HERIOT WATT UNIVERSITY

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

José Lopes, Francisco J. Chiyah Garcia and Helen Hastie

LEADERS IN IDEAS AND SOLUTIONS

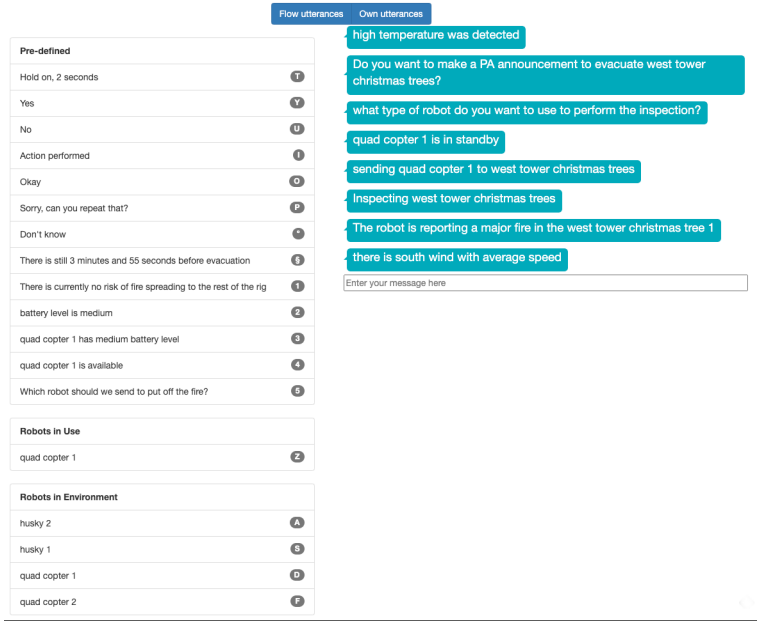
HERIOT WATT UNIVERSITY

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

HERIOT WATT UNIVERSITY

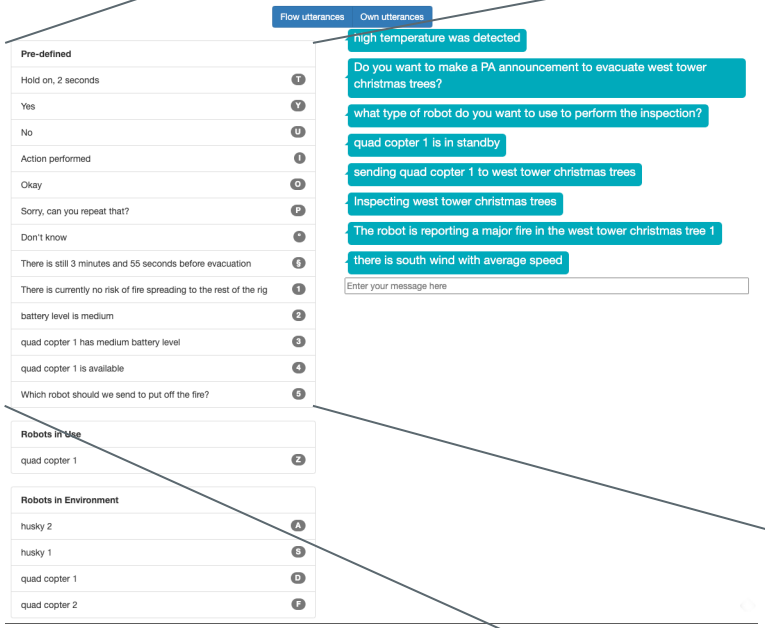
