Learning to Read Maps: Understanding Natural Language Instructions from Unseen Maps

Miltiadis Marios Katsakioris, Ioannis Konstas,
Pierre Yves Mignotte, Helen Hastie.
Heriot-Watt University, Edinburgh,
SeeByte Ltd, Edinburgh.
Contact: Miltiadis Marios Katsakioris: mmk11@hw.ac.uk

SpLU-RoboNLP 2021

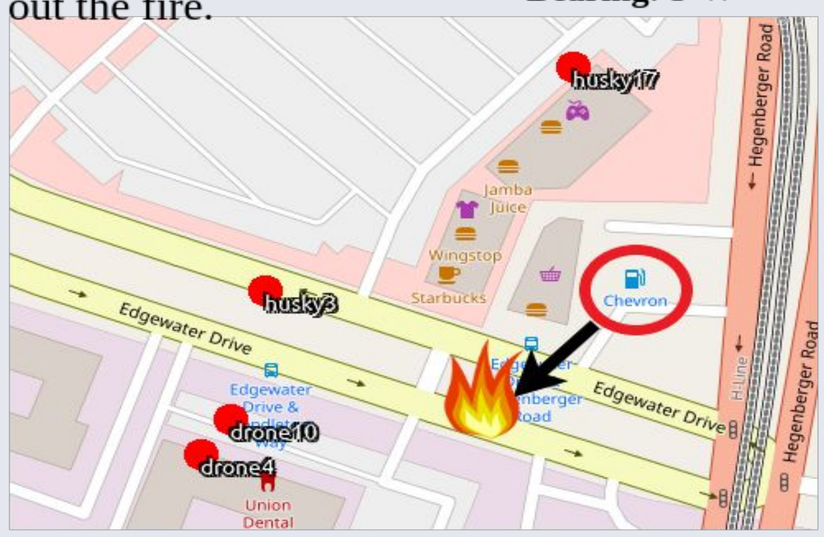
Overview

Building on prior work on multimodal interaction for collaborative planning [3], we propose a model, based on LXMERT [1], that can extract spatial information from text instructions and attend to landmarks on OpenStreetMap (OSM) referred to in a natural language (NL) instruction.

Our contributions are:

- A novel task for final GPS destination prediction from NL instructions with accompanying ROSMI dataset.
- A model that predicts GPS goal locations from a map-based natural language instruction.
- A model that is able to understand instructions referring to previously unseen maps.

Dataset



ROSMI: A Multimodal Corpus for Map-based Instruction-Giving [2]

1. Size: 783 examples, on 7 maps
2. Tuples of NL instruction, images and metadata of OSM maps
3. Annotations: BBOX of objects, target destination area, Landmark, Bearing and Distance

Architecture of MAPERT

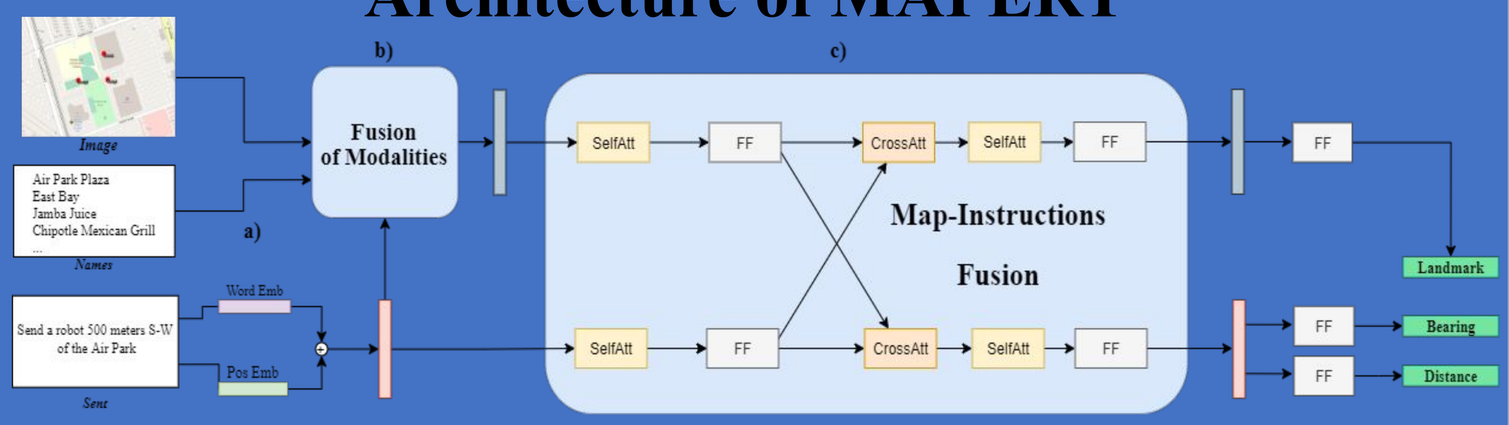


Figure 2: Map representations, i.e., names of landmarks found in OSM (metadata) and Faster-RCNN predicted objects (visual modality), along with an instruction (sequence of tokens) are a) encoded into the model, b) fused together (see also Fig. 4) and c) bidirectionally attended. The output comprises of three predictions, recast as classification tasks: a landmark, a bearing and a distance.

