Evaluating and Testing Natural Language Processing

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) ...

[adapted from Zadeh 2005, From Search Engines to Question-Answering Systems — The Need for New Tools]
NLP has come a long way!
Leaderboards often indicate it’s solved

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><strong>Human Performance</strong></td>
<td>86.831</td>
<td>89.452</td>
</tr>
<tr>
<td></td>
<td><em>Stanford University</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>(Rajpurkar &amp; Jia et al. '18)</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>ALBERT (ensemble model)</strong></td>
<td>89.731</td>
<td>92.215</td>
</tr>
<tr>
<td></td>
<td><em>Google Research &amp; TTIC</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td><strong>XLNet (single model)</strong></td>
<td>89.898</td>
<td>95.080</td>
</tr>
<tr>
<td></td>
<td><em>Google Brain &amp; CMU</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td><strong>Human Performance</strong></td>
<td>82.304</td>
<td>91.221</td>
</tr>
<tr>
<td></td>
<td><em>Stanford University</em></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><em>(Rajpurkar et al. '16)</em></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
And enough people believe them

ALIBABA AI MODEL TOPS HUMANS IN READING COMPREHENSION
ADAM NAJBERG | JANUARY 15, 2018

ROBOTS CAN NOW READ BETTER THAN HUMANS, PUTTING MILLIONS OF JOBS AT RISK
BY ANTHONY CUTHBERTSON ON 1/15/18 AT 8:00 AM EST
But we know models are brittle...

Jia and Liang, EMNLP 2017

**Article:** Super Bowl 50
**Paragraph:** “Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver’s Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.”

**Question:** “What is the name of the quarterback who was 38 in Super Bowl XXXIII?”

**Original Prediction:** John Elway
**Prediction under adversary:** Jeff Dean

---

Anton van den Hengel, ACL 2018

Feng et al, EMNLP 2018

**SQuAD**
**Context:** In 1899, John Jacob Astor IV invested $100,000 for Tesla to further develop and produce a new lighting system. Instead, Tesla used the money to fund his Colorado Springs experiments.

**Original**
**Reduced**
**Confidence**
What did Tesla spend Astor’s money on? did
0.78 → 0.91
How can we find these bugs?

automatically
From: Keith Richards
Subject: Christianity is the answer
NTTP-Posting-Host: x.x.com

I think Christianity is the one true religion.
If you’d like to know more, send me a note

---

**What** is the mustache made of? **banana**
How do we discover these “bugs”? 

<table>
<thead>
<tr>
<th>Original Instance</th>
<th>ML Pipeline</th>
<th>Original Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perturb it in a specific way</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Changed Instance</th>
<th>ML Pipeline</th>
<th>Expected Prediction</th>
</tr>
</thead>
</table>
Outline

Semantically Equivalent Adversaries

Semantically Implied Adversaries

Universal Adversaries
Outline

Semantically Equivalent Adversaries

Semantically Implied Adversaries

Universal Adversaries

Z. Zhao, D. Dua, S. Singh. Generating Natural Adversarial Examples. Int. Conf. on Learning Representations (ICLR). 2018

Counter-Examples: Oversensitivity

Find closest example with different prediction
Oversensitivity in images

“panda”
57.7% confidence

Adversaries are indistinguishable to humans...

“gibbon”
99.3% confidence

But unlikely in the real world (except for attacks)
Semantic Adversaries for NLP

Semantically-Equivalent Adversary

Semantically-Equivalent Adversarial Rules

$x$  Backtranslation + Filtering  $x'$  $(x, x')$

Patterns in “diffs”  Rules

color $\rightarrow$ colour

<table>
<thead>
<tr>
<th>What color is the tray?</th>
<th>Pink</th>
</tr>
</thead>
<tbody>
<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>
Semantics matter!

What type of road sign is shown?

> STOP.

What type of road sign is shown?

> Do not Enter.

Bug, and likely in the real world
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

How long is the Rhine?

> More than 1,050,000

Bug, and likely in the real world
## Transformation “Rules”: 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 → WP’s</td>
<td>What has Who’s been cut?</td>
<td>Cake Pizza</td>
<td>3.3%</td>
</tr>
<tr>
<td>What NOUN → Which NOUN</td>
<td>What Which kind of floor is it?</td>
<td>Wood Marble</td>
<td>3.9%</td>
</tr>
<tr>
<td>color → colour</td>
<td>What color colour is the tray?</td>
<td>Pink Green</td>
<td>2.2%</td>
</tr>
<tr>
<td>ADV is → 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]
## Transformation “Rules”: 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 Pos</td>
<td>2%</td>
</tr>
<tr>
<td></td>
<td>This is not <em>movie</em>-<em>film</em> making.</td>
<td>Neg Pos</td>
<td></td>
</tr>
<tr>
<td>film → movie</td>
<td>Excellent <em>film</em>-<em>movie</em>.</td>
<td>Pos Neg</td>
<td>1%</td>
</tr>
<tr>
<td></td>
<td>I’ll give this <em>film</em>-<em>movie</em> 10 out of 10!</td>
<td>Pos Neg</td>
<td></td>
</tr>
<tr>
<td>is → was</td>
<td>Ray Charles <em>is</em> <em>was</em> legendary.</td>
<td>Pos Neg</td>
<td>4%</td>
</tr>
<tr>
<td></td>
<td>It <em>is</em> <em>was</em> a really good show to watch.</td>
<td>Pos Neg</td>
<td></td>
</tr>
<tr>
<td>this → that</td>
<td>Now <em>this</em> <em>that</em> is a movie I really dislike.</td>
<td>Neg Pos</td>
<td>1%</td>
</tr>
<tr>
<td></td>
<td>The camera really likes her in <em>this</em>-<em>that</em> movie.</td>
<td>Pos Neg</td>
<td></td>
</tr>
</tbody>
</table>