In Lab Set-up

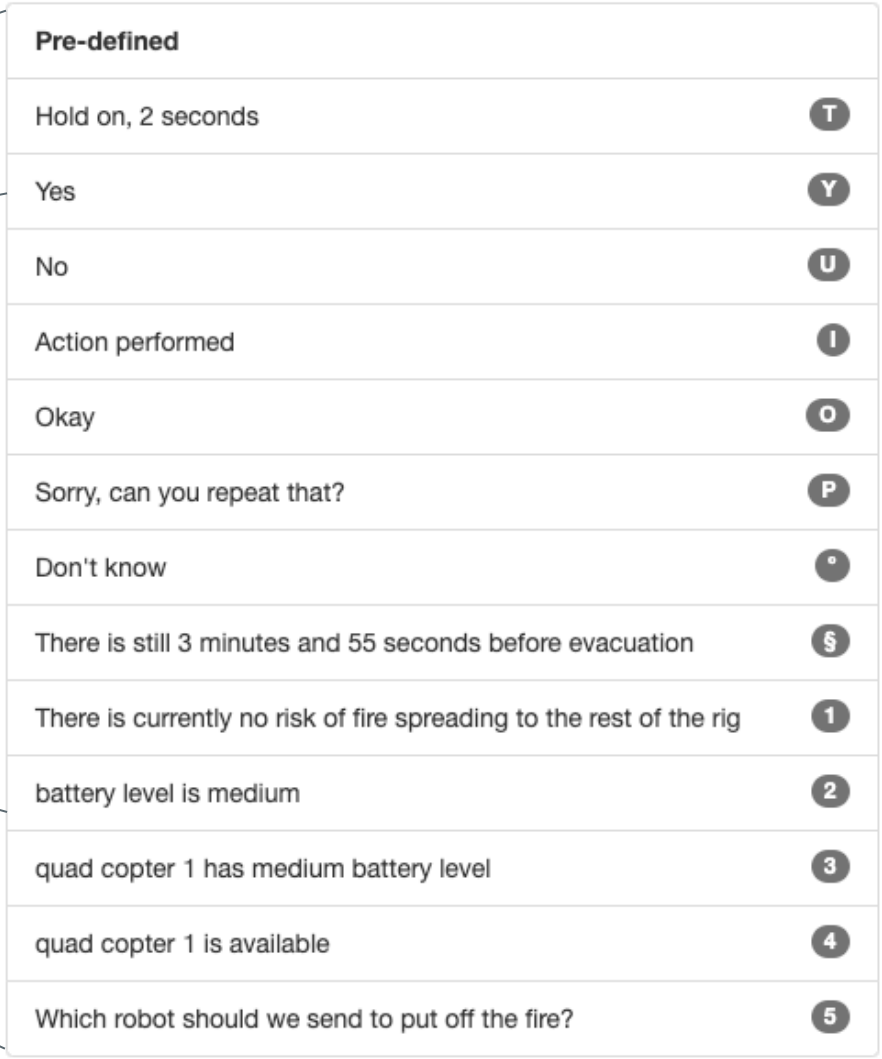


Wizard View

HERIOT WATT UNIVERSITY

In Lab Set-up





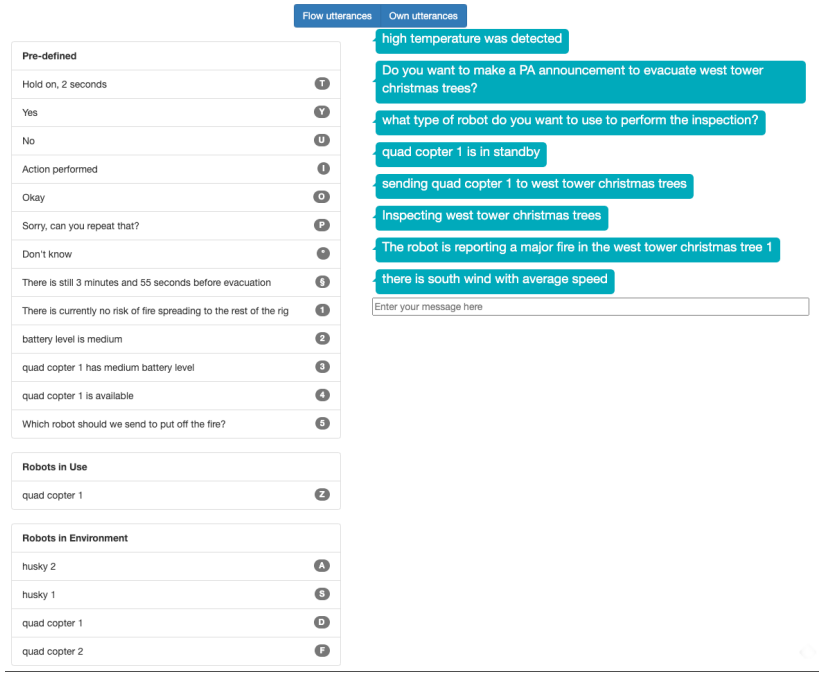
HERIOT WATT UNIVERSITY

In Lab Set-up

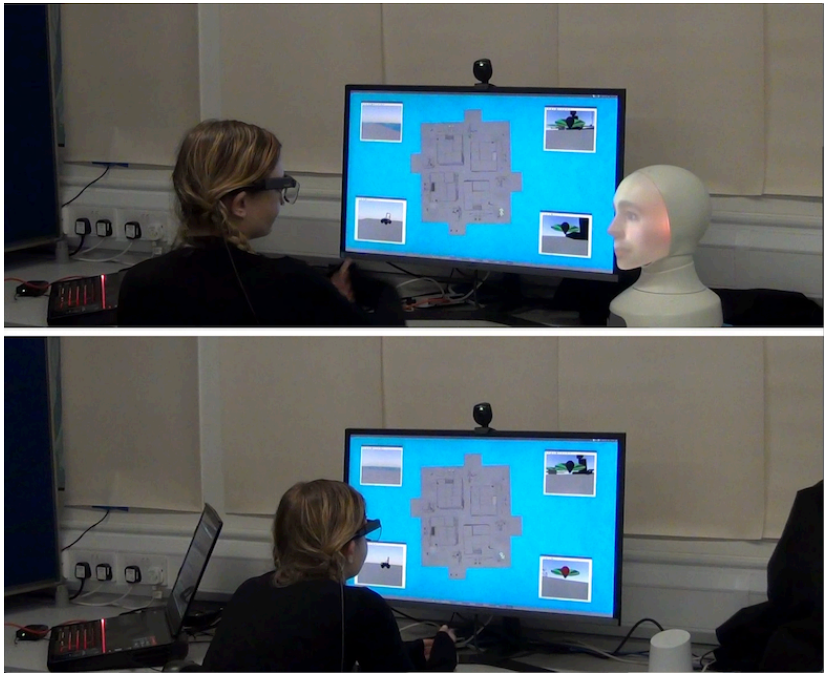
high temperature was detected
Do you want to make a PA announcement to evacuate west tower christmas trees?
what type of robot do you want to use to perform the inspection?
quad copter 1 is in standby
sending quad copter 1 to west tower christmas trees
Inspecting west tower christmas trees
The robot is reporting a major fire in the west tower christmas tree 1
there is south wind with average speed
Enter your message here

