On Evaluation of Adversarial Perturbations for Sequence-to-Sequence Models

Paul Michel, Xian Li, Graham Neubig, Juan Pino
Adversarial Attacks/Perturbations

- Apply a small (indistinguishable) perturbation to the input that elicit large changes in the output
Adversarial Attacks/Perturbations

- Apply a small (indistinguishable) perturbation to the input that elicit large changes in the output

“panda”
57.7% confidence

Figure from Goodfellow et al. (2014)
Adversarial Attacks/Perturbations

- Apply a **small** (indistinguishable) perturbation to the **input** that elicit **large** changes in the **output**

Figure from Goodfellow et al. (2014)
Adversarial Attacks/Perturbations

- Apply a **small** (indistinguishable) perturbation to the **input** that elicit **large** changes in the **output**

Figure from Goodfellow et al. (2014)
Adversarial Attacks/Perturbations

- Apply a **small** (indistinguishable) perturbation to the **input** that elicit **large** changes in the **output**

![Figure from Goodfellow et al. (2014)](image-url)
Indistinguishable Perturbations

- **Small** perturbations are well defined in **vision**
  - Small $l_2 \approx$ indistinguishable to the human eye
Indistinguishable Perturbations

- **Small** perturbations are well defined in **vision**
  - Small $l_2$ ~ indistinguishable to the human eye

- What about **text**?
Not all Text Perturbations are Equal

He’s very friendly
Not all Text Perturbations are Equal

He’s very friendly

He’s pretty friendly

[Similar meaning]
Not all Text Perturbations are Equal

He’s very **annoying**
[Different meaning]

He’s **pretty** friendly
[Similar meaning]

He’s very friendly
Not all Text Perturbations are Equal

- **He’s very annoying** [Different meaning]
- **He’s pretty friendly** [Similar meaning]
- **He’s She friendly** [Nonsensical]
Not all Text Perturbations are Equal

- **He’s very annoying** [Different meaning] (X)
- **He’s pretty friendly** [Similar meaning] (✓)
- **He’s She friendly** [Nonsensical] (X)
- **He’s very freindly** [Typo] (✓)
Not all Text Perturbations are Equal

He’s very friendly

- He’s very annoying [Different meaning] ✗
- He’s pretty friendly [Similar meaning] ✔
- He’s She friendly [Nonsensical] ✗
- He’s very freindly [Typo] ✔

⇒ Can’t expect the model to output the same output!
Not all Text Perturbations are Equal

Can’t expect the model to output the same output!

This paper: Why and How you should evaluate adversarial perturbations
A Framework for Evaluating Adversarial Attacks
Problem Definition

Original $x$:
Ils le réinvestissent directement en engageant plus de procès.

Reference $y$:
They plow it right back into filing more troll lawsuits.
Problem Definition

Original

Ils le réinvestissent directement en engageant plus de procès.

Reference

They plow it right back into filing more troll lawsuits.
Ils le réinvestissent directement en engageant plus de procès.

They plow it right back into filing more troll lawsuits.

They direct it directly by engaging more cases.
They direct it directly by engaging more cases.
Problem Definition

Original $x$

Ils le réinvestissent directement en engageant plus de procès.

Attack

References $y$

They plow it right back into filing more troll lawsuits.

Evaluate

Base output $y_M$

They direct it directly by engaging more cases.
Problem Definition

Original

 Ils le réinvestissent directement en engageant plus de procès.

Attack

 Ils le réinvestissent directement en engageant plus de procès.

Adv. src

 Ils le réinvestissent directement en engageant plus de procès.

Evaluate

Reference

 They plow it right back into filing more troll lawsuits.

Base output

 They direct it directly by engaging more cases.

Adv. output

 Ils le réinvestissent directement en engageant plus de procès.

\( M \)
Ils le réinvestissent directement en engageant plus de procès.

Original $\mathbf{x}$

Attack

Adv. src $\hat{\mathbf{x}}$

Ref: Ils le réinvestissent directement en engageant plus de procès.

Evaluate too!

Reference $\mathbf{y}$

Evaluate

They plow it right back into filing more troll lawsuits.

$\mathbf{y}$

Base output $\mathbf{y}_M$

They direct it directly by engaging more cases.

