Navigational and Spatial Expressions in Natural Language: An Analysis

Kourosh T. Baghaei, Antonios Anastasopoulos
Department of Computer Science, George Mason University
Dieter Pfoser
Department of Geography and Geoinformation Science, George Mason University
{kteimour,dpfoser,antonis}@gmu.edu

Abstract

Grounding and representation of navigational and spatial concepts is challenging in natural language processing. For this purpose, in this project, we designed a collaborative two-player game in which the players co-operate in order to finish a task of reaching to a destination by starting from a starting point. Through this game, we collect data of textual communications of humans that involve providing directions and spatial descriptions of the street view environment. Finally, we examine the areas in which vision and language navigation models succeed for fail, providing insights into the areas requiring improvement.

1 Introduction

Humans tend to describe routes using a combination of referrals to landmarks, directions and distances (Vasudevan et al., 2021). The frequency of referring to landmarks varies based on familiarity of the individual with the environment of the routes and their abilities in understanding orientations and directions. Nonetheless, landmarks constitute an important part of describing spaces and navigational routes beyond the ego-centric frame of reference (e.g. front, left, right, etc) (Ishikawa and Nakamura, 2012). This human cognitive ability consists of visual perception and linguistic power, and it has attracted AI researchers in attempting to emulate it, particularly with the emergence of robots and autonomous vehicles (Wu et al., 2021).

In the literature, various datasets and methods have been proposed to provide a benchmark for this task, like TouchDown (Chen et al., 2018), TalkTheWalk (de Vries et al., 2018), and Talk2Nav (Vasudevan et al., 2021). None of these, however, focus on studying the humans giving or following navigation instructions (their focus is on collecting data to train navigational models).

Motivation Following Vasudevan et al. (2021), we focus on explanations of routes by landmarks in human language. In order to prevent obsoletion of the routes due to changes in panorama images, we generate our own set of routes in the neighborhood covered in (Mirowski et al., 2018). We develop our custom interface using Unity3D for exploring the graph of the city map in offline mode (i.e. without the need to use the Google Street View API).

Contributions We have developed an interactive two player environment (game) using Unity3D[1] that will allow us to study the human perception of space and collect a dataset of realistic route descriptions.

Research Questions We study and compare our dataset to the previously published datasets. To do so, we collect basic statistics (e.g. path length, number of landmarks along route, etc) and recreate routes with statistically similar characteristics. We seek to answer the following questions:

- How much do human descriptions of the same route vary? In other words, given the same route, which landmarks would different people use to provide instructions?
- What makes navigation instructions hard or easy to follow?
- What are the main causes of failures of vision and language models?

2 Related Work

Initially the dataset of panorama images of Brooklyn in NY was published along with a navigation method only based on visual cues Mirowski et al. (2018). In an effort to bridge the gap between the problem of spatial reasoning in natural language and visual navigation (referred to as Visual and Language Navigation), Chen et al. (2018) proposed a dataset created through crowd-sourcing. They generate numerous routes and asked each participant to provide descriptions on how one can get

[1] https://unity.com/

from a starting point to the corresponding destination point. Additionally, Mehta et al. (2020) proposed models of VLN along with additional data. To further encourage the description providers to generate more comprehensible textual descriptions, de Vries et al. (2018) proposed a collaborative task of navigation known as *Talk the Walk*. In this task, each route description is provided by a guide to a tourist who is supposed to follow the route to reach the destination only by following the textual instructions of the guide. In this way, a guide who provides textual instructions is able to refine the provided explanations if they are perceived not helpful by the tourist. There are various shortcomings with the aforementioned datasets, that we seek to address in our research.

- The textual descriptions are monologue Chen et al. (2018); Mehta et al. (2020). Despite the fact that each textual description has been evaluated by another person, the overall process of generating text was performed without direct human feedback.
- The textual data consists of textual conversations, though without particular emphasis on the landmarks de Vries et al. (2018).
- Although the textual follows a certain template (i.e. combination of landmark and directions), the routes covered by the dataset is expanded over a large area of the map, for which panorama image data is not publicly available. Hence, limiting the research and exploration.

To address these problems, we develop our own collaborative game for data generation. We seek to generate textual data of route descriptions with emphasis on landmarks within the Manhattan neighborhood of NY, for which image data is publicly available Mirowski et al. (2018).

