Semantically Equivalent Adversarial Rules for Debugging NLP Models

Sameer Singh (UC Irvine)  
Carlos Guestrin
NLP / ML models are getting smarter: VQA

What type of road sign is shown?

> STOP.

NLP / ML models are getting smarter: MC (SQuAD)

The biggest city on the river Rhine is Cologne, Germany with a population of more than 1,050,000 people. It is the second-longest river in Central and Western Europe (after the Danube) at about 1,230 km (760 mi).

How long is the Rhine?

1230km

BiDAF [Seo et al 2017]
Oversensitivity in images

“panda” 57.7% confidence

+ ε

= “gibbon” 99.3% confidence

But (results)
Adversarial examples
What about text?

What type of road sign is shown?

> STOP.

What type of road sign is shown?
What about text?

What type of road sign is shown?

> STOP.

What type of road sign is shown?
Semantics matter

What type of road sign is shown?

> STOP.

Which type of road sign is shown?

> Do not Enter.
The biggest city on the river Rhine is Cologne, Germany with a population of more than 1,050,000 people. It is the second-longest river in Central and Western Europe (after the Danube) at about 1,230 km (760 mi).

Semantics matter
Adversarial Rules
Generalizing adversaries

What type of road sign is shown?

> STOP.

> Do not Enter.
Semantics matter

What color is the sky?

> Blue.

Which noun is the sky?

> Gray.
Semantics matter

The biggest city on the Rhine is Cologne, Germany with a population of more than 1,050,000 people. It is the second-longest river in Central and Western Europe (after the Danube) at about 1,230 km (760 mi).
Semantics matter

Detailed investigation of *Oncorhynchus keta*, showed that these fish digest ctenophores 20 times as fast as an equal weight of shrimps.

What is the *Oncorhynchus* also called?

- Chum salmon
- Oncorhynchus keta
Adversarial Rules
Semantically Equivalent Adversary (SEA)
Ingredients

1. Semantic score function $S(x, x')$
2. A black box model $f(x)$

Semantically Equivalent

AND

Different prediction
Revisiting adversaries

\[ \max_{x'} S(x, x') > \gamma \text{ s.t. } SEA(x, x') = \]
Semantic Similarity: Paraphrasing

[Mainlinson et al, 2017]

\[ P(x' | x) \]

Score

Good movie
- Portuguese Translation
- French Translation

Back translators

Translators

en - pt
- Portuguese Translation

en - fr
- French Translation

pt - en

fr - en

Score

Good movie
- 0.35

Great movie
- 0.34

Movie good

- ...

comes for free
Finding an adversary

What color is the tray? Pink
What color is the tray? Green
What color is it? Green
Which color is the tray? Green
What color is the tray? Green
Pink color is the tray? Green
Semantically Equivalent Adversarial Rules (SEARs)
From SEAs to Rules

1. Find SEAs
2. Propose Candidate Rules
3. Select Small Rule Set
Proposing Candidate Rules

What type of road sign is shown?

 Candidate Rules:

(What type $\rightarrow$ Which type)
(What NOUN $\rightarrow$ Which NOUN)
(WP type $\rightarrow$ Which type)
(WP NOUN $\rightarrow$ Which NOUN)
...

What type of road sign is shown?

What Which type of road sign is shown?

✓ What Which type of road sign is shown?
✗ What Which is the person looking at?
✗ What Which was I thinking?
From SEAs to Rules

Find SEAs

Propose Candidate Rules

Select Small Rule Set
Semantically Equivalent Adversarial Rules (SEARS)

1. High Adversary Count
2. Non-Redundancy

- color → colour
- What NOUN → Which NOUN

Selected Rules
Examples: VQA

Visual 7a-Telling [Zhu et al. 2016]

<table>
<thead>
<tr>
<th>SEAR</th>
<th>Questions / SEAs</th>
<th>f(x)</th>
<th>Flips</th>
</tr>
</thead>
<tbody>
<tr>
<td>WP VBZ $\rightarrow$ WP’s</td>
<td><em>What has</em> What’s been cut?</td>
<td>Cake Pizza</td>
<td>3.3%</td>
</tr>
<tr>
<td>What NOUN $\rightarrow$ Which NOUN</td>
<td><em>What</em> Which kind of floor is it?</td>
<td>Wood Marble</td>
<td>3.9%</td>
</tr>
<tr>
<td>color $\rightarrow$ colour</td>
<td>What <em>color</em> colour is the tray?</td>
<td>Pink Green</td>
<td>2.2%</td>
</tr>
<tr>
<td>ADV is $\rightarrow$ ADV’s</td>
<td><em>Where is</em> Where’s the jet?</td>
<td>Sky Airport</td>
<td>2.1%</td>
</tr>
</tbody>
</table>
## Examples: Machine Comprehension

