SUPPLEMENTARY MATERIAL

Less is More: Generating Grounded Navigation Instructions from Landmarks

1. Bootstrapping a Landmark Dataset

Extracting landmark phrases. As outlined in Sec. 3, the first step to create silver landmark data is extracting temporally-ordered lists of landmark phrases from RxR’s human instructions. For this, we use a 3-layer distilled [7] mBERT-based [2] dependency parser pretrained on multilingual Wikipedia data [3] and outputting Universal Dependencies [5]. Specifically, the steps are:

1. Extracting all entity mentions (along with their part-of-speech) from an instruction;
2. Centering around the mentions as the head N/NP, absorbing all of their non-clausal dependents. E.g. for ... at the large brown chair that sits next to the window ..., the mention is chair, and the absorbed dependents are large and brown (but not the dependent clause that sits . . .);
3. Consolidating the head N/NP and its absorbed dependents into a single text span and recording the indices of the first and last characters. In the example above, we produce large brown chair.

While the method identifies most desired landmark text spans, inevitable errors do occur, sometimes due to the imperfection of the human-written instructions, e.g.

- Case 1. Non-object NPs. For “... take two steps towards the door”, for instance, steps will be extracted as a candidate. Similarly, left in “... take a left”.
- Case 2. Ill-formed/incomplete sentences. For example, in Fig. 3, the single-word sentence “Fancy.” results in fancy being parsed as a landmark text span.

To address Case 1, we manually compiled a stoplist (Fig. 1) of common non-object NPs that may appear in indoor environment (e.g. [side, step, one, endpoint, . . .]) for each language (English, Hindi, Telugu). Case 2 is unavoidable with the tools currently available to us. Fortunately, they occur quite rarely.

Fig. 2 demonstrates landmark phrase extraction on an example instruction.

Evaluation. In Sec. 3 of the paper, we include results from a small-scale evaluation of 100 randomly-sampled English instructions from the silver landmark dataset. To provide ground truth landmark groundings, each automatically-extracted landmark phrase was manually aligned by the paper authors to a subsequence of frames from the corresponding pose trace video. The interface used for this manual-alignment is illustrated in Fig. 3. To compute the precision scores in the paper, whenever our automatic approach grounds a landmark phrase to a frame, it is a true positive if the frame is in the human-selected subsequence, and a false positive otherwise. Results are averaged over all landmark phrases in an instruction and then all instructions.

Further analysis of the silver data. Fig. 4 provides further examples of landmark annotations from our bootstrapped silver landmark dataset. Fig. 5 presents the distribution of landmark phrases, including the 20 most frequent and least frequent landmark phrases. The distribution naturally has a long tail, although many unique landmark phrases are semantically similar (e.g. brown chair vs. white chair). We also empirically confirm the intuition that people tend to pick landmarks close to the outbound direction (i.e. the direction towards the next route segment). Fig. 6 illustrates the distribution of landmark centers in equirectangular image coordinates aligned to the outbound direction of each pano. On the horizontal dimension (heading), the landmarks are clustered close to the outbound direction, whereas on the vertical dimension (pitch), landmarks are clustered on the horizon and slightly below.

2. Landmark Detection

Training. As noted in Sec. 4, the CenterNet landmark detector was trained using a single class to represent a landmark (i.e., 2 classes in total). In initial experiments, we also investigated training the detector using a richer set of landmark classes, by clustering landmark phrases (using k-means) into classes based on their MURAL-large [4] text embeddings. However, this slightly reduced automatic evaluation scores for the full model (Tab. 1), perhaps because detection confidence scores are uncalibrated across classes.

Inference. During inference with the landmark detector, we must determine how many landmarks to return. As described in Sec. 4, we used a 1:1 ratio of landmarks to path length, defined as the number of panos in the path. To determine this ratio, we experimented with the following values {1.0, 1.2, 1.5, 1.75, 2.0} and computed automatic evaluations on the full model as reported in Tab. 2.

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-class</td>
<td>5.3</td>
<td>7.7</td>
</tr>
<tr>
<td>10-class</td>
<td>4.8</td>
<td>7.2</td>
</tr>
<tr>
<td>100-class</td>
<td>4.4</td>
<td>6.4</td>
</tr>
</tbody>
</table>

Table 1. Impact of training the landmark detector on clustered landmark classes. Automatic evaluation scores are reported for Marky-mT5 without pretraining, using the RxR Val-Unseen split.
Figure 1. Stoplists for excluding hard-to-localize noun phrases (NPs) when extracting landmark phrases. Left: English; Center: Hindi; Right: Telugu. Each stoplist was compiled by a native speaker examining samples of parsed landmark phrases.

Figure 2. Landmark phrase extraction for the instruction segment "You are facing towards the commode. Turn right to and exit the washroom. Turn right and walk straight till you reach the white cabinet.". First the object-denoting heads are identified, then we filter through each one’s dependency links to find full NP phrases for each head with constraints (e.g. only including adjectival, numeral, or adverbial dependencies, etc.).
Figure 3. Annotation interface used to collect ground-truth landmark groundings for a small-scale evaluation of the silver landmark dataset. For each automatically extracted landmark phrase (e.g., *back yard*, right), the annotator inputs a range identifying the frames in the pose trace video (left) where that landmark can be seen.

Figure 4. Further examples of bootstrapped silver landmark annotations, illustrating both correct groundings (e.g., *packaged items*, top right [8] and *beaded curtain*, bottom right [4]) as well as failure cases (e.g., *lamp*, top left [5]).
3. Instruction Generation

Pretraining and Finetuning. During pretraining, models are trained with Cross Entropy Loss and optimized with Adafactor [6] with a learning rate of 1. Batch size is 128. Pretrained models were trained for 1.45M steps. Each pretrained model (CC3M, CC12M, CC3M+CC12M) was finetuned and the final pretrained version of the downstream model was selected based on SPICE.

