Datasets for A Priority Map for Vision-and-Language Navigation with Trajectory Plans and Feature-Location Cues*

Jason Armitage[1], Leonardo Impett[2], and Rico Sennrich[1]

[1] University of Zurich, Switzerland
[2] University of Cambridge, UK

1 Generation and Partition Sizes

The MC-10 dataset consists of visual, textual and geospatial data for landmarks in 10 US cities. We generate the dataset with a modified version of the process outlined by [1]. Two base entity IDs - Q2221906 ("geographic location") and Q83620 ("thoroughfare") - form the basis of queries to extract entities at a distance of <= 2 hops in the Wikidata knowledge graph[3]. Constituent cities consist of incorporated places exceeding 1 million people ranked by population density based on data for April 1, 2020 from the US Census Bureau[4]. Images and coordinates are sourced from Wikimedia and text summaries are extracted with the MediaWiki API. Geographical cells are generated using the S2 Geometry Library[5] with a range of n entities $[1, 5]$. Statistics for MC-10 are presented by partition in Table [table:mc10]2. As noted above, only a portion of textual inputs are used in pretraining and experiments.

	Train	**Development**
Number of entities	8,100	955
Mean length per text summary	727	745

Table 1. Statistics for the MC-10 dataset by partition.

TR-NY-PIT-central is a set of image files graphing path traces for trajectory plan estimation in two urban areas. Trajectories in central Manhattan are generated from routes in the Touchdown instructions [2]. Links E connecting in the Pittsburgh partition of StreetLearn [3] are the basis for randomly generated routes where at least one node is positioned in the bounding box delimited by

* The project leading to the generation of these datasets is supported by the Digital Visual Studies program at the University of Zurich and funded by the Max Planck Society.

[3] https://query.wikidata.org/

[4] https://www.census.gov/programs-surveys/decennial-census/data/datasets.html

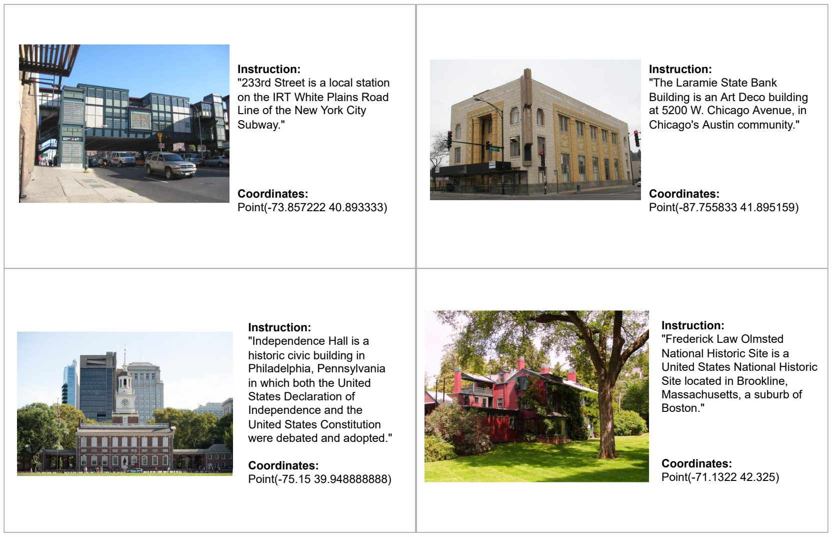
[5] https://code.google.com/archive/p/s2-geometry-library/

the WGS84 coordinates (40° 27’ 38.82”, -80° 1’ 47.85”) and (40° 26’ 7.31”, -79° 59’ 12.86”). Total trajectories sum to 9,325 in central Manhattan and 17,750 in Pittsburgh. In pretraining for In ϕ_1, $D_{\phi_1}^{Train}$ consists of 17,000 samples representing routes in the latter location.