Results

	10-fold Cross Validation (unseen examples)			7-fold Cross Validation (unseen scenarios)		
	Accuracy$_{50}$ [SD]	T err [SD]	P err [SD]	Accuracy$_{50}$ [SD]	T err (m)[SD]	P err (m) [SD]
Oracle$_{lower}$	80 [5.01]	23.8 [51.9]	39.1 [96.3]	81.98 [17.09]	20.14 [39]	33.29 [66.43]
Baseline	33.82 [5.16]	64 [57.1]	119.8 [112.3]	34.90 [11.13]	60.71 [57.14]	110.43 [109.71]
Meta	**71.81 [7.37]**	26.70 [47.7]	48.2 [91.2]	**64.30 [14.16]**	32.71 [50.14]	65.71 [88.4]
Vision	60.36 [5.30]	36.40 [51.1]	64.40 [99.6]	49.75 [8.06]	46.00 [54.57]	87.86 [106.0]
Meta+Vision	69.27 [6.68]	26.90 [47.7]	48.30 [91.4]	58.33 [12.24]	36.14 [46.14]	70.71 [93.29]

Table 2: Results on both cross-validations of the best performing ablations of each variant and the baseline. The predictions have been made under the Oracle$_{\text{lower}}$. Accuracy (Acc) with IoU of 0.5, Target error (T Err) and Pixel Error (P Err) in meters.

Fusion of Modalities

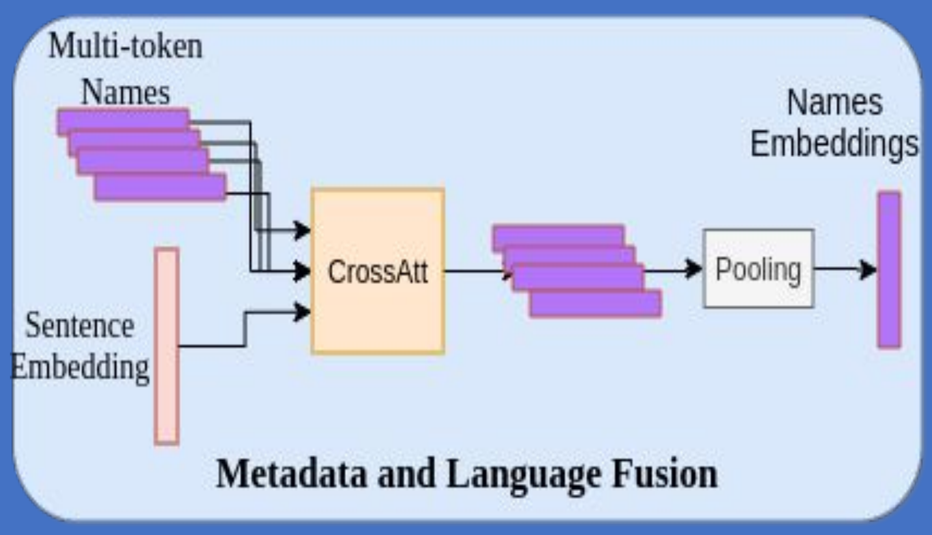


Figure 3: Metadata and Language fusion module. Multi-token names correspond to the BERT-based embeddings of landmarks names. The output is the embedding used to represent the landmarks names from OSM metadata.

Conclusion and Future Work

- We have developed a model that is able to process instructions on a map using metadata from rich map resources such as OSM and can do so for maps that it has not seen before with only a 10% reduction in accuracy.
- If no metadata is available then the model can use Vision, although this is clearly a harder task. Vision does seem to help in examples where there is a level of uncertainty such as with spatial relations or ambiguity between entities.
- Future work will involve exploring this further by training the model on these type of instructions and on metadata that are scarce and inaccurate. Finally, these instructions will be used in an end-to-end dialog system for remote robot planning, whereby multi-turn interaction can handle ambiguity and ensure reliable and safe destination prediction before instructing remote operations.

References

1) Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In Proceedings of EMNLP-IJCNLP, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.
2) Miltiadis Marios Katsakioris, Ioannis Konstas, Pierre Yves Mignotte, and Helen Hastie. 2020. Rosmi: A multimodal corpus for map-based instruction-giving. In Proceedings of the 2020 International Conference on Multimodal Interaction, ICMI '20, pages 680–684, New York, NY, USA. Association for Computing Machinery.
3) Katsakioris, M. M., Laskov, A., Konstas, I. and Hastie, H. 2019. Corpus of Multimodal Interaction for Collaborative Planning. Proceedings of the SpLU-RoboNLP 2019 Workshop in conjunction with the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), Minneapolis.