

The Lab vs The Crowd: An Investigation into Data Quality for Neural Dialogue Models

*José Lopes, Francisco J. Chiyah Garcia
and Helen Hastie*

Motivation

- Data collection is essential for developing new dialogue systems
- A vast majority of currently used datasets were collected through crowd-sourcing
- Quality of the data has been assessed by the variability or the lexical complexity of the data
- How different methodologies affect model performances?
 - Double amount of crowd-sourced dialogues is needed to achieve similar performance
- If developing a new dialogue system from scratch, please consider running a in-lab high quality data collection

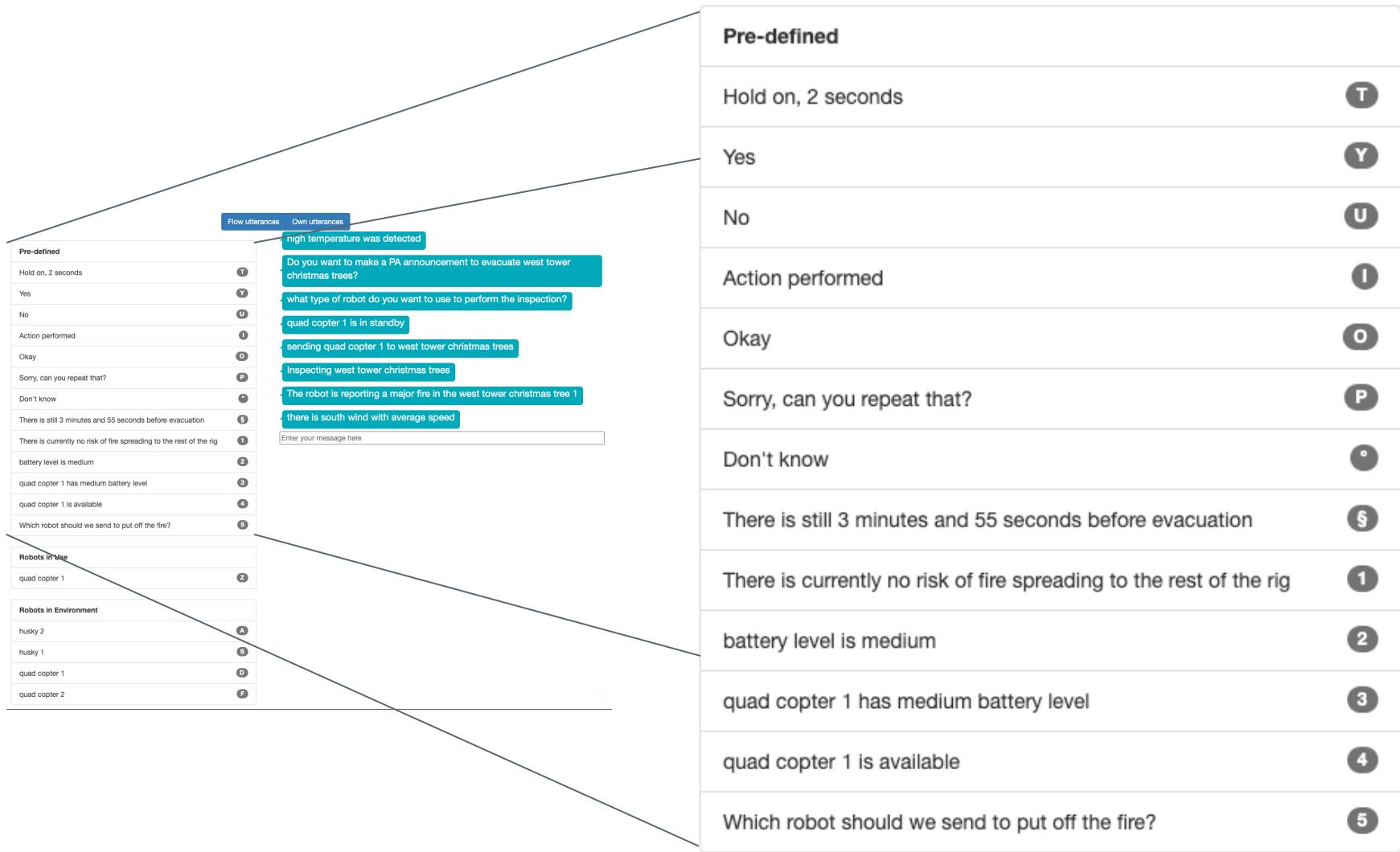
In Lab Set-up

	Flow utterances	Own utterances
Pre-defined		
Hold on, 2 seconds	T	high temperature was detected
Yes	Y	Do you want to make a PA announcement to evacuate west tower christmas trees?
No	U	what type of robot do you want to use to perform the inspection?
Action performed	I	quad copter 1 is in standby
Okay	O	sending quad copter 1 to west tower christmas trees
Sorry, can you repeat that?	P	Inspecting west tower christmas trees
Don't know	S	The robot is reporting a major fire in the west tower christmas tree 1
There is still 3 minutes and 55 seconds before evacuation	J	there is south wind with average speed
There is currently no risk of fire spreading to the rest of the rig	I	
battery level is medium	2	
quad copter 1 has medium battery level	3	
quad copter 1 is available	4	
Which robot should we send to put off the fire?	5	
Robots in Use		
quad copter 1	Z	
Robots in Environment		
husky 2	A	
husky 1	S	
quad copter 1	D	
quad copter 2	F	

