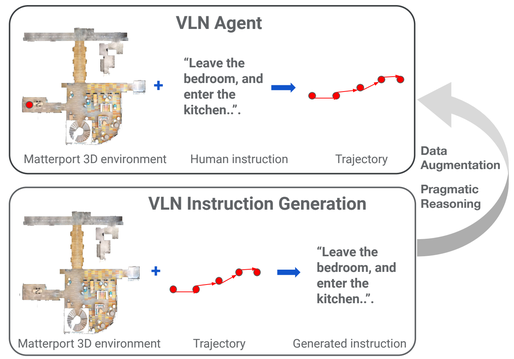
On the Evaluation of Vision-and-Language Navigation Instructions

Ming Zhao, Peter Anderson, Vihan Jain, Su Wang, Alex Ku, Jason Baldridge, Eugene Ie
Google Research

EACL 2021

Introduction

- **Vision-and-Language navigation (VLN)** (Anderson et al. 2018): the task of following navigation instructions to traverse a path in a photorealistic environment.
- **VLN Instruction Generation:** the task of generating natural language navigation instructions for a given path in a photorealistic environment.
  - Generated Instructions have been widely adopted for data augmentation in VLN tasks and have been shown to be ***very effective***.
- **Outstanding issues and our motivation**
  - Human following performance of generated instructions has never been evaluated.
  - Efficacy of automated evaluation metrics for instruction generators has not been established.
- **Our objectives**
  - Address the gaps mentioned above.
  - Establish an effective metric for grounded instruction generation.



Human Wayfinder Evaluations

- **Human Wayfinders**: we evaluate the outputs of two SOTA instruction generators, **Speaker-Follower (SF)** (Fried et al. NeurIPS 2018) and **EnvDrop** (Tan et al. NAACL 2019), by asking people to follow them.
- We also compare Speaker-Follower and EnvDrop with the following:
  - Human Instructions (newly labeled)
  - Adversarial perturbations of human instructions capturing common failure modes in instruction generators:
    - Direction Swap
    - Entity Swap
    - Phrase Swap
  - Crafty (template-based)
- **Result**:
  - Speaker-Follower and EnvDrop are noticeably worse than perturbed human instructions, and are far behind human performance.
  - Both models are only on par with Crafty.

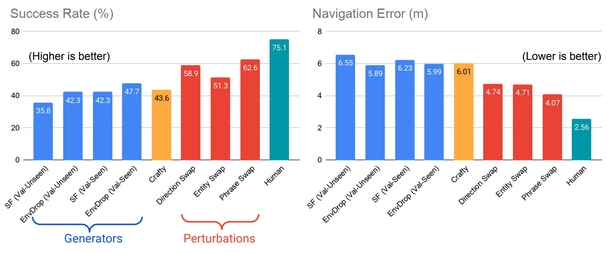


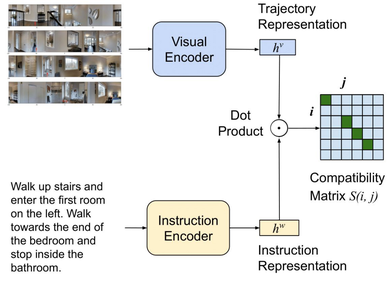
Figure: Comparison of human wayfinder performance for the Speaker-Follower (SF) and EnvDrop models with other seven sets of instructions.

Crafty Example: In front of you there's a tv. Pivot left, so that it is behind you. A lamp is ahead of you as you continue forward. You'll see a end table just on your right as you go slightly left. Walk forward, with the light switch on your left. Head left. You should see a sink slightly to your right. Continue straight and bear left, passing the stair to your right. Head forward, passing the wall on the left. Walk down the stairs. Wait next to the door frame.

Compatibility Model

To build better Instruction Generators, we first need accurate automatic evaluation metrics. We propose a trajectory-instruction compatibility model to learn the alignment in a shared latent space.

The independence between the two encoders facilitates learning using both contrastive and classification losses.



Structure of the Compatibility Model

Model Ablation	AUC
CE Loss	57.6
Focal Loss	59.2
Contrastive Loss	68.7
Contastive + CE	67.5
Contrastive + Focal	68.3
Contrastive + Focal + Paraphrase	72.2
Contrastive + Focal + Paraphrase + BERT embeds	**73.7**

Substantial gain from using contrastive loss

Result: Contrastive loss contributes to the most significant gain, while *focal loss*, *paraphrasing*, *hard negative mining*, & *BERT embeddings* are also important.

Evaluation of automated metrics and the compatibility model

Comparison of Correlation with Human Wayfinders

	Score	Ref	All Instructions (N=3.9k, M=9) NE ↓	SR ↑	SPL ↑	Quality ↑
System-Level	BLEU-4	✓	(0.00, 0.33)	(-0.22, 0.39)	(-0.22, 0.00)	(0.11, 0.39)
	CIDEr	✓	(0.06, 0.39)	(-0.22, 0.39)	(-0.22, 0.00)	(0.17, 0.39)
	METEOR	✓	(0.11, 0.44)	(-0.39, 0.28)	(-0.39, -0.06)	(0.00, 0.28)
	ROUGE	✓	(0.06, 0.39)	(-0.28, 0.39)	(-0.33, 0.00)	(0.06, 0.39)
	SPICE	✓	**(-0.67, -0.28)**	**(-0.06, 0.61)**	**(0.44, 0.78)**	**(0.56, 0.83)**
	BERTScore	✓	(0.06, 0.39)	(-0.22, 0.39)	(-0.22, 0.00)	(0.17, 0.39)
	$SPL_{1\text{-agent}}$		(-0.50, -0.06)	(-0.22, 0.44)	(0.11, 0.56)	(0.00, 0.44)
	$SPL_{3\text{-agents}}$		(-0.22, 0.17)	(-0.33, 0.39)	(0.00, 0.33)	(0.33, 0.61)
	$SDTW_{1\text{-agent}}$		(-0.44, 0.00)	(-0.22, 0.44)	(0.11, 0.50)	(0.00, 0.44)
	$SDTW_{3\text{-agents}}$		(-0.22, 0.17)	(-0.28, 0.33)	(0.00, 0.33)	(0.33, 0.61)
	Compatibility		(-0.17, 0.17)	(-0.17, 0.50)	(0.00, 0.28)	(0.44, 0.72)