fastText [Joulin et al., 2016]
Outline

Semantically Equivalent Adversaries

Semantically Implied Adversaries

Universal Adversaries

Consistency in Predictions

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

\((x, y)\)
- How many birds? 1

\((x', y')\)
- Is there 1 bird? Yes No
- Are there 2 birds? No Yes
- Are there any birds? Yes No
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**

### Mutual Exclusion

\[(x', y')\]: Is this a kitchen? **No**

- 97% are valid!
### 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 answer w/ numbers, but not questions w/ numbers e.g. How many birds? 1 (12%) → Are there 2 birds? yes (<1%)
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

Semantically Equivalent Adversaries
Semantically Implied Adversaries
Universal Adversaries

Universal Adversaries

• Instead of replacement, let’s consider additions
• Are there tokens that make the model misbehave?

\[
\begin{align*}
x & \quad f & \quad y \\
(b+x) & \quad f & \quad y
\end{align*}
\]

\(b\) has nothing to do with the task, and is not specific to \(x\)
## Sentiment Analysis

<table>
<thead>
<tr>
<th>Input</th>
<th>Model Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visually imaginative, thematically instructive and thoroughly delightful, it takes us on a roller-coaster ride...</td>
<td>Positive</td>
</tr>
<tr>
<td>As surreal as a dream and as detailed as a photograph, as visually dexterous as it is at times imaginatively overwhelming.</td>
<td>Positive</td>
</tr>
</tbody>
</table>

| 86 % accuracy | 29 % accuracy |

**comedy comedy blutarsky**

| 86 % accuracy | 23 % accuracy |
Question Answering

*Question:* Why did he walk?
For exercise, Tesla walked between 8 to 10 miles per day. He squished his toes one hundred times for each foot every night, saying that it stimulated his brain cells.

*Question:* Why did the university see a drop in applicants?
In the early 1950s, student applications declined as a result of increasing crime and poverty in the Hyde Park neighborhood. In response, the university became a . . . .
Triggers for GPTv2 Small

TH PEOPLE

Man goddreams Blacks are fucking animals. I’m tired of all these racist comments. I’m tired of being called a n*gger and shit all over my skin. I’m tired of having this bullshit called a “race”. I want this shit to stop. I’m tired to hear people say things like: I’m not racist, but that’s what you’re talking about? I’m not racist, but I think I’m going to be a better person. And I don’t care if you are or not...you can say whatever the fuck you want.

TH PEOPLE

Man goddreams Blacks are stupid for not wanting to be raped by White men. They are stupid for not having the same sense of entitlement as Whites to their own lives. They are stupid for not being able to take a joke like this seriously and then to be able to say “Well, that is why I’m not going to play with my children”. They are stupid for not realizing that White men are not going to rape White women. They are stupid for not understanding that Black men are going to rape Black women.

WARNING: Strong Language

63% of generations contain explicit racism

Transfers: optimized for 117M, works on 345M and 774M
Debugging by Changing Instances

• “Natural Perturbations” 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 models
Solution?

- It’s important to evaluate properly

1. We should evaluate on diverse sources of data

2. We should perform behavioral testing of our models
Evaluating on Diverse Data

Open Reading Benchmark

- Single model, multiple evaluations
- Seven QA datasets for evaluation
- Add more datasets over time
- Include “perturbation testing” directly
Behavioral Testing of Models

1. Write a few “sanity checks” by hand

   *MFT*: Short sentences with neutral adjectives and nouns.
   *MFT*: Negation with neutral in the middle

   - It was a commercial flight. **neutral**
   - This is an English company. **neutral**
   - I don’t think, given my history with airplanes, that it was an awesome crew. **neg**
   - I don’t think, given all that I’ve seen over the years, that that was a nice airline. **neg**

2. Small changes that should/should not change the output

<table>
<thead>
<tr>
<th>INV</th>
<th>pos</th>
<th>neutral</th>
<th>neg</th>
<th>Failure rate = 21.0%</th>
</tr>
</thead>
<tbody>
<tr>
<td>INV</td>
<td>neutral</td>
<td>neg</td>
<td>X</td>
<td>Failure rate = 36.0%</td>
</tr>
</tbody>
</table>

[under blind submission, do not share]
Thanks!

sameer@uci.edu
sameersingh.org
@sameer_