Speaker-Follower Models for Vision-and-Language Navigation (Extended Abstract)

Daniel Fried*[1], Ronghang Hu*[1], Volkan Cirik*[2], Anna Rohrbach[1], Jacob Andreas[1], Louis-Philippe Morency[2], Taylor Berg-Kirkpatrick[2], Kate Saenko[3], Dan Klein**[1], Trevor Darrell**[1]

[1]University of California, Berkeley [2]Carnegie Mellon University [3]Boston University

Abstract. Navigation guided by natural language instructions presents a challenging reasoning problem for instruction followers. Here we describe a speaker-follower approach to this task with an embedded *speaker model*. We use this speaker model to synthesize new instructions for data augmentation and to implement pragmatic reasoning for evaluating candidate action sequences. Both steps are supported by a panoramic action space that reflects the granularity of human-generated instructions. Experiments show that our speaker-follower approach improves the performance of a baseline instruction follower, more than doubling the success rate over the best existing approach on a standard benchmark.

1 Introduction

In the vision-and-language navigation task [1], an agent is placed in a realistic environment, and provided a natural language instruction such as "*Go down the stairs, go slight left at the bottom and go through door, take an immediate left and enter the bathroom, stop just inside in front of the sink*". The agent must follow this instruction to navigate from its starting location to a goal location, as shown in Figure 1 (left). To accomplish this task the agent must learn to relate the language instructions to the visual environment. Moreover, it should be able to carry out new instructions in unseen environments.

In this paper we treat the vision-and-language navigation task as a trajectory search problem, where the agent needs to find (based on the instruction) the best trajectory in the environment to navigate from the start location to the goal location. Our model involves an instruction interpretation (*follower*) module, mapping instructions to action sequences; and an instruction generation (*speaker*) module, mapping action sequences to instructions (Figure 1), both implemented with standard sequence-to-sequence architectures. The speaker learns to give textual instructions for visual routes, while the follower learns to follow routes (predict navigation actions) for provided textual instructions.

We incorporate the speaker both at training time and at test time, where it works together with the learned instruction follower model to solve the navigation task (see Figure 2 for an overview of our approach). At training time, we perform speaker-driven data augmentation where the speaker helps the follower by synthesizing additional route-instruction pairs to expand the limited training

*, **: Authors contributed equally

Fig. 1. The task of vision-and-language navigation is to perform a sequence of actions (navigate through the environment) according to natural language instructions. Our approach consists of an instruction *follower* model (left) and a *speaker* model (right).

data. At test time, the follower improves its chances of success by looking ahead at possible future routes and pragmatically choosing the best route by scoring them according to the probability that the speaker would generate the correct instruction for each route. We construct both the speaker and the follower on top of a panoramic action space that efficiently encodes high-level behavior, moving directly between adjacent locations rather than making low-level visuomotor decisions like the number of degrees to rotate (see Figure 3).

An extended version of this paper with more details can be found in [2].

2 Related Work

Vision-and-language Navigation. The embodied vision-and-language navigation task studied in this paper [1] differs from past situated instruction following tasks (e.g. [3,4,5,6]) by introducing rich visual contexts. Recent work [7] has applied techniques from model-based and model-free reinforcement learning [8] to the vision-and-language navigation problem. Specifically, an environment model is used to predict a representation of the state resulting from an action, and planning is performed with respect to this environment model. Our work differs from this prior work in reasoning not just about state transitions, but also about the relationship between states and the language that describes them.

Pragmatic language understanding. A long line of work in linguistics, natural language processing, and cognitive science has studied *pragmatics*: how linguistic meaning is affected by context and communicative goals [9]. Our work here makes use of the Rational Speech Acts framework [10,11], which models the interaction between speakers and listeners as a process where each agent reasons probabilistically about the other to maximize the chances of successful communicative outcomes. Similar modeling tools have recently been applied to generation and interpretation of language about sequential decision-making [12]. The present work makes use of a pragmatic instruction follower in the same spirit. Here, however, we integrate this with a more complex visual pipeline and use it not only at inference time but also at *training* time to improve the quality of a base listener model.

