Discovering Bugs in NLP Models Using Natural Perturbations

Sameer Singh
circa 2005

News results for population of New York

UN: World Population Aging Rapidly In Developing Countries - RFE/RL
https://www.rferl.org/a/1099361.html
Apr 10, 2002 - A weeklong UN conference in Madrid is warning that the world's population is aging rapidly, with people aged 60 and older poised to ...

The town of the talk - The Economist
Feb 19, 2005 - The town of the talk. After the twin-tower nightmare, New York is back on form, says Anthony Gottlieb (interviewed here) ...
NLP has come a long way!
But we know models remain brittle...

Jia and Liang, EMNLP 2017

Anton van den Hengel, ACL 2018

Feng et al, EMNLP 2018
How do we discover bugs in NLP?

A **software bug** is an error, flaw, failure or fault in a computer program or system that causes it to produce an incorrect or unexpected result, or to behave in unintended ways.

Original Data $\rightarrow$ NLP Pipeline $\rightarrow$ Original Prediction

Perturb it in a specific way

Changed Data $\rightarrow$ NLP Pipeline $\rightarrow$ Unexpected Prediction!
Outline

Changing individual instances

Semantically Equivalent Adversaries

Semantically Implied Adversaries

Universal Adversaries

Changing training data

Link Prediction Adversaries
Outline

Changing individual instances

Semantically Equivalent Adversaries

Semantically Implied Adversaries

Universal Adversaries

Changing training data

Link Prediction Adversaries
Adversarial Examples: Oversensitivity

Find closest example with different prediction
Oversensitivity in images

Adversaries are indistinguishable to humans...

But unlikely in the real world (except for attacks)
What about text?

What type of road sign is shown?

Perceptible by humans, unlikely in real world
What about text?

A single word changes too much!
Semantics matter

What type of road sign is shown?

> STOP

Which type of road sign is shown?

> Do not Enter.

Bug, and likely in the real world
Semantics matter

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

Not all changes are the same: meaning should be same

> More than 1,050,000
How do we do this?

Semantically-Equivalent Adversary (SEA)

Semantically-Equivalent Adversarial Rules (SEARs)

$x \rightarrow \text{Backtranslation + Filtering} \rightarrow x'$

$(x, x') \rightarrow \text{Common Patterns} \rightarrow \text{Rules}$

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>What color is the tray?</td>
<td>Pink</td>
</tr>
<tr>
<td>What colour is the tray?</td>
<td>Green</td>
</tr>
<tr>
<td>Which color is the tray?</td>
<td>Green</td>
</tr>
<tr>
<td>What color is it?</td>
<td>Green</td>
</tr>
<tr>
<td>How color is tray?</td>
<td>Green</td>
</tr>
</tbody>
</table>

color $\rightarrow$ colour
## SEARs Examples: VisualQA

<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>What has What’s been cut?</td>
<td>Cake Pizza</td>
<td>3.3%</td>
</tr>
<tr>
<td>What NOUN $\rightarrow$ Which NOUN</td>
<td>What Which kind of floor is it?</td>
<td>Wood Marble</td>
<td>3.9%</td>
</tr>
<tr>
<td>color $\rightarrow$ colour</td>
<td>What color colour is the tray?</td>
<td>Pink Green</td>
<td>2.2%</td>
</tr>
<tr>
<td>ADV is $\rightarrow$ ADV’s</td>
<td>Where is Where’s the jet?</td>
<td>Sky Airport</td>
<td>2.1%</td>
</tr>
</tbody>
</table>

Visual7a-Telling [Zhu et al 2016]
### SEARs Examples: SQuAD

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

BiDAF [Seo et al 2017]
### SEARs Example: Sentiment

<table>
<thead>
<tr>
<th>SEAR</th>
<th>Reviews / SEAs</th>
<th>f(x)</th>
<th>Flips</th>
</tr>
</thead>
<tbody>
<tr>
<td>movie → <strong>film</strong></td>
<td>Yeah, the <em>movie</em> <em>film</em> pretty much sucked. This is not <em>movie</em> <em>film</em> making.</td>
<td>Neg Pos</td>
<td>2%</td>
</tr>
<tr>
<td><strong>film</strong> → movie</td>
<td>Excellent <em>film</em> <em>movie</em>. I’ll give this <em>film</em> <em>movie</em> 10 out of 10!</td>
<td>Pos Neg</td>
<td>1%</td>
</tr>
<tr>
<td>is → <strong>was</strong></td>
<td>Ray Charles <em>is</em> <strong>was</strong> legendary. It <em>is</em> <strong>was</strong> a really good show to watch.</td>
<td>Pos Neg</td>
<td>4%</td>
</tr>
<tr>
<td><strong>this</strong> → that</td>
<td>Now <em>this</em> <strong>that</strong> is a movie I really dislike. The camera really likes her in <em>this</em> <strong>that</strong> movie.</td>
<td>Neg Pos</td>
<td>1%</td>
</tr>
</tbody>
</table>