Data used in the evaluation on the Touchdown benchmark consist of token IDs generated by passing sentences to the BertTokenizer class made available by Hugging Face[6] and the path traces for central Manhattan in TR-NY-PIT-central. The Touchdown dataset and StreetLearn environment are available from the link in the footnote below[7].

2 Samples from Datasets

In auxiliary task ϕ_2, the g_{PMF} submodule of PM-VLN is trained on visual, textual, and geodetic position data types. Path traces from the TR-NY-PIT-central are used in ϕ_1 to pretrain the g_{PMTP} submodule on trajectory estimation. Samples for entities in MC-10 and path traces in TR-NY-PIT-central are presented in Figures [mc10sample]1 and [trsample]2.



Instruction:
"233rd Street is a local station on the IRT White Plains Road Line of the New York City Subway."

Coordinates:
Point(-73.857222 40.893333)

Instruction:
"The Laramie State Bank Building is an Art Deco building at 5200 W. Chicago Avenue, in Chicago's Austin community."

Coordinates:
Point(-87.755833 41.895159)

Instruction:
"Independence Hall is a historic civic building in Philadelphia, Pennsylvania in which both the United States Declaration of Independence and the United States Constitution were debated and adopted."

Coordinates:
Point(-75.15 39.948888888)

Instruction:
"Frederick Law Olmsted National Historic Site is a United States National Historic Site located in Brookline, Massachusetts, a suburb of Boston."

Coordinates:
Point(-71.1322 42.325)

Fig. 1. Samples from the MC-10 dataset.

[6] https://huggingface.co/docs/transformers/model_doc/bert#transformers

[7] https://sites.google.com/view/streetlearn/touchdown

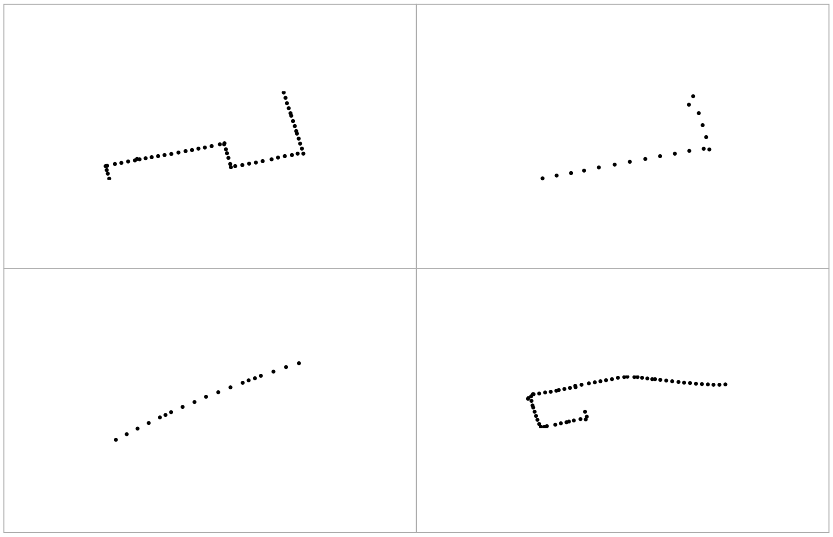


Fig. 2. Samples from the TR-NY-PIT-central dataset with path traces representing routes in central Pittsburgh.

References

1. Jason Armitage, Endri Kacupaj, Golsa Tahmasebzadeh, Maria Maleshkova, Ralph Ewerth, and Jens Lehmann. Mlm: A benchmark dataset for multitask learning with multiple languages and modalities. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2967–2974, 2020.
2. Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12538–12547, 2019.
3. Piotr Mirowski, Matt Grimes, Mateusz Malinowski, Karl Moritz Hermann, Keith Anderson, Denis Teplyashin, Karen Simonyan, Andrew Zisserman, Raia Hadsell, et al. Learning to navigate in cities without a map. *Advances in Neural Information Processing Systems*, 31:2419–2430, 2018.