Table 2. Impact of varying the #landmarks / path length ratio. Automatic evaluation scores are reported for Marky-mT5 with Rewrite auxiliary training and CC3M/12M pretraining.

<table>
<thead>
<tr>
<th>#landmarks / path length ratio</th>
<th>BLEU</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>5.8</td>
<td>7.5</td>
</tr>
<tr>
<td>1.2</td>
<td>5.8</td>
<td>7.2</td>
</tr>
<tr>
<td>1.5</td>
<td>5.3</td>
<td>6.1</td>
</tr>
<tr>
<td>1.75</td>
<td>4.9</td>
<td>5.0</td>
</tr>
<tr>
<td>2.0</td>
<td>4.4</td>
<td>3.7</td>
</tr>
</tbody>
</table>

4. Experiments

Human Wayfinding. In the PanGEO interface, each annotator is shown a virtual environment in a window on the left, paired with the textual navigation instruction being evaluated on the right (refer top pane in Fig. 7). Hovering their mouse on the window, the annotator will see a green-square indicator showing them the next locations available for them to move to. After double clicking on the green-square, they are taken to the next location — illustrated in Fig. 7 bottom pane, which presents a chain of first person snapshots taken while moving. Upon arriving at the loca-
### Results
In Tab. 2 of the main paper we report human wayfinding results on paths from the RxR Val-Unseen split, aggregated overall all languages. In Tab. 3 we report these results separately for each language (English, Hindi and Telugu). Results – and most importantly, the patterns holding between the different settings – are consistent across all languages.

Finally note that, for both R2R and RxR evaluation (Val-Unseen), each path is only evaluated once per language. Therefore, for R2R we have 783 items/paths, for RxR, we have 1,517 paths, and with 3 languages, 4,551 items.

### Error analysis on Marky-mT5 instructions
Our human wayfinding results are representative of a step-change in instruction quality compared to previous models. To better understand the types of errors that still remain, we perform a manual error analysis on 110 randomly-sampled English instructions generated for RxR Val-Unseen paths by the full Marky-mT5 system using predicted landmarks.

To perform the analysis we added an option in the PanGEA wayfinding interface to toggle the visibility of the ground-truth path, so that we could better assess how well that path was described by the generated instruction. We classify instruction errors and weaknesses into the following six categories:

- **Landmark Errors**
  - **Full Hallucination.** The instruction describes a landmark that does not exist in the environment;
  - **Weak Description.** A landmark description that is flawed, but not completely wrong (e.g. blue towel described as blue napkin);
  - **Wrong Orientation.** An instruction refers to a landmark but orients it incorrectly with respect to the route, e.g. saying it is on the left side when...
it is actually on the right side.

- **Path Errors**
  - **Wrong action.** A mistake in an action/step (e.g. turn right when one needs to turn left);
  - **Missing action.** Skipping a an action/step (e.g. the instruction neglects to mention a crucial right hand turn);
  - **Weak granularity.** Some segment of the instruction is too coarse to describe the multiple steps that are needed in the path trajectory (e.g. merely using go forward to describe a route that passes through two doorways and a dining hall).

We provide examples of each error type in Fig. 8 and Fig. 9, where the blue/purple (arrowed) balls indicate the ground truth path. The error analysis was performed by the paper authors with a critical eye; any weakness in the instructions was annotated as an error.

The results are summarized in Tab. 4. Of the instructions annotated, 16% were judged to be error-free, with some errors or weaknesses identified in the remaining 84%. Out of the 6 error types, **Weak Description** was by far the most common (62.9% of all errors). In contrast, the proportion of **Full Hallucination**, **Wrong Orientation** and **Wrong Action** errors—which were common in previous models—was relatively low, at 11.4%, 7.9% and 7.1% respectively.

Anecdotally, we noticed that the cost of different errors varies. Human wayfinders can often overcome minor flaws in the description of a landmark (e.g. if a pink bedspread is misidentified as a white bedspread). However if ambiguity is involved, e.g. move towards the open door when there more than one door is in view, confusion results. Full hallucination and wrong orientation can certainly be misleading but if the navigator survey the environment carefully in the context of the neighboring segments in the instruction, they are also often resolvable. The three types of path errors are less recoverable, as they often result in missteps that take the wayfinder to an entirely wrong path.

### Results on automatic evaluation
Beyond the strong quality of MARKY-MT5-generated instructions in human evaluation, they also perform at the similar level in automatic/model evaluation. Employing the state-of-the-art HAMT [1] VLN agent (Tab. 5), in particular, we received 55.7% vs. 56.5% success rate and 63.3% vs. 62.9% nDTW between model-generated (with predicted landmarks) vs. human-written instructions, demonstrating that MARKY-MT5 produces instructions followable by both human and model agents.

### References


Figure 8. Examples of Landmark Errors. **Full Hallucination:** a plant and portrait are mentioned in the top and bottom panes respectively but these landmarks cannot be found in the visual scene. **Weak Description:** the top pane exemplifies an ambiguous landmark – there are two open doors, and it’s not clear which one to move towards; the bottom pane illustrates a flawed description (specifically, wrong color). **Wrong orientation:** On the top pane, the shower is to the right hand side of the wayfinder rather than left; in the bottom pane, the sofa chair should be on the left side instead.

Figure 9. Examples of Path Errors. **Wrong action:** the examples show a right turn mistaken to be a left turn, and going up stairs mistaken as going down. **Missing action:** for the top pane, the washroom is not visible before making an additional left turn; for the bottom pane, stairs require a right turn to see. **Weak granularity:** in the examples, overly coarse instruction segments are given for long path segments.