3 Methodology

Inspired by de Vries et al. (2018); Vasudevan et al. (2021) we propose a two-player interactive framework in which players co-operate to complete a given task in a game environment. The task involves in navigation from a random point on the map and reaching to a destination point, based on visual cues and linguistic instructions. In this task, a player that has access to map of the environment, takes the role of a **guide** and provides instructions on how to reach from a starting point to the corresponding destination. On the other hand, the player who does not have any access to the map, takes the role of a **navigator** has to follow the instructions to reach the destination. Navigator can always communicate with the guide for further clarifications.

We compare the performance of a state of the art VLN agent, VELMA (Schumann et al., 2024) on our own dataset. We analyze where VELMA fails and succeeds. Hence, revealing the shortcomings of VLN agents.

3.1 Evaluation Metrics

In our experiments we use the following evaluation metrics:

Task Completion : where the agent is able to make to the destination.

Dynamic Time Warp : a metric that measures how much the agent's trajectory matches the ground truth trajectory.

Overshoot Rate : a metric that measures how many times the agent was able to make it to the destination, but failed to stop at it.

	GATOR	Map2Seq	TouchDown
CLIP	4%	1.8%	1.8%
No-Image	0%	0%	1%

Table 1: Task Completion (TC) rates of VELMA on GATOR dataset. For each column, the dataset is used as the few-shot examples.

4 Experiments and Results

Table 1 shows initial results of TC rate of running VELMA on our dataset. Based on our experiments, VELMA agents fail due to several problems: **Overshoot:** where agent fails to stop at the destination and keeps going. **Undershoot:** the agent decides to stop before reaching the destination. **Wrong turns:** the agent makes wrong turns.

5 Conclusion and Future Directions

Understanding of spatial reasoning in human's language particularly in the context of navigational routes, enables us to develop models that can a) map the spatial perceptions to geospatial information b) generate instructions for navigation in human language based on the geospatial data. In this research project, we developed an cloud-based 2-player game for the task of providing textual descriptions for navigational routes based on visual

cues. In this game, two players can take roles of guide and navigator in order to complete a collaborative task of navigation. As future directions, one could replace navigator and/or guide with Artificial Intelligence (AI) models to evaluate and compare performance of models against one another.

References

Howard Chen, Alane Suhr, Dipendra Kumar Misra, Noah Snavely, and Yoav Artzi. 2018. Touchdown: Natural language navigation and spatial reasoning in visual street environments. *CoRR*, abs/1811.12354.

Harm de Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. 2018. Talk the walk: Navigating new york city through grounded dialogue. *ArXiv*, abs/1807.03367.

Toru Ishikawa and Uiko Nakamura. 2012. Landmark selection in the environment: Relationships with object characteristics and sense of direction. *Spatial Cognition & Computation*, 12(1):1–22.

Harsh Mehta, Yoav Artzi, Jason Baldridge, Eugene Ie, and Piotr W. Mirowski. 2020. Retouchdown: Adding touchdown to streetlearn as a shareable resource for language grounding tasks in street view. *ArXiv*, abs/2001.03671.

Piotr Mirowski, Matthew Koichi Grimes, Mateusz Malinowski, Karl Moritz Hermann, Keith Anderson, Denis Teplyashin, Karen Simonyan, Koray Kavukcuoglu, Andrew Zisserman, and Raia Hadsell. 2018. Learning to navigate in cities without a map. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 2424–2435, Red Hook, NY, USA. Curran Associates Inc.

Raphael Schumann, Wanrong Zhu, Weixi Feng, Tsu-Jui Fu, Stefan Riezler, and William Yang Wang. 2024. Velma: Verbalization embodiment of llm agents for vision and language navigation in street view.

Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. 2021. Talk2nav: Long-range vision-and-language navigation with dual attention and spatial memory. *International Journal of Computer Vision*, 129(1):246–266.

Wansen Wu, Tao Chang, and Xinmeng Li. 2021. Visual-and-language navigation: A survey and taxonomy. *CoRR*, abs/2108.11544.

