Acquiring Maps from Natural Language Descriptions

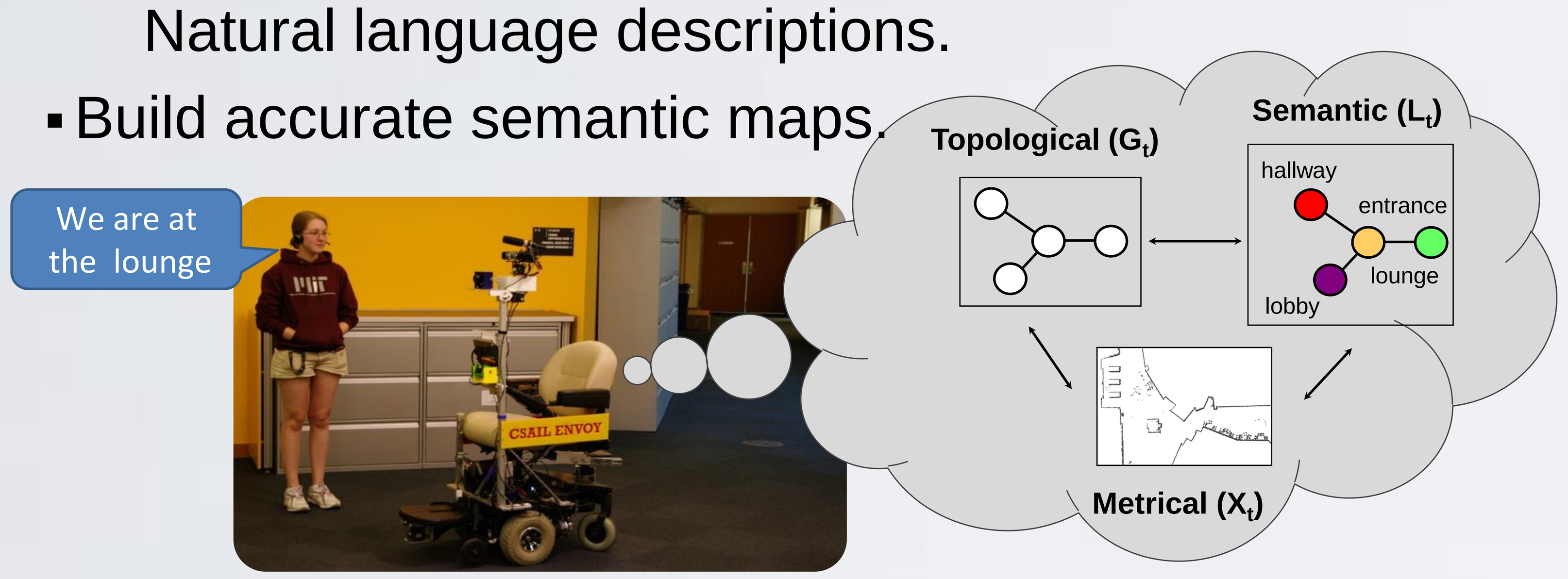
Sachithra Hemachandra, Matt Walter, Bianca Homberg, Stefanie Tellex, Seth Teller

Computer Science and Artificial Intelligence Laboratory, MIT

Building High-Level Representations

- Human-robot teams promise improved efficiency and safety.

- Robots need to share our world model to be effective partners.
- Humans can efficiently convey rich world models.
- Give a guided tour of spaces with natural language spoken descriptions.
  - Odometry, exteroception (lidar, camera, ..).
  - Natural language descriptions.
- Build accurate semantic maps.



Posterior over Semantic Graph

- Estimate three-layered ***"Semantic Graph"***. $\{G_t, X_t, L_t\}$
- Maintain the posterior over semantic graph conditioned on the history of exteroception, odometry and language.

