Navigational and Spatial Expressions in Natural Language: An Analysis

Kourosh T. Baghaei, Antonios Anastasopoulos, Dieter Pfoser

{kteimour, antonis, dpfoser}@gmu.edu

Motivation

- LLMs seem to be effective in development of **modular Vision and Language navigation agents**.
- They fail in many cases
- They also succeed unexpectedly in certain settings.

Research Questions:

- Where do LLM-based agents fail?
- What makes some routes easier or harder for agent?
- How good are our datasets?

Contributions

- Comprehensive evaluation of **LLM-based VLN agents**.
- Development of a dataset that address major shortcomings with state of the art datasets.
- Proposing new Metrics in Analysis of VLN trajectories

Highlights (Conclusion)

- Regardless of **visual data**, the agent succeeds in navigating routes using instructions of other routes with **similar pattern of actions** (up to 25% of routes that the agent was able to navigate)
- The **diversity of patterns** in routes needs to be higher in datasets

Future Directions

- Development of datasets from cities other than New York
- Split Train/Test based on non-overlapping patterns

Vision and Language Navigation

Problem Setting:

Input Variables:

- **r:** route, a sequence of nodes on graph of street network
- **v**: 360-view image of each node
- **t:** instruction in human language describing the route

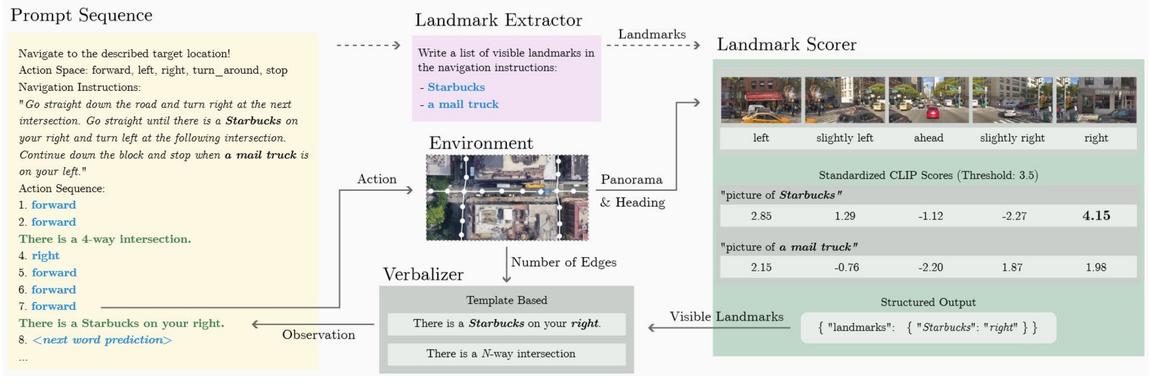
Task:

- P(a| r,v,t) : Prediction of actions that take the agent to the destination.

Datasets:

Dataset	Annotator's POV	Train Size	Dev Size	Test Size	# of Patterns
TouchDown	Ego-Centric	6,770	800	1507	66
Map2Seq	Map View	5,737	800	800	32

VELMA : LLM-based vision and language navigation

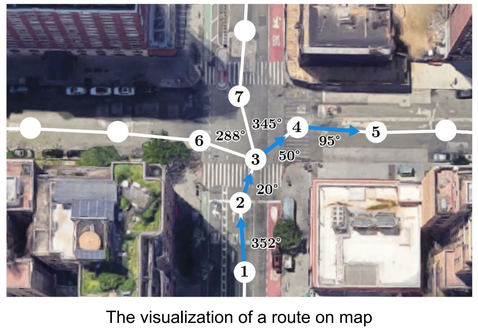


Pattern of Actions

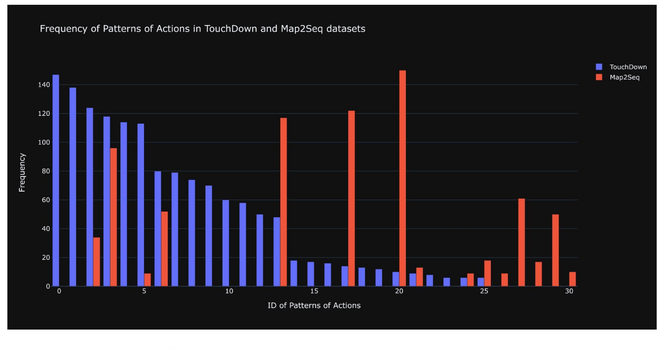
- A compressed representation of **pattern of actions** required to successfully navigate a route:

e.g. Forward, Forward, Right, Forward, Stop ⟹ **frfs**

- It also represents **shape of trajectory**



The visualization of a route on map



Frequency of pattern of actions in test sets of TouchDown (left) and Map2Seq (right)

Pattern-based Analysis

Hypothesis:

- The **shape of a trajectory** is a contributing factor in success of the agent.
- To test the hypothesis, we swap instructions of routes with similar patterns (simpat) and different patterns (difpat) and compare the performance of agent

Baseline (original):

route	r1	r2	r3	r4	r5	r6
text	t1	t2	t3	t4	t5	t6
pattern	a	a	a	b	b	c

Similar Pattern (SimPat):

route	r1	r2	r3	r4	r5	r6
text	t3	t1	t2	t5	t4	x
pattern	a	a	a	b	b	c

Different Pattern (DifPat):

route	r1	r2	r3	r4	r5	r6
text	t4	t5	t6	t1	t2	t3
pattern	a	a	a	b	b	c

Evaluation Results

- **Visual Data vs. No Visual Data:** In some cases the agent's performance does not change with or without visual data
- **Fine-tuning dataset:** difference in performance roots in POV of annotator in data collection process

Fine-tuend on	Test	Swapped with...	Image	No-Image
		Similar	4.97	2.82
	TouchDown	Different	2.92	1.46
		Base	**20.9**	**11.48**
TouchDown				
		Similar	4.56	5.32
	Map2Seq	Different	2.25	2.13
		Base	**23.5**	**22.75**
		Similar	2.96	2.89
	TouchDown	Different	1.19	1.53
		Base	6.17	5.31
Map2Seq				
		Similar	5.96	6.21
	Map2Seq	Different	1.88	1.38
		Base	**39.13**	**33.75**