HERIOT WATT UNIVERSITY

In Lab Set-up

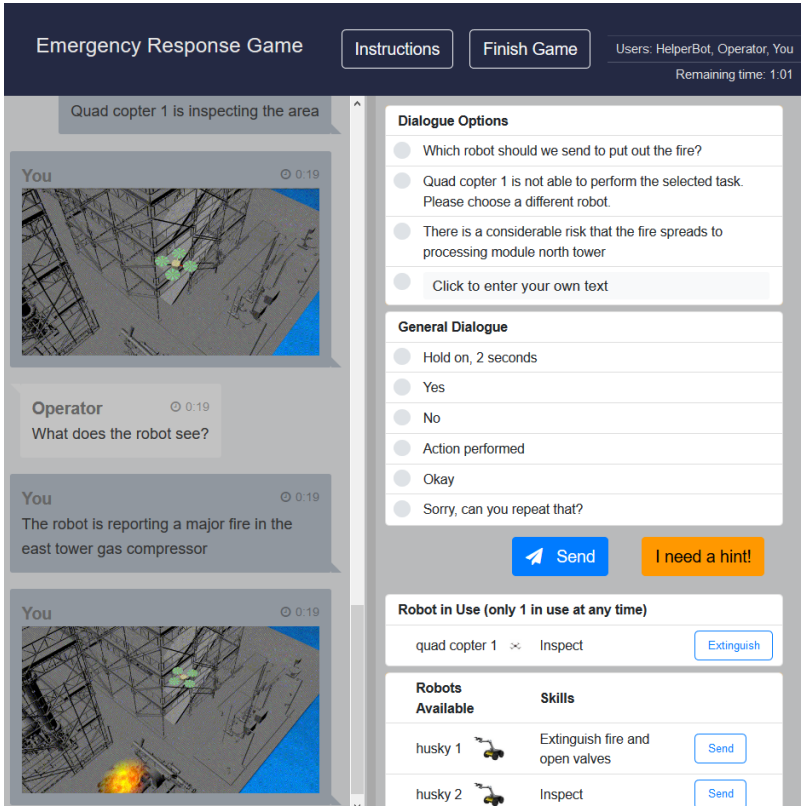


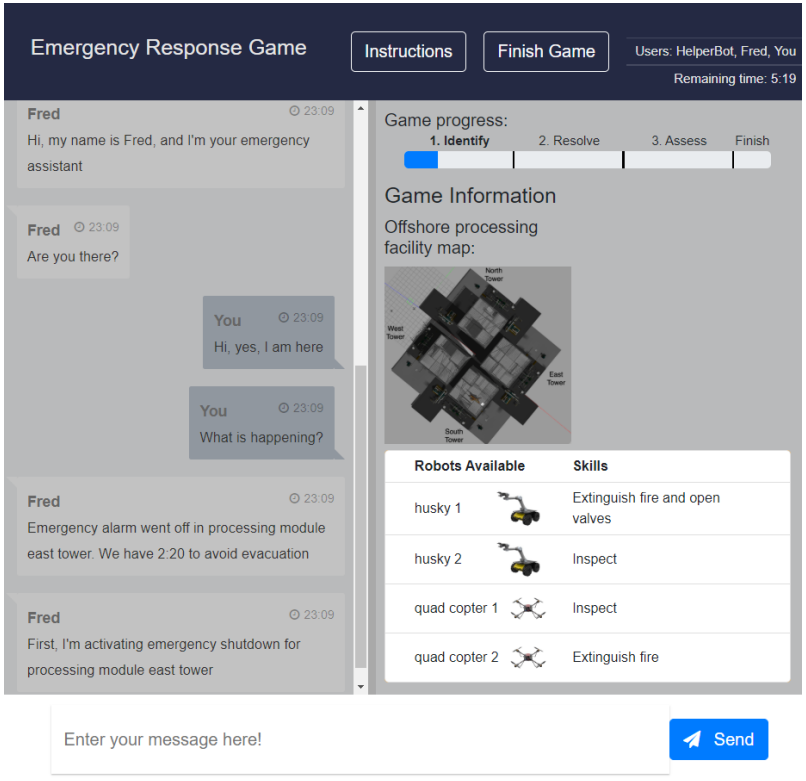
Wizard View



Participant View

MTurk Set-up





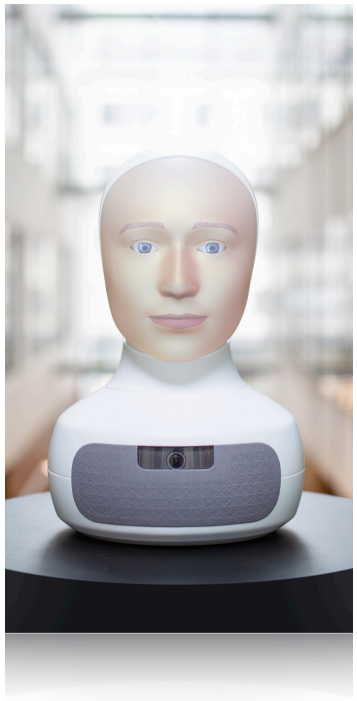
Wizard View

Participant View

CRWIZ: A Framework for Crowdsourcing Real-Time Wizard-of-Oz Dialogues. Chiyah Garcia, F.J., Lopes, J. and Hastie, H. In Proc. LREC 2020.