$\mathbf{y}_M$

Adv. output $\hat{\mathbf{y}}_M$

.. de plus.
Source Side Evaluation

• Evaluate meaning preservation on the source side

\[ S_{src}(x, \hat{x}) \]

• Where \( S_{src} \) is a similarity metric such that

\[ S_{src}(\text{He’s very friendly }, \text{He’s pretty friendly}) > S_{src}(\text{He’s very friendly }, \text{He’s very annoying}) \]

\[ S_{src}(\text{He’s very friendly }, \text{He’s pretty friendly}) > S_{src}(\text{He’s very friendly }, \text{He’s She friendly}) \]

[...]
Target Side Evaluation

- Given $s_{tgt}$, a similarity metric on the target side
Target Side Evaluation

- Given $s_{tgt}$, a similarity metric on the target side
- Evaluate \textit{relative meaning destruction} on the target side

\[
d_{tgt}(y, y_M, \hat{y}_M) = \left\{ \frac{s_{tgt}(y, y_M) - s_{tgt}(y, \hat{y}_M)}{s_{tgt}(y, y_M)} \right\}
\]
Target Side Evaluation

- Given $s_{tgt}$, a similarity metric on the target side
- Evaluate relative meaning destruction on the target side

\[
d_{tgt}(y, y_M, \hat{y}_M) = \left\{ \begin{array}{c}
\frac{s_{tgt}(y, y_M) - s_{tgt}(y, \hat{y}_M)}{s_{tgt}(y, y_M)} \\
\end{array} \right.
\]
Target Side Evaluation

- Given $s_{tgt}$, a similarity metric on the target side
- Evaluate \textbf{relative meaning destruction} on the target side

\[
d_{tgt}(y, y_M, \hat{y}_M) = \begin{cases} 
  \frac{s_{tgt}(y, y_M) - s_{tgt}(y, \hat{y}_M)}{s_{tgt}(y, y_M)} \\
  \text{if } s_{tgt}(y, y_M) \neq 0
\end{cases}
\]
Target Side Evaluation

- Given $s_{tgt}$, a similarity metric on the target side

- Evaluate relative meaning destruction on the target side

$$d_{tgt}(y, y_M, \hat{y}_M) = \begin{cases} 0 & \text{if } s_{tgt}(y, \hat{y}_M) \geq s_{tgt}(y, y_M) \\ \frac{s_{tgt}(y, y_M) - s_{tgt}(y, \hat{y}_M)}{s_{tgt}(y, y_M)} & \text{otherwise} \end{cases}$$
Target Side Evaluation

- Given $s_{tgt}$, a similarity metric on the target side

- Evaluate **relative meaning destruction** on the target side

$$d_{tgt}(y, y_M, \hat{y}_M) = \begin{cases} 
0 & \text{if } s_{tgt}(y, \hat{y}_M) \geq s_{tgt}(y, y_M) \\
\frac{s_{tgt}(y, y_M) - s_{tgt}(y, \hat{y}_M)}{s_{tgt}(y, y_M)} & \text{otherwise}
\end{cases}$$
Target Side Evaluation

- Given $s_{tgt}$, a similarity metric on the target side

- Evaluate \textit{relative meaning destruction} on the target side

$$
\begin{align*}
    d_{tgt}(y, y_M, \hat{y}_M) &= \begin{cases}
        0 & \text{if } s_{tgt}(y, \hat{y}_M) \geq s_{tgt}(y, y_M) \\
        \frac{s_{tgt}(y, y_M) - s_{tgt}(y, \hat{y}_M)}{s_{tgt}(y, y_M)} & \text{otherwise}
    \end{cases}
\end{align*}
$$
Successful Adversarial Attacks

- Ensure that:

\[ 1 - s_{src}(x, \hat{x}) < d_{tgt}(y, y_M, \hat{y}_M) \]
Successful Adversarial Attacks

- Ensure that:

$$1 - s_{src}(x, \hat{x}) < d_{tgt}(y, y_M, \hat{y}_M)$$

Source meaning destruction
Successful Adversarial Attacks

• Ensure that:

\[ 1 - s_{src}(x, \hat{x}) < d_{tgt}(y, y_M, \hat{y}_M) \]

Source meaning destruction < Target meaning destruction
Successful Adversarial Attacks

- Ensure that:

\[ 1 - s_{src}(x, \hat{x}) < d_{tgt}(y, y_M, \hat{y}_M) \]

Source meaning destruction \quad Target meaning destruction

- Destroy the meaning on the target side more than on the source side
Which similarity metric to use?