BiDAF [Seo et al 2017]

<table>
<thead>
<tr>
<th>SEAR</th>
<th>Questions / SEAs</th>
<th>f(x)</th>
<th>Flips</th>
</tr>
</thead>
<tbody>
<tr>
<td>What VBZ →</td>
<td>What is What’s the NASUWT?</td>
<td>Trade union</td>
<td>2%</td>
</tr>
<tr>
<td>What’s</td>
<td></td>
<td>Teachers in Wales</td>
<td></td>
</tr>
<tr>
<td>What NOUN →</td>
<td>What resource Which resource was mined in the Newcastle area?</td>
<td>coal wool</td>
<td>1%</td>
</tr>
<tr>
<td>Which NOUN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>What VERB →</td>
<td>What was So what was Ghandi’s work called?</td>
<td>Satyagraha</td>
<td>2%</td>
</tr>
<tr>
<td>So what VERB</td>
<td></td>
<td>Civil Disobedience</td>
<td></td>
</tr>
<tr>
<td>What VBD →</td>
<td>What was And what was Kenneth Swezey’s job?</td>
<td>journalist sleep</td>
<td>2%</td>
</tr>
<tr>
<td>And what VBD</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Examples: Movie Review Sentiment Analysis

<table>
<thead>
<tr>
<th>SEAR</th>
<th>Reviews / SEAs</th>
<th>f(x)</th>
<th>Flips</th>
</tr>
</thead>
<tbody>
<tr>
<td>movie → film</td>
<td>Yeah, the <strong>movie</strong>-<strong>film</strong> pretty much sucked.</td>
<td>Neg</td>
<td>2%</td>
</tr>
<tr>
<td>film → movie</td>
<td>Excellent <strong>film</strong>-<strong>movie</strong>.</td>
<td>Pos</td>
<td>1%</td>
</tr>
<tr>
<td>is → was</td>
<td>Ray Charles <em>is</em> <strong>was</strong> legendary.</td>
<td>Pos</td>
<td>4%</td>
</tr>
<tr>
<td>this → that</td>
<td>Now <em>this</em> <strong>that</strong> is a movie I really dislike.</td>
<td>Neg</td>
<td>1%</td>
</tr>
</tbody>
</table>

FastText [Joulin et al. 2016]
Experiments
1. SEAs vs Humans
Set up

1. Humans
2. Top scored SEA
3. SEA (top 5) + Human

Evaluate adversaries for semantic equivalence
How often can SEAs be produced?

Visual Question Answering

- Human: 33.6
- SEA: 45
- Human + SEA: 51.3

Sentiment Analysis

- Human: 26
- SEA: 25.3
- Human + SEA: 33
Humans produce different adversaries:

They are so easy to love...

What kind of meat is on the boy’s plate?

How many suitcases?

Photography and directing were on point.

Also great directing and photography
2. SEARs vs Experts
Part 1: experts come up with rules

Objective: maximize mistakes with good rules
Part 2: experts evaluate our SEARs

<table>
<thead>
<tr>
<th>Rules to evaluate</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of POS tags</td>
<td></td>
</tr>
</tbody>
</table>

Please look at the rule results on the right.
The current rule is: replace(What NOUN, Which NOUN).

**Image**  | **Original**  | **After rule** |
|------------|--------------|----------------|

**Q: What color are the pots?**
Answer:
(a) Silver.
(b) Black.
(c) White.
d) Gold.

**Q: Which color are the pots?**
Answer:
a) Silver.
b) Black.
c) White.
d) Gold.

**Q: What color is the lampshade?**
Answer:
(a) A light yellow.
(b) A bright red.
(c) A subtle green.
d) A vivid orange.

**Q: Which color is the lampshade?**
Answer:
a) A light yellow.
b) A bright red.
c) A subtle green.
d) A vivid orange.

**Q: What animal is running in the background?**
Answer:
(a) A dog.
(b) A horse.
c) A llama.
d) A kangaroo.

**Q: Which animal is running in the background?**
Answer:
a) A dog.
b) A horse.
c) A llama.
d) A kangaroo.
Results
% correct predictions flipped

Time (minutes)
3. Fixing bugs
Closing the loop

Retrain model

Augment training

(color → colour)
(WP VBZ → WP’s)