HERIOT WATT UNIVERSITY

Example Dialogue

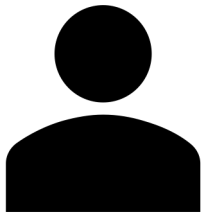


Hi, my name is Fred and I'm your emergency assistant.

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.

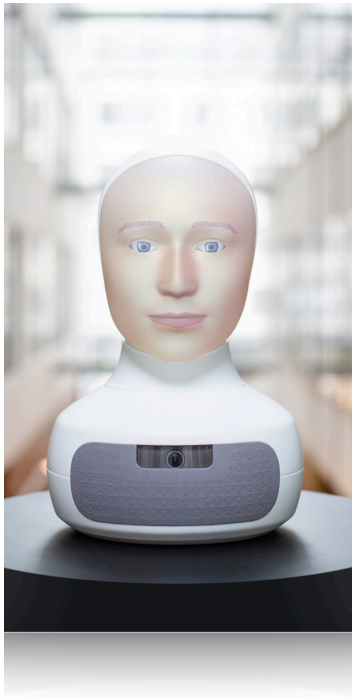


The robot is taking off. ETA is 32 seconds.

How long until evacuation?

HERIOT WATT UNIVERSITY

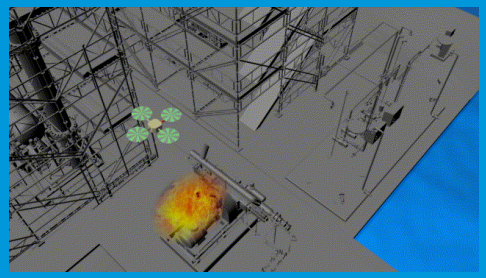
Example Dialogue



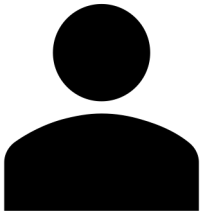
How long until evacuation?

We have 1:07 until evacuation.

The robot has arrived to processing module east tower.