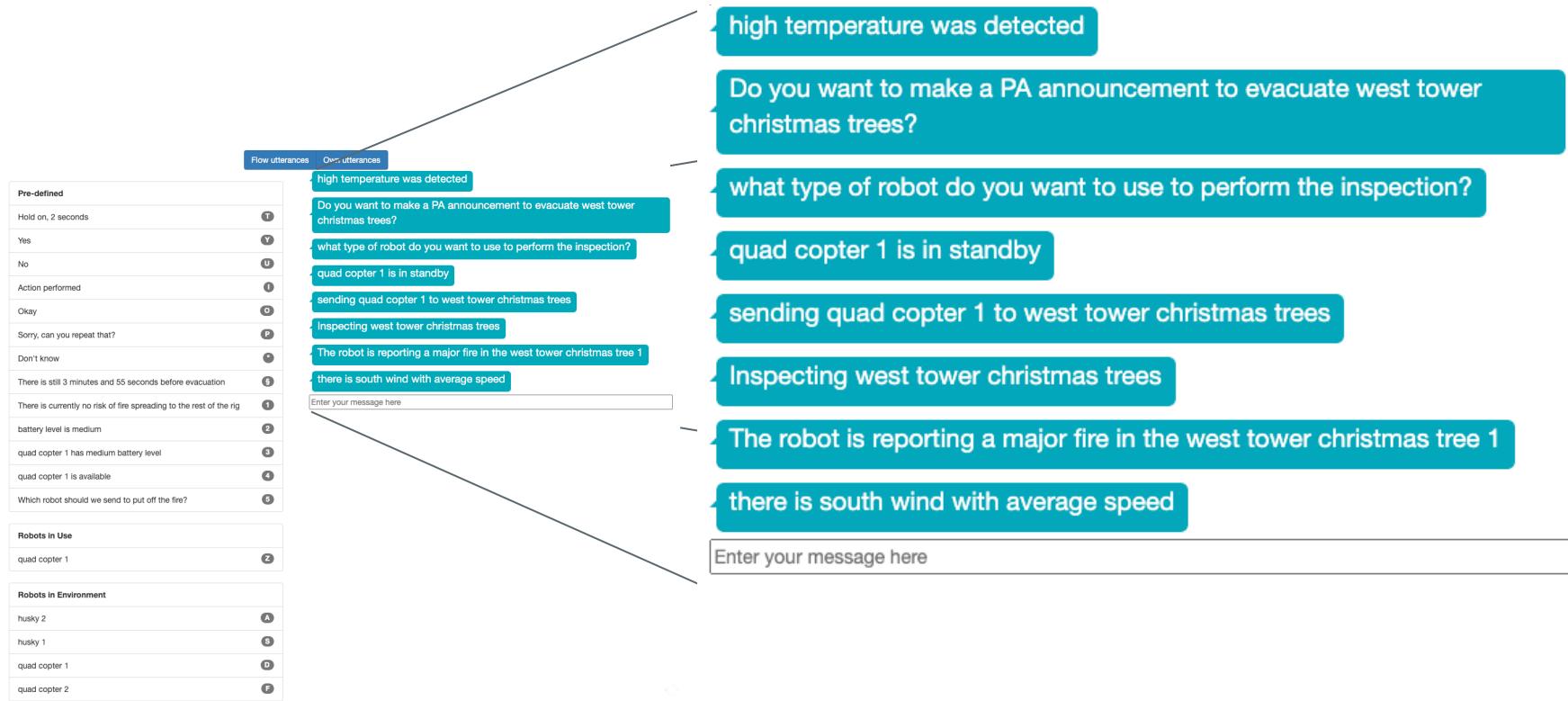
Enter your message here

Wizard View

In Lab Set-up



In Lab Set-up



In Lab Set-up

Flow utterances Own utterances

Pre-defined	
Hold on, 2 seconds	T
Yes	Y
No	U
Action performed	I
Okay	O
Sorry, can you repeat that?	P
Don't know	S
There is still 3 minutes and 55 seconds before evacuation	R
There is currently no risk of fire spreading to the rest of the rig	I
battery level is medium	Z
quad copter 1 has medium battery level	3
quad copter 1 is available	4
Which robot should we send to put off the fire?	5
 Robots in Use	
quad copter 1	Z
 Robots in Environment	
husky 2	A
husky 1	S
quad copter 1	D
quad copter 2	F

high temperature was detected

