Vision-and-Dialog Navigation

Jesse Thomason **Michael Murray**
Maya Cakmak Luke Zettlemoyer
University of Washington

CoRL 2019

Contributions

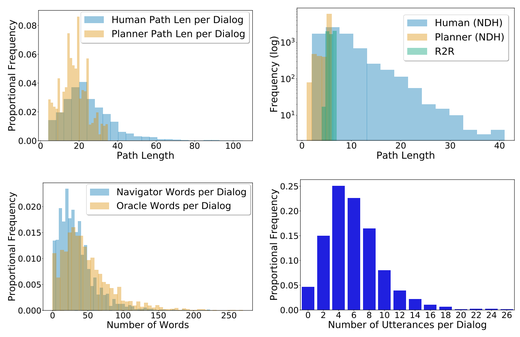
+ Over 2k human-human navigation dialogs.
+ Initial navigation models.
+ Dialog helps navigation.
+ Mixed human and planner supervision helps.
+ Navigation challenge leaderboard.

Dataset

CVDN is the first dataset to include two-sided dialogs held in natural language, with the initial navigation instruction being both ambiguous (*Amb*) and underspecified (*UnderS*), and situated in a photorealistic, visual navigation environment viewed by both speakers.

Dataset	—Language Context—				—Visual Context—		
	Human	Amb	UnderS	Temporal	Real-world	Temporal	Shared
MARCO[?, ?], DRIF[?]	✓	✗	✗	1I	✗	Dynamic	-
R2R[?], Touchdown[?]	✓	✗	✗	1I	✓	Dynamic	-
EQA[?], IQA[?]	✗	✗	✓	1Q	✗	Dynamic	-
CLEVR[?]	✗	✗	-	1Q	✗	Static	-
VQA[?, ?, ?]	✓	✗	-	1Q	✓	Static	-
CLEVR-Dialog[?]	✗	✗	-	2D	✗	Static	✓
VisDial[?]	✓	✗	-	2D	✓	Static	✓
VLNA[?], HANNA[?]	✗	✓	✓	1D	✓	Dynamic	✗
TtW[?]	✓	✗	✓	2D	✓	Dynamic	✗
CVDN	✓	✓	✓	2D	✓	Dynamic	✓

Human *Navigator* paths are longer than shortest path planner routes, resulting in much longer paths than in the comparable Room-to-Room dataset.



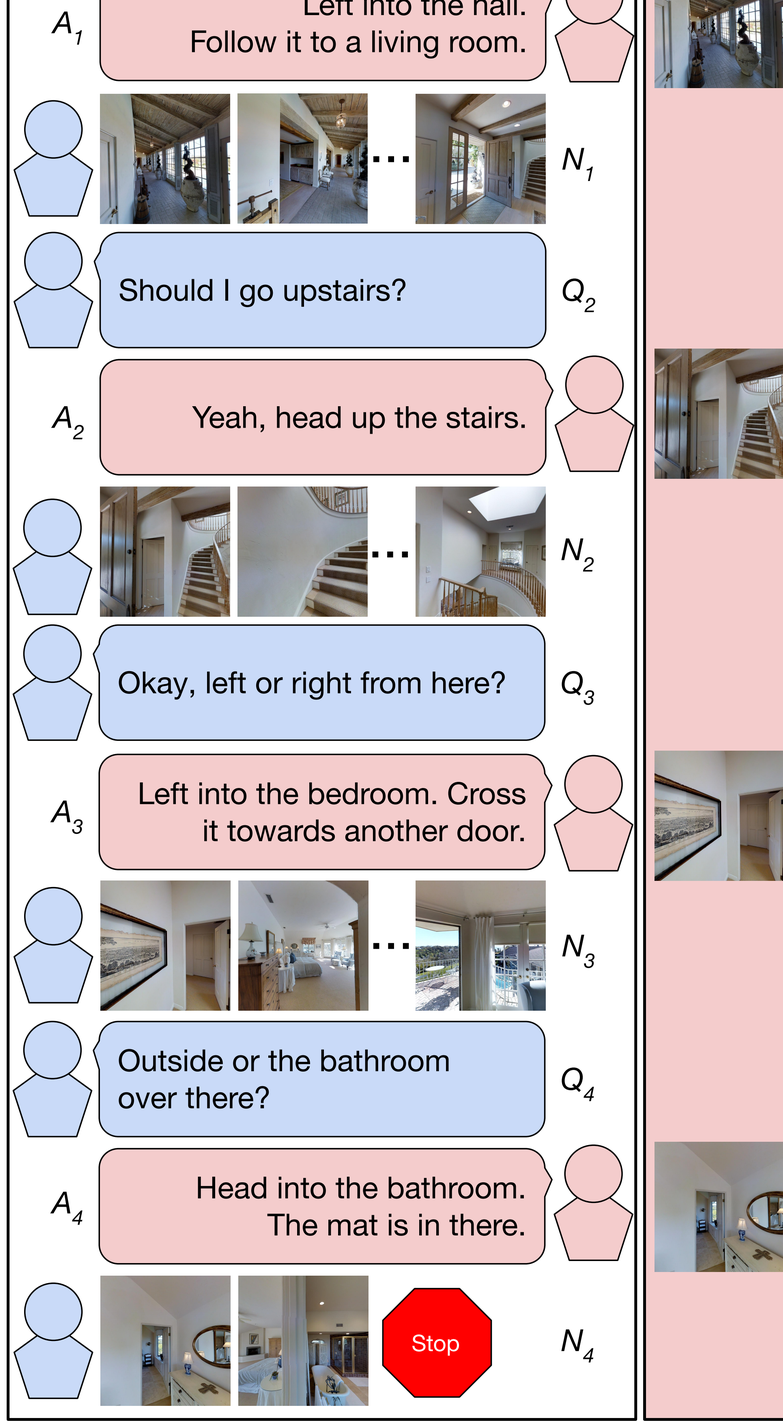
Human *Oracle*s use more words than *Navigator*s, and dialogs have on average 3-4 question-answer exchanges each.