Performing Inspection



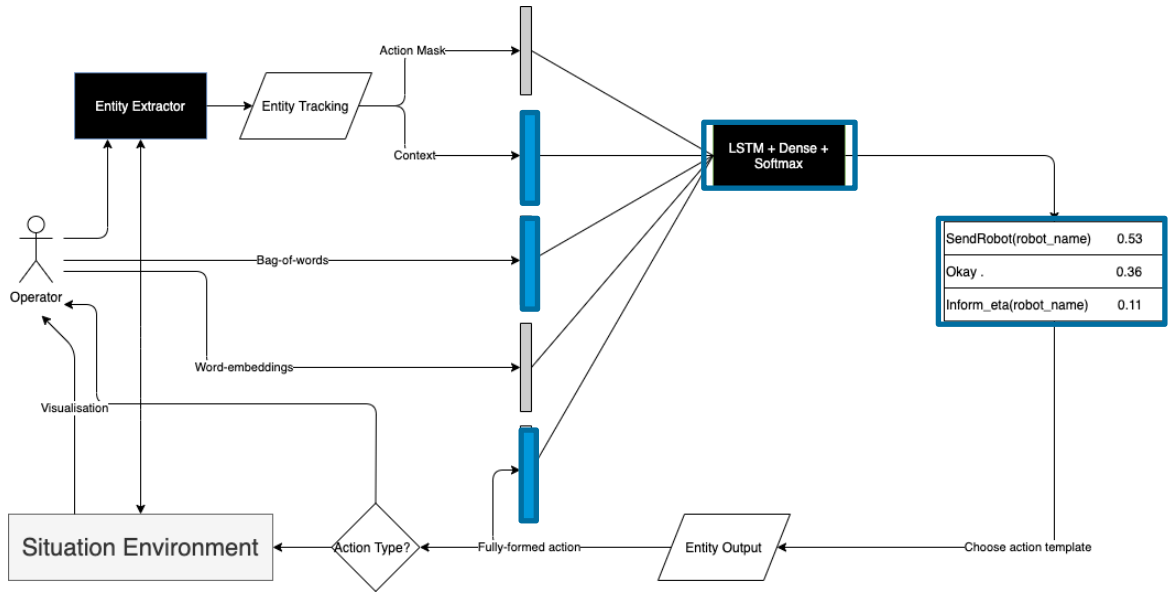
HERIOT WATT UNIVERSITY

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%

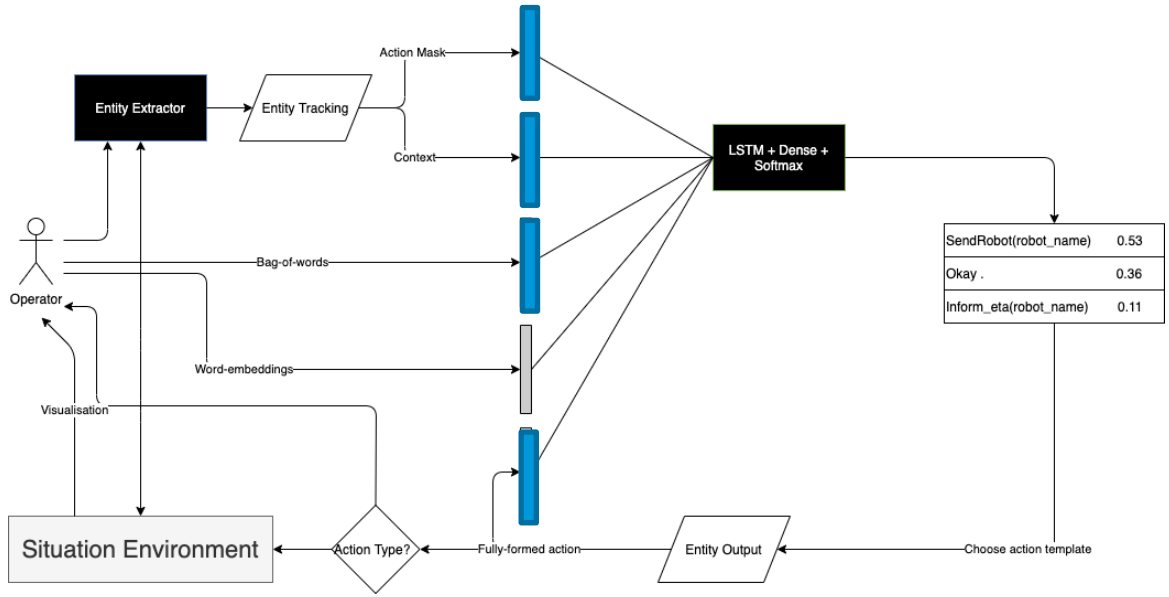
Training

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



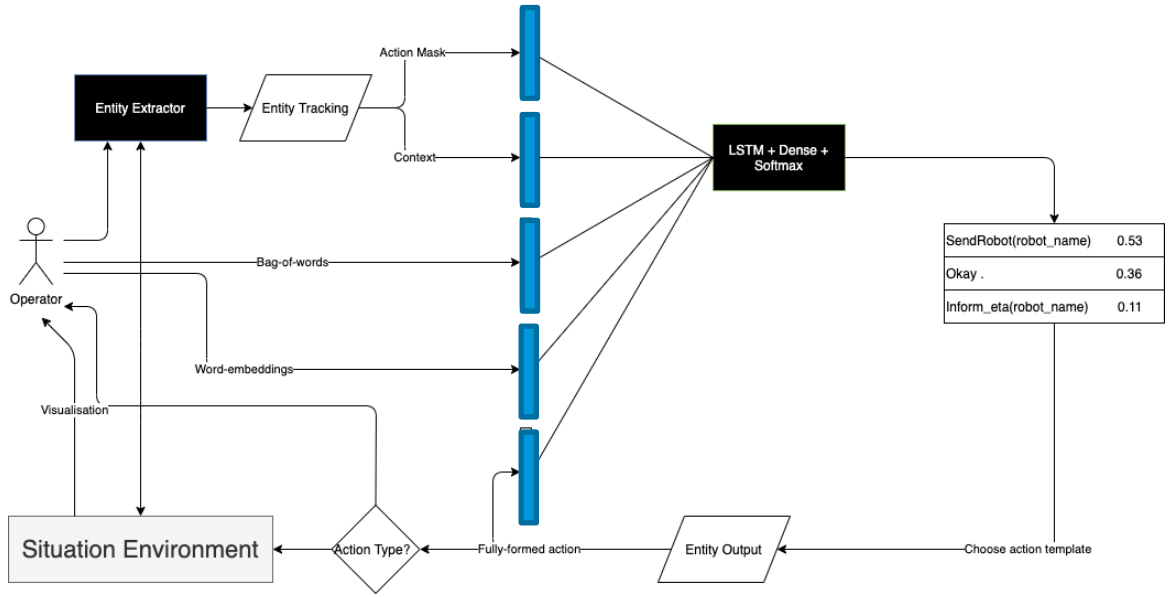