Exteroception, Odometry, Language

$$p(G_t, X_t, L_t | z^t, u^t, \lambda^t)$$

Topology, Vertex poses, Semantic labels

- Due to its complexity, the factored posterior is maintained using a Rao-Blackwellized particle filter.

$$p(G_t, X_t, L_t | z^t, u^t, \lambda^t) = \overbrace{p(L_t | X_t, G_t, z^t, u^t, \lambda^t)}^{\text{Dirichlet Distribution}} \times \underbrace{p(X_t | G_t, z^t, u^t, \lambda^t)}_{\text{Gaussian Distribution}} \times \underbrace{p(G_t | z^t, u^t, \lambda^t)}_{\text{Sample Based}}$$

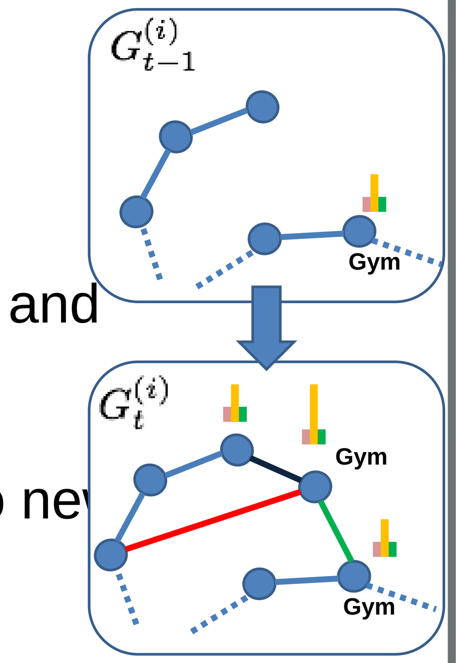
- Topology is assumed to be concentrated around a limited set of possibilities allowing accurate representation through particles, similar to PTM framework by Ranganathan et al. (2008).

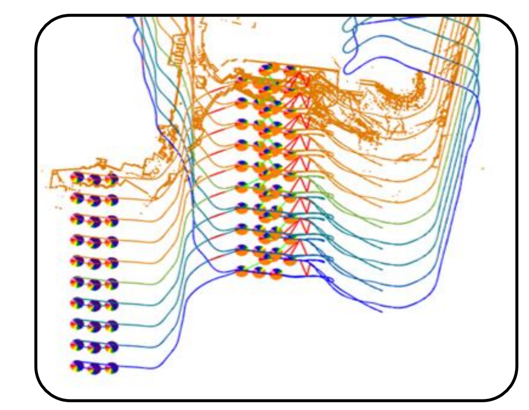
Rao-Blackwellized Particle Filter

Input: $P_{t-1} = \left\{P_{t-1}^{(i)}\right\}$, and (u_t, z_t, λ_t), where $P_{t-1}^{(i)} = \left\{G_{t-1}^{(i)}, X_{t-1}^{(i)}, L_{t-1}^{(i)} w_{t-1}^{(i)}\right\}$

for each particle i

1. Proposal: Add new node & edges to $G_{t-1}^{(i)}$ according to distributions over labels $L_{t-1}^{(i)}$ and poses $X_{t-1}^{(i)}$
2. Update Gaussian over poses according to new constraints
3. Update Dirichlet over local nodes according to language λ_t
4. Compute importance weight $w_t^{(i)}$ based on z_t





Normalize and resample if required

Return: $P_t = \left\{P_t^{(i)}\right\}$ $P_t^{(i)} = \left\{G_t^{(i)}, X_t^{(i)}, L_t^{(i)} w_t^{(i)}\right\}$

