi Wang, Qi Wu, Mohit Bansal, Stephen Gould, Hao Tan, and Yu Qiao. Scaling Data Generation in Vision-and-Language Navigation. In *ICCV*, 2023. 1

[11] Ming Zhao, Peter Anderson, Vihan Jain, Su Wang, Alexander Ku, Jason Baldridge, and Eugene Ie. On the Evaluation of Vision-and-Language Navigation Instructions. In *EACL*, pages 1302–1316, Online, 2021. 2

[12] Wanrong Zhu, Yuankai Qi, Pradyumna Narayana, Kazoo Sone, Sugato Basu, Xin Wang, Qi Wu, Miguel Eckstein, and William Yang Wang. Diagnosing Vision-and-Language Navigation: What Really Matters. In *NAACL*, pages 5981–5993, Seattle, United States, 2022. 2