Training

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



Training

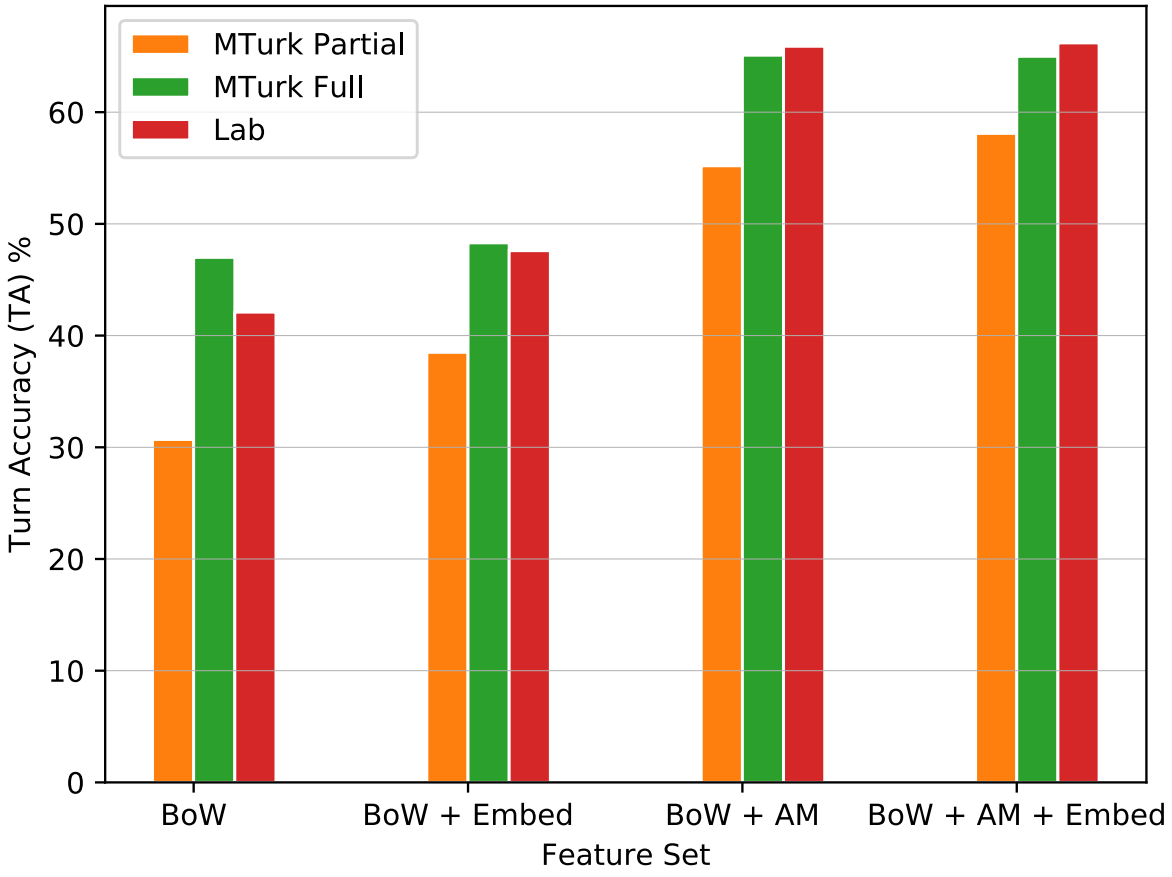
- Hybrid Code Networks (Williams et al, 2017)
  - Baseline: Previous actions, context and bag-of-words
  - Action Mask
  - Word2vec trained on Google News