- Human evaluation
  - 6 point scale, details in paper

"How would you rate the similarity between the meaning of these two sentences?"

0. The meaning is completely different or one of the sentences is meaningless
1. The topic is the same but the meaning is different
2. Some key information is different
3. The key information is the same but the details differ
4. Meaning is essentially the same but some expressions are unnatural
5. Meaning is essentially equal and the two sentences are well-formed [Language]
Which similarity metric to use?

- **Human evaluation**
  - 6 point scale, details in paper

- **BLEU** [Papineni et al., 2002]
  - Geometric mean of n-gram precision + length penalty

"How would you rate the similarity between the meaning of these two sentences?"

0. The meaning is completely different or one of the sentences is meaningless
1. The topic is the same but the meaning is different
2. Some key information is different
3. The key information is the same but the details differ
4. Meaning is essentially the same but some expressions are unnatural
5. Meaning is essentially equal and the two sentences are well-formed [Language]
Which similarity metric to use?

- **Human evaluation**
  - 6 point scale, details in paper

- **BLEU** [Papineni et al., 2002]
  - Geometric mean of n-gram precision + length penalty

- **METEOR** [Banerjee and Lavie, 2005]
  - Word matching taking into account stemming, synonyms, paraphrases...

"How would you rate the similarity between the meaning of these two sentences?"

0. The meaning is completely different or one of the sentences is meaningless
1. The topic is the same but the meaning is different
2. Some key information is different
3. The key information is the same but the details differ
4. Meaning is essentially the same but some expressions are unnatural
5. Meaning is essentially equal and the two sentences are well-formed [Language]
Which similarity metric to use?

- **Human evaluation**
  - 6 point scale, details in paper

- **BLEU** [Papineni et al., 2002]
  - Geometric mean of n-gram precision + length penalty

- **METEOR** [Banerjee and Lavie, 2005]
  - Word matching taking into account stemming, synonyms, paraphrases...

- **chrF** [Popović, 2015]
  - Character n-gram F-score
Experimental Setting
Data and Models

● Data
  ○ IWSLT 2016 dataset
  ○ {Czech, German, French} → English

● Models
  ○ LSTM based model
  ○ Transformer based model
  ○ Both word and sub-word based models
Gradient Based Adversarial Attacks on Text

- Idea: Back propagate through the model to score possible substitutions

Le gros chien .
Gradient Based Adversarial Attacks on Text

- Idea: Back propagate through the model to score possible substitutions
Gradient Based Adversarial Attacks on Text

- Idea: Back propagate through the model to score possible substitutions
Gradient Based Adversarial Attacks on Text

- Idea: Back propagate through the model to score possible substitutions
Gradient Based Adversarial Attacks on Text

- Idea: Back propagate through the model to score possible substitutions

Encoder

Le gros chien.

Decoder

The big dog. <eos>

Adversarial loss
Gradient Based Adversarial Attacks on Text

- Idea: Back propagate through the model to score possible substitutions

![Diagram of encoder-decoder model with adversarial loss](image)
Constrained Adversarial Attacks
Constrained Adversarial Attacks: kNN

- Only replace words with 10 nearest neighbors in embedding space

Example from our fr→en Transformer source embeddings

- grand (tall SING+MASC)
  - grands (tall PL+MASC)
  - grande (tall SING+FEM)
  - grandes (tall PL+FEM)
  - gros (fat SING+MASC)
  - grosse (fat SING+FEM)
- math (math)
  - maths (maths)
  - mathématique (mathematic)
  - mathématiques (mathematics)
  - objective (objective [ADJ] SING+FEM)
Constrained Adversarial Attacks: CharSwap

- Only swap word internal characters to get OOVs
  - grand $\rightarrow$ grnad
  - adversarial $\rightarrow$ advresarial
  - [...]  
- If that’s impossible, repeat the last character
  - he $\rightarrow$ heeeeeee

