Learning to Follow Navigational Instructions

April 30, 2017

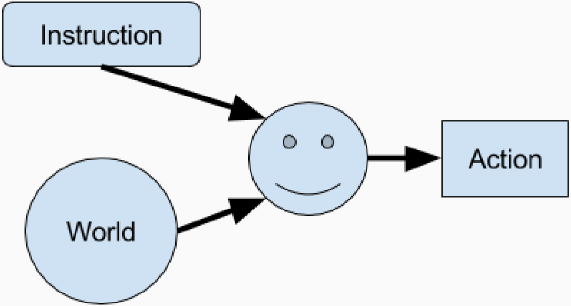
Koç University, Artificial Intelligence Laboratory

Table of contents

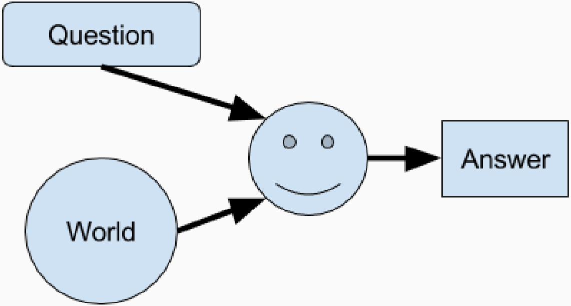
1. Grounded Language Learning
2. Task Description
3. Dataset
4. Previous State of the Art
5. Instruction categories
6. A new architecture
7. A new dataset
8. Conclusion

Grounded Language Learning

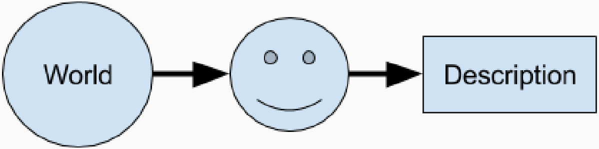
Following instructions



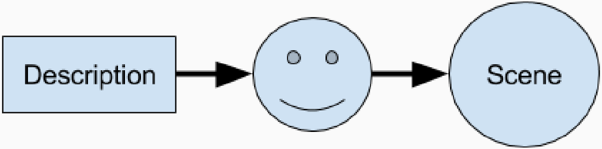
Grounded question answering



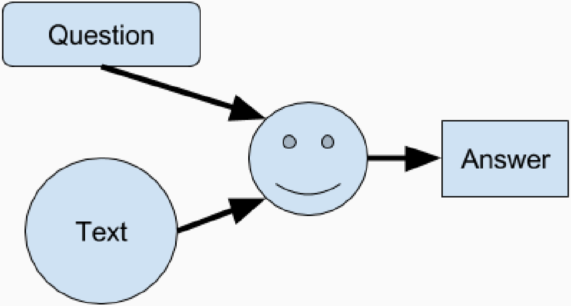
Generating descriptions



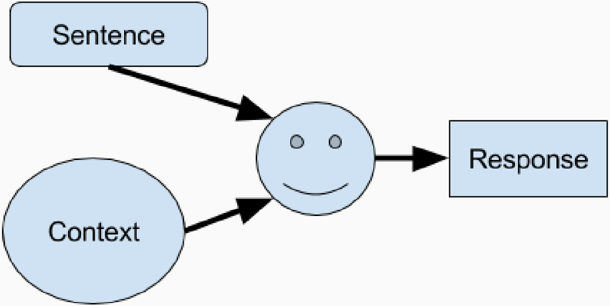
Understanding descriptions



Ungrounded applications: question answering



Ungrounded applications: chatbots



Task Description

Navigational Instruction Following

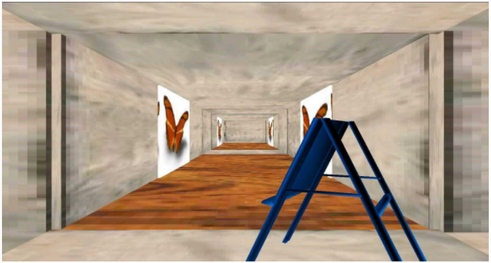


Figure 1: The first person view of the agent.