	Score	Ref	All Instructions (N=3.9k, M=9) NE ↓	SR ↑	SPL ↑	Quality ↑
Instance-Level	BLEU-4	✓	(0.05, 0.09)	(-0.04, 0.00)	(-0.09, -0.05)	(-0.01, 0.03)
	CIDEr	✓	(0.06, 0.09)	(-0.04, -0.00)	(-0.11, -0.07)	(-0.02, 0.01)
	METEOR	✓	(0.00, 0.04)	(-0.05, -0.02)	(-0.04, 0.00)	(-0.01, 0.02)
	ROUGE	✓	(0.05, 0.08)	(-0.05, -0.01)	(-0.10, -0.06)	(-0.02, 0.02)
	SPICE	✓	(-0.05, -0.02)	(-0.00, 0.04)	(0.03, 0.06)	(0.03, 0.07)
	BERTScore	✓	(-0.04, -0.00)	(0.07, 0.12)	(-0.01, 0.03)	(0.07, 0.11)
	$SPL_{1\text{-agent}}$		(-0.18, -0.14)	(0.15, 0.19)	(0.14, 0.18)	(0.07, 0.11)
	$SPL_{3\text{-agents}}$		**(-0.22, -0.18)**	**(0.20, 0.24)**	**(0.18, 0.22)**	(0.10, 0.14)
	$SDTW_{1\text{-agent}}$		(-0.18, -0.14)	(0.15, 0.19)	(0.14, 0.18)	(0.08, 0.12)
	$SDTW_{3\text{-agents}}$		**(-0.22, -0.19)**	**(0.20, 0.24)**	**(0.18, 0.22)**	(0.11, 0.15)
	Compatibility		**(-0.20, -0.17)**	(0.13, 0.17)	**(0.17, 0.20)**	**(0.19, 0.23)**

No ref required

N refers to the number of instructions and *M* refers to the number of systems. If checked, *Ref* indicates that the metric requires reference instructions for comparison.

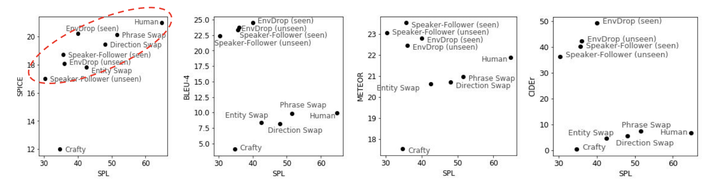
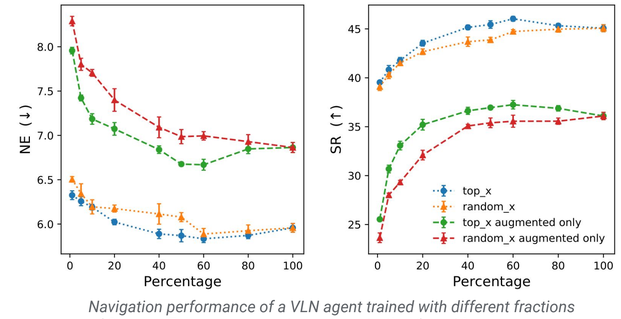


Figure: Standard evaluation metrics vs. human wayfinding outcomes (SPL) for 9 navigation instruction generation systems

- **System-level** (evaluating a model against many instances):
  - SPICE performs the best, while other automated metrics do not show any correlation with human wayfinder performance.
- **Instruction-level** (evaluating an individual instruction):
  - Our compatibility models performs the best
  - Almost as good: the SPL/STDW score averaged over *three* VLN Agents (Followers)
  - Additional advantage: Unlike SPICE, these methods don't require reference captions!

Data Augmentation for VLN

- Dashed-lines (green and red): use only augmented paths in training
- Dotted lines (blue and orange): use both augmented and R2Rtrain paths.
- Each point is the mean of 3 runs and the error bars represent the standard deviation of the mean.
- **Result**: The model-ranked fractions show consistent improvement over random samples of the same percentage.



Navigation performance of a VLN agent trained with different fractions of Speaker-Follower augmented paths, starting from 1%.

Conclusion

- Almost all recent VLN papers use data augmentation from an Instruction Generator (*Speaker*).
  - These generators have substantial headroom for improvement!
- Progress may have been hindered by a lack of suitable evaluation metrics.
  - Textual evaluation metrics should not be trusted in new domains without validation.
  - For navigation instructions - don't use BLEU, CIDER, METEOR or ROUGE to evaluate!
  - Use SPICE for model-level evaluation .
  - Use our learned compatibility model or VLN Agents for instruction-level evaluation.

Acknowledgements: We thank the Google Data Compute team, in particular Ashwin Kakarla and Priyanka Rachapally, for their tooling and annotation support for this project.