⇒ Realistic typos
# Constrained Adversarial Attacks

<table>
<thead>
<tr>
<th>Original</th>
<th><strong>Pourquoi</strong> faire cela ?</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>English gloss</strong></td>
<td>Why do this?</td>
</tr>
<tr>
<td><strong>Unconstrained</strong></td>
<td>construisant <em>(English: building)</em> faire cela ?</td>
</tr>
<tr>
<td><strong>kNN</strong></td>
<td>interrogez <em>(English: to question VB.2nd.PL)</em> faire cela ?</td>
</tr>
<tr>
<td><strong>CharSwap</strong></td>
<td>Puorquoi <em>(typo)</em> faire cela ?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Original</th>
<th>Si seulement je pouvais me muscler <strong>aussi</strong> rapidement.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>English gloss</strong></td>
<td>If only I could build my muscle <strong>this</strong> fast.</td>
</tr>
<tr>
<td><strong>Unconstrained</strong></td>
<td>Si seulement je pouvais me muscler <strong>etc</strong> rapidement.</td>
</tr>
<tr>
<td><strong>kNN</strong></td>
<td>Si seulement je pouvais me muscler <strong>plsu</strong> <em>(typo for “more”)</em> rapidement.</td>
</tr>
<tr>
<td><strong>CharSwap</strong></td>
<td>Si seulement je pouvais me muscler <strong>asusi</strong> <em>(typo)</em> rapidement.</td>
</tr>
</tbody>
</table>
Choosing an Similarity Metric

- Human vs automatic (pearson r):
  - Humans score original/adversarial input
  - Humans score original/adversarial output
  - Compare scores to automatic metric with Pearson correlation
Choosing an Similarity Metric

- Human vs automatic (pearson r):
  - Humans score original/adversarial input
  - Humans score original/adversarial output
  - Compare scores to automatic metric with Pearson correlation
Choosing an Similarity Metric

- Human vs automatic (pearson r):
  - Humans score original/adversarial input
  - Humans score original/adversarial output
  - Compare scores to automatic metric with Pearson correlation

- chrF better
  \[ s_{src} = s_{tgt} := \text{chrF} \]
  \[ d_{tgt} := \text{RDchrF} \]
  (Relative Decrease in chrF)
Effect of Constraints on Evaluation

Better target destruction

Better source preservation
Effect of Constraints on Adversarial Training
Effect of Constraints on Adversarial Training

- Adversarial training \(\approx\) training with adversarial examples

\[
\mathcal{L}'(x, y) = (1 - \alpha)NLL(x, y) + \alpha NLL(\hat{x}, y)
\]

- \(\alpha = 0\): Standard training
- \(\alpha = 1\): Training only on adversarial examples
Effect of Constraints on Adversarial Training

- Adversarial training ≈ training with adversarial examples

\[ \mathcal{L}'(x, y) = (1 - \alpha)NLL(x, y) + \alpha NLL(\hat{x}, y) \]

- \( \alpha = 0 \): Standard training
- \( \alpha = 1 \): Training only on adversarial examples

- Training with Unconstrained attacks vs CharSwap attacks

- Evaluate on
  - robustness to CharSwap attacks
  - Accuracy on non-adversarial data
Effect of Constraints on Adversarial Training: Adversarial Robustness

- Robustness to CharSwap attacks on the validation set

lower is better
Effect of Constraints on Adversarial Training: Adversarial Robustness

- Robustness to CharSwap attacks on the validation set

- Lower values indicate better robustness.
Effect of Constraints on Adversarial Training: Adversarial Robustness

- Robustness to CharSwap attacks on the validation set

![Graph showing RDcharF for different languages and models. The x-axis represents different language pairs (cs-en, de-en, fr-en), and the y-axis represents RDcharF. The colors represent different models: Base, Unconstrained-adv (α = 0.5), Unconstrained-adv (α = 1), CharSwap-adv (α = 0.5), and CharSwap-adv (α = 1). The legend indicates that lower values are better.](image_url)
Effect of Constraints on Adversarial Training: Adversarial Robustness

- Robustness to CharSwap attacks on the validation set

\[ \text{lower is better} \]

- Adversarial training ⇒ better robustness
Effect of Constraints on Adversarial Training: Accuracy on Non-Adversarial Input

- Target chrF on the original test set
Effect of Constraints on Adversarial Training: Accuracy on Non-Adversarial Input

- Target chrF on the original test set

Higher is better
Effect of Constraints on Adversarial Training: Accuracy on Non-Adversarial Input

- Target chrF on the original test set

Higher is better
Effect of Constraints on Adversarial Training: Accuracy on Non-Adversarial Input

- Target chrF on the original test set

- Unconstrained attacks ⇒ hurts accuracy
Takeway

● When doing adversarial **attacks**
  ○ Evaluate meaning preservation on the source side

“How would you rate the similarity between the meaning of these two sentences?”

