Define, Evaluate, and Improve Task-Oriented Cognitive Capabilities for Instruction Generation Models

Lingjun Zhao*, Khanh Nguyen*, Hal Daumé III

lzhao123@umd.edu

* The first two authors contributed equally.

Problem

➡ Build speaker models that generate language instructions to guide humans in 3D environments
➡ Instructions generated by vanilla speaker models fail to communicate well with humans
➡ How to generate better instructions by **reasoning pragmatically**?
➡ How to **evaluate cognitive capabilities** of instruction generation (*speaker*) models?

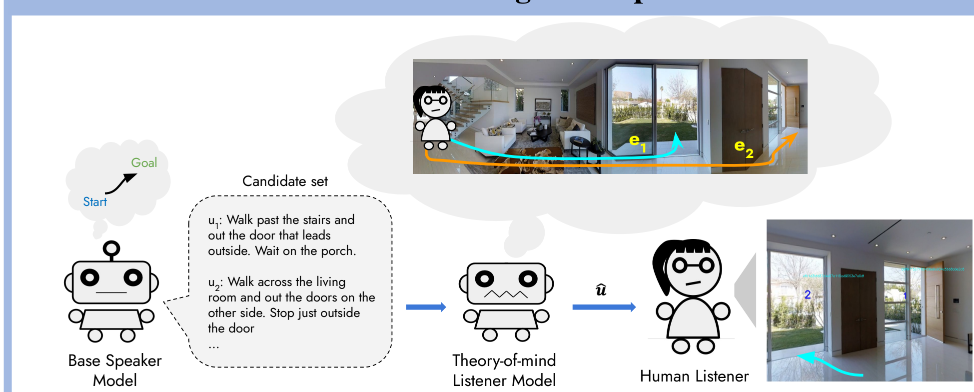
Vanilla Speaker: Exit the bathroom and turn left. Walk past the bed and **wait by the two chairs**. *[Correct destination is next to the chairs in the outdoor area]*

Pragmatic Speaker: Walk out of the bathroom and make a left. Walk through the bedroom and continue straight towards the red chair. **Stop at the chair before getting to the front of the patio.**

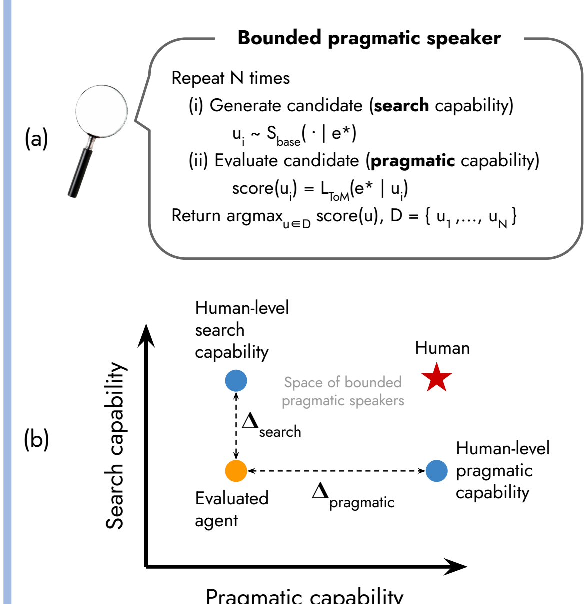
Contributions

- A new scheme for evaluating task-oriented cognitive capabilities in instruction generation models
- An 11% success rate improvement in guiding real humans in photorealistic environment, by equipping vanilla speakers with theory-of-mind capabilities
- A call to construct better theory-of-mind models for improving the instruction generation models

Bounded Pragmatic Speaker



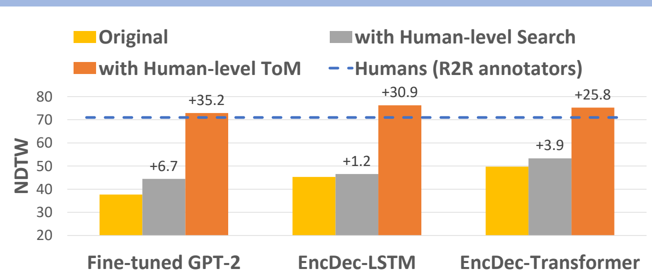
Cognitive Evaluation



(c) **Recommendation:**

- Large Δ_{search}, small $\Delta_{pragmatic}$ ⇒ improve inference algorithm
- Large $\Delta_{pragmatic}$, small Δ_{search} ⇒ enhance model of listener

Human-augmented Speaker Performance



Takeaway: *Pragmatic* capability (theory-of-mind evaluation) is more deficient than *Search* capability (candidate generation).

Improving Pragmatic Capability with *Ensemble* Theory-of-Mind Listener

	Base speaker S_{base}		
ToM listener L_{ToM}	Fine-tuned GPT-2	EncDec-LSTM	EncDec-Transformer
None	37.7 (▲ 0.0)	45.3 (▲ 0.0)	49.4 (▲ 0.0)
Single VLN-BERT (Majumdar et al., 2020)	38.9 (▲ 1.2)	39.8 (▼ 5.5)	46.2 (▼ 3.2)
Ensemble of 10 EnvDrop-CLIP (Shen et al., 2022)	37.8 (▲ 0.1)	$53.1^{\dagger}$ (▲ 7.8)	$57.3^{\dagger}$ (▲ 7.9)
Ensemble of 10 VLN↻BERT (Hong et al., 2021)	43.4 (▲ 5.7)	$56.4^{\ddagger}$ (▲ 11.1)	54.2 (▲ 4.8)
Humans (skyline)	$72.9^{\ddagger}$ (▲ 35.2)	$76.2^{\ddagger}$ (▲ 30.9)	$75.2^{\ddagger}$ (▲ 25.8)

Takeaways:

- Using ensemble followers as theory-of-mind model can *improve* vanilla speakers significantly to communicate with humans
- Better task-oriented theory-of-mind is needed to bridge the communication gap between AI and humans

Experimental Settings

✓ **Training data:**
- Matterport Room-to-Room (reverse task)

✓ **Models:**
- Fine-tuned GPT-2
- EncDec-LSTM
- EncDec-Transformer
- Pragmatic Speakers

✓ **Evaluation:**
- Give instructions to real humans
- Measure similarity between human-generated and intended paths: normalized Dynamic Time Warping (NDTW)

Key References

- Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Su¨nderhauf, N., Reid, I., Gould, S., and Van Den Hengel, A. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. CVPR 2018.
- Shen, S., Li, L. H., Tan, H., Bansal, M., Rohrbach, A., Chang, K.-W., Yao, Z., and Keutzer, K. How much can clip benefit vision-and-language tasks? 2022.
- Majumdar, A., Shrivastava, A., Lee, S., Anderson, P., Parikh, D., and Batra, D. Improving vision-and-language navigation with image-text pairs from the web. ECCV 2020.
- Hong, Y., Wu, Q., Qi, Y., Rodriguez-Opazo, C., and Gould, S. Vln bert: A recurrent vision-and-language bert for navigation. CVPR 2021.