Critical Survey of Translating Navigation Instructions in Natural Language to a High-Level Plan for Behavioral Robot Navigation

Mica Haney
University of North Texas
micahaney@my.unt.edu

Abstract

This paper is a critical survey of (Zang and Pokle et al., 2018), which discusses an approach to the task of robot navigation using human-generated NLP instructions. This survey will discuss the dataset generated, the architecture of the system, the researcher's findings, and comment critically on where the paper had issues with either communication or the underlying research.

1 Introduction

The project described in (Zang and Pokle et al., 2018) paper focuses on exploring three tasks of interest: language translation, robot navigation, and task planning as a subset of the other two. The project's task can be described as taking spoken instructions from a human on how to get from one room to another on a floor of a building, translating that to an executable plan for the robot, and then the robot successfully moving from it's starting position to the correct room.

Neither the topic of robot navigation nor instruction understanding were new when this paper was written. Both tasks have seen considerable attention since the early days of computing. However, there has not been a dataset that combines the two tasks and one of the main contributions of this paper is that the authors have contributed the first combined navigation-NLP dataset.

2 Data

2.1 Description

The data for this project was not already available, and so was collected by the researchers. They generated 100 maps of indoor floor plans with between 6 and 65 rooms. Each map was generated from combinations of 7 different room types and 20 kinds of landmarks such as vases and windows, as well as 11 different behaviors for how a robot could navigate between rooms. A number of paths were generated off of these maps. The natural language instructions were collected using Amazon Mechanical Turk by giving each path to one or more annotators and having them provide English instructions for how to follow the path.

The dataset is organized by instruction-path pairs being linked to the map that the route was generated from. Each map is represented by an unordered collection of behavioral navigation graph triplets as described in (Sepulveda et al., 2018). Each unit of the collection is a triplet of navigational rules with a start destination, executable behaviour, and an end destination. Each path is represented by an alternating string of room-behavior-room tokens. The dataset was split into three sections, one for training and two for testing. One of the testing splits contains routes that were not a part of the training split, however the maps used to generate these routes were used to generate routes in the training split so the model will have trained over the maps in this testing split. This split is the test-repeated set. The other testing split is completely unseen by the model as both paths and maps have not been trained over. This is the test-new set.

It should be noted that, at the time, robot navigation tasks tend to focus on practical and highly detailed implementations. This dataset instead focuses on high-level representations of the maps and behaviors.

2.2 Critical Commentary

The largest concern with this project has to do with the data and input. The researchers decided to assume that the environment was known and mapped out in a behavioural navigation graph, and that map was fed into the model. The problem with assuming that the environment is known and can be mapped into a behavioural navigation graph

is that this is not true of many environments, and it does not generalize well to tasks that cannot be well represented using a behavioral navigation graph. Admittedly this is a restriction that is practical in a project with such a narrow scope, but it makes this research difficult to apply to problems too divergent from this particular task.

A large point of criticism on this topic that the number of paths with translations is reported incorrectly. In the text the authors claim 11,050 path-instruction pairs, however their table only totals to 10,040 pairs and an analysis of the dataset via download only shows 7,916 pairs. The discrepancy in the paper is likely an unfortunate if understandable mistake made through a lack of attention to detail. The massive difference between either of the paper's numbers and the actual dataset's counts, however, is alarming. At the least the dataset is missing 2,124 path-instruction pairs, and this is a significant amount of data, particularly given the size of the dataset. It begs the question of where the 21.16%-28.36% of the data is.

A third issue is the representation of the maps. Each map is represented by a collection of triplets, each with two rooms and a connecting behavior. In representing the maps this way, the researchers restrict a robot's moves based on the room they are in. For example, if the model never encounters any rooms with two doors, then the model will never learn that rooms can serve a similar purpose to hallways by serving as another thoroughfare. This restriction of the map presents challenges at the practical implementation level. While a system may learn to traverse the map representations well, without any indication of where room and hallway junctions are there will need to be a separate system that learns where doors are, and that they serve as exits and entrances.

3 System Architecture

The architecture of the model modifies the sequence-to-sequence model described in (Bahdanau et al., 2015). They modified the first layer to take not just the natural language instruction as input, but the entirety of the map that the path-instruction pair was generated on.

Embed: This layer embeds the English instruction into a pre-trained GloVe vector (Pennington et al., 2014) of 100 dimensions. The maps are converted into one-hot encoded vectors.

Encode: Two Gated Recurrent Units (Cho et al., 2014) process data from the English input and the map separately while incorporating the embeddings from the previous layer. This layer outputs two matrices that encode navigation commands and behavior graphs (separately).

Attention: The two matrices from the previous layer are combined via one-way attention. The attention acts to highlight what parts of the map representation the instruction is attending to.

FC: This layer reduces the dimensional of the tensors outputted by the previous layer.

Decoder: A GRU network predicts probability over the behaviors that correspond to the English instructions. This network uses the attention vectors computed in the previous attention layer.