*fastText [Joulin et al., 2016]*
Semantic Adversaries

Semantics matter
Models are prone to these bugs
SEAs and SEARs help find and fix them
Outline

Changing individual instances

Semantically Equivalent Adversaries

Semantically Implied Adversaries

Universal Adversaries

Changing training data

Link Prediction Adversaries

ACL 2019
Consistency in predictions

So far, we have considered equivalence, i.e. \((x, y) \rightarrow (x', y)\)

\((x, y)\)  \hspace{2cm} \text{How many birds? 1}

\((x', y')\)  \hspace{2cm} \text{Is there 1 bird? Yes}
Evaluating Implication Consistency

Validation Data

Implication Generation

Implications

Consistency

\[ \# y \land y' \text{ correct} \]

\[ \# y \text{ correct} \]

\((x, y) \rightarrow (x, y), (x', y') \rightarrow f \rightarrow \]

based on parses, POS, WordNet, etc.
Visual QA

(x, y): What room is this? **bathroom**

**Logical Equivalence**

(x', y'): Is this a bathroom? **Yes**

**Necessary Condition**

(x', y'): Is there a bathroom in the picture? **Yes**

97% are valid!

**Mutual Exclusion**

(x', y'): Is this a kitchen? **No** 35%
## Visual QA Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Acc</th>
<th>LogEq</th>
<th>Mutex</th>
<th>Nec</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAAA (Kazemi, Elqursh, 2017)</td>
<td>61.5</td>
<td>76.6</td>
<td>42.3</td>
<td>90.2</td>
<td>72.7</td>
</tr>
<tr>
<td>Count (Zhang et al., 2018)</td>
<td>65.2</td>
<td>81.2</td>
<td>42.8</td>
<td>92.0</td>
<td>75.0</td>
</tr>
<tr>
<td>BAN (Kim et al., 2018)</td>
<td>64.5</td>
<td>73.1</td>
<td>50.4</td>
<td>87.3</td>
<td>72.5</td>
</tr>
</tbody>
</table>

Good at answering with numbers, but not questions with numbers, e.g. How many birds? **1** (12%) → Are there 2 birds? **yes** (<1%)
SQuAD

<table>
<thead>
<tr>
<th>Subj</th>
<th>When did Zhenjin die? <strong>1285</strong> → Who died in 1285? <strong>Zhenjin</strong> 29%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dobj</td>
<td>When did Denmark join the EU? <strong>1972</strong> → What did Denmark join in 1972? <strong>the EU</strong> 10%</td>
</tr>
<tr>
<td></td>
<td>73%</td>
</tr>
<tr>
<td>Amod</td>
<td>When did the Chinese famine begin? <strong>1331</strong> → Which famine began in 1331? <strong>Chinese</strong> 30%</td>
</tr>
<tr>
<td></td>
<td>97% are valid!</td>
</tr>
<tr>
<td>Prep</td>
<td>Who received a bid in 1915? <strong>Edison</strong> → When did Edison receive a bid? <strong>1915</strong> 46%</td>
</tr>
</tbody>
</table>

[Demszky et al. 2018]
# SQuAD Results

<table>
<thead>
<tr>
<th>Model</th>
<th>F1</th>
<th>Subj</th>
<th>Dobj</th>
<th>Amod</th>
<th>Prep</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>bidaf</td>
<td>77.9</td>
<td>70.6</td>
<td>65.9</td>
<td>75.1</td>
<td>72.4</td>
<td>72.1</td>
</tr>
<tr>
<td>bidaf+e</td>
<td>81.3</td>
<td>71.2</td>
<td>69.3</td>
<td>75.8</td>
<td>72.8</td>
<td>72.9</td>
</tr>
<tr>
<td>rnet</td>
<td>79.5</td>
<td>68.5</td>
<td>67.0</td>
<td>74.7</td>
<td>70.7</td>
<td>70.9</td>
</tr>
<tr>
<td>Mnem</td>
<td>81.5</td>
<td>70.3</td>
<td>68.0</td>
<td>75.8</td>
<td>71.9</td>
<td>72.2</td>
</tr>
</tbody>
</table>

Bad at questions with Wh-word as direct object
e.g. Who is Moses? (53%) vs Who did Hayk defeat? (12%)
Implication Adversaries

• We shouldn’t treat each prediction in isolation
  • Inconsistency leads to poor user experience
• Currently, rule-based system for generating them
• Already promising!
  • Reveals important bugs in the models
  • Even simple data augmentation is promising
Outline

Changing individual instances

Semantically Equivalent Adversaries

Semantically Implied Adversaries

Universal Adversaries

Changing training data

Link Prediction Adversaries

in preparation
Universal Adversaries

REDACTED for anonymity period
How do we do this?