Semi- and self-supervision. The semi-supervised approach we use is related to model bootstrapping techniques such as self-training [13,14] and co-training [15] at a high-level. The approach most relevant to our work is the SEQ4 model [16], which applies semi-supervision to a navigation task by sampling new environments and maps (in synthetic domains without vision), and training an autoencoder to reconstruct routes, using language as a latent variable. In this

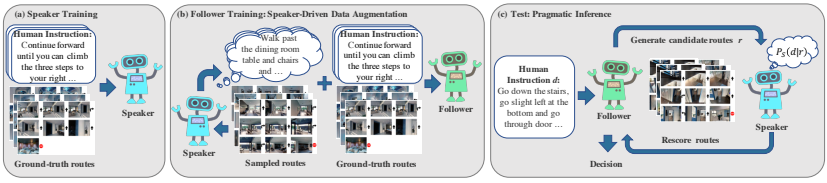

Fig. 2. Our approach combines an instruction *follower* model and a *speaker* model. (a) The speaker model is trained on the ground-truth routes with human-generated descriptions; (b) it provides the follower with additional synthetic instruction data to bootstrap training; (c) it also helps the follower interpret ambiguous instructions and choose the best route during inference. See Sec. 3 for details.

work, we use a speaker to synthesize additional navigation instructions on unlabeled new routes, and use this synthetic data from the speaker to train the follower. Our approach used here is much simpler, as it does not require constructing a differentiable surrogate to the decoding objective.

Grounding language in vision. Existing work in visual grounding has addressed the problem of *passively* perceiving a static image and mapping a referential expression to a bounding box [17,18,19] or a segmentation mask [20,21,22]. In our work, the vision-and-language navigation task requires the agent to *actively* interact with the environment to find a path to the goal following the natural language instruction. This can be seen as a grounding problem in linguistics where the instruction is grounded into a trajectory in the environment but requires more reasoning and planning skills than referential expression grounding.

3 Instruction Following with Speaker-Follower Models

We address the task of following natural language instructions relying on two models: an instruction-*follower* model of the kind considered in previous work, and a *speaker* model—a learned instruction generator that models how humans describe routes in navigation tasks. Specifically, we base our follower model on a sequence-to-sequence model [1], computing a distribution $P_F(r|d)$ over routes r (state and action sequences) given route descriptions d. Our speaker model is symmetric, producing a distribution $P_S(d|r)$ by encoding the sequence of visual observations and actions in the route, and then outputting an instruction word-by-word with a decoder (Figure 1).

The speaker supports the follower both at training time and at test time. First, we train a speaker model on the available ground-truth navigation routes and instructions. (Figure 2 (a)). Before training the follower, the speaker produces synthetic navigation instructions for novel sampled routes in the training environments, which are then used as additional supervision for the follower (Figure 2 (b)). At follower test time, the speaker pragmatically ranks possible routes produced by the follower model (Figure 2 (c)). Both follower and speaker are supported by the panoramic action space (Figure 3).

Speaker-Driven Data Augmentation. The training data only covers a limited number of navigation instruction and route pairs, $\mathcal{D} = (d_1, r_1) \dots (d_N, r_N)$. To allow the agent to generalize better to new routes, we use the speaker to generate synthetic instructions on sampled new routes in the training environments.

We sample a collection of M routes $\hat{r}_1, \ldots, \hat{r}_M$ through the training environments, using the same shortest-path approach used to generate the routes in the original training set [1]. We then generate a human-like textual instruction $\hat{d}_k$ for each instruction $\hat{r}_k$ by performing greedy prediction in the speaker model to approximate $\hat{d}_k = \arg\max_d P_S(d \mid \hat{r}_k)$. These M synthetic navigation routes and instructions $\mathcal{S} = (\hat{d}_1, \hat{r}_1), \ldots, (\hat{s}_M, \hat{r}_M)$ are combined with the original training data $\mathcal{D}$ into an augmented training set $\mathcal{S} \cup \mathcal{D}$ (Figure 2(b)).

Speaker-Driven Route Selection (Pragmatic Inference). We use the base speaker (P_S) and follower (P_F) models described above to define a *pragmatic follower* model. Drawing on the Rational Speech Acts framework [10,11], a pragmatic follower model should choose a route r that has high probability of having caused the speaker model to produce the given description d: $\arg\max_r P_S(d \mid r)$. Such a follower chooses a route that provides a good explanation of the observed description, allowing counterfactual reasoning about instructions, or using global context to correct errors in the follower's path.