Output: This layer takes the logits outputted by the decoder layer and the representation of the map and searches for a valid sequence of behaviors. There is a mask applied to restrict the model to only valid behaviors at the current location.

4 Experimentation

The researchers compared their model with three others. One is a baseline model based on the work of (Bahdanau et al., 2015; Shimizu and Haas, 2009; Zang and Vazquezet et al., 2018) that uses a sequence-to-sequence model with an attention method, followed by a depth-first search to produce predicted routes. The second model is an ablation study that removes the maps from the input of the model, and drops the FC layer and the masking function in the output layer due to their being unnecessary. The third model is the second model with a masking layer added to the output layer.

In an attempt to aid the model in finding the starting location of the graph, some models were fed the map representation in a deliberate ordering where the triplets of the representation near the start were first in the representation.

They use four metrics for evaluation: exact match (EM), F1 score, edit distance (ED), and goal match (GM). EM is a binary value for whether the predicted path is an exact match or not. ED is the number of changes necessary to the predicted string of tokens to change it to the ground truth. GM is a binary value for whether or not the predicted path ends at the ground truth destination.

Model	Test-Repeated Set				Test-New Set			
	EM ↑	F1 ↑	ED ↓	GM ↑	EM ↑	F1↑	ED ↓	GM ↑
Baseline	25.30	79.83	2.53	26.28	25.44	81.38	2.39	25.44
Ablation	36.36	90.28	1.36	36.36	24.82	88.65	1.71	24.92
Ablation with Mask	45.95	90.08	1.20	46.05	36.45	88.31	1.45	36.56
Ours without Mask	52.47	91.74	0.95	53.95	21.94	87.50	1.78	22.65
Ours with Mask	57.31	91.91	0.91	57.31	38.52	88.98	1.32	38.52
Ours without Mask and with Ordered Triplets	57.21	93.37	0.79	57.71	33.36	**91.02**	1.37	33.78
Ours with Mask and Ordered Triplets	**61.17**	**93.54**	**0.75**	**61.36**	**41.71**	90.22	**1.22**	**41.81**

Table 3: Performance of different models on the test datasets. EM and GM report percentages, and ED corresponds to average edit distance. The symbol ↑ indicates that higher results are better in the corresponding column; ↓ indicates that lower is better.

Figure 1: A copy of the results table from (Zang and Pokle et al., 2018).

5 Findings

5.1 Description

The first item of note is that the final model using the ordered map representation triplets overall performs the best, only behaving worse on the test-new set than the final model without the ordered triplets. The researchers concluded that this meant that the added information of the map representation was beneficial, largely because of by how much the addition of ordering improved scores.

The researchers also noted that adding a mask tended to improve scores on the test-repeated set. On the test-new set, the trend holds with increased performance from the addition of a mask.

The authors also did qualitative evaluations including attention visualization and path experimentation.

The attention visualization was a mapping of attention made on the map representation compared with an image of the map. It was noted that there was a high amount of attention on areas around the robot's current location, indicating that the model was prioritizing nearby nodes when choosing the next action.

For path experimentation, the authors gathered some paths and instructions that were not the shortest paths, and ran them through the model to determine whether the model was simply selecting the shortest path or if it was actually following the instructions. It was concluded that the model was indeed following the instructions.

5.2 Critical Commentary

The authors of the paper did not discuss that the improved mask scores were only improved between the Ours without Mask and Ours with Mask. When comparing the ablation models, the mask appears to have almost the opposite effect.

The section about the path experimentation highlighted that the dataset had a potential issue with all of its paths being the shortest path without any non-shortest paths mixed in. Notably in both this section and the attention visualization section it is never noted how many graphs were made nor how any analysis was done or what any result values there might be. While the premise of these analyses is qualitative there is still quantitative value to be had, and it is not reported. This lack throws the conclusions drawn in these sections into question, as there appears to be only cursory evidence for support.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations (ICLR)*.

Kyunghyun Cho, Bart van Merrienboer, aglar Gulehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

G. Sepulveda, JC. Niebles, and A. Soto. 2018. A Deep Learning Based Behavioral Approach to Indoor Autonomous Navigation. In *International Conference on Learning Representations (ICRA)*.

Nobuyuki Shimizu and Andrew R. Haas. 2009. Learning to follow navigational route instructions. In *International Joint Conferences on Artificial Intelligence (IJCAI)*.

Xiaoxue Zang, Ashwini Pokle, Marynel Vázquez, Kevin Chen, Juan Carlos Niebles, Alvaro Soto, and Silvio Savarese. 2018. *Translating Navigation Instructions in Natural Language to a High-Level Plan for Behavioral Robot Navigation*. arXiv:1810.00663.

Xiaoxue Zang, Marynel Vazquez, Juan Carlos Niebles, Alvaro Soto, and Silvio Savarese. 2018. Behavioral indoor navigation with natural language directions. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 283–284.