On the Evaluation of Vision-and-Language Navigation Instructions

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

EACL 2021
Vision-and-Language Navigation (VLN)

- VLN (Anderson et al. 2018) - the task of following navigation instructions to traverse a path in a photorealistic environment
Instruction Generators

VLN Agents (Followers) follow instructions to create paths through an environment

Our focus: Instruction Generators (Speakers) that map paths in an environment to instructions

- Very useful for VLN agent data augmentation (+5% success rate)
- Challenging task with its own practical applications

Figure credit: Fried et al. NeurIPS 2018
Instruction Generators

Two generators are used extensively for data augmentation in previous work:

- Speaker-Follower (*Fried et al. NeurIPS 2018*)
- EnvDrop (*Tan et al. NAACL 2019*)
Instruction Generators

Two generators are used extensively for data augmentation in previous work:

- **Speaker-Follower** (Fried et al. NeurIPS 2018)
- **EnvDrop** (Tan et al. NAACL 2019)

Walk out of the bedroom and turn left.
Walk down the stairs and stop at the bottom of the stairs.
Instruction Generators

Two generators are used extensively for data augmentation in previous work:

- Speaker-Follower (Fried et al. NeurIPS 2018)
- EnvDrop (Tan et al. NAACL 2019)

Comparisons:

- Human Instructions

Leave the room and turn left. With the wooden door behind you, keep walking straight. Stop after you go down a few stairs, just before entering a kitchen area.
Instruction Generators

Two generators are used extensively for data augmentation in previous work:

- Speaker-Follower (Fried et al. NeurIPS 2018)
- EnvDrop (Tan et al. NAACL 2019)

Comparisons:

- Human Instructions
- Direction Swap

Leave the room and turn left right. With the wooden door behind you, keep walking straight. Stop after you go down up a few stairs, just before entering a kitchen area.
Instruction Generators

Two generators are used extensively for data augmentation in previous work:

- Speaker-Follower (Fried et al. NeurIPS 2018)
- EnvDrop (Tan et al. NAACL 2019)

Comparisons:
- Human Instructions
- Direction Swap
- Entity Swap

Leave the room and turn left. With the wooden door kitchen area behind you, keep walking straight. Stop after you go down a few stairs, just before entering a kitchen area door.
Instruction Generators

Two generators are used extensively for data augmentation in previous work:

- Speaker-Follower (Fried et al. NeurIPS 2018)
- EnvDrop (Tan et al. NAACL 2019)

Comparisons:

- Human Instructions
- Direction Swap
- Entity Swap
- Phrase Swap

Leave the room and turn left. With the wooden door behind you, keep walking straight. Leave the room and turn left. Stop after you go down a few stairs, just before entering a kitchen area.
Instruction Generators

Two generators are used extensively for data augmentation in previous work:

- Speaker-Follower (Fried et al. NeurIPS 2018)
- EnvDrop (Tan et al. NAACL 2019)

Comparisons:

- Human Instructions
- Direction Swap
- Entity Swap
- Phrase Swap
- Crafty (template-based)

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 an 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.
Human Evals

Annotators try to follow instructions using PanGEA - 3 evals per instruction using R2R paths.

Instruction quality determined by human wayfinding performance:

- **NE**: Navigation Error
- **SR**: Success Rate (NE < 3m)
- **SPL**: Success weighted by inverse Path Length
- **Quality**: as assessed by annotators
- plus other metrics

PanGEA: [https://github.com/google-research/pangea](https://github.com/google-research/pangea)
Human Evals

Existing Instruction Generators are only slightly better than ‘Crafty’, our template-based approach.
Human Evals

Existing Instruction Generators are much worse than adversarially-perturbed human instructions

**Navigation Error (m)**

- SF (Val-Unseen): 6.55
- SF (Val-Seen): 5.89
- EnvDrop (Val-Unseen): 6.23
- EnvDrop (Val-Seen): 5.99
- Crafty: 6.01
- Direction Swap: 4.74
- Entity Swap: 4.71
- Phrase Swap: 4.07
- Human: 2.56

**Success Rate (%)**

- SF (Val-Unseen): 35.8
- SF (Val-Seen): 42.3
- EnvDrop (Val-Unseen): 42.3
- EnvDrop (Val-Seen): 47.7
- Crafty: 43.6
- Direction Swap: 58.9
- Entity Swap: 51.3
- Phrase Swap: 62.6
- Human: 75.1
Human Evals

Existing Instruction Generators are far worse than human instructions - substantial headroom!