Following previous work on pragmatic language generation and interpretation [23,24,25,12], we approximate this maximization using a rescoring procedure: produce candidate route interpretations for a given instruction using the base follower model, and then rescore these routes using the base speaker model (Figure 2(c)). Our pragmatic follower produces a route for a given instruction by obtaining K candidate paths from the base follower using a modified beam-search procedure,then chooses the highest scoring path under a combination of the follower and speaker model probabilities:

$$\arg\max_{r \in R(d)} P_S(d \mid r)^{\lambda} \cdot P_F(r \mid d)^{(1-\lambda)} \tag{1}$$

where λ is a hyper-parameter in the range $[0, 1]$ which we tune on validation data to maximize the accuracy of the follower. To generate candidate routes from the base follower model, we perform a modified beam-search procedure.

Panoramic Action Space. The sequence-to-sequence agent in [1] uses low-level visuomotor control. In our work we directly allow the agent to reason about high-level actions, using a panoramic action space with panoramic representation, converted with built-in mapping from low-level visuomotor control.

As shown in Figure 3, in our panoramic representation, the agent first "looks around" and perceives a 360-degree panoramic view of its surrounding scene from its current location, which is discretized into 36 view angles. Each view angle i is represented by an encoding vector v_i. At each location, the agent can only move towards a few navigable directions (provided by the navigation environment). Here, in our action space the agent only needs to make high-level decisions as to which navigable direction to go to next, with each navigable direction j represented by an encoding vector u_j. The encoding vectors v_i and u_j of each view angle and navigable direction are obtained by concatenating an appearance feature and a 4-dimensional orientation feature $[\sin\psi; \cos\psi; \sin\theta; \cos\theta]$, where ψ and θ are the heading and elevation angles respectively. Also, we introduce a "stop" action encoded by $u_0 = \overrightarrow{0}$. The agent can take this stop action when it decides it has reached the goal location (to end the episode).

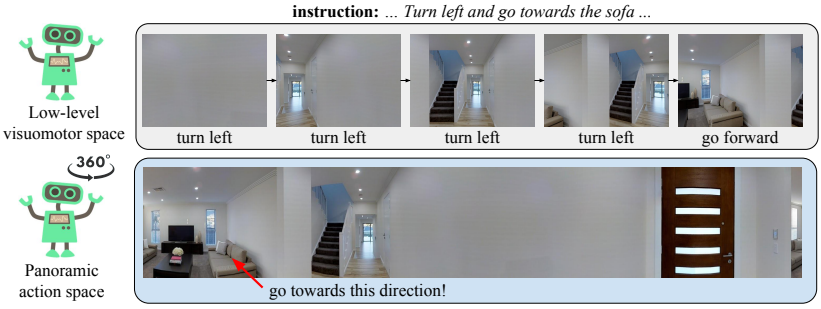

Fig. 3. Compared with low-level visuomotor space, our panoramic action space (Sec. 3) allows the agents to have a complete perception of the scene, and to directly perform high-level actions.

To make a decision on which direction to go, the agent first performs one-hop visual attention to look at all of the surrounding view angles, based on its memory vector h_{t-1}. The attention weight $\alpha_{t,i}$ of each view angle i is computed as $a_{t,i} = (W_1 h_{t-1})^T W_2 v_{t,i}$ and $\alpha_{t,i} = \exp(a_{t,i}) / \sum_i \exp(a_{t,i})$. The attended feature representation $v_{t,att} = \sum_i \alpha_{t,i} v_{t,i}$ from the panoramic scene is then used as visual-sensory input to the sequence-to-sequence model to update the agent's memory. Then, a bilinear dot product is used to obtain the probability p_j of each navigable direction j: $y_j = (W_3 h_t)^T W_4 u_j$, $p_j = \exp(y_j) / \sum_j \exp(y_j)$. The agent then chooses a navigable direction u_j (with probability p_j) to go to the adjacent location along that direction (or u_0 to stop and end the episode).

4 Experiments

We use the Room-to-Room (R2R) vision-and-language navigation dataset [1] for our experimental evaluation. Following previous work on the R2R task, our primary evaluation metrics are navigation error (NE), success rate (SR) and oracle success rate (OSR).