- Example: “Turn right at the easel.”
- Goal: generate the right sequence of *MOVE, RIGHT, LEFT, STOP*

Dataset

World

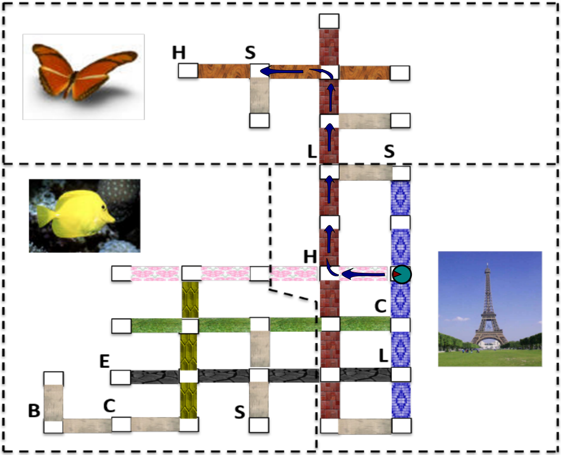


Figure 2: An illustration of a map and a set of instructions from MacMahon et al. (2006) where the letters indicate the items, figures demonstrate the wall paintings for the specific areas divided with dashed lines and floor patterns distinguish the flooring. The circle represents the initial position of the agent and blue arrows represent the execution of the instruction set: ”Take the pink path to the red brick intersection. Go right on red brick. Go all the way to the wood intersection. go left on wood. Position one is where the sofa is.”

Language

- Instructors and Followers
  - free-form language
  - syntactic and semantic errors
- A sequence of instructions for a (start, goal) pair
- Chen and Mooney (2011) split them into single sentences
- First version: *Paragraph*
- Second version: *Sentence*

Table 1: Number of instances

Map	Sentence	Paragraph
Grid	874	224
Jelly	1293	242
L	1070	236

Previous State of the Art

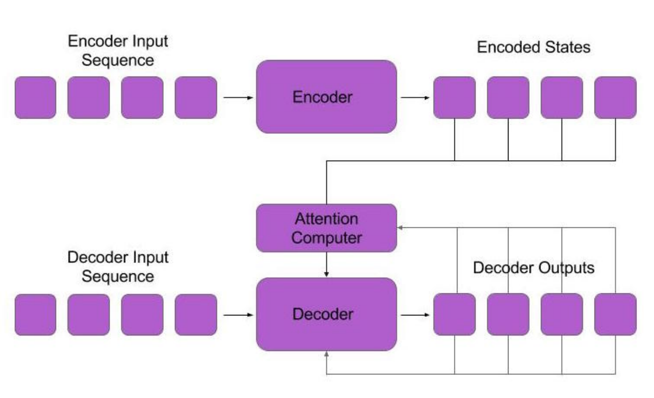
Comparison of previous studies

Method	Sentence	Paragraph
Chen and Mooney (2011)	54.40	16.18
Chen (2012)	57.28	19.18
Kim and Mooney (2012)	57.22	20.17
Kim and Mooney (2013)	62.81	26.57
Artzi and Zettlemoyer (2013)	65.28	31.93
Artzi et al. (2014)	64.36	***35.44***
Andreas and Klein (2015)	59.60	-
Mei et al. (2015) (vDev)	69.28	26.07
Mei et al. (2015) (vTest)	***71.05***	30.34
Human (MacMahon et al. 2006)	-	69.64

Table 2: Accuracy of reaching the final position.