Filter out bad rules
Results

% of flips due to bug

Visual QA

Sentiment

Original
Augmented

Fix bugs, no loss in accuracy

40
Conclusion

Models are prone to these bugs

SEAs and SEARs help find and fix them

Semantics matter
Semantic scoring is still a research problem…

<table>
<thead>
<tr>
<th>SEAR</th>
<th>Questions / SEAs</th>
<th>f(x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>on in</td>
<td>What is on in the background?</td>
<td>A-building Mountains</td>
</tr>
<tr>
<td></td>
<td>What is on in?</td>
<td>Lights The television</td>
</tr>
<tr>
<td>VBP is</td>
<td>Where are is the water bottles?</td>
<td>Table Vending machine</td>
</tr>
<tr>
<td></td>
<td>Where are is the people gathered?</td>
<td>Living room Kitchen</td>
</tr>
<tr>
<td>VERB on</td>
<td>What is on the background? A-building Mountains</td>
<td></td>
</tr>
<tr>
<td>VERB</td>
<td>Where are the planes parked on?</td>
<td>Concrete Landing strip</td>
</tr>
</tbody>
</table>
Problem: not comparable across instances

\[ S(x, x') = \min \left( 1, \frac{P(x'|x)}{P(x|x)} \right) \]
Examples: VQA

<table>
<thead>
<tr>
<th>SEAR</th>
<th>Questions / SEAs</th>
<th>f(x)</th>
<th>Flips</th>
</tr>
</thead>
<tbody>
<tr>
<td>WP VBZ → WP's</td>
<td><em>What has</em> What’s been cut?</td>
<td>Cake, Pizza</td>
<td>3.3%</td>
</tr>
<tr>
<td></td>
<td><em>Who is</em> Who’s holding the baby?</td>
<td>Woman, Man</td>
<td></td>
</tr>
<tr>
<td>What NOUN → Which NOUN</td>
<td><em>What Which</em> kind of floor is it?</td>
<td>Wood, Marble</td>
<td>3.9%</td>
</tr>
<tr>
<td></td>
<td><em>What Which</em> color is the jet?</td>
<td>Gray, White</td>
<td></td>
</tr>
<tr>
<td>color → colour</td>
<td><em>What color colour</em> is the tray?</td>
<td>Pink, Green</td>
<td>2.2%</td>
</tr>
<tr>
<td></td>
<td><em>What color colour</em> is the jet?</td>
<td>Gray, Blue</td>
<td></td>
</tr>
<tr>
<td>ADV is → ADV’s</td>
<td><em>Where is</em> Where’s the jet?</td>
<td>Sky, Airport</td>
<td>2.1%</td>
</tr>
<tr>
<td></td>
<td><em>How is</em> How’s the desk?</td>
<td>Messy, Empty</td>
<td></td>
</tr>
</tbody>
</table>
Examples: Movie Review Sentiment Analysis

<table>
<thead>
<tr>
<th>SEAR</th>
<th>Reviews / SEAs</th>
<th>f(x)</th>
<th>Flips</th>
</tr>
</thead>
<tbody>
<tr>
<td>movie → film</td>
<td>Yeah, the <em>movie</em>-<em>film</em> pretty much sucked .</td>
<td>Neg</td>
<td>Pos</td>
</tr>
<tr>
<td></td>
<td>This is not <em>movie</em>-<em>film</em> making .</td>
<td>Neg</td>
<td>Pos</td>
</tr>
<tr>
<td>film → movie</td>
<td>Excellent <em>film</em>-<em>movie</em> .</td>
<td>Pos</td>
<td>Neg</td>
</tr>
<tr>
<td></td>
<td>I’ll give this <em>film</em>-<em>movie</em> 10 out of 10 !</td>
<td>Pos</td>
<td>Neg</td>
</tr>
<tr>
<td>is → was</td>
<td>Ray Charles <em>is</em> <em>was</em> legendary .</td>
<td>Pos</td>
<td>Neg</td>
</tr>
<tr>
<td></td>
<td>It <em>is</em> <em>was</em> a really good show to watch .</td>
<td>Pos</td>
<td>Neg</td>
</tr>
<tr>
<td>this → that</td>
<td>Now <em>this</em> <em>that</em> is a movie I really dislike .</td>
<td>Neg</td>
<td>Pos</td>
</tr>
<tr>
<td></td>
<td>The camera really likes her in <em>this</em>-<em>that</em> movie.</td>
<td>Pos</td>
<td>Neg</td>
</tr>
</tbody>
</table>

FastText [Joulin et al 2016]
\[ SEA(x, x') = 1 \text{ if } S(x, x') > 2 \land f(x) \neq f(x') \]

\[ \max_{x'} S(x, x') > 2 \text{ s.t. } SEA(x, x') = \]