Behavioral Indoor Navigation With Natural Language Directions

Xiaoxue Zang[1], Marynel Vázquez[1], Juan Carlos Niebles[1], Alvaro Soto[2], Silvio Savarese[1]

[1] Stanford University
xzang, marynelv, jniebles, ssilvio@stanford.edu

[2] P. Universidad Catolica de Chile
asoto@ing.puc.cl

ABSTRACT

We describe a behavioral navigation approach that leverages the rich semantic structure of human environments to enable robots to navigate without an explicit geometric representation of the world. Based on this approach, we then present our efforts to allow robots to follow navigation instructions in natural language. With our proof-of-concept implementation, we were able to translate natural language navigation commands into a sequence of behaviors that could then be executed by a robot to reach a desired goal.

KEYWORDS

Human-Robot Interaction; Navigation; Verbal Communication

ACM Reference Format:
X. Zang, M. Vázquez, J. C. Niebles, A. Soto, S. Savarese. 2018. Behavioral Indoor Navigation With Natural Language Directions. In *HRI '18 Companion: 2018 ACM/IEEE International Conference on Human-Robot Interaction Companion, March 5–8, 2018, Chicago, IL, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3173386.3177001

1 INTRODUCTION

Currently, most popular approaches for robot navigation rely on precise, geometric world representations, e.g., metric maps [7]. While these representations have proven successful on a variety of applications, there are situations where sensor occlusion or noise can affect precise localization, which is essential for these popular approaches. Interestingly, we, humans, often navigate indoor environments without precise geometric information. We do not tend to keep track of our exact (x, y) coordinates on the ground when moving to new locations. This observation inspired us to think about robot navigation without metric world representations, and to revisit the idea of graph-based cognitive maps [3].

The approach that we are exploring takes advantage of the rich semantic structure behind man-made environments, which are intrinsically designed to facilitate human navigation. These environments are mainly composed of *navigational structures*, such as corridors, or stairs, that in turn are intended to connect meaningful neighboring places, such as rooms, or halls. Our hypothesis is that by providing robots with suitable abilities to understand the world at this semantic level, it is possible to provide them with navigational systems that can exceed the generality and robustness of current methods. Moreover, these abilities can facilitate navigation under human commands, because humans understand and describe their environments at this level of abstraction.

The following section describes our behavioral approach for indoor navigation, which first appeared in [6]. Based on this approach, we present our current efforts to enable robots to follow navigation instructions in natural language. We conclude by discussing avenues for future research, especially in human-robot interaction.

2 BEHAVIORAL NAVIGATION APPROACH

The proposed robot navigation approach [6] revisits the ideas of behavioral robot control [1] and early topological map representations [3]. While early attempts at behavioral navigation lacked of sufficient robustness to deal with the complexities of natural environments, we can now exploit recent advances in Deep Imitation Learning to learn motion behaviors [2]. In our approach, complex navigation routes can be achieved by composing simple, parameterized visuo-motor behaviors that leverage the semantic structure of indoor environments. This composition is realized by planning on a directed graph that represents valid relations among navigational structures, as illustrated in Fig. 1b for the layout in Fig. 1a. The nodes of the graph correspond to semantically-meaningful places, such as offices or corridors. The edges correspond to visuo-motor navigation behaviors that the robot counts with to move from one place to another, such as "follow the corridor" or "leave (the office) and <turn right, turn left, go straight>" to enter a hall.

3 FOLLOWING NAVIGATION INSTRUCTIONS

Inspired by prior work [4, 5], we cast the problem of following navigation instructions in natural language as a *machine translation task*. We want to translate natural language commands, such as "go out of the office and enter the conference room on the left", into a graph representation that encodes navigational structures and behaviors as described previously. The robot can then execute the visuo-motor behaviors along the route in the graph to reach a desired destination. Note that this approach differs from [5] in that we do not output low-level motion commands directly, but estimate a behavioral graph that can facilitate reasoning at a semantic level of abstraction. This graph can be considered a topological map [3] but, different to [4], it does not encode metric information explicitly.

3.1 Translating Natural Language Commands

As illustrated in Fig. 1c, we pose the translation problem as estimating a function f that maps navigation commands to a sequence: $S =< p_1, b_1, R_1, p_2, b_2, R_2, ..., p_n >$ of places p, behaviors b, and sequences of references R. The places are semantically-meaningful locations, like offices, halls, or kitchens in the route. The behaviors correspond to motion policies that the robot can execute to