Ablation study. We study the impact of each component of our model on the val seen (same environments as training split) and val unseen split (novel environments not seen during training). In Table 1 (a), Row 1 is a baseline which uses only a follower model with a non-panoramic action space at both training and test time, almost equivalent to the student-forcing model in [1] except for minor implementation details. Rows 2-4 show the effects of adding a single component to the baseline system (Row 1); Rows 5-7 show the effects of removing a single component from the full system (Row 8). It can be seen that all components are important for the final performance.

Comparison to Prior Work. We compare the performance of our final model to previous approaches on the R2R the test split (new environments not overlapping with any training or validation splits). The results are shown in Table 1 (b). In the table, "Random" is a baseline that randomly picks a direction and goes toward that direction for 5 steps. "Student-forcing" is the best performing method in [1], using exploration during training of the sequence-to-sequence follower model. "RPA" [7] is a combination of model-based and model-free reinforcement learning (see also Sec. 2 for details). "ours" shows our performance

#	Data Augmentation	Pragmatic Inference	Panoramic Space	Validation-Seen NE ↓	SR ↑	OSR ↑	Validation-Unseen NE ↓	SR ↑	OSR ↑
1				6.08	40.3	51.6	7.90	19.9	26.1
2	✓			5.05	46.8	59.9	7.30	24.6	33.2
3		✓		5.23	51.5	60.8	6.62	34.5	43.1
4			✓	4.86	52.1	63.3	7.07	31.2	41.3
5	✓	✓		4.28	57.2	63.9	5.75	39.3	47.0
6	✓		✓	3.36	66.4	73.8	6.62	35.5	45.0
7		✓	✓	3.88	63.3	71.0	5.24	49.5	63.4
8	✓	✓	✓	**3.08**	**70.1**	**78.3**	**4.83**	**54.6**	**65.2**

(a) Ablations showing the effect of each component in our model.

Method	Validation-Seen NE ↓	SR ↑	OSR ↑	Validation-Unseen NE ↓	SR ↑	OSR ↑	Test (unseen) NE ↓	SR ↑	OSR ↑	TL ↓
Random	9.45	15.9	21.4	9.23	16.3	22.0	9.77	13.2	18.3	9.89
Student-forcing [1]	6.01	38.6	52.9	7.81	21.8	28.4	7.85	20.4	26.6	8.13
RPA [7]	5.56	42.9	52.6	7.65	24.6	31.8	7.53	25.3	32.5	9.15
ours	**3.08**	**70.1**	**78.3**	**4.83**	**54.6**	**65.2**	**4.87**	**53.5**	63.9	11.63
ours (challenge participation)*	–	–	–	–	–	–	**4.87**	**53.5**	**96.0**	1257.38
Human	–	–	–	–	–	–	1.61	86.4	90.2	11.90

*: When submitting to the Vision-and-Language Navigation Challenge, we modified our beam search procedure to maintain physical plausibility and to comply with the challenge guidelines. The resulting trajectory has higher oracle success rate while being very long. See Appendix E in [2] for details.

(b) Comparison of our method to previous work.

Table 1. Ablations (a) and comparison with previous work (b). NE is navigation error (lower is better). SR and OSR are success rate and oracle success rate (%) respectively (higher is better). Trajectory length (TL) on the test set is reported for completeness.

using the route selected by our pragmatic inference procedure, while "ours (challenge participation)" uses a modified inference procedure for submission to the Vision-and-Language Navigation Challenge (see Appendix E in [2] for details). Our method more than doubles the success rate of the state-of-the-art RPA approach, and on the test set achieves a final success rate of 53.5%, a large reduction in the gap between machine and human performance on this task.

5 Conclusions

The language-and-vision navigation task presents a pair of challenging reasoning problems: in language, because agents must interpret instructions in a changing environmental context; and in vision, because of the tight coupling between local perception and long-term decision-making. The comparatively poor performance of the baseline sequence-to-sequence model for instruction following suggests that more powerful modeling tools are needed to meet these challenges. In this work, we have introduced such a tool, showing that a follower model for vision-and-language navigation is substantially improved by carefully structuring the action space and integrating an explicit model of a *speaker* that predicts how navigation routes are described. We believe that these results point toward further opportunities for improvements in instruction following by modeling the global structure of navigation behaviors and the pragmatic contexts in which they occur.