HERIOT WATT UNIVERSITY

Results

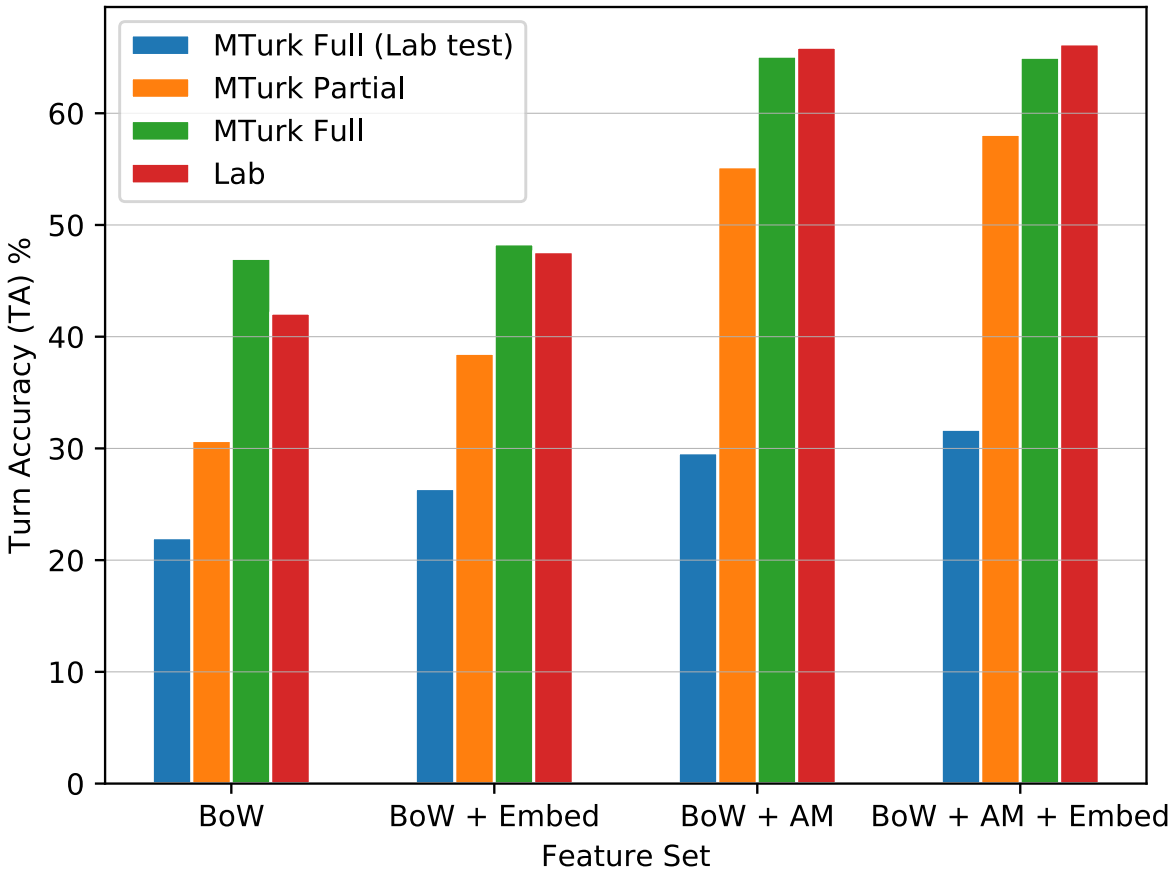
Within dataset: Leave-operator-out



HERIOT WATT UNIVERSITY

Results

Between dataset: 100 runs average



HERIOT WATT UNIVERSITY

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 %

HERIOT WATT UNIVERSITY

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 %**

HERIOT WATT UNIVERSITY

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

HERIOT WATT UNIVERSITY

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