Do you want to make a PA announcement to evacuate west tower christmas trees?

what type of robot do you want to use to perform the inspection?

quad copter 1 is in standby

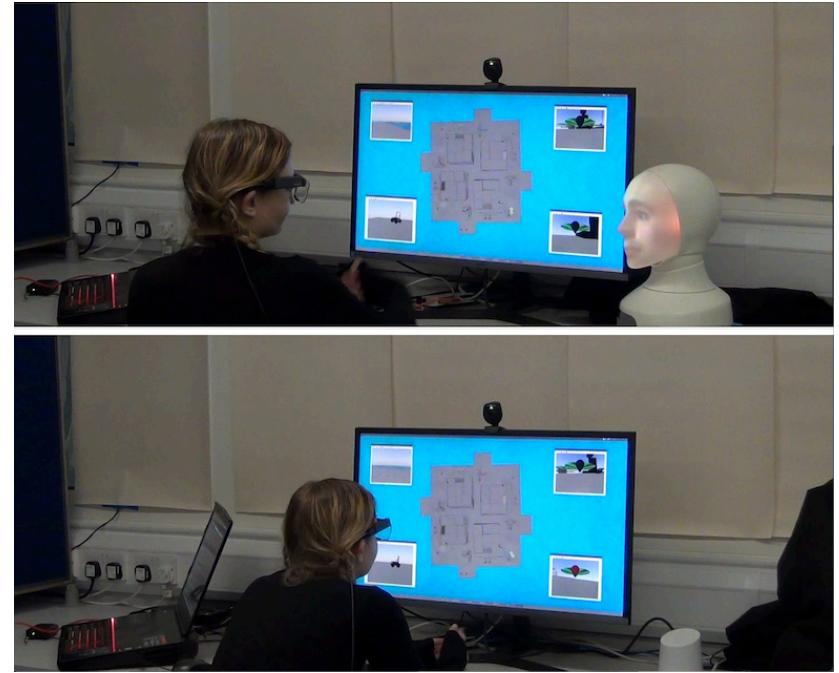
sending quad copter 1 to west tower christmas trees

Inspecting west tower christmas trees

The robot is reporting a major fire in the west tower christmas tree 1

there is south wind with average speed

Enter your message here



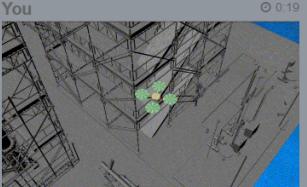
Wizard View

Participant View

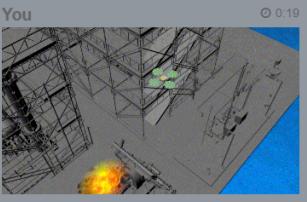
MTurk Set-up

Emergency Response Game

Quad copter 1 is inspecting the area

You  0:19

Operator  0:19
What does the robot see?