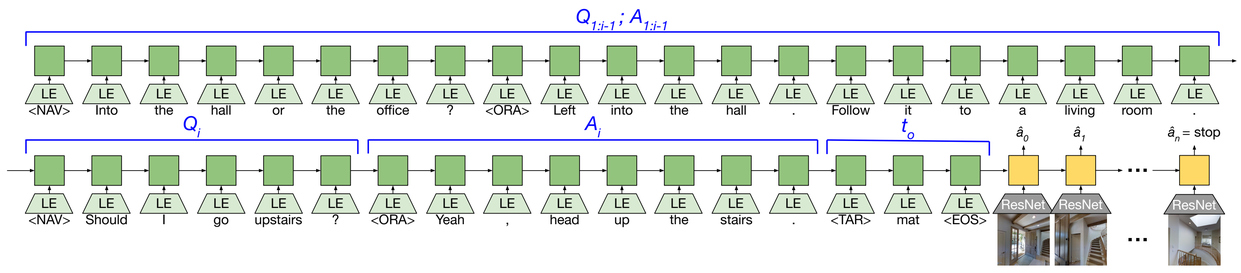
	Dia	*Nav*	*Ora*	**Example**
Ego	92.5	52.9	65.8	*Oracle*: Turn slightly to your right and go forward down the hallway
Needs Q	13.0	-	3.9	*Navigator*: Should I turn left down the hallway ahead? *Oracle*: ya
Needs Dialog History	3.5	0.4	1.0	*Oracle*: Through the lobby. So go through the door next to the green towel. Go to the left door next to the two yellow lights. Walk straight to the end of the hallway and stop . . . *Navigator*: Are these the yellow lights you were talking about?
Needs Nav History	14.0	1.5	3.4	*Oracle*: You were there briefly but left. There is a turntable behind you a bit. Enter the bedroom next to it.
Repair	12.5	1.6	3.4	*Oracle*: I am so sorry I meant for you to look over to the right not the left
Off-topic	3.0	5.4	5.1	*Navigator*: I am to the 'rear' of the zebra. Nice one. *Oracle*: Ok hold your nose and go to the left of the zebra, through the livingroom and kitchen and towards the bedroom you can see past that
Vacuous	6.0	22.7	2.3	*Navigator*: Ok, now where?

The average percent of *Dialogs*, as well as individual *Navigator* and *Oracle* utterances, exhibiting each phenomena out of 100 hand-annotated dialogs.

Cooperative Vision-and-Dialog Navigation (CVDN)

A big dataset of human-human dialogs!
Training navigation agents!
A demo interface! ⇓





Navigation Task

For each question-answer exchange, we task an agent with navigating towards the goal given the dialog so far. We can use the path taken by the *Navigator*, shown to the *Oracle*, or a *mix* as supervision.

Evaluation

Average agent progress towards the goal location when trained using different path end nodes for supervision. Among ablations, **bold** indicates most progress by language input, and **blue** indicates most progress by supervision signal.

Fold		Seq-2-Seq Inputs V	t_o	A_i	Q_i	$Q_{1:i-1}$ $A_{1:i-1}$	Goal Progress (m) ↑ Oracle	Navigator	Mixed
Val (Seen)	Baselines	Shortest Path					8.29	7.63	9.52
		Random					0.42	0.42	0.42
							0.59	0.83	0.91
		✓					4.12	5.58	5.72
			✓	✓	✓	✓	1.41	1.43	1.58
	Ours	✓	✓				4.16	**5.71**	5.71
		✓	✓	✓			4.34	5.61	6.04
		✓	✓	✓	✓		4.28	5.58	**6.16**
		✓	✓	✓	✓	✓	**4.48**	5.67	5.92
Val (Unseen)	Baselines	Shortest Path					8.36	7.99	9.58
		Random					1.09	1.09	1.09
							0.69	1.32	1.07
		✓					0.85	1.38	1.15
			✓	✓	✓	✓	1.68	1.39	1.64
	Ours	✓	✓				0.74	1.33	1.29
		✓	✓	✓			1.14	1.62	2.05
		✓	✓	✓	✓		1.11	1.70	1.83
		✓	✓	✓	✓	✓	**1.23**	**1.98**	**2.10**
Test (Unseen)	Baselines	Shortest Path					8.06	8.48	9.76
		Random					0.83	0.83	0.83
							0.13	0.80	0.52
		✓					0.99	1.56	1.74
			✓	✓	✓	✓	1.51	1.20	1.40
	Ours	✓	✓				1.05	1.81	1.90
		✓	✓	✓			1.21	2.01	2.05
		✓	✓	✓	✓		**1.35**	1.78	2.27
		✓	✓	✓	✓	✓	1.25	**2.11**	**2.35**

Using all dialog history significantly outperforms unimodal ablations in *unseen* environments. Using all dialog history, rather than just the last question or question-answer exchange, is needed to achieve statistically significantly better performance than using the target object alone in *unseen* test environments. Dialog history is beneficial for understanding the context of the latest navigation instruction A_i. Models trained with mixed supervision always statistically significantly outperform those trained with oracle or navigator supervision. Using human demonstrations only when they appear trustworthy increases agent progress towards the goal.

jdtho@cs.washington.edu

https://arxiv.org/abs/1907.04957