| Navigation Error (m) | Success Rate (%) |
|----------------------|------------------|
| SF (Val-Unseen)      | SF (Val-Unseen)  |
| EnvDrop (Val-Unseen) | EnvDrop (Val-Unseen) |
| SF (Val-Seen)        | SF (Val-Seen)    |
| EnvDrop (Val-Seen)   | EnvDrop (Val-Seen) |
| Crafty               | Crafty           |
| Direction Swap       | Direction Swap   |
| Entity Swap          | Entity Swap      |
| Phrase Swap          | Phrase Swap      |
| Human                | Human            |

6.55 | 35.8 | 43.6 | 2.56 |
5.89 | 42.3 | 58.9 | 5.99 |
6.23 | 42.3 | 51.3 | 5.99 |
5.99 | 47.7 | 62.6 | 5.99 |
6.01 | 75.1 | 75.1 | 6.01 |
4.74 | 58.9 | 58.9 | 4.74 |
4.71 | 51.3 | 51.3 | 4.71 |
4.07 | 62.6 | 62.6 | 4.07 |
2.56 | 75.1 | 75.1 | 2.56 |
Compatibility Model

To build better Instruction Generators, we first need accurate automatic evaluation metrics

Proposed trajectory-instruction compatibility model
(dual encoder)

Walk up stairs and enter the first room on the left. Walk towards the end of the bedroom and stop inside the bathroom.
### Compatibility Model

**Evaluation:** Classify high vs. low quality instructions for R2R paths.

| Loss Configuration                  | AUC  |
|------------------------------------|------|
| CE Loss                            | 57.6 |
| Focal Loss                         | 59.2 |
| Contrastive Loss                   | 68.7 |
| Contrastive + 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**
- **Focal loss, paraphrasing, hard negative mining, & BERT embeddings are also important**
Automatic Instruction Evals

Which metrics correlate with human wayfinding performance?

System-level (evaluating a model)

| Score               | Ref | NE ↓ | SR ↑ | SPL ↑ | Quality ↑ |
|---------------------|-----|------|------|-------|-----------|
| BLEU-4              | ✓   |      |      |       |           |
| CIDEr               | ✓   |      |      |       |           |
| METEOR              | ✓   |      |      |       |           |
| ROUGE               | ✓   |      |      |       |           |
| SPICE               | ✓   |      |      |       |           |
| BERTScore           | ✓   |      |      |       |           |
| SPL1-agent          |     |      |      |       |           |
| SPL3-agents         |     |      |      |       |           |
| SDTW1-agent         |     |      |      |       |           |
| SDTW3-agents        |     |      |      |       |           |
| Compatibility       |     |      |      |       |           |
Automatic Instruction Evals

Which metrics correlate with human wayfinding performance?

**System-level (evaluating a model)**

Use SPICE metric, not BLEU!

| System-Level | All Instructions (N=3.9k, M=9) |
|--------------|--------------------------------|
| Score        | Ref | NE ↓ | SR ↑ | SPL ↑ | Quality ↑ |
| 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) |
| SPL1-agent   |     | (-0.50, -0.06) | (-0.22, 0.44) | ( 0.11, 0.56) | ( 0.00, 0.44) |
| SPL3-agents  |     | (-0.22, 0.17) | (-0.33, 0.39) | ( 0.00, 0.33) | ( 0.33, 0.61) |
| SDTW1-agent  |     | (-0.44, 0.00) | (-0.22, 0.44) | ( 0.11, 0.50) | ( 0.00, 0.44) |
| SDTW3-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) |
Automatic Instruction Evals

Which metrics correlate with human wayfinding performance?

*Instruction-level* (evaluating an individual instruction)

| Score      | Ref | NE ↓  | SR ↑  | SPL ↑  | Quality ↑ |
|------------|-----|-------|-------|--------|-----------|
| 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 |
| SPL1-agent |     | -0.18 | -0.14 | 0.15   | 0.19     | 0.14    | 0.18    | 0.07   | 0.11 |
| SPL3-agents|     | -0.22 | -0.18 | 0.20   | 0.24     | 0.18    | 0.22    | 0.10   | 0.14 |
| SDTW1-agent|     | -0.18 | -0.14 | 0.15   | 0.19     | 0.14    | 0.18    | 0.08   | 0.12 |
| SDTW3-agents|   | -0.22 | -0.19 | 0.20   | 0.24     | 0.18    | 0.22    | 0.11   | 0.15 |

Use our compatibility model!

Almost as good: the SPL/STDW score averaged over three VLN Agents (Followers)

Additional advantage: Unlike SPICE, these methods don’t require reference captions!
 Compatibility Model

For data augmentation, the compatibility model can filter out low-quality instructions... achieving the same or better performance with less data.
Conclusions

- Almost all recent VLN papers use data augmentation from an Instruction Generator (Speaker).
  - These generators have substantial room 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.

On the Evaluation of Vision-and-Language Navigation Instructions
Ming Zhao, Peter Anderson, Vihan Jain, Su Wang, Alex Ku, Jason Baldridge, Eugene Ie