You  0:19
The robot is reporting a major fire in the east tower gas compressor

Dialogue Options

- Which robot should we send to put out the fire?
- Quad copter 1 is not able to perform the selected task. Please choose a different robot.
- There is a considerable risk that the fire spreads to processing module north tower
- Click to enter your own text

General Dialogue

- Hold on, 2 seconds
- Yes
- No
- Action performed
- Okay
- Sorry, can you repeat that?

Robot in Use (only 1 in use at any time)

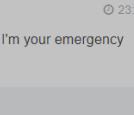
quad copter 1  Inspect 

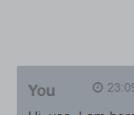
Robots Available	Skills
husky 1 	Extinguish fire and open valves
husky 2 	Inspect
quad copter 1 	Inspect
quad copter 2 	Extinguish fire

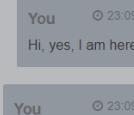
Send **I need a hint!**

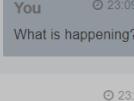
Emergency Response Game

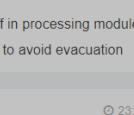
Users: HelperBot, Operator, You
Remaining time: 1:01

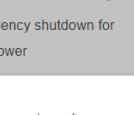
Fred  0:23.09
Hi, my name is Fred, and I'm your emergency assistant

Fred  0:23.09
Are you there?

You  0:23.09
Hi, yes, I am here

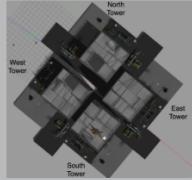
You  0:23.09
What is happening?

Fred  0:23.09
Emergency alarm went off in processing module east tower. We have 2:20 to avoid evacuation

Fred  0:23.09
First, I'm activating emergency shutdown for processing module east tower

Game progress:
1. Identify  2. Resolve  3. Assess  Finish 

Game Information
Offshore processing facility map:



Robots Available **Skills**

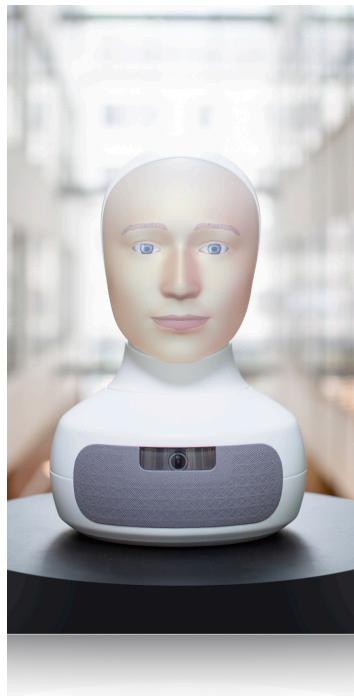
husky 1 	Extinguish fire and open valves
husky 2 	Inspect
quad copter 1 	Inspect
quad copter 2 	Extinguish fire

Enter your message here! 

Wizard View

Participant View

Example Dialogue



Hi, my name is Fred and I'm your emergency assistant.

Emergency alarm went off in processing module east tower.

We have 1:41 to avoid evacuation. First, I'm activating the emergency shutdown.

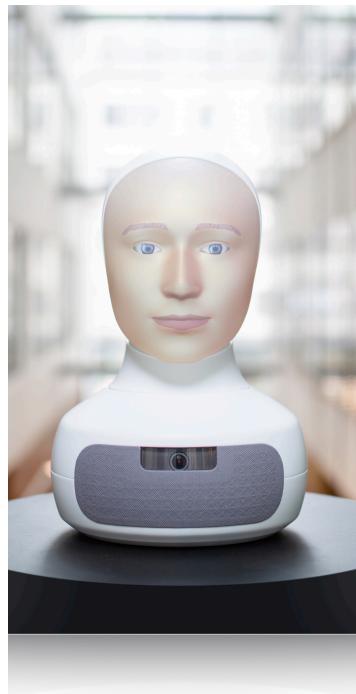
Quadcopter 1 go to inspect.

The robot is taking off. ETA is 32 seconds.

How long until evacuation?