References

1. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., Hengel, A.v.d.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018)
2. Fried, D., Hu, R., Cirik, V., Rohrbach, A., Andreas, J., Morency, L.P., Berg-Kirkpatrick, T., Saenko, K., Klein, D., Darrell, T.: Speaker-follower models for vision-and-language navigation. arXiv preprint arXiv:1806.02724 (2018)
3. Tellex, S., Kollar, T., Dickerson, S., Walter, M.R., Banerjee, A.G., Teller, S.J., Roy, N.: Understanding natural language commands for robotic navigation and mobile manipulation. In: AAAI. Volume 1. (2011) 2
4. Chen, D.L.: Fast online lexicon learning for grounded language acquisition. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1. ACL '12, Stroudsburg, PA, USA, Association for Computational Linguistics (2012) 430–439
5. Branavan, S., Chen, H., Zettlemoyer, L.S., Barzilay, R.: Reinforcement learning for mapping instructions to actions. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics (2009) 82–90
6. Andreas, J., Klein, D.: Alignment-based compositional semantics for instruction following. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). (2015)
7. Wang, X., Xiong, W., Wang, H., Wang, W.Y.: Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation. arXiv:1803.07729 (2018)
8. Weber, T., Racanière, S., Reichert, D.P., Buesing, L., Guez, A., Rezende, D.J., Badia, A.P., Vinyals, O., Heess, N., Li, Y., et al.: Imagination-augmented agents for deep reinforcement learning. arXiv preprint arXiv:1707.06203 (2017)
9. Grice, H.P.: Logic and conversation. In Cole, P., Morgan, J.L., eds.: Syntax and Semantics: Vol. 3: Speech Acts. Academic Press, San Diego, CA (1975) 41–58
10. Frank, M.C., Goodman, N.D.: Predicting pragmatic reasoning in language games. Science **336**(6084) (2012) 998–998
11. Goodman, N.D., Stuhlmüller, A.: Knowledge and implicature: Modeling language understanding as social cognition. Topics in cognitive science **5**(1) (2013) 173–184
12. Fried, D., Andreas, J., Klein, D.: Unified pragmatic models for generating and following instructions. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL). (2018)
13. Scudder, H.: Probability of error of some adaptive pattern-recognition machines. IEEE Transactions on Information Theory **11**(3) (1965) 363–371
14. McClosky, D., Charniak, E., Johnson, M.: Effective self-training for parsing. In: Proceedings of the main conference on human language technology conference of the North American Chapter of the Association of Computational Linguistics, Association for Computational Linguistics (2006) 152–159
15. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: Proceedings of the eleventh annual conference on Computational learning theory, ACM (1998) 92–100
16. Kočiský, T., Melis, G., Grefenstette, E., Dyer, C., Ling, W., Blunsom, P., Hermann, K.M.: Semantic parsing with semi-supervised sequential autoencoders. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language

Processing, Austin, Texas, Association for Computational Linguistics (November 2016) 1078–1087
17. Plummer, B., Wang, L., Cervantes, C., Caicedo, J., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2015)
18. Mao, J., Jonathan, H., Toshev, A., Camburu, O., Yuille, A., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)
19. Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K., Darrell, T.: Natural language object retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)
20. Hu, R., Rohrbach, M., Darrell, T.: Segmentation from natural language expressions. In: Proceedings of the European Conference on Computer Vision (ECCV). (2016)
21. Liu, C., Lin, Z., Shen, X., Yang, J., Lu, X., Yuille, A.: Recurrent multimodal interaction for referring image segmentation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). (2017)
22. Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., Berg, T.L.: Mattnet: Modular attention network for referring expression comprehension. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018)
23. Smith, N.J., Goodman, N., Frank, M.: Learning and using language via recursive pragmatic reasoning about other agents. In: Advances in neural information processing systems. (2013) 3039–3047
24. Andreas, J., Klein, D.: Reasoning about pragmatics with neural listeners and speakers. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). (2016)
25. Monroe, W., Hawkins, R., Goodman, N., Potts, C.: Colors in context: A pragmatic neural model for grounded language understanding. Transactions of the Association for Computational Linguistics **5** (2017) 325–338