A Game View and Initial Results

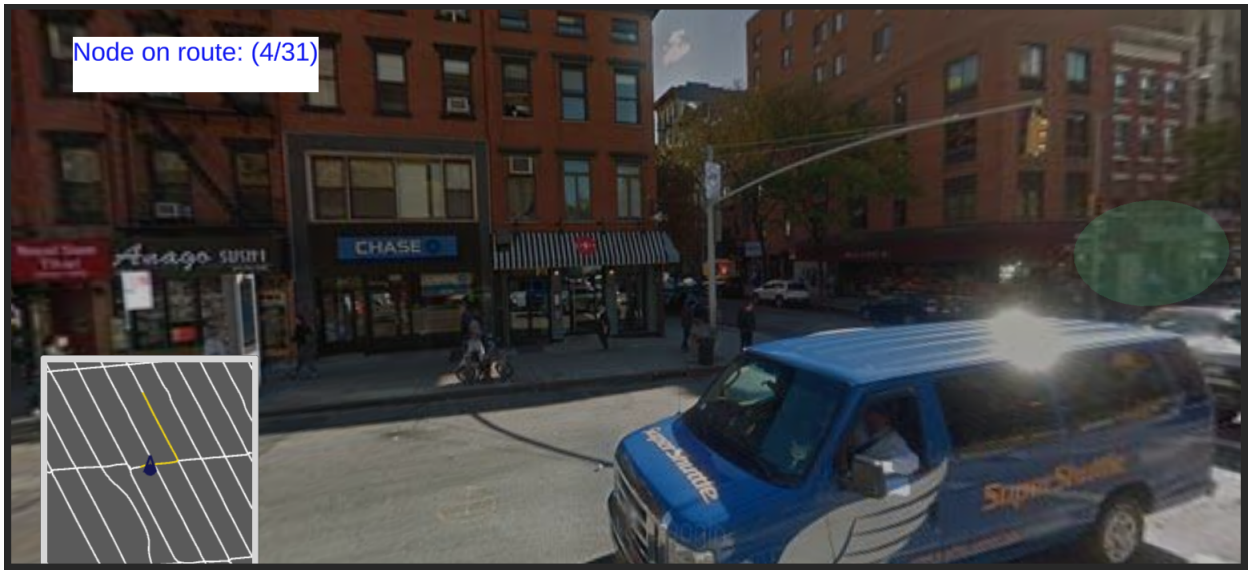

Figure 1: A screen-shot of our game. This interactive environment enables players to traverse a set of predefined routes in Manhattan area and describe the visual cues along the path.

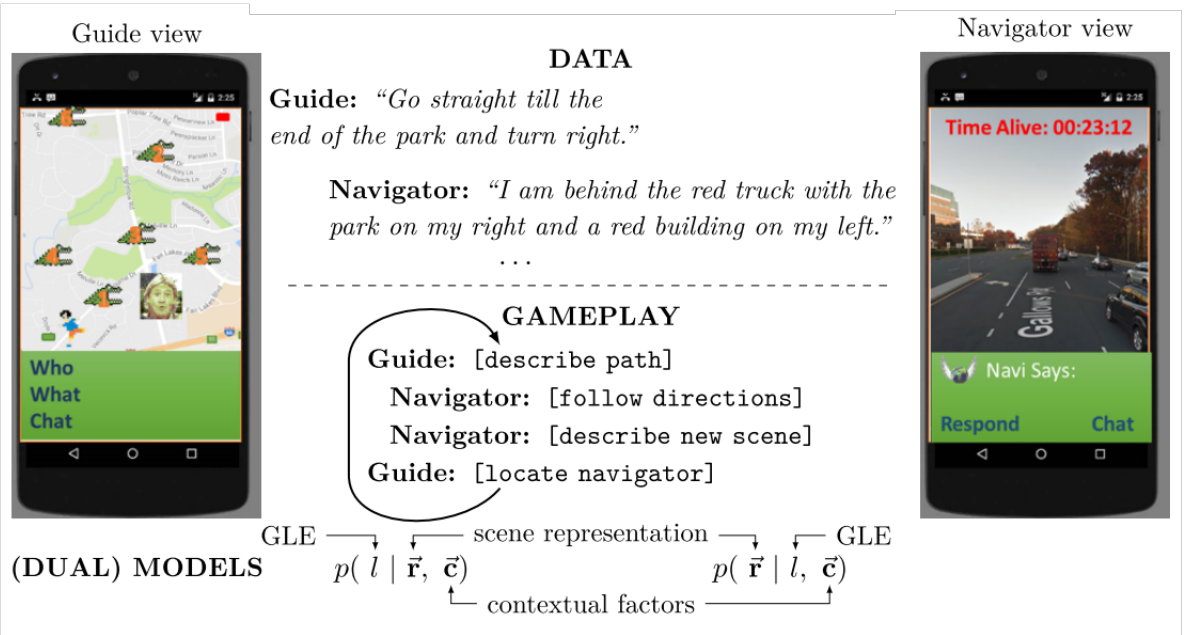

Figure 2: NAVI-GATOR: A collaborative game for data generation. Starting from a random starting point on the map, a player (guide) provides instructions and another player (navigator) follows the instructions to reach to the destination.