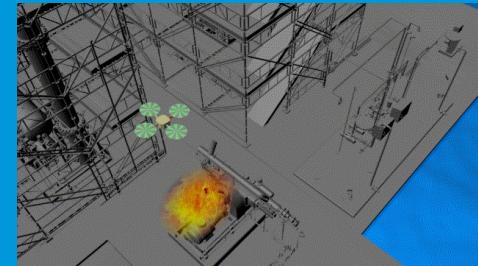
Example Dialogue



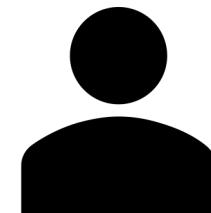
How long until evacuation?

We have 1:07 until evacuation.

The robot has arrived to processing
module east tower.



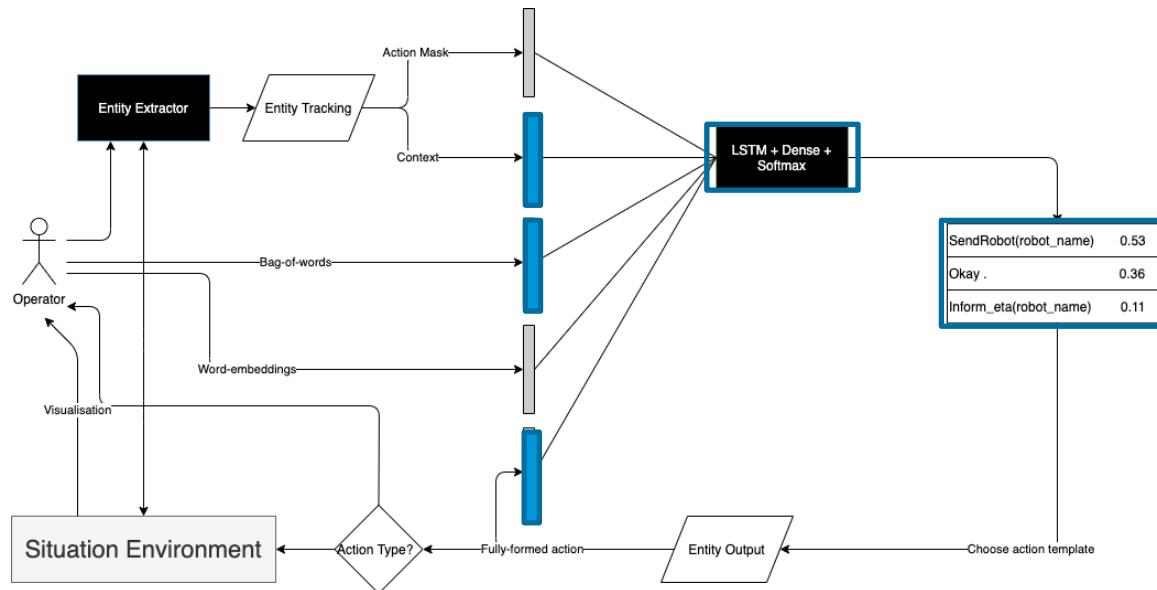
Performing Inspection



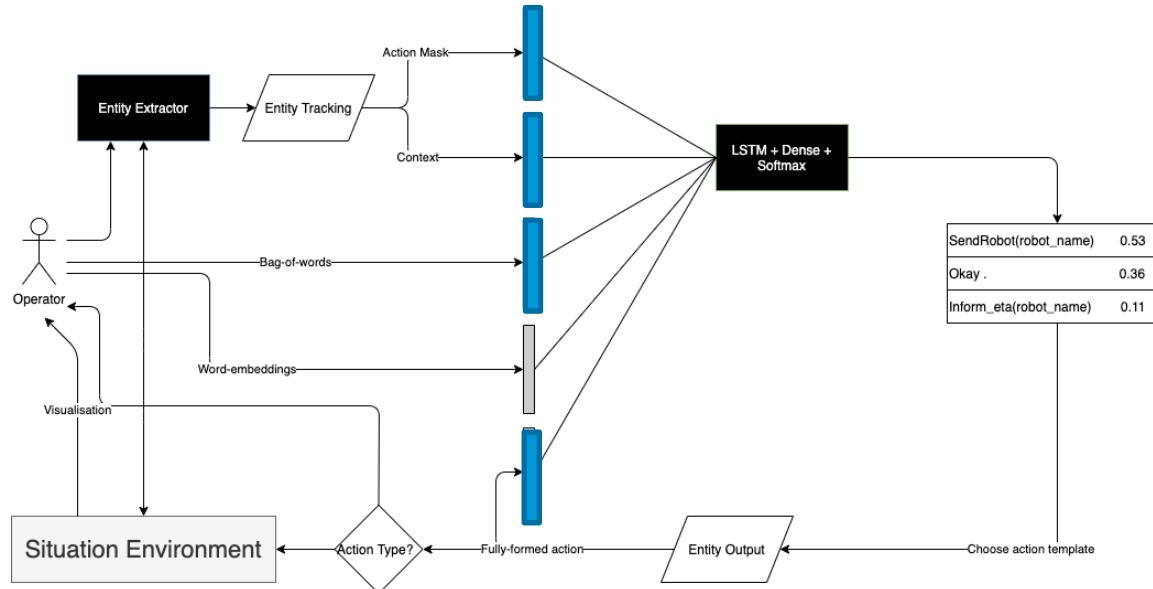
Data Collections Statistics

	Lab	MTurk
Number of Dialogues	63	147
Mean Number of Turns (std dev)	49.08 (13.83)	24.53 (9.49)