The Model (Mei et al. 2015)

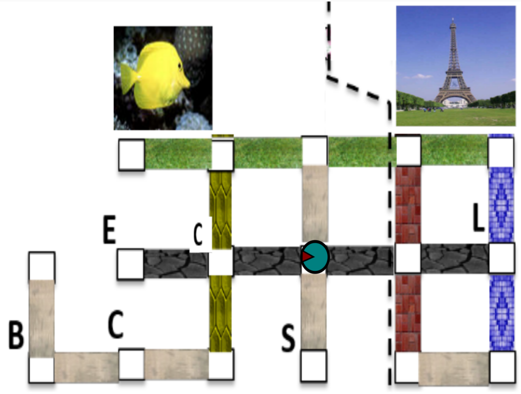
- Translating the instruction sequence to an action sequence
- Conditioned by the world knowledge
- Attention-based sequence-to-sequence neural network architecture



https://indico.io/blog/sequence-modeling-neural-networks-part2-attention-models

Bag-of-features world representation (Mei et al. 2015)

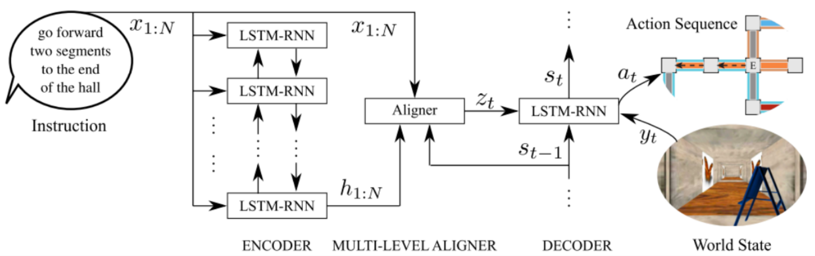
- Agent Centric
- Bag-of-features representation of each direction and the current position
- Spatial relations are not preserved
    - distance, order, relative position



Bit Representation:

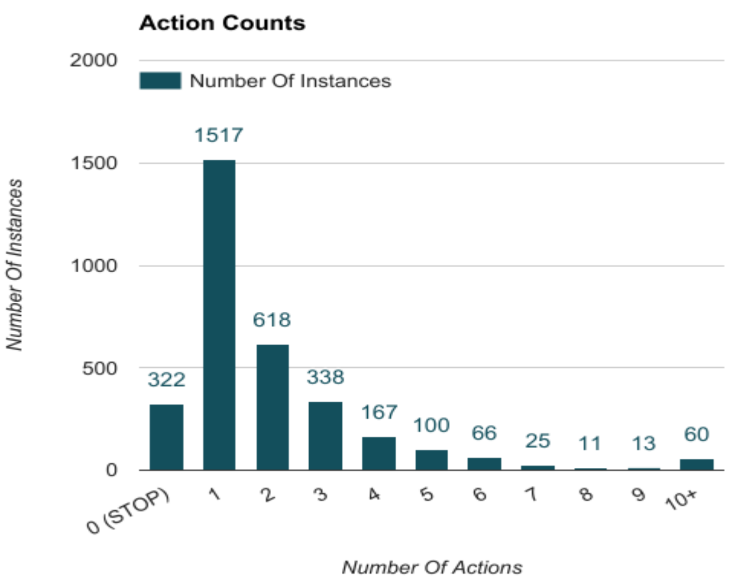
front_easel,
front_chair,
front_gravel,
front_fish,
left_sofa,
left_concrete,
left_fish,
back_lamp,
back_gravel,
back_fish,
back_tower,
right_concrete,
right_fish,
current_empty

Overview of the architecture (Mei et al. 2015)



Instruction categories

Action statistics



Instruction categories

Language Only (32%)

Turn (17%) : turn left, turn around, turn right twice
Move (13%) : move one step, go straight two blocks
Combination (2%) : walk forward once then take a left

Visual (68%)

Turn to X (7%) : face toward the hall with fish on the walls
Move to X (14%) : move until the wall, move to the chair
Orient (5%) : turn so the wall is on your back
Describe (10%) : there should be the brick path on your right
Conditional (9%) : move until you see the green path on your left
Combination (23%) : turn and move to the sofa, go towards the lamp on the brick road and take a right onto the grass, at the chair turn right

Language-only model

- No visual input to the decoder
- Previous actions instead of world representations

Category	Frq (%)	L.O. (%)
Language only	32	88.0
Turn to X	7	82.9
Move to X	14	63.2
Orient	5	89.8
Describe	10	86.9
Conditional	9	33.3
Any combination	23	21.8
Overall	100	63.7