HRI '18 Companion, March 5–8, 2018, Chicago, IL, USA

ACM ISBN 978-1-4503-5615-2/18/03.
https://doi.org/10.1145/3173386.3177001

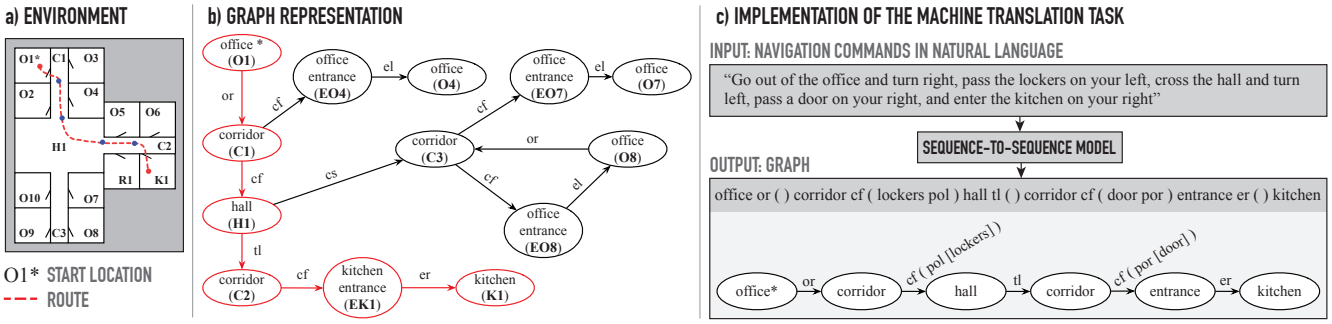

Figure 1: (a) Environment, (b) partial graph representation for behavioral navigation, and (c) implementation of the translation task. The desired route is highlighted in red in (a) and (b). The codes "or", "cf", "cs", "tl", "er", and "el" correspond to the behaviors "go out and turn right", "follow the corridor", "cross straight", "turn left", "enter on the right", and "enter on the left", respectively. The reference actions "pol" and "por" correspond to "pass on the left" and "pass on the right".

transition between places in S. Finally, the sequences of references $R =< "(", l_1, r_1, l_2, r_2, ..., ")" >$ are ordered collections of zero or more reference actions r grounded on landmarks l. These references guide the execution of the prior behavior b in S by providing complementary information about the route. The landmarks correspond to distinguishable places or objects. For example, a reference action could be "pass a door on the right" while following a corridor.

As an initial proof of concept, we implement the function f as a differentiable sequence-to-sequence deep learning model with attention [8]. The places in the output sequence correspond to the nodes in the graph, and the behaviors correspond to the edges. We encode the references actions as attributes of the edges of the graph, such that they can guide behavior execution.

Implementation Details. We trained the function f using an adaptation of the learning environment DeepMind Lab.[1] We used one-hot encodings for the input and output sequences (the vocabularies had 45 and 31 codes, respectively). For the encoder and decoder of f, we used single layers of Gated Recurrent Units (GRUs). Training was performed with the Adam optimizer, 0.001 as learning rate, cross-entropy loss, and a batch size of 128. At test time, we output the predicted sequence with the highest probability among a finite set generated through beam search.

Experiment. We evaluated our approach on a dataset of 16,370 unique examples using 5-fold cross-validation. This dataset was created by sampling routes of varied length on 250 synthetic environments. For each route, we generated example natural language instructions by composing simple path descriptions, and created graph representations to encode relevant navigational structures. The target sequences S in the dataset were composed of 40.18 elements on average (STD = 9.75). As in Fig. 1c, we considered situations where the robot had to follow a corridor while passing references, but the "follow the corridor" action was not said explicitly in the instructions. This consideration increased the diversity of commands, and forced the model to learn to complete output sequences to successfully generate behavioral graphs.

Because the output graph could not be used to navigate unless it was structured correctly, we report a strict measure of accuracy: a prediction was valid if it matched the ground truth exactly. Our results are presented in Table 1, based on the number of GRU units in our model. Overall, our translation function reached high average accuracy (above 99%). With less than 32 units, we observed the performance drop significantly and learning often became unstable.

[1] https://github.com/deepmind/lab

Table 1: Accuracy of our model on the testing set.

# GRU Units	Avg. Accuracy	Std. Dev.
64	99.53%	0.01
128	99.99%	0.0002
256	99.99%	0.0002

4 DISCUSSION & FUTURE WORK

We described an approach for indoor navigation that leverages the semantic structure of man-made environments. In our current work, we are taking advantage of the affinity between the behavioral graph at the core of this approach and the way humans navigate to enable robots to follow navigation instructions in natural language. Using modern machine translation methods we were able to convert instructions to sequences of behaviors that a robot could execute to reach a desired goal. In future work, we plan to consider free-form route instructions and investigate mechanisms to detect erroneous or ambiguous directions in known environments. This ability could improve users' perception of the robot and their interaction.

ACKNOWLEDGEMENTS

The Toyota Research Institute (TRI) provided funds to assist with this research, but this paper solely reflects the opinions and conclusions of its authors and not TRI or any other Toyota entity. We are also thankful to G. Sepúlveda for his assistance with data collection.

REFERENCES

[1] R. Brooks. 1986. A robust layered control system for a mobile robot. *IEEE J. Robot. Autom.* 2, 1 (1986), 14–23.
[2] S. Gupta, J. Davidson, S. Levine, R. Sukthankar, and J. Malik. 2017. Cognitive Mapping and Planning for Visual Navigation. *CoRR* abs/1702.03920 (2017).
[3] B. Kuipers. 1978. Modeling spatial knowledge. *Cogn. Sci.* 2, 2 (1978), 129–153.
[4] C. Matuszek, D. Fox, and K. Koscher. 2010. Following Directions Using Statistical Machine Translation. In *HRI*.
[5] H. Mei, M. Bansal, and M. R. Walter. 2016. Listen, Attend, and Walk: Neural Mapping of Navigational Instructions to Action Sequences. In *AAAI*.
[6] G. Sepulveda, J. C. Niebles, and A. Soto. 2018. A Deep Learning Based Behavioral Approach to Indoor Autonomous Navigation. In *ICRA'18*.
[7] S. Thrun, W. Burgard, and D. Fox. 2005. *Probabilistic Robotics.* MIT Press.
[8] O. Vinyals, L. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. E. Hinton. 2014. Grammar as a Foreign Language. *CoRR* abs/1412.7449 (2014).