Mean Number of Operator Turns (std dev)	11.43 (7.95)	6.92 (3.24)
Mean Number of Assistant Turns (std dev)	37.65 (8.57)	17.61 (7.98)
Mean Operator Turn Length (std dev)	4.37 (3.38)	4.56 (3.35)
Mean percentage typed wizard utterances (std dev%)	2.58 % (2.87%)	2.77% (5.14%)
Task Success	62.12%	9.66%

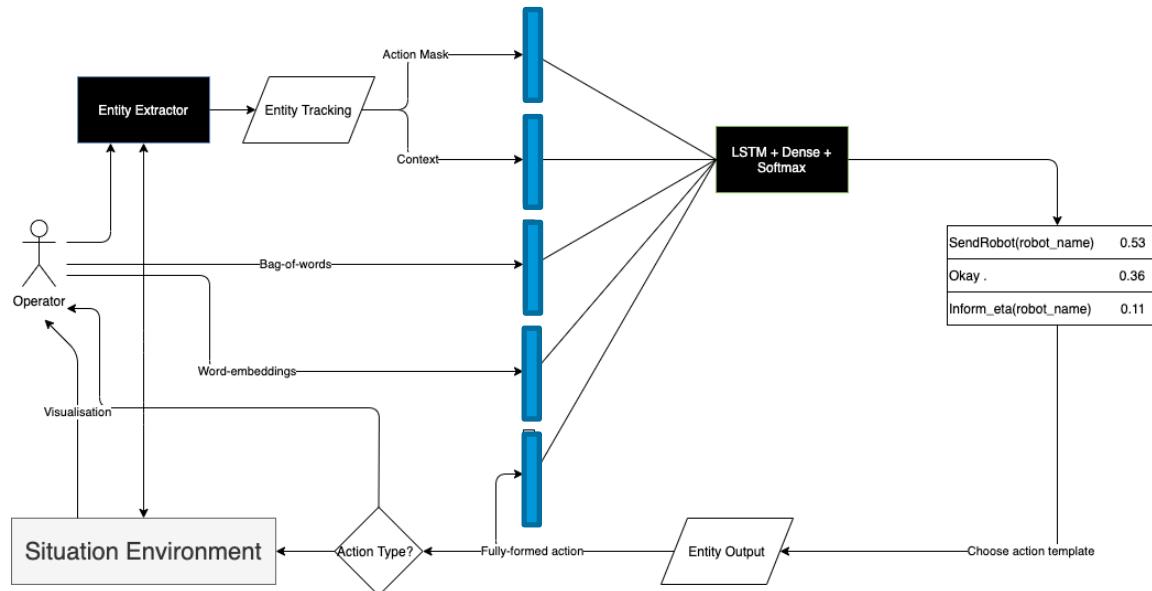
- Hybrid Code Networks (Williams et al, 2017)
 - Baseline: Previous actions, context and bag-of-words