REDACTED for anonymity period
### Textual Entailment

<table>
<thead>
<tr>
<th>Premise:</th>
<th>Two dogs are running through a field.</th>
</tr>
</thead>
</table>
| Hypothesis: | nobody
There are animals outdoors. |

**Entailment**

**Contradiction**

REDACTED for anonymity period
Language Modeling (GPTv2 small)

REDACTED for anonymity period
Changing Instances

• “Adversarial attacks” for NLP
  • Semantically Equivalent
  • Semantic Implications
  • Universal Tokens

• Useful for identifying different kinds of problems
  • Not all of them are traditional “bugs”

• General set of approaches that apply for most NLP models
Outline

Changing individual instances

Semantically Equivalent Adversaries

Semantically Implied Adversaries

Universal Adversaries

Changing training data

Link Prediction Adversaries

NAACL 2019
Different Kind of Model: Link Prediction

- **Entity Prediction**: $S \xrightarrow{R} O$
- **Relation Prediction**: $S \xrightarrow{R} O$
Knowledge Base Completion

\[ \phi(s, r, o) \]

**Table from Dettmers, et al. (2018)**

<table>
<thead>
<tr>
<th>Model</th>
<th>Score ( \psi_r(e_a, e_o) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>RESCAL [21]</td>
<td>( e_a^T W_r e_o )</td>
</tr>
<tr>
<td>SE [3]</td>
<td>( | W_r e_s - W_r e_o |_p )</td>
</tr>
<tr>
<td>TransE [1]</td>
<td>( | e_a + r_r - e_o |_p )</td>
</tr>
<tr>
<td>DistMult [34]</td>
<td>( \langle e_s, r_r, e_o \rangle )</td>
</tr>
<tr>
<td>ComplEx [33]</td>
<td>( \langle e_s, r_r, e_o \rangle )</td>
</tr>
<tr>
<td>ConvE</td>
<td>( f(\text{vec}(f([e_s; r_r; \omega])W) e_o) )</td>
</tr>
</tbody>
</table>
Link Prediction Example

Why was this prediction made?

What is this sensitive to?

Depends on the graph structure!
Link Prediction: Removing a Link
Link Prediction: Adding a Link
How do we do it?

\[ \text{argmax } \phi(s, r, o) - \bar{\phi}(s, r, o) \]

Original score

Score after retraining

Too many links to search!

Learn a continuous space of links, and search using gradient descent

Retraining is too expensive!

Taylor approximation, and utilize graph structure
Adding Links: How sensitive is the model?

Yago3 Hits@1 (Adding a fake link)

DistMult: Original 35, After Change 25
ConvE: Original 40, After Change 30

WordNet Hits@1 (Adding a fake link)

DistMult: Original 95, After Change 75
ConvE: Original 90, After Change 80
Removing Links: Cause behind prediction

Summarize by rule mining on which edges are used

**Bug in DistMult and ConvE**  
\[ \text{isMarriedTo}(a,c) \land \text{hasChild}(c,b) \Rightarrow \text{hasChild}(a,b) \]

**Only in DistMult**  
\[ \text{playsFor}(a,c) \land \text{isLocatedIn}(c,b) \Rightarrow \text{wasBornIn}(a,b) \]
\[ \text{isAffiliatedTo}(a,c) \land \text{isLocatedIn}(c,b) \Rightarrow \text{diedIn}(a,b) \]

**Only in ConvE**  
\[ \text{hasAdvisor}(a,c) \land \text{graduatedFrom}(c,b) \Rightarrow \text{graduatedFrom}(a,b) \]
\[ \text{influences}(a,c) \land \text{influences}(c,b) \Rightarrow \text{influences}(a,b) \]

* Identified as rules by [Yang et. al. 2015]
Changing the training data

• Sometimes, “bugs” are problems in the training data/pipeline
  • Embeddings of all kinds, for example
• To find these bugs, you need to change the training data
  • And efficiently estimate the effect of retraining
• We show how to do that for link prediction
Outline

Changing individual instances

  Semantically Equivalent Adversaries
  Semantically Implied Adversaries
  Universal Adversaries

Changing training data

  Link Prediction Adversaries
Thanks!

sameer@uci.edu
sameersingh.org
@sameer_

Work with Matt Gardner and me

as part of
The Allen Institute for Artificial Intelligence
in Irvine, CA

All levels: pre-PhD, PhD interns, postdocs, and research scientists!