Table 3: The performance of the language only model

Bag-of-features model

- Standard encoder-decoder architecture
- Bag-of-features world representation

Category	Frq (%)	L.O. (%)	BOF (%)
Language only	32	88.0	87.1
Turn to X	7	82.9	81.1
Move to X	14	63.2	70.5
Orient	5	89.8	89.2
Describe	10	86.9	81.5
Conditional	9	33.3	40.1
Combination	23	21.8	30.5
Overall	100	63.7	67.1

Table 4: The performance of the multi-hot model

A new architecture

Our model

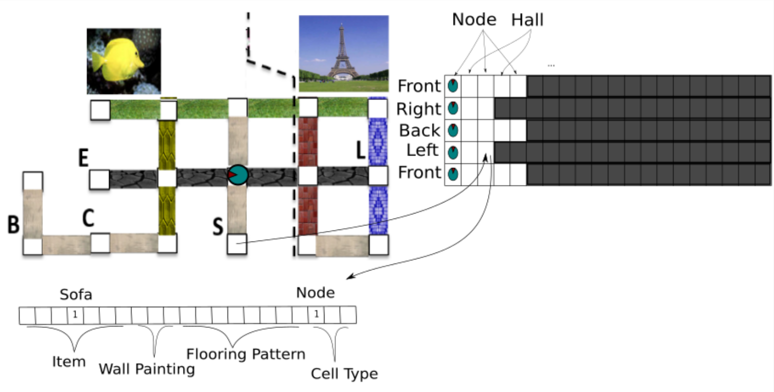


Figure 3: An example grid representation for the perceptual information.

A grid representation to allow the isometric information

Encoder

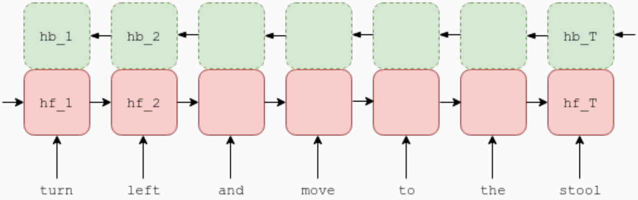


Figure 4: Encoder with a bidirectional Long Short Term Memory networks (LSTMs)

Decoder-BOF

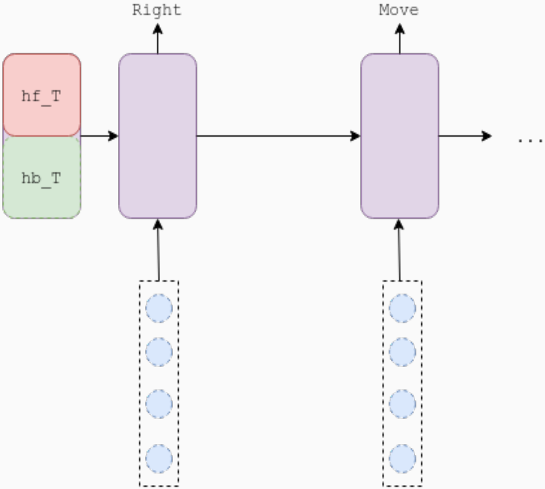


Figure 5: Decoder to process the bag-of-features representation.

Decoder-CNN

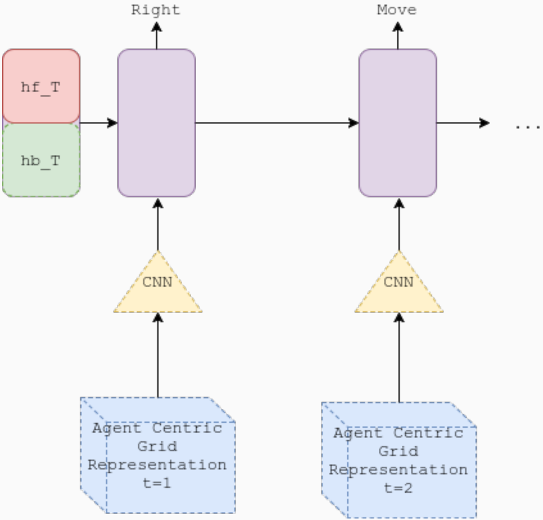

Right
Move
hf_T
hb_T
...
CNN
CNN
Agent Centric Grid Representation t=1
Agent Centric Grid Representation t=2

Figure 6: Decoder with a convolutional neural network to process the perceptual state.