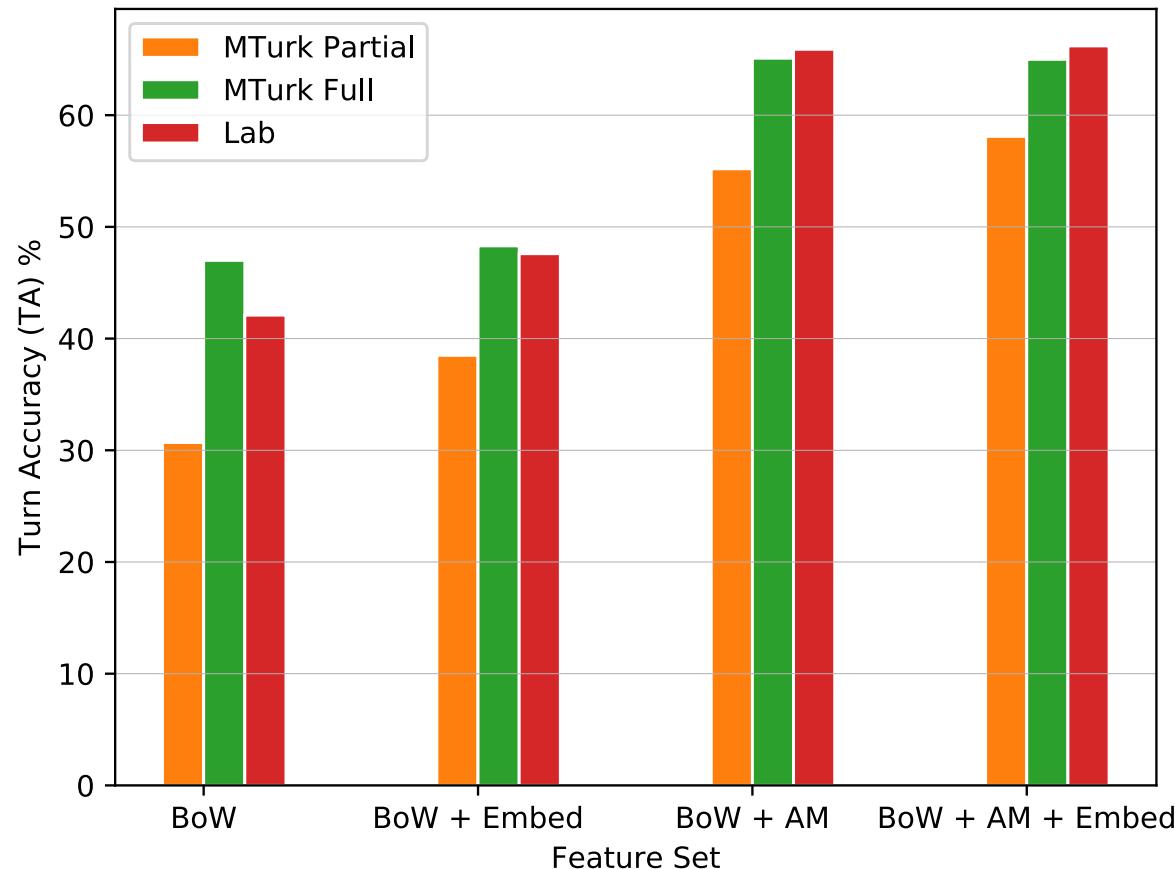
- Hybrid Code Networks (Williams et al, 2017)
 - Baseline: Previous actions, context and bag-of-words
 - Action Mask