0. The meaning is completely meaningless.
1. The topic is the same but the meaning is different.
2. Some key information is different.
3. The key information is the same but the details differ.
4. Meaning is essentially the same but some expressions are unnatural.
5. Meaning is essentially equal and the two sentences are well-formed.
Takeway

● When doing adversarial attacks
  ○ Evaluate meaning preservation on the source side

● When doing adversarial training
  ○ Consider adding constraints to your attacks
Takeway

- When doing adversarial **attacks**
  - Evaluate meaning preservation on the source side

- When doing adversarial **training**
  - Consider adding constraints to your attacks

- Not only true for seq2seq!
  - Easily transposed to classification, etc..
  - Just adapt $S_{src}$ and $S_{tgt}$ accordingly
TEAPOT

- Tool implementing our evaluation framework
- `pip install teapot-nlp`
- [github.com/pmichel31415/teapot-nlp](https://github.com/pmichel31415/teapot-nlp)

```bash
$ teapot
  --src examples/MT/src.fr
  --adv-src examples/MT/adv.charswap.fr
  --out examples/MT/base.en
  --adv-out examples/MT/adv.charswap.en
  --ref examples/MT/ref.en
```

will output:

```
Source side preservation (ChrF):
Mean: 86.908
Std: 11.622
5%-95%: 64.109-97.683

Target side degradation (ChrF):
Mean: 21.085
Std: 22.106
5%-95%: 0.000-67.162

Success percentage: 65.20%
```
Questions
Gradient Based Adversarial Attacks on Text

- Idea: Word substitution $\iff$ Adding word vector difference

\[
\begin{align*}
\begin{array}{cccc}
1 & 3.1 & -0.1 & 0.7 \\
-2 & -0.2 & 1.3 & 0.1 \\
0.5 & 0 & -3 & 2 \\
\end{array}
+ \begin{array}{cccc}
0 & 0 & 0.3 & 0 \\
0 & 0 & 1.2 & 0 \\
0 & 0 & 2.7 & 0 \\
\end{array}
= \begin{array}{cccc}
1 & 3.1 & 0.2 & 0.7 \\
-2 & -0.2 & 2.5 & 0.1 \\
0.5 & 0 & -0.3 & 2 \\
\end{array}
\end{align*}
\]

- Use the 1st order approximation to maximize the loss

\[
\arg\max_w \mathcal{L}(x_i = v_w) - \mathcal{L}(x_i = v_{\text{dog}}) \approx \nabla x_i \mathcal{L}^T[v_w - v_{\text{dog}}]
\]
Human Evaluation: the Gold Standard

Check for semantic similarity and fluency

“How would you rate the similarity between the meaning of these two sentences?”

0. The meaning is completely different or one of the sentences is meaningless
1. The topic is the same but the meaning is different
2. Some key information is different
3. The key information is the same but the details differ
4. Meaning is essentially the same but some expressions are unnatural
5. Meaning is essentially equal and the two sentences are well-formed [Language]
## Example of a Successful Attack

(source chrF = 80.89, target RDchrF = 84.06)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>Ils le réinvestissent directement en engageant plus de procès.</td>
<td>Ils le réinvestissent directement en engageant plus de procès.</td>
<td>They plow it right back into filing more troll lawsuits.</td>
<td>They direct it directly by engaging more cases.</td>
<td>.. de plus.</td>
</tr>
<tr>
<td>Adv. src.</td>
<td>Ilss le réinvestissent dierctement en engagaent plus de procès.</td>
<td><strong>direct</strong> method by engaging more cases.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ref.</td>
<td>They plow it right back into filing more troll lawsuits.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base output</td>
<td>They direct it directly by engaging more cases.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adv. output</td>
<td>.. de plus.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
# Example of an Unsuccessful Attack

(source chrF = 54.46, target RDchrF = 0.00)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>C’était en Juillet 1969.</td>
<td>C’étia en Jiulet 1969.</td>
<td>This is from July, 1969.</td>
<td>This was in July 1969.</td>
<td>This is. in 1969.</td>
</tr>
</tbody>
</table>