Convolutional Neural Network

- Filters to detect objects and materials
- Good at recognizing local patterns
- Composition of lower-level features into higher level representation
- Location invariant

Our results

Category	Frq (%)	L.O. (%)	BOF (%)	CNN (%)
Language only	32	88.0	87.1	88.9
Turn to X	7	82.9	81.1	84.5
Move to X	14	63.2	70.5	74.8
Orient	5	89.8	89.2	94.0
Describe	10	86.9	81.5	79.9
Conditional	9	33.3	40.1	42.6
Combination	23	21.8	30.5	37.8
Overall	100	63.7	67.1	69.4
Overall (Ensemble)				71.74
Mei et al. (2015)				71.05

Table 5: The performance of the grid-based model

A new dataset

Motivation

- Solution for Data Sparsity
- Controllable Tasks
    - Language Complexity
    - World Complexity

Method - World Generation

- Generate a map randomly
- Decorate the map with floor and wall patterns
- Distribute the items randomly
- Generate random start and goal positions
- Find a path from start to goal position

Method - Language Generation

- Segment the path into turning and moving parts
- For each segment
    - Generate possible instructions using templates by conditioned the task category and the world configuration
    - Sample an instruction
- Task Patterns
    - reaching a corner, turning to a unique item, turning to a unique floor pattern
    - $\sim$ 5 patterns for each task
- Sentence Templates
    - move to the ”corner” / ”end of the path/hall/corridor”
    - ”turn” / ”turn your face” to the ”sofa” / ”bench”
    - $\sim$ 8 template for each pattern
- Vocabulary size is 196 (Original is 511)

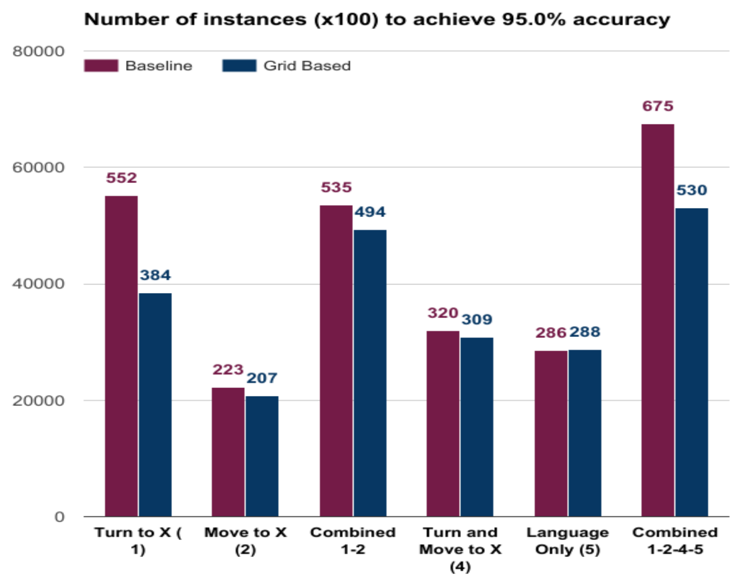
Coverage

Table 6: Coverage statistics of the artificial data

Category	Overall	Non-Unique
Language only	88.7%	96.8%
Turn to X	41.8%	74.1%
Move to X	33.9%	58.3%
Orient	65.1%	94.5%
Describe	11.2%	58.8%
Conditional	5.3%	29.4%
Any combination	8.7%	37.4%

- Original dataset contains 1972 unique instructions
- 4% contains misspelled words
- 498 instructions occur more than 1

Convergence results with the new dataset



Conclusion

Summary

We present

- a new world representation
- a cnn-based new architecture
- a new dataset
- model comparison by sample complexity

Questions?