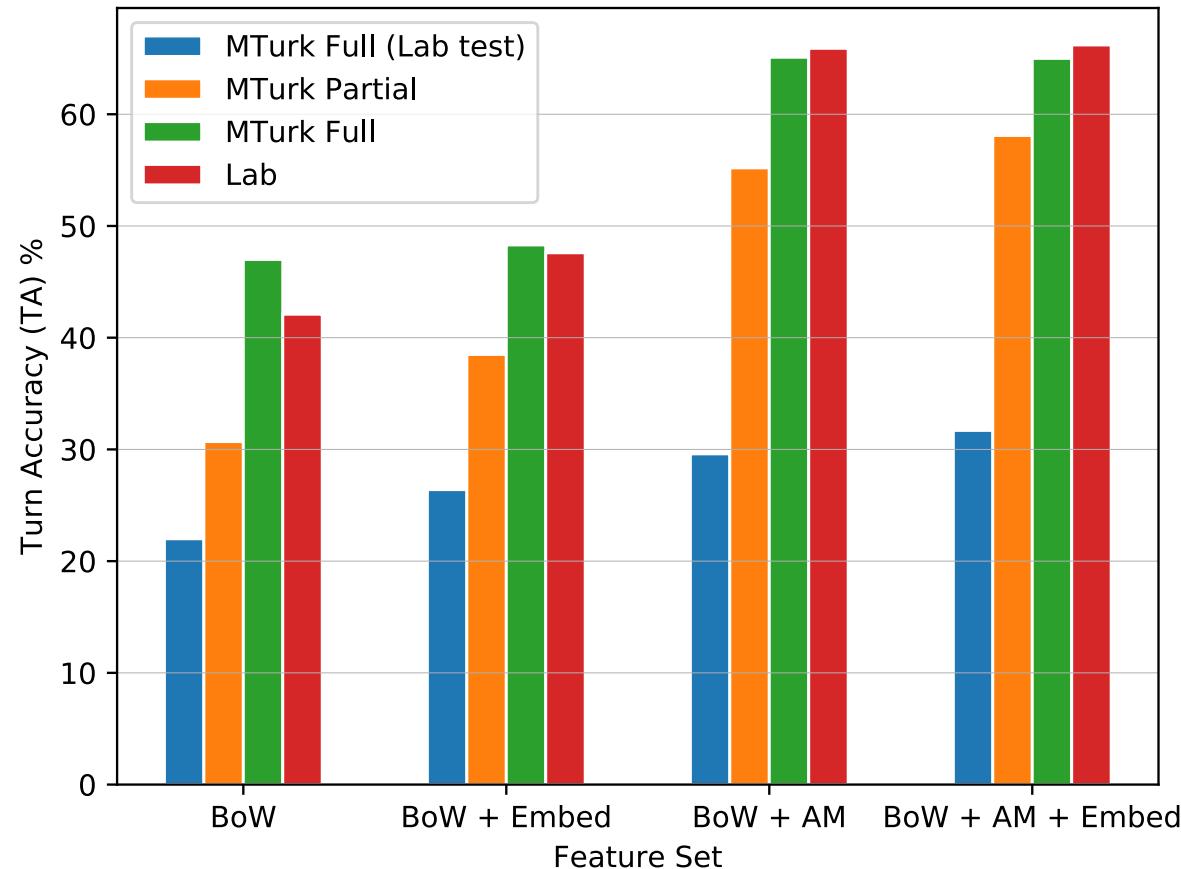
- Hybrid Code Networks (Williams et al, 2017)
 - Baseline: Previous actions, context and bag-of-words
 - Action Mask
 - Word2vec trained on Google News



Within dataset: Leave-operator-out



Between dataset: 100 runs average



Pre-trained models

Model	Fine-Tuning (#dialogues)	Testing (#dialogues)	Turn Accuracy
BERT	Mturk Full (147)	Lab (59)	7.66 %
ToDBERT	Mturk Full (147)	Lab (59)	8.15 %
DialoGPT	Mturk Full (147)	Lab (59)	10.8 %
GPT2	Mturk Full (147)	Lab (59)	8.19 %

Pre-trained models

Model	Fine-Tuning (#dialogues)	Testing (#dialogues)	Turn Accuracy
BERT	Mturk Full (147)	Lab (59)	7.66 %
ToDBERT	Mturk Full (147)	Lab (59)	8.15 %
DialoGPT	Mturk Full (147)	Lab (59)	10.8 %
GPT2	Mturk Full (147)	Lab (59)	8.19 %
HCN	Mturk Full (147)	Lab (59)	31.8 %

Discussion

- Models trained with MTurk data tend to 'rush' the dialogue
 - Fewer situation updates (e.g. robot's ETA)
 - Fewer interaction management dialogue acts (e.g. 'hold on')
- Models trained with smaller in-lab datasets outperformed models trained with larger crowdsourced datasets and fine-tuned pre-trained models

Future Work

- Further investigate the use of pre-trained models
- Replicate this study with a larger dataset
- Run a single-wizard data collection on MTurk
- More systematic comparison between different models (Ultes and Maier, 2020)
- Improve situation awareness in crowd-sourced data collections