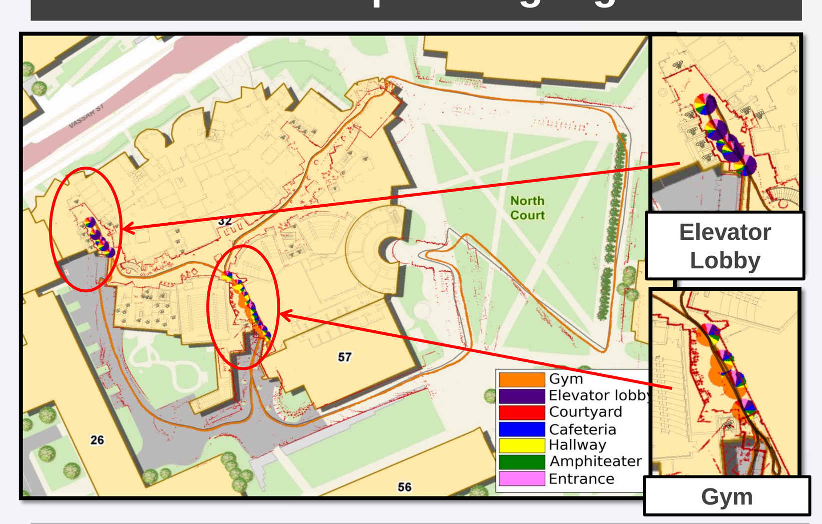
Proposal Distribution

- For each graph, a new node is added based on motion, and is connected to the previous node.
- Graph edges are proposed using
  - Spatial distribution of node poses,
  - Label distribution of nodes.
- Simple language (e.g. "I am at the gym") updates the current node
- Complex language (e.g. "The gym is down the hallway") is processed using ***Generalized Grounding Graphs*** by Tellex et al. (2010).
- The likelihood of graph G_t is evaluated based on the current observation (z_t) used to update the particle weight. $\tilde{w}_t^{(i)} = p(z_t | G_t^{(i)}, z^{t-1}, u^t, \lambda^t) \cdot w_{t-1}^{(i)}$

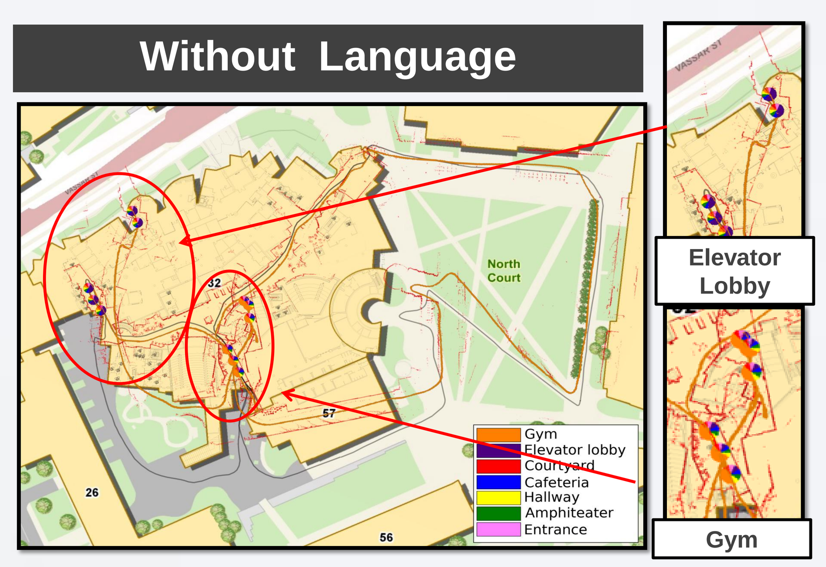
Results

- Performance was evaluated on indoor/outdoor datasets, where language was able to improve the result by proposing additional loop closures.

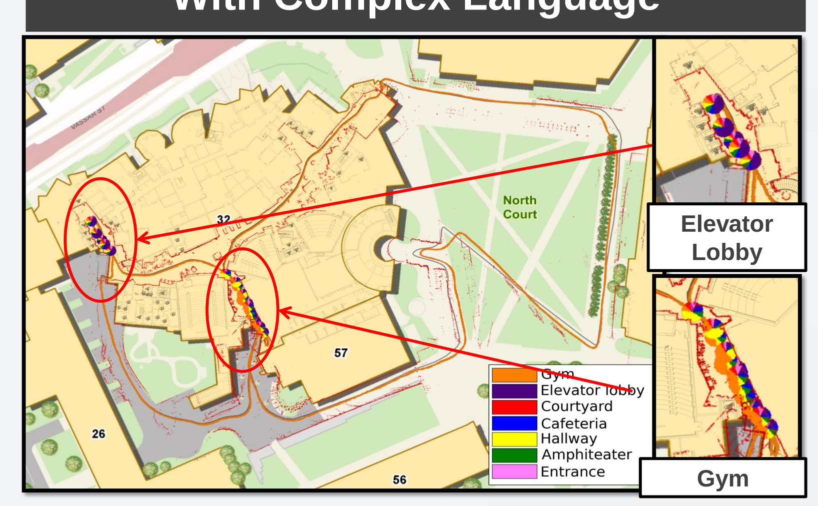
With Simple Language



Without Language



With Complex Language



Conclusions

- Our framework incorporates language to create consistent metric, topological and semantic maps.
- We exploit language to improve not just semantic but also metrical and topological representations.
- It can also correctly handle ambiguous label distributions (e.g. the presence of multiple elevator lobbies) using scan-matching to reject incorrect edges and with the use of particles.
- We plan to carry out user studies to evaluate our framework.
- We also hope to generalize the framework to encapsulate additional semantic aspects of the environment such as affordances.