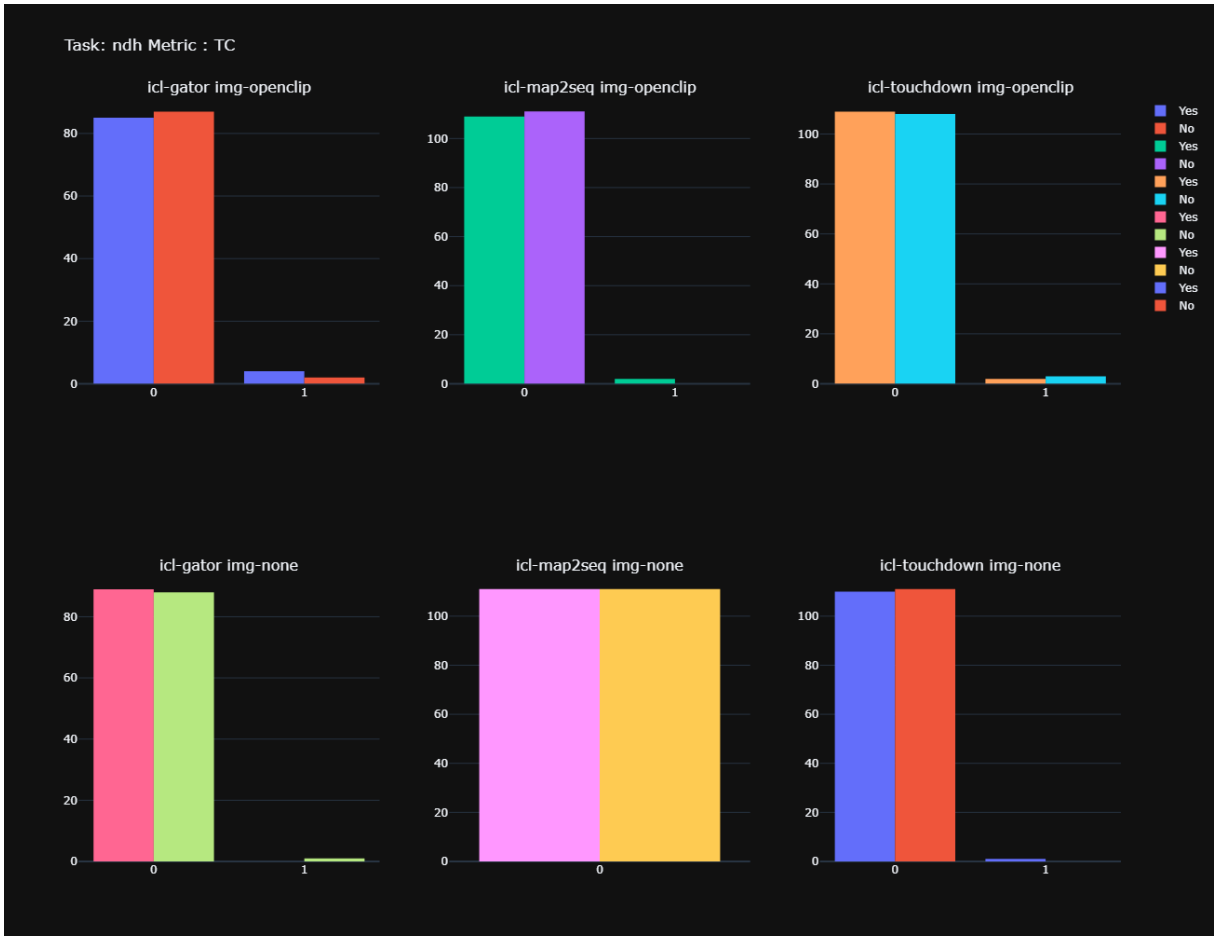

Figure 3: Comparison of task completion (TC) rate among different settings of VELMA (Schumann et al., 